UNIVERSITY OF SOUTHAMPTON

Engineering and the Environment

Institute of Sound and Vibration Research

SPEECH ENHANCEMENT BASED ON NEURAL NETWORKS FOR IMPROVED SPEECH PERCEPTION IN NOISE BY PEOPLE WITH HEARING LOSS

by

Tobias Goehring

Thesis for the degree of
Doctor of Philosophy

December 2016

UNIVERSITY OF SOUTHAMPTON

Engineering and the Environment

Auditory Research and Signal Processing

Thesis for the degree of Doctor of Philosophy

SPEECH ENHANCEMENT BASED ON NEURAL NETWORKS FOR IMPROVED SPEECH PERCEPTION IN NOISE BY PEOPLE WITH HEARING LOSS

Tobias Goehring

ABSTRACT

Hearing loss can lead to problems with communication, affect the psychological wellbeing and decrease the quality of life of an affected person. One of the main challenges for people with hearing loss is speech perception in noisy environments. Whereas hearing devices such as hearing aids and cochlear implants successfully provide high levels of speech understanding in quiet acoustic conditions, they still fail to recover speech intelligibility in situations with background noise to a level obtained with the healthy auditory system. This thesis is about the development and evaluation of a speech enhancement algorithm for hearing devices to improve speech perception in noise by people with hearing loss. The proposed algorithm applies an artificial neural network algorithm to the task of speech enhancement. The algorithm decomposes the noisy input signal into time-frequency units, extracts a set of auditory-inspired features and feeds them to the neural network to produce an estimate of the frequency channels that contain more perceptually important information in terms of the energy ratio between the speech and the background noise. This estimate is used to retain only the frequency channels with high speech energy by attenuating the noise-dominated frequency channels. It is hypothesized that this processing leads to improved speech intelligibility in noise provided that the estimate of the energy ratio is accurate. The neural network is optimized for this task using significant amounts of acoustic training data and has been evaluated in several listening experiments with people with hearing loss including users of hearing aids and users of cochlear implants. Several aspects of the proposed speech enhancement framework based on neural networks have been investigated. The temporal window that can be used to extract auditory-inspired features for the processing has been evaluated by measuring the tolerance of processing delay by people with hearing loss. It was shown that the occurrence of hearing loss significantly increased the tolerance of processing delay and thus may allow for the use of longer temporal windows that are required for the processing of more complex speech enhancement algorithms. The results of a second listening study on processing delay indicated that further increases in the tolerable length of processing delay may occur based on long-term acclimatisation effects. The increased tolerance to processing delay by people with hearing loss and effects of long-term acclimatisation may allow for longer temporal windows for the processing that enable complex algorithms to run in real time without causing disturbance for the user of a hearing device. The speech enhancement framework was evaluated in terms of its benefits for understanding speech in challenging environments by users of hearing aids. Speech intelligibility and quality scores were obtained for subjects with mild to moderate hearing loss listening to sentences in speech-shaped noise and multi-talker babble following processing with the algorithm. Intelligibility and quality scores were significantly improved by the proposed approach using an auditory-inspired feature set. Results indicated advantages in performance over a more classical Wiener filter algorithm. Furthermore, the neural network based approach appeared more promising than dictionary-based sparse coding in terms of performance and ease of implementation. In order to evaluate the algorithm for users of cochlear implants, two listening studies were performed to measure speech intelligibility in background noise. Firstly, normal hearing subjects listening to vocoded stimuli to simulate CI speech perception obtained significant improvements in speech intelligibility in stationary and fluctuating noise over both unprocessed and Wiener filter processed conditions. Secondly, a listening study with 14 CI users obtained improvements in speech-in-noise performance for three types of background noise. Two neural network based algorithms were compared: a speaker-dependent algorithm, that was trained on the target speaker used for testing, and a speaker-independent algorithm, that was trained on different speakers. Significant improvements in the intelligibility of speech in stationary and fluctuating noises were found relative to the unprocessed condition for the speaker-dependent algorithm in all noise types and for the speaker-independent algorithm in 2 out of 3 noise types. The neural network based algorithms used noise-specific neural networks that generalized to novel segments of the same noise type and worked over a range of SNRs. The proposed algorithm has the potential to improve the intelligibility of speech in noise for users of hearing aids and cochlear implants while meeting the requirements of low computational complexity and processing delay for real-time application. The last investigation in this thesis was concerned with the individual preferences by potential users of speech enhancement algorithms. A study was performed to obtain user-controlled parameters for the strength of noise reduction processing by normal hearing and hearing impaired subjects and the choice in parameters was evaluated in terms of their efficacy for improving speech understanding in background noise, the awareness of background sounds and perceptual quality ratings. Interestingly, the group with hearing loss chose similar parameters compared to the normal hearing group and was as good or better in terms of the obtained benefits for speech understanding in noise with the individualized noise reduction processing. However, hearing impaired listeners seemed to be less robust to variations in the parameters and were significantly less aware of the background sounds after noise reduction processing. Overall, the results of this thesis provide further evidence for the promising approach of neural network based speech enhancement for potential application in hearing devices to obtain benefits in speech perception in noise for people with hearing loss.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Declaration of Authorship

I, TOBIAS GOEHRING declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

SPEECH ENHANCEMENT BASED ON NEURAL NETWORKS FOR IMPROVED SPEECH PERCEPTION IN NOISE BY PEOPLE WITH HEARING LOSS

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Goehring, T., Bolner, F., Monaghan, J. J., van Dijk, B., Zarowski, A., & Bleeck, S. (2017). Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Research*, 344, (pp. 183-194).

Monaghan, J. J., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C., & Bleeck, S. (2017). Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 141(3), (pp. 1985-1998).

Goehring, T., Yang, X., Monaghan, J., Bleeck, S. (2016). Speech enhancement for hearing-impaired listeners using deep neural networks with auditory-model based features. In *2016 EURASIP 24th European Signal Processing Conference (EUSIPCO)* (pp. 2300-2304). IEEE.

Bolner, F., Goehring, T., Monaghan, J., van Dijk, B., Wouters, J. and Bleeck, S. (2016). Speech enhancement based on neural networks applied to cochlear implant coding strategies. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6520-6524). IEEE.

Signed:

Date:

Acknowledgements

First of all, I would like to thank my supervisor Prof Stefan Bleeck for his advice, continuous motivation and support. I appreciate the trust and freedom to explore my ideas in a self-dependent way and the open door and endless discussions that helped me along the way. I also want to thank Dr Steven Bell, Dr Bas van Dijk, Prof Norbert Dillier and Prof Mahesan Niranjan for their help and support during my studies.

I would like to acknowledge my colleagues Jessica Monaghan, Xin Yang and Federico Bolner for the successful collaboration. It has been a great experience to work with you. Furthermore, I want to thank my colleagues within the ISVR: Mark Fletcher, Martin Spencer, Frederique Vanheusden, Zoe Bevis, Hannah Shutt and Matt Blyth for the helpful discussions and the good vibe in the office. My family and friends back in Germany have provided great support during the years and I appreciate that the larger distance did not affect our relationships too much. I want to thank my buddies Michele Iodice and Giorgos Dritsakis for their friendship and my flatmates Dirk Banholzer and Michele Ziletti for the successful flatsharing. Thanks to the students that passed through our lab: Clement Gaultier, Josie Chapman, Enrique Marquez, Sakeenah Kanji, Shangquo Wang, Varut Wiseschinda and Marina Forbes for your help with the studies and I hope that I have been a useful addition to your supervisory team.

I want to acknowledge the colleagues within the ICanHear project. There are too many names to mention (all on www.icanhear.eu) but I very much appreciate that I had the opportunity of being part of this great team. The scientific workshops (8 in total) were all inspiring, scientifically and fun-loaded trips that added a lot of knowledge, pleasure and motivation to these three years. My PhD research was funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. PITN-GA-2012-317521 (ITN ICanHear).

Finally, I am very grateful to Kira Liebert for her motivation throughout and the initial momentum that made me find this opportunity in the first place. It was one of the best decisions in my life and I think we both never regretted this step. Love you and hope we will continue taking on new challenges together.

It has been a great journey and I highly appreciate the opportunity.

LIST OF ABBREVIATIONS

ACE	advanced combination encoder
AI	articulation index
AIM	auditory image model
AMS	amplitude modulation spectrum
CI	cochlear implant
DCT	discrete cosine transform
DNN	deep neural network
DSP	digital signal processor
FA	false alarm rate
FFT	fast fourier transform
GMM	gaussian mixture model
HA	hearing aid
HI	hearing impaired
HIT	hit rate
HL	hearing loss
IBM	ideal binary mask
IRM	ideal ratio mask
IWF	ideal wiener filter
LC	local criterion
MFCC	mel frequency cepstral coefficients
NCM	normalized covariance metric
NH	normal hearing
NN	neural network
NNSE	neural network based speech enhancement
NR	noise reduction
PTA	pure-tone audiometry
RMS	root mean square
SII	speech intelligibility index
SE	speech enhancement
SNR	signal-to-noise ratio
SPL	sound pressure level
SRT	speech reception threshold
STFT	short-time fourier transform
STOI	short-term objective intelligibility measure
SSN	speech shaped noise
SWN	speech weighted noise
UN	unprocessed (without speech enhancement processing)
VAD	voice activity detection
WF	Wiener filter

1 INTRODUCTION

Speech perception in noise is still one of the main challenges for people with hearing loss (Arehart et al., 2015). This problem can have a strong impact on the life of an affected person since it may lead to avoidance of social situations or disadvantages at work due to communication difficulties. One in six people in the UK are affected by hearing loss rendering this problem a major one for society. Hearing devices such as hearing aids and cochlear implants successfully provide high levels of speech understanding in quiet acoustic conditions but still fail to recover speech intelligibility in noise to a level obtained with the healthy auditory system (*"I can hear you but I do not understand.")*. This leads to an investigation of novel approaches to speech enhancement that try to overcome this limitation by improving speech perception in noise for people with hearing loss.

The main objective of this thesis is to investigate a novel approach for improving speech perception in noise for people with hearing loss. This goal is tackled from different angles:

- Firstly, *processing delay*, one of the main limitations for the development of speech enhancement algorithms, was investigated by measuring listeners' tolerance of processing delay for speech using a real time setup. Short processing delay constitutes a vital requirement to ensure that an algorithm can potentially be used in hearing devices that operate in real time and determines the temporal window that can be used for speech processing in those devices.
- Secondly, a recent and very promising approach to *speech enhancement* was followed and optimized further for the application to hearing devices. This approach makes use of insights into the human auditory system that constitutes an ideal prototype for detecting speech in noisy environments. This auditory-inspired frontend was combined with a powerful neural network algorithm that can adaptively learn to predict a target signal, e.g. speech in noise, based on characteristic patterns in the training data. The algorithm was evaluated with users of hearing aids and users of cochlear implants to investigate whether listeners with hearing loss benefit in terms of speech perception in noise.
- Thirdly, the individual preferences of users of speech enhancement algorithms were investigated by using a parameterisation procedure to determine individually-preferred processing parameters for noise reduction and an evaluation was performed to investigate if the *user-controlled parameters* were chosen adequately to obtain benefits in understanding speech in noise while maintaining an awareness of the background sounds. This compromise between reduction of background noise and awareness of the acoustic environment is of great importance for applications in hearing devices in practice.

This thesis begins with a thorough introduction of the background topics in chapter 2 that build the fundamental basis for the research conducted in this thesis. This includes an overview on how humans hear sound, in particular speech, the causes and effects of sensorineural hearing loss and the technical working principles of hearing devices such as hearing aids and cochlear implants. Focus is put on the intelligibility of speech in background noise perceived by humans and how algorithms function that attempt to enhance speech perception in noise.

Chapter 3 deals with the tolerance of *processing delay* - the time that passes until a sound is presented to the user of a hearing device. This delay is restricted to only a few milliseconds by the perceptual requirements of the user and thus limits the development of speech enhancement algorithms by determining the maximum temporal window that can be used for the processing. The first part of chapter 3 investigates if and by how much the occurrence of *hearing loss affects the tolerance of processing delay* by comparing normal hearing and hearing impaired listener groups' annoyance ratings of stimuli with processing delay. The second part of chapter 3 explores the question if *long-term acclimatization to processing delay* increases tolerance of processing delay.

Chapter 4 and 5 contain the core of this thesis and cover the work on speech enhancement algorithms that try to improve speech perception in noise by listeners with hearing loss. The algorithm framework is based on two main aspects: the modelling of the human auditory system to extract auditory-inspired features from noise-corrupted speech sounds and the use of an artificial neural network algorithm that is trained to identify speech patterns in background noise. The combination of an auditory-inspired frontend with a powerful and adaptive backend is intended to overcome the limitations of state-of-the-art speech enhancement algorithms that struggle to provide improvements in speech intelligibility in background noise. Chapter 4 deals with the evaluation of the algorithm framework for *speech perception in noise by hearing aid users* and the comparison between auditory-inspired and more conventional feature sets. Chapter 5 describes the optimization steps that have been taken to apply the algorithm to cochlear implant processing strategies and evaluates the algorithm with normal hearing listeners using vocoder simulations of cochlear implant processing which led to an investigation of the benefits in terms of *speech perception in noise by users of cochlear implants*.

Chapter 6 investigates *user-controlled parameters for speech enhancement processing*. Large individual differences in the benefits obtained from speech enhancement processing in the other chapters lead to the question if listeners do require more individually-optimized parameters to access the full potential of speech enhancement algorithms. The influence of hearing loss on the compromise between the required strength of noise reduction and the awareness of background sounds is of interest for the development of speech enhancement algorithms that follow the approach taken in chapters 4 and 5.

A *general discussion* in chapter 5 and the *concluding part* in chapter 7 complete this thesis.

2 BACKGROUND

Overview

This chapter covers the background to this thesis and is based entirely on the existing literature. Thus no claims of authority or completeness are made by the author for the intellectual content of this chapter.

First, the basic structure of the human auditory system is presented and followed by the physiological changes and psychological consequences of sensorineural hearing loss. The prevailing treatments of hearing loss in form of hearing prostheses such as hearing aids and cochlear implants are introduced and their potential benefits for the users in terms of speech understanding in noise are reviewed. The important characteristics of speech that increase its robustness to background sounds are presented. Finally, the research field of speech enhancement is introduced by a brief overview on its main ideas and working principles. Emphasis is put on a review of single-microphone noise reduction algorithms that constitute the main approach of this thesis.

2.1 THE HUMAN AUDITORY SYSTEM

The human auditory system receives acoustic waves from the outside world that carry information in the audible range from 20 to 20.000 Hz and translates them into auditory sensations (Plack, 2013). The sound pressure waves pass through sequential stages of the auditory system to finally result in an auditory sensation in the human brain. The first stage in the auditory system is the *auditory periphery* that consists of the outer and middle ear (the *conductive hearing system*) before the sound wave enters the inner ear (the *cochlea*) that connects to the auditory nerve. From there, auditory signals pass, now as electrical activity, through the auditory brainstem with its numerous stages of auditory nuclei and the auditory midbrain up to the primary auditory cortex of the brain (Figure 2.1).

Figure moved to Appendix

Figure 2.1 Basic structure of the human auditory system that generates an auditory sensation from an incoming acoustic sound wave after passing several stages (image from Lily et al., 2013, doi:10.1038/nrendo.2013.58).

The conductive stages of the auditory system (the auditory periphery) filter and amplify the incoming sound waves and match the impedance to the inner ears' cochlear fluids. Hereby, the pinna causes spectral modifications depending on the direction that the sound is coming from and the ear canal acts as a band-pass filter to increase the sensitivity to frequencies in the range between 1000 and 6000 Hz (Plack, 2013). The middle ear picks up the air-conducted sound wave from the outer ear through the tympanic membrane and propagates the pressure wave through the auditory ossicles (malleus, incus and stapes) to the oval window at the cochlea (inner ear). The pressure is increased by a factor of 20-30 by projection onto a smaller area and using a lever motion of the tympanic membrane and the ossicles to match the higher impedance of the cochlear fluids. Spectral modifications by the outer ear and the impedance matching of the middle ear can be described by transfer functions that are applied to acoustic signals to simulate their effects (Pickles, 2008).

The cochlea (the inner ear) performs the transduction process that translates acoustic pressure waves into electrical neural activations in the auditory nerve. The cochlea comprises a complex structure of three membranes (Reissner's membrane, RM, in Fig. 2.1, tectorial and basilar membrane) and three fluid-filled compartments (scala vestibuli, SV, scala media, SM, and scala tympani, ST) to form a snail-shaped bony tube. Sensory hair cells make up a vital part of the organ of Corti on the basilar membrane and are connected with their stereocilia (hairs) to the tectorial membrane. There are two types of hair cells placed along the basilar membrane in the human cochlea: the inner hair cells, which convert the vibrations of the basilar membrane into electrical activity, and the outer hair cells, that adjust the properties of the mechanical connection between the basilar and the tectorial membrane. The interplay between the continuously changing mechanical properties of the basilar and tectorial membrane with the activation of inner and outer hair cells along the cochlea resembles a crucial mechanism for human hearing – the frequency-to-place mapping that decodes the incoming acoustic wave into its frequency components. Each position along the basilar membrane is tuned to a specific frequency at which it is most excitable (so called *tonotopic* organization, with high frequencies at the base to low frequencies at the apex) and thus acts as one stage of a filter bank that performs a spectral analysis with overlapping frequency bands (Moore, 2012; Plack, 2013).

The inner hair cells are activated by the shearing motion of the basilar membrane in relation to the tectorial membrane. Hereby, the stretching of the stereocilia on top of the hair cells open up ion channels that cause an inflow of electrically charged potassium ions. This leads to a depolarization of the hair cells and the release of chemical neurotransmitter into the synaptic cleft between the hair cell body and the spiral ganglion cell of the auditory nerve. In parallel, the outer hair cells are activated in a similar manner, but instead of transmitting information to the auditory nerve, they change their length with the help of motor proteins (prestin) embedded in their cell membrane. This expansion and reduction of outer hair cells in response to the incoming sound wave leads to a non-linear amplification behaviour that is of vital importance to the auditory system as it enables a dynamic range of more than 100 dB SPL while maintaining high sensitivity to weak sounds (Schnupp et al., 2012).

The dendrites of the spiral ganglion cells (in their entirety with 30,000 cells called *auditory nerve*) detect the release of neurotransmitter and react with the initiation of an electrical potential that travels along the neuron to the cell body. The cell body integrates the incoming information and as soon as its excitation threshold is surpassed, generates an action potential ('firing') that excites the following stages of neural circuits. As each ganglion cell is connected to an inner hair cell at a specific location along the cochlea, the frequency-specific information about the incoming sound is still encoded within the auditory nerve fibers (with inner fibers, at the centre of the auditory nerve, tuned to low and outer fibers tuned to high frequencies). Also without excitatory input, auditory nerve fibers still show spontaneous activity in form of action potentials which is used to classify them into different classes of fibers depending on their spontaneous rate of firing. This is important for the *rate-level coding* of the auditory nerve that translates input signals with higher levels into higher firing rates. The auditory nerve fibers differ in their dynamic range and their sensitivity to the level of the incoming sounds (with high spontaneous rate fibers with a small dynamic range and being sensitive to weak sounds and low spontaneous rate fibers with a larger dynamic range and being sensitive to higher levels of sound). The rate-level code in combination with the tonotopic organization within the cochlea and the auditory nerve results in a (*rate-) place code* by which the incoming sound wave is transmitted. This code conveys frequency information by place and level information by firing rate within the bundle of auditory nerve fibers. The already mentioned shearing motion between the basilar and the tectorial membrane leads to an additional coding mechanism to convey frequency-information in the inner ear. The stereocilia of the inner hair cells open ion channels only in one direction of their movement restricting their firing activity to that direction only and leading to a behaviour called *phase-locking*.

The next stage in the auditory system is the auditory brainstem where the auditory nerve fibers connect to the auditory neurons in the cochlear nucleus. There are several types of neurons in the cochlear nucleus which differ in their firing activity after receiving an excitation stimulus (Pickles, 2008). Some neurons act similarly to the auditory nerves' ganglion cells (primary-like neurons), while others only respond to specific cues of the incoming sound wave (onset neurons) or encode temporal information (chopper neurons). Even though these neurons react in different ways to the input stimulus, they are still organized tonotopically. From here, the auditory pathway splits up into several parallel and sequential stages of auditory nuclei (superior olivary complex, lateral lemniscus, inferior colliculus) before the information is passed to the medial geniculate body in the thalamus as the last stage before the auditory cortex. Within these auditory nuclei, numerous low-level processes take place that extract or analyse more basic characteristics of the input sound (as with the different types of neurons in the cochlear nucleus). For example, certain acoustic properties of the input stimulus such as the input level, temporal and spectral characteristics (e.g. on- and offsets, phase and frequency, bandwidth and spectral complexity) are detected and encoded by different cell types before they reach the cortical regions of the auditory system. Within those stages, information from the contralateral side gets integrated to form the basis of binaural processing of sound. Finally, the auditory cortex receives and analyses the neural projections from the medial geniculate body and generates auditory sensations. Still, a tonotopic organization as initiated in the cochlea can be found in several areas of the auditory cortex.

2.2 Sensorineural hearing loss

There are two main types of hearing loss depending on which stage of the auditory pathway is affected. The first type is a conductive hearing loss that is present when the peripheral stages before the cochlea are not functioning correctly, for example when the ossicles are immobilized or the transmission of pressure waves that enter the ear canal is compromised before the cochlea. The second and most common type of hearing loss affects the inner ear and the auditory nerve, and is called sensorineural or cochlear hearing loss.

Sensorineural hearing loss occurs naturally with the ageing process for people of about 40 years and older (age-related hearing loss or presbycusis). The second most important reason is repeated exposure to loud noises as they occur in some work places, music venues or traffic situations (noise induced hearing loss) or when being exposed to a sudden exceptionally loud noise (acoustic trauma). Furthermore, the genetic constitution and various diseases, such as Ménière's disease, meningitis or an acoustic neuroma (besides others) may result in permanent sensorineural hearing loss. Ototoxic medication may lead to a reversible or permanent impairment of the sensorineural auditory stages or increase their vulnerability to noise exposure. Severe health incidents such as strokes or physical trauma can also impair the functioning of the sensorineural stages of the auditory system.

Physiological changes

The most obvious physiological reason for sensorineural hearing loss is the damage to or the complete loss of hair cells in the cochlea. Especially damage to the outer hair cells, which seem to be more vulnerable than the inner hair cells (Moore, 1995), is a common physiological change that occurs with sensorineural hearing loss. Typically, the outer hair cells' stereocilia are partly damaged or even completely removed (see Figure 2.2 for an example electron micrograph). This leads to an impairment of the active mechanism between the outer hair cells and the basilar membrane and compromises their ability to actively influence the mechanical properties of the membrane structure. In more severe types of hearing loss, when also the inner hair cells loose their function due to damage to their stereocilia, so called *dead regions* occur for segments in the cochlea for which the sensitivity to sound is completely diminished. Besides the inner and outer hair cells, another critical part that may cause malfunction of the inner ear transduction process is the stria vascularis, which is responsible for maintaining the ion composition of the fluids in the organ of Corti (Moore, 2007). Furthermore, the auditory nerves' spiral ganglion neurons (especially those with low spontaneous rates, i.e. high-threshold fibers) degenerate as a consequence of loud noise exposure (Furman et al., 2013) and compromise the synaptic connection between the inner hair cells and the auditory nerve (Kujawa and Liberman, 2015). As a result of the damage to the synaptic connection, the decrease in stimulation leads to a slow degeneration of the affected auditory nerve fibers over time.

Figure moved to Appendix

Figure 2.2 Scanning electron micrographs of the normal (a) and damaged (b) cochlear sensory epithelium. In the normal cochlea, the stereocilia of a single row of inner hair cells (IHCs) and three rows of outer hair cells (OHCs) are present in an orderly array. In the damaged cochlea, hair cells are missing, and stereocilia are abnormal, leading to hearing loss (Ryan, 2000).

Psychophysical changes

Sensorineural hearing loss typically leads to several detrimental effects for the perception of acoustic signals that can be observed using psychophysical measurements. The damage to one or several parts of the transduction process leads to a *loss in sensitivity* to sound, shown by a shift in the hearing thresholds in an audiogram (pure-tone audiometry, PTA) (Pickles, 2008). Usually, sensitivity to higher frequencies is affected before shifts occur in the lower frequency range. Furthermore, the loss of the outer hair cell function may lead to a *loss in frequency selectivity*, with less sharp and wider tuning curves for a specific frequency that can be detected by measuring the physical tuning curves (PTC). This effect leads to more overlapping auditory filters and compromises the ability to distinguish between different frequencies in complex sounds. This pronounces several types of *spectral masking* effects in hearing-impaired subjects compared to normal-hearing subjects (e.g. upward spread of masking - lower frequency components mask higher ones). The broader and less steep shape of tuning curves in combination with damaged auditory nerve fibers leads to an effect called *loudness recruitment*. It describes the dependence of the sensitivity to sound in respect to its loudness. Whereas weak sounds need to be amplified in order to be perceived, loud sounds might be perceived normally or as uncomfortably loud as a consequence of normal or even increased sensitivity to sound at the upper range of loudness (*hyperacusis*). This leads to a smaller range of loudness for comfortable hearing and can be described as a frequency-dependent reduction in dynamic range. The higher the hearing thresholds at a specific frequency, the smaller the dynamic range will be because of the fixed upper limit of comfortable levels. In the healthy auditory system, the frequency selectivity depends on the signal level, with sharper tuning curves at lower frequencies with decreasing level (Moore, 1995). In the impaired system, damage to the active mechanism of the outer hair cells diminishes this level-dependent behaviour, and thus leads to broadened filter shapes at all sensation levels. Temporal masking effects are found to a greater extent with sensorineural hearing loss in comparison to normal hearing (Dillon, 2001). This is a result of *reduced temporal resolution* and adversely affects speech understanding in background noise, as listeners benefit less from short temporal gaps in the background sounds (*listening in the dips*).

Overall, sensorineural hearing loss leads to impaired speech understanding in background noise, and a loss in the ability to benefit from *masking release* in fluctuating background noise (Festen and Plomp, 1990). This is shown in Fig. 2.3: hearing-impaired listeners achieved good (>60%) speech understanding only in positive signal-to-masker ratios for all three interfering noises. In contrast, normal-hearing listeners were able to understand steady-state noise (>60%) at about -5 dB SNR and benefitted from *masking release* in modulated background noises (further 4 and 6 dB improvement for modulated noise and an interfering voice, respectively).

Figure moved to Appendix

Figure 2.3 Average speech discrimination curves for sentences presented in steady-state noise, two-band-modulated noise, and with an interfering voice, for normal-hearing and hearing-impaired listeners (Plomp and Festen, 1990).

The shifts in speech reception thresholds (SRT, the SNR at which 50% of the speech is intelligible) for hearing-impaired listeners range from 2.5 up to 7 dB in stationary, speech-shaped noise and depend on additional factors such as the individual hearing impairment and presentation levels of the stimuli (Moore, 1995). It should be noted, that even a comparably small increase in SRT by about 2.5 dB, as it may occur with mild hearing loss, can lead to problems understanding speech in challenging listening situations (e.g. in a restaurant) (Moore, 2007). This is because the ability to understand speech in noise varies strongly with SNR (1 dB change in SNR may change percentage correct scores between 7 and 19%). For competing voice scenarios and other modulated background noises, SRT values are 9 to 25 dB higher than with normal hearing - a huge difference. Furthermore, people with hearing loss show less benefit from spatial separation between the target speech and noise, which leads to a further disadvantage in terms of SRT by up to 7 dB (Moore, 2007). This means that speech understanding in background noise may be severely affected with sensorineural hearing loss and thus constitutes one of the most important problems facing people with hearing loss in everyday life (Henshaw et al., 2015).

2.3 Characteristics of speech sounds

Humans perceive the acoustic environment as a mixture of overlapping sound components from individual sources: *the auditory scene*. The auditory system is able to separate the individual components that originate from different sound sources into auditory objects and streams. Hereby, the detection and recognition of human speech are among the most important functions of the auditory system to enable communication.

Speech is produced by an air flow from the lungs and the controlled vibration of the vocal folds, which define the pitch of the sound (for voiced speech) (Plack, 2013). The shape and the acoustic properties of the following cavities of the vocal tract are adjusted with different articulators (such as the tongue and lips) and influence the harmonic structure of the speech sound. The change of the harmonics enables the production of different vowels. Other alterations of the air flow such as restriction and release by the articulators, without the vibration of the vocal folds (unvoiced sounds), account for the production of consonants. The basic low-level structure of speech is formed by a sequence of alternating vowels, consonants and pauses. With longer segments of speech, successively more high-level elements such as syllables, words and sentences build up the informational content of speech.

Physical properties

In physical terms, speech can be considered as a wide-band complex signal modulated in time (Rosen, 1992). Speech evolved certain physical characteristics that increase its effectiveness and robustness to environmental and other interfering sounds. The physical properties of speech can be described as acoustic cues with different spectral and temporal features and generally incorporate a high level of redundancy. The acoustic cues and their relative importance may change according to the acoustic environment and informational context of the situation, leading to a multidimensional signal that allows for high levels of distortion before a loss of information occurs (Moore, 2012).

Spectral cues

The long-term spectrum of speech resembles a band-pass filtered signal with a typical bandwidth between 100 Hz and 10000 Hz. Most of the energy is concentrated in the lower frequency bands up to about 1000 Hz. The fundamental frequency (F0) is produced by the vibration of the vocal folds (Plack, 2013) and is an important cue for distinguishing different speakers (different voices) (Loizou, 2013). Other spectral peaks at multiples of F0 define the formants (F1, F2 etc.) of the speech sound. The controlled change of resonances in the vocal tract are used to alter the formants and produce vowel, glide and stop-consonant sounds. The concentration of energy in certain frequency locations makes speech stand out from broadband background sounds (Loizou, 2013). Even though the speech spectrum has a characteristic shape that changes over the audible frequency range, it is highly robust to distortions or missing parts of the spectrum. Several listening experiments have shown, that even when just a small portion of the spectral information of speech is available (through band-pass filtering, e.g. cutting off the spectrum at 1800 Hz (Moore, 2012), high intelligibility scores (about 90%) can still be achieved. Consequently, in addition to spectral cues humans also rely on other cues for speech understanding, such as temporal features.

Temporal cues

Rosen (1992) classified the temporal features of speech into three main categories according to the frequency of their temporal fluctuations: the envelope (ENV), periodicity (PER) and temporal fine structure (TFS) cues. Fluctuation rates from 2 up to 50 Hz are classified as envelope information and used to convey segmental cues (boundaries between speech components), manner of articulation, voicing and vowel identity. ENV has been further defined by its main modulation frequency (average of 4 Hz for the syllable rate of speech) and accounts for several acoustic features such as intensity, duration and rise- and fall-time which correspond to loudness, length and attack/decay perception in psychoacoustic terms. It has been shown in quiet acoustic conditions, that near-to-perfect speech intelligibility can be achieved by using only envelope information in four different frequency bands (Shannon et al., 1995). Hereby, a *vocoder* was used to extract speech ENV information in certain frequency bands (usually between 1 and 16) that is then used to amplitude modulate a carrier signal in the respective frequency band (either a noise or tone based approach has been used). Whereas vocoded speech is highly intelligible in quiet acoustic conditions with just 5 frequency channels and no more than 8 channels being necessary to reach the asymptotic limit of intelligibility (Loizou et al., 1999), this situation changes drastically when background sounds occur (Stickney et al., 2004). With steady background noise and especially with fluctuating background sounds such as interfering talkers, speech intelligibility drops significantly in comparison with unprocessed, natural speech and the number of channels necessary for achieving good speech understanding (>60%) increases to at least 8 channels. Even with a large number of vocoder channels (=24), there was a significant difference of more than 10 dB in SRT in comparison to natural speech (Qin and Oxenham, 2003). This indicates that, especially in noisy listening situations, other temporal features are used to improve the robustness of speech in addition to the information conveyed by spectral and ENV cues. Faster temporal fluctations in the range of 50 to 500 Hz are referred to as periodicity information and are predominantly present during voiced parts of speech (Rosen, 1992). Thus they can be used to distinguish between voiced and unvoiced speech sounds. Furthermore, PER information is used to perceive the pitch of a voice, which is especially important in tonal languages such as chinese and for speech intonation and prosody, but also to distinguish different speakers from each other (with F0 of 80 Hz for low male voices up to 280 Hz for children), for example in multi-talker scenarios (Loizou, 2013). Environmental sounds that occur in nature often present more stochastic and non-harmonic characteristics (e.g. rain, wind), whereas speech and animal vocalizations comprise higher amounts of PER information. It can be speculated that PER cues developed within speech to stand out from such natural environmental sounds.

Rosen (1992) classifies fluctuations in the range from 600 Hz to 10 kHz as temporal fine structure information. TFS describes the change in the pressure wave between single periods of voiced sounds and conveys psychoacoustic correlates of timbre and sound quality. TFS is responsible for segmental cues to separate voiced sounds from each other, place of articulation and voice quality. It gives information for distinguishing words that differ in their frequency spectrum and formant transitions at higher frequencies. TFS information is used to segregate speech in a sound mixture and thus contributes to speech intelligibility (Moore, 2014). In Lorenzi et al. (2006), hearing impaired (HI) listeners achieved good speech understanding (comparable to normal hearing, NH, listeners) with speech that conveyed ENV cues only, but speech intelligibility was greatly reduced with speech that conveyed only TFS cues (NH listeners were not affected by this). For the HI group, intelligibility scores with TFS-only speech was highly correlated to masking release by temporal dips in background noise. Another study by Hopkins et al. (2008) confirmed this problem for HI listeners to make use of TFS information in a competing voice scenario. Thus, reduced sensitivity to TFS due to sensorineural hearing loss or the use of ENV-only speech (for example with a Vocoder or CI speech processor) can lead to deteriorated speech intelligibility, and is likely one of the most prominent reasons that accounts for speech perception difficulties in noisy environments.

2.4 HEARING DEVICES

Hearing devices constitute the main treatment for sensorineural hearing loss due to the lack of pharmaceutical therapies. Depending on the type and degree of hearing loss, different hearing devices are available that include various types of traditional acoustic hearing aids (HA, different forms include behind-the-ear, in-the-canal or open-fit/closed-fit types), bone-anchored hearing aids (BAHAs, transmission of sound via the skull bones), middle ear implants (direct excitation of the ossicles), cochlear implants (CI, electrical stimulation of the auditory nerve) and auditory brainstem implants (ABI, electrical stimulation of the auditory brainstem).

Hearing aids

Hearing aids are by far the most common hearing device and do usually not require a surgery (some types of hearing aids may require minor treatments, such as the placement of the device into the ear canal). HAs are electroacoustic devices that amplify the sound wave reaching the ear of the user. HAs can be specifically tailored to the user's hearing loss during a fitting procedure according to the perceptual requirements by using psychophysical measurements such as PTA.

HAs successfully compensate for various aspects of sensorineural hearing loss, such as loss in audibility (by using amplification) and loudness recruitment (by using non-linear compression) (Dillon, 2001). In quiet listening situations, HAs deliver near-to-normal speech understanding and can recover a large amount of acoustic information for mild up to severe degrees of hearing loss.

State-of-the-art HAs employ digital signal processing to enhance perceptual aspects of the listening experience (Hamacher et al., 2005; Wood and Lutman, 2004), such as improved speech and music perception and reduced intrusiveness of interfering background noises (e.g. wind noise). Numerous digital processing techniques are implemented in the digital signal processor (DSP) of the hearing aid and include directional microphones (beamformer), frequency-analysis filter banks (or fast fourier transform, FFT), frequency-dependent gain and compression (spectral fitting), automatic gain control, automatic feedback cancellation (adaptive filtering), and several other signal enhancement techniques (noise reduction, sound scene adaptation etc.). Current HAs nearly exclusively use digital processing in both time and frequency-domain based schemes. Figure 2.4 shows a simplified schematic of an exemplified HA signal path: the sound wave is picked up by the microphones, passed through several processing stages and finally presented acoustically to the user.

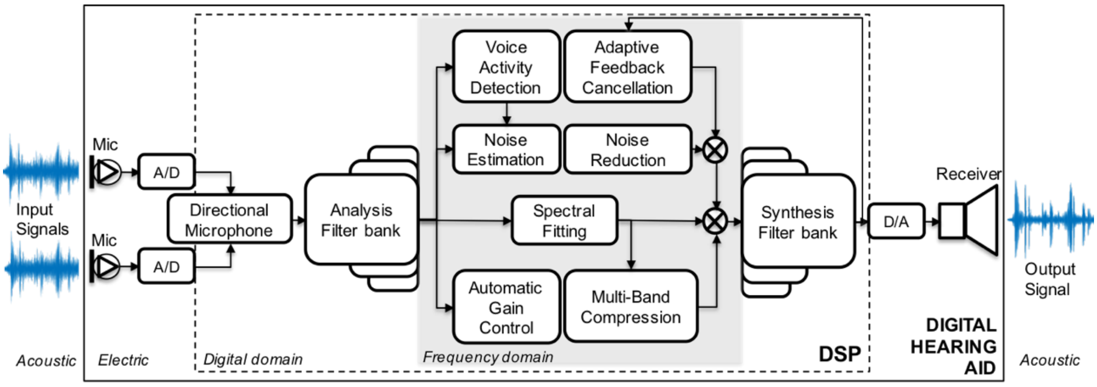


Figure 2.4 Exemplified schematic of a digital hearing aid showing the processing stages from acoustic input sound, electric to digital conversion, several digital signal processing techniques and the transformation back to an acoustic signal for the presentation to the user (figure generated by the author).

While the introduction of digital signal processing has provided significant improvements over analog HAs in several domains: more accurate spectral fitting, higher levels of amplification without acoustic feedback (Dillon, 2001), improved speech understanding in background sounds using compression systems (Moore, 2012) and directional microphones (Magnusson et al., 2012), and potential benefits in speech perception from frequency transposition (Kuk et al., 2009); there are still many acoustic situations where users of HAs do not obtain benefits (Levitt, 2007) and future developments are needed to achieve improvements (Edwards, 2007) in challenging real-world listening situations (Henshaw et al., 2015).

Especially in listening situations with higher levels of background noise, users of HAs struggle to understand speech better - a situation that did not significantly change with the introduction of digital signal processing. Improvements in speech understanding have been reported for approaches using directional microphones in combination with background sounds that were spatially separated from the speech source in laboratory setups (McCreery et al., 2012; Nordrum et al., 2006; Ricketts and Hornsby, 2005). Despite several decades of research and the development of numerous single-microphone noise reduction algorithms, that are designed to work without spatial information, only marginal and mostly not significant improvements have been achieved in terms of speech understanding in interfering background noise so far (Arehart et al., 2003; Bentler et al., 2008; Dahlquist et al., 2005; Killion, 1997; Loizou, 2013; Loizou and Kim, 2011; McCreery et al., 2012; Mueller et al., 2006; Nordrum et al., 2006). However, conventional single-microphone noise reduction approaches implemented in current HAs have been reported to achieve significant improvements in subjective measures of listening effort and tolerance to background noise (Arehart et al., 2003; Brons et al., 2014b; Hu and Loizou, 2007a) and were strongly preferred over unprocessed sounds in terms of perceived sound quality (Bentler et al., 2008; Boymans and Dreschler, 2000; Loizou, 2013; Loizou and Kim, 2011; Mueller et al., 2006; Nordrum et al., 2006; Ricketts and Hornsby, 2005).

Cochlear implants

For severe to profound sensorineural hearing loss, auditory prostheses such as cochlear implants (CIs) can be used to partially restore the sensation of hearing by electrical stimulation of the auditory nerve's ganglion cells in the cochlea. Thus even with a complete loss of hair cells, CIs can deliver sound to the user, given that the auditory nerve is at least partially intact. Since CI development started in the late 1950s and the first multi-channel device was approved in 1985, hundreds of thousands of people have received a CI until now (worldwide approx. 324.000 in 2012, NIDCD). CI devices consist of an external (called *speech processor*) and internal part (the implanted parts: an electro-magnetic receiver in the skull bone and the stimulation electrode in the cochlea). Similar to a HA, the *speech processer* picks up sound waves with one or more microphones, performs several processing stages with a DSP and transmits the signal along with the required electro-magnetic energy to the internal part for electrical stimulation of the auditory nerve (Zeng et al., 2008). Whereas the first part of CI processing is similar to HA processing (as shown in Fig. 2.5: directional microphone, spectral analysis, automatic gain control and noise reduction), the second part is different from a HA as it translates the acoustic information into an electrode output pattern (electric pulse train with a few hundreds of pulses per second) that is used to stimulate the auditory nerve directly. Thus, instead of the synthesis of an acoustic signal, a mapping procedure is performed to match the spectral energy of the signal to the electrode channels and dynamic range of the user.

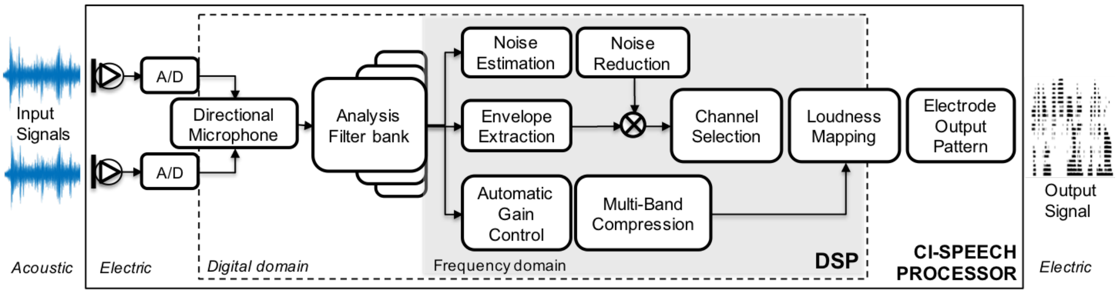


Figure 2.5 Exemplified schematic of a cochlear implant speech processor showing the processing stages from acoustic input sound, electric to digital conversion, several digital signal processing algorithms and the conversion to an electrode stimulation pattern for presentation to the user (figure generated by the author).

CIs present mainly ENV cues to the user (PER or TFS information is mostly discarded, or conveyed via spectral cues only, due to *place coding*), and can deliver good speech understanding in quiet acoustic conditions (Fetterman and Domico, 2002; Zeng et al., 2008). CI technology is a remarkable success story that enables deaf people to regain the ability to hear speech to communicate with others or enjoy and perform musical activities. However, because of the limited amount of spectral (usually the number of independent spectral channels is about 6-12; Friesen et al., 2001) and temporal information (no PER and TFS cues) conveyed by a CI, the speech-in-noise performance of CI users is dramatically reduced in comparison to normal-hearing listeners. It has been shown that CI users require a 10 to 25 dB higher SNR than NH listeners to achieve the same level of speech understanding

(Spriet et al., 2007; Wouters and Vanden Berghe, 2001). Especially in fluctuating background noise with high amounts of spectro-temporal modulations, CI users do not show benefits from masking-release (Cullington and Zeng, 2008; Nelson et al., 2003; Stickney et al., 2004; Zeng et al., 2008). This can be explained by the largely reduced frequency selectivity and temporal resolution of CI processing, in combination with the electric current spread in the cochlea - the so called *electrode-nerve bottleneck*. The structure of the cochlea with its ion-fluid filled compartments, interconnected with tissue, facilitates an unfocussed spread of stimulation current, and further reduces the spectral and temporal details that can be perceived by the user. Another disadvantage of CIs is the insertion depth of the electrode. The snail-shaped cochlea is most sensitive to low frequency sounds at the basal end which is not reachable with the electrode due to its stiffness and mechanical properties. This leads to reduced low-frequency sound perception and affects the perception of the fundamental frequency and first formants in the case of speech sounds (Schnupp et al., 2012). Again, this reduction in low-frequency components of speech sounds may not impair speech understanding in quiet acoustic conditions, but it can have much more dramatic effects on speech comprehension in background noise, when the fundamental frequency serves as an important cue to pick out a speaker from interfering background sounds. For users that have residual hearing at low frequencies (<1000 Hz), a combination of HA and CI technology can be applied (electro-acoustic stimulation, EAS) and may provide large improvements up to 10 dB in SNR for speech perception in noise (Turner et al., 2004; Zeng et al., 2008).

In contrast to HA users, improvements in understanding speech in background noise have been reported for CI users with several single-microphone noise reduction algorithms based on noise estimation techniques (Dawson et al., 2011; Hu and Loizou, 2007b; Mauger et al., 2012; Verschuur et al., 2006; Yang and Fu, 2005; Ye et al., 2013). One reason for this success could be the better performance (reduced estimation errors) of the algorithms in positive SNR conditions (>= 0 dB SNR), where usually CI users are being tested, in comparison to hearing-impaired or normal-hearing listeners, that are usually being tested at smaller or even negative SNRs. Most of the studies reported improvements in stationary background noise (so called speech-weighted noise, SWN), but no improvements were found, or they were much reduced in magnitude, in fluctuating background noises such as multi-talker babble noise. Two main reasons may account for the decrease in performance in fluctuating background noise: firstly, the algorithms produce more estimation errors in non-stationary conditions leading to artefacts and speech distortions that may impair the intelligibility of speech (Yang and Fu, 2005); secondly, CI users do not benefit from release from masking (*listening in the dips*) as a result of reduced spectral resolution and lack of temporal fine structure information and perform worse in fluctuating background noise than in stationary conditions also when using noise reduction algorithms (Mauger et al., 2012a). Mauger *et al.* (2012a; 2012b) showed that CI users generally preferred stronger noise reduction settings than usually applied for normal hearing listeners, suggesting that CI users might be more resistant to speech removal distortions (type-II errors) and less resistant to noise addition errors (type-I) (also reported by Qazi et al., 2013). Using a more aggressive noise reduction setting for CI users, maximum benefits of about 2 dB in SRT were found in stationary noise, but again the benefit was much reduced when the interfering noise was non-stationary, as in the case of competing talkers (Dawson et al., 2011; Mauger et al., 2012a). The large difference between NH and CI performance in speech understanding in background noise requires further developments in noise reduction techniques with the goal to close this gap in future.

Both classes of hearing devices, HAs and CIs, are restricted by several technical requirements that limit the development of more sophisticated, complex algorithms (Schweitzer, 1997). These mobile devices need to run continuously over several hours (>12 hours) to enable the user to take part in full-day activities. This restricts the computational complexity of the algorithms that are executed by the DSP (usually a highly efficient and optimized application specific integrated circuit, ASIC, is used). Another constraint arises from the small size of the devices (restricting the space that can be used for technical parts such as the battery, processors and microphones). With regards to the sound perception of the user, the HA/CI device has to process and output the signal within a few milliseconds of delay to ensure that no disrupting echoes or disturbing artefacts occur (Stone and Moore, 1999). Taken together these requirements constitute a big challenge for the development of noise reduction algorithms that have to work in real time, be computationally and energy efficient and satisfy the perceptual requirements of the users. In addition to the difficult task *per se* - to distinguish speech sounds from background sounds - a potential algorithm has to satisfy all of these requirements for succesful application in HA and CI devices.

2.5 THE INTELLIGIBILITY OF SPEECH IN NOISE

The intelligibility of speech is crucially dependant on the audibility of the signal. If parts of the speech spectrum are below the detection threshold or masked by a background noise, speech intelligibility is decreased (Moore, 2007). Traditionally, the aspect of audibility has been investigated with the *Articulation Index* measure (AI, French and Steinberg, 1947), which is nowadays incorporated in the *speech intelligibility index* (SII) metric. The basic concept of the AI or SII metric is to sum up the SNRs in individual spectral subbands of the signal (called *local* SNR, in contrast to the *global* SNR) to form an overall score of intelligibility between 0 and 1, where each subband is weighted according to its contribution to speech intelligibility (determined also by auditory masking properties of the auditory system). A frequency range between 200 and 9000 Hz is used, as it contributes the most to speech intelligibility, and the weighting factors applied to the individual subbands are adjusted according to the speech material. Speech is assumed to cover a dynamic range of 30 dB and the intelligibility contribution from individual bands is mapped to a score between 0 and 1 (with -15 dB or lower than the root mean square, RMS, of the speech, yielding a score of 0 and +15 dB and above yielding a score of 1). The AI/SII metric was adapted to predict intelligibility for fluctuating background noise, by calculating the scores over short-time periods to account for the change in local SNRs over time (Rhebergen and Versfeld, 2005). The SNR- or audibility-based concept of the AI/SII metric is of great importance for the speech processing approach followed in this thesis. However, for predicting the intelligibility of nonlinearly processed signals, that include nonlinear distortions and distortions in the time domain (e.g. for signals that have been processed with a noise reduction algorithm or contain reverberation), AI/SII based metrics fail to predict the outcome in speech intelligibility obtained by human listeners (Goldsworthy and Greenberg, 2004).

The limitation of the AI/SII based metrics, that operate mainly in the spectral domain, have been addressed with the development of the *speech transmission index* (STI, Steeneken and Houtgast, 1980). The STI is based on a test signal that is passed through the system under test (e.g. a sound production system, or a speech processing algorithm) to detect the amount of nonlinear distortions and the reduction in modulation magnitude using a *modulation transfer function* (MTF). Similar to the AI approach, the contributions of individual frequency bands are summed up to form the final STI score (based on individual transmission indices per frequency bands). The STI metric suffers from reduced validity because an artificial test signal (MTF) is used and variations have been proposed to employ speech signals as test signals (Goldsworthy and Greenberg, 2004). Another variation of the STI metric calculates the covariance between input and processed signals and is called the *normalized covariance metric* (NCM, Holube and Kollmeier, 1996). This technique computes the Hilbert envelope in each frequency band, which was downsampled to 25 Hz and filtered to contain modulations in the range of 0 up to 12.5 Hz, for both the unprocessed and enhanced signals and calculates the normalized covariance between the two. The result is then bandlimited to a dynamic range of 30 dB and mapped to a local SNR estimate for each frequency band before summing up the weighted values to form a global score between 0 and 1 (exact description provided by Ma et al., 2009), similar to the AI approach. Another intelligibility prediction model is called *short-time objective intelligibility* measure (STOI) that calculates the correlation coefficient between the temporal envelopes of the two signals in short-time regions (384-ms long segments) and was proposed for the evaluation of time-frequency weighted and nonlinearly processed speech signals (Taal et al., 2011).

Most of these metrics have been developed for the prediction of intelligibility outcomes for normal hearing listeners. When applied to hearing-impaired listeners, or even CI users, the accuracy of the predictions decreases and their variability increases (Chen and Loizou, 2011; Holube and Kollmeier, 1996). Suprathreshold effects, that occur with sensorineural hearing loss, are thought to be responsible for the variability in speech perception of hearing impaired listeners in addition to the effects explained by audibility (Moore, 2007). With mild degrees of hearing loss, models based on audibility predict more or less accurately the data obtained with human listeners, however with higher degrees of hearing loss, models based on audibility (AI/SII) are not sufficient anymore. For CI listeners, the NCM measure has proven to be more useful and yielded high correlation scores to CI users' speech perception in comparison with other measures (Chen and Loizou, 2011; Santos et al., 2013). In this case, the processing applied by the NCM measure (or other STI based measures) is relatively similar in its working principle to the processing of a CI (reduced number of overlapping frequency channels and the computation of envelope information, while discarding temporal fine structure information; Goldsworthy and Greenberg, 2004).

In addition to the instrumental measures for objective prediction of speech intelligibility in noise obtained with human listeners, listening tests using human listeners are still of importance, especially when evaluating nonlinear processing techniques or testing whether potential benefits occur for hearing-impaired listeners. Two main approaches are used to compare speech processing systems with human listeners: firstly, methods evaluating the percentage of intelligible speech for the unprocessed and enhanced signals at a *fixed* SNR condition to control the amount of masking effects and evaluate the algorithms under test at comparable acoustic conditions; and secondly, methods using *adaptive* techniques to determine a specific *speech reception threshold* for each listener (the SNR where 50% of the speech is intelligible) and to compare the algorithms under test at comparable performance levels in the perceptual space of the listeners. For listener groups with large variabilites in the individual performances in speech-in-noise understanding (e.g. CI users), the adaptive procedure provides a better tool to test the benefit of a processing scheme, as it avoids ceiling effects due to the variability in performance between the listeners.

How to increase the intelligibility of speech in noise?

In relation to the fundamental concept of the AI, that the weighted sum of local SNRs over all frequency channels determines a "total intelligibility" of the speech signal, the concept of the ideal Wiener filter (IWF) has been developed to obtain an enhanced signal in terms of speech intelligibility

(Lim and Oppenheim, 1979). The IWF defines an "optimum" filter that is applied to the noise-corrupted speech in the temporal or spectral domain to increase its intelligibility (see Figure 2.6 for a visualisation in the cochleagram domain). The ideal Wiener filter M_{IWF} formula in the frequency domain is derived mathematically based on the assumption that the speech S and noise N components are uncorrelated, stationary random processes that have been added together:

$$M_{IWF}(t,f) = \frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} = \frac{\xi(t,f)}{1 + \xi(t,f)}$$

with t denoting the time frame and f denoting the frequency channel in a time-frequency representation of the signals (obtained with a FFT- or filter bank analysis). The ideal Wiener filter is computed with *a priori* information about the speech and noise signals and applies a spectral weight to the spectral magnitudes of the noise-corrupted speech signal in each time frame to estimate the magnitudes of the clean speech signal. This computation achieves a near to perfect intelligibility of the enhanced speech signal, despite the fact that the spectral phase information is still noise-corrupted (Madhu et al., 2013). The basic paradigm of the ideal Wiener filter attenuates the frequency channels that are dominated by noise components as determined by a low local SNR and to retain the frequency channels with a high local SNR containing important speech information. This principle has also been called *ideal ratio mask* (IRM), because a ratio between the speech and noise-corrupted speech (values between 0 and 1, based on the local SNR) defines the spectral gain values (over time the filter can be viewed as a *mask* that is applied to the noisy spectrogram to cut out the speech).

An adapted version of the IWF/IRM approach has been developed in the field of *computational auditory scene analysis* (CASA; Brown and Cooke, 1994) that goes a step further and applies an *ideal binary mask* (IBM, values of either 0 or 1). The values of the IBM gain function M_{IBM} are determined by a thresholding for the local SNR (*local criterion*, LC) in each frequency channel:

$$M_{IBM}(t,f) = \begin{cases} 1, & if\ \xi(t,f) > LC \\ 0, & otherwise \end{cases}$$

The *LC* value is usually fixed to 0 dB SNR, but sometimes a relative *LC* is chosen according to the global SNR of the mixture between speech and noise signals (e.g. *LC* = -5 dB for mixtures at a global SNR of 0 dB). The benefit of the IBM in terms of speech intelligibility depends on the choice of an appropriate *LC*, but it has been shown that the benefit is relatively robust to changes in *LC* values (at least over a range of up to 10 dB, Kjems et al., 2009). Even though the gain values of the IBM are of binary form, which may negatively affect the perceived quality of the speech signal (Brons et al., 2012; Madhu et al., 2013), large increases in speech intelligibility have been reported for normal hearing listeners (Brons et al., 2012; Brungart et al., 2006; Li and Loizou, 2008; Madhu et al., 2013), hearing impaired listeners (Anzalone et al., 2006; Wang et al., 2009) and CI users (Hu and Loizou, 2008; Koning et al., 2015; Mauger et al., 2012b; Qazi et al., 2013).

The example of the IBM indicates that the global SNR is an important factor for speech intelligibility. For IBM-processed speech, the SNR in each frequency channel and time frame is not altered (only gains of 0 or 1 are applied), but the global SNR is improved after the summation of the local SNRs of all frequency channels since frequency channels below the SNR threshold (*LC*) have been discarded. This leads back to the calculation of the AI that is based on a sum of local SNRs and supports the assumption that the audibility of the speech signal is the main factor for intelligibility. A slightly different approach was proposed by Anzalone et al. (2006), that retains the energy of frequency channels with high speech energy (based on an analysis of the target speech alone) and regardless of the SNR in that particular frequency channel. Anzalone et al. reported large improvements in the intelligibility for speech in noise by hearing impaired listeners. Kjems et al. (2009) followed up this approach and showed that the pattern of the underlying target speech signal is the determining factor instead of the local SNR in individual frequency channels. Kjems et al. propose a *target binary mask* (TBM) that was solely based on the energy of the target speech signal. The application of speech-energy or SNR-based masking paradigms is regarded as the most successful way to increase the intelligibility of noise-corrupted speech signals, at least for the ideal case when *a priori* information about the speech signal is available.

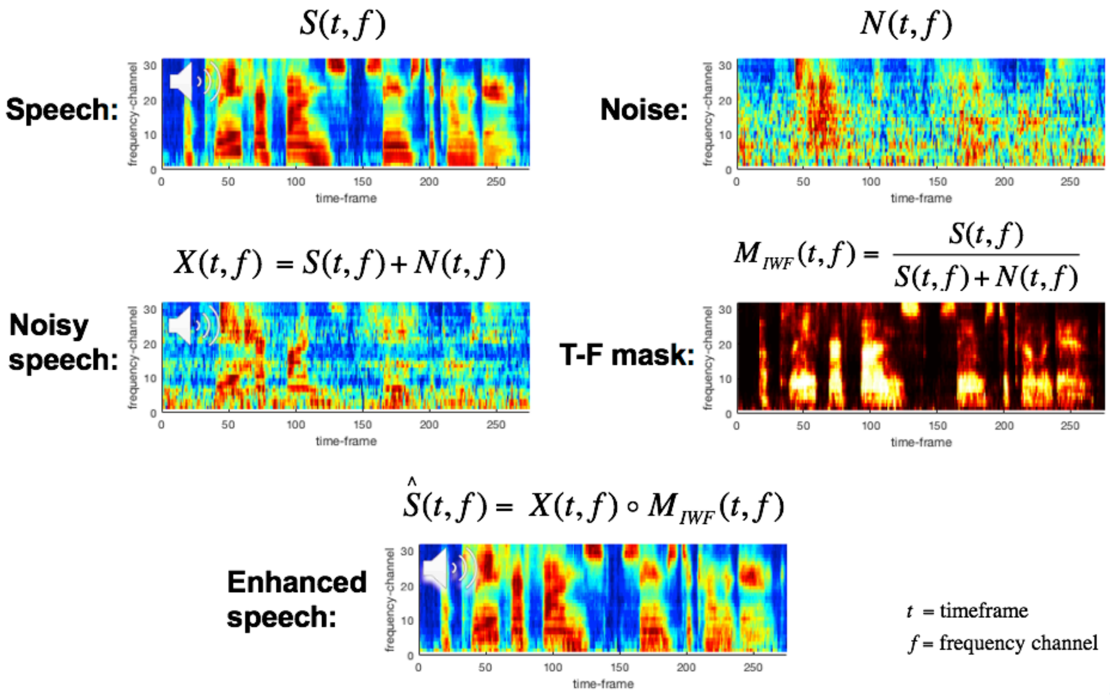


Figure 2.6 Comparison of cochleagram plots of a noise-free speech signal (top left), noise-corrupted speech signal (middle left, noise signal: top right), ideal filter mask (IRM, middle right) and enhanced speech signal after application of the filter (bottom) (generated by the author).

Alternative ways for increasing speech intelligibility in noise have been proposed that include spectral phase manipulation techniques (Krawczyk and Gerkmann, 2014) or spectral change enhancement of speech signals (Chen et al., 2013; Koning and Wouters, 2012). These methods indicated improvements in intelligibility but to a smaller degree than for SNR-based enhancement techniques. The success of SNR-based enhancement techniques in the ideal case, such as the IWF or IBM approaches, motivates for the development of speech enhancement algorithms that try to estimate these target functions and are introduced in the next chapter.

2.6 SPEECH ENHANCEMENT ALGORITHMS

Speech enhancement (SE) algorithms try to alleviate the detrimental effects of background noise and other distortions to speech signals (e.g. reverberation). SE algorithms are designed with the goal to improve the intelligibility and perceived quality of speech (Loizou, 2013) and can be classified into single-microphone (also called *single-channel*) and multi-microphone (*multi-channel*) algorithms, depending on how many input signals they process and whether they make use of spatial information (in the case of multi-microphone algorithms). Multi-microphone algorithms, also called *directional microphone* or *beamformer*, have shown significant improvements in the intelligibility of speech in noise for conditions when the speech signal was spatially separated from the noise (McCreery et al., 2012; Nordrum et al., 2006; Ricketts and Hornsby, 2005). Multi-microphone algorithms do provide benefits in situations where the listener faces the speaker directly, but do not improve speech perception in situations where only one microphone signal is available, the speech is not spatially separated from the noise or the attended speech signal is coming from a different direction than the direction selected by the directional algorithm (in situations when listening to a person that can not be directly faced, e.g. when sitting in a car and talking with a rear passenger). In all of these situations, single-microphone algorithms can still be applied, because they do not rely on spatial information, and are still of high importance for hearing devices (Hamacher et al., 2005). In practice, state-of-the-art hearing devices employ a combination of both multi- and single-microphone algorithms, to improve the perception of speech in adverse listening situations and it has been reported that this combination improves performance further (Dillon, 2001).

Single-microphone noise reduction

Several studies have shown that the perceptual quality of speech sounds can be improved with single-microphone noise reduction algorithms (Brons et al., 2014; Hu and Loizou, 2007; Loizou, 2013; Ricketts and Hornsby, 2005). However, the same studies have also shown that the intelligibility of speech in noise was not improved by the processing with the SE algorithms. This is still the case, even after several decades of research and many SE algorithms that have been developed over the years (Loizou, 2013). A comprehensive overview of conventional single-microphone SE algorithms is given by Loizou (2013) and covers traditional techniques such as spectral subtraction (Boll, 1979), Wiener filtering (Lim and Oppenheim, 1979), statistical-model based (Ephraim and Malah, 1984) and subspace approaches (Schmidt, 1986). In the next sections, the basic working principle of both traditional and modern approaches to single-microphone noise reduction algorithms will be introduced, with the focus on "rule-based" approaches using Wiener filtering and adaptive noise estimation techniques and "data-based" approaches using acoustic feature extraction methods together with machine learning techniques. These two approaches were compared in several listening experiments with various listener groups in the main part of this thesis.

Traditional approach: "rule-based"

Traditional approaches are based on assumptions about the statistical properties of speech and noise signals. The subclasses of Spectral Subtraction (SS), Wiener Filtering (WF) and statistical-model based (MMSE) SE algorithms have several basic principles in common, most importantly, the implementation in the spectral domain using a fast-fourier transform (FFT) analysis and synthesis scheme (using the overlap-and-add technique). This scheme is used to analyze the spectral energy of the signal, to calculate and apply a spectral gain function for each spectral component and to synthesize an acoustic signal with the inverse FFT for the presentation to the user (at least for hearing aids, for cochlear implants the synthesis stage is omitted as described in the previous chapter).

Generally, single-microphone noise reduction algorithms use the noisy phase information for the synthesis of the acoustic output signal, firstly because of its simplicity and efficiency and secondly because of the lack of a better alternative (even though there are recent results that provide evidence against the traditional method, Krawczyk and Gerkmann, 2014). This approach introduces noise characteristics to the enhanced output signal via the noisy phase information and is non-optimal in terms of the perceived quality of the enhanced speech (yielding a "rougher" sound quality the lower the SNR). However, detrimental effects on intelligibility are negligible in this case, which is supported by studies using noise reduction systems in the ideal case that have been shown to achieve perfect speech understanding in negative SNR conditions with the noise-corrupted phase information (Madhu et al., 2013).

The main challenge of SE algorithms (being of type SS, WF or MMSE) is the estimation of the noise variance and *a priori* SNR that is used to calculate the spectral gain function. In the mathematical derivation of the spectral gain function, these parameters are assumed to be known, but in practice these have to be estimated from the noise-corrupted speech signal. The simplest solution for this uses a voice activity detection (VAD) algorithm to classify time segments of the noisy speech into either speech-active or speech-absent periods and then calculates a running average of the spectral magnitudes during speech-absent periods to estimate the spectrum of the background noise (Loizou, 2013). This approach has been shown to work in stationary background noise, under the assumption that the noise spectrum does not change over short time periods and that the VAD does not falsely detect speech-active periods as speech-absent periods. In non-stationary noise types, this approach is likely to fail because the noise spectrum that is estimated during speech-absent frames changes during speech-active periods leading to estimation errors and artefacts such as speech distortions and underestimation of the background noise spectrum.

In realistic environments with fluctuating background noise, the noise spectrum changes rapidly over time and has to be estimated in a continuous manner also during speech-active periods. Several noise estimation algorithms have been proposed to solve this problem and make use of different tracking principles to obtain an estimate of the background noise, with the most important being minima tracking and recursive averaging through time in individual spectral channels (Cohen and Berdugo, 2002; Gerkmann and Hendriks, 2012; Martin, 2001; Rangachari and Loizou, 2006). These algorithms yield continuous estimates of the spectral energy of the background noise and can be used to calculate the *a priori* SNR. The calculation of the *a priori* SNR is often performed in the way described by Ephraim and Malah (1985) (and called *decision directed approach*) that uses the estimated noise spectrum, the noise-corrupted speech signal and the enhanced output signal of the previous time frame to estimate the *a priori* SNR.

Single-microphone noise reduction algorithms that made use of these advanced noise estimation techniques achieved good results in terms of the perceived quality of speech in noise in combination with different (SS-, WF- and MMSE-based) noise reduction techniques for stationary background noise and constitute the *state-of-the-art* for current hearing devices in terms of single-microphone SE approaches (used for example in Hu and Loizou, 2007; Dawson et al., 2011; Mauger et al., 2012b; Harlander et al., 2012). As described earlier, improvements in speech intelligibility were found for stationary background noise for CI users but not for normal-hearing and hearing-impaired listeners and for non-stationary background noise. These approaches successfully reduce the level of the background noise which leads to reduced listening effort and improved perceived quality of the speech but comes at the cost of introducing distortions to the speech that prevent improvements in terms of intelligibility (Loizou and Kim, 2011). Figure 2.7 shows a comparison between several single-microphone noise reduction algorithms that were evaluated for two SNR levels (0 and 5 dB) and four background noises (Babble, Street, Car, Train) containing different levels of modulations in terms of percentage correct scores by 40 normal-hearing listeners (Hu and Loizou, 2007). The data shows that the algorithms fail to improve the intelligibility for speech in noise relative to the reference signal ("noisy" speech). Instead, decreases in speech intelligibility occured relative to the reference signal.

Figure moved to Appendix

Figure **2.7** Comparison between the noise-corrupted speech signal condition ("noisy") and eight "rule-based" single-microphone noise reduction algorithms in terms of percentage of correctly identified key words for four noise types (Babble, Street, Car and Train) and two SNRs (0 and 5 dB). The data are group scores of 40 normal-hearing, native American listeners (Hu and Loizou, 2007).

This disappointing outcome for normal hearing listeners does not necessarily mean that hearing impaired listeners or users of cochlear implants do not benefit from the processing with the algorithms (indeed, benefits for CI users have been shown by Hu et al., 2007; Dawson et al., 2011; Mauger et al., 2012b), but it indicates that there seems to be a fundamental problem with "rule-based" algorithms that prevents improvements in speech intelligibility in more realistic background noises. What could be the reason for this failure?

- Several "hand-crafted" assumptions about the statistical properties of the target speech signal and the background noise are used for the processing steps of rule-based SE algorithms: voice activity detection, noise estimation and SNR estimation. For example, one principal assumption is that the speech is non-stationary and that the background sound is stationary and that both signals are uncorrelated. This assumption may lead to poor performance when evaluating the algorithms in more realistic non-stationary background noise with speech-like characteristics. Noise estimation algorithms based on minima tracking or temporal smoothing are usually not able to follow fast changes of the spectrum of a fluctuating background noise and introduce estimation errors in terms of underestimation of the noise spectrum. This leads to an overestimation of the SNR in negative SNR regions that has been shown to harm speech intelligibility (Chen and Loizou, 2012).
- Conventional SE algorithms analyze the noise-corrupted input signal in the spectral domain. The estimation of signal properties is based on a narrowband analysis in individual frequency channels and does not integrate information across spectral channels (broadband analysis). This integration across spectral channels is considered as highly beneficial (speech conveys information via several channels to increase its robustness to background noise) and one of the reasons why human listeners perform so well in background noise (Moore, 2012). Human listeners exploit regions with high SNR in the spectro-temporal domain to estimate information about speech components that are hidden by noise energy in regions with low SNR by integrating information across different frequency channels (*listening in the dips*, or *glimpsing*). The narrowband processing paradigm of rule-based algorithms fails to accomplish this.
- SE algorithms for application in hearing devices are designed to work in real time and thus restricted to use current and past information (causal processing). The algorithm has to calculate a spectral gain nearly instantly (within a few ms) to increase intelligibility in background noise (15 ms or less as reported by Anzalone et al., 2006). Typically, rule-based algorithms incorporate more slow-acting estimation techniques because of the temporal smoothing process in the calculation of the spectral estimate of the background noise and thus are to slow to improve speech intelligibility on a frame by frame basis.

Modern approach: "data-based"

A recent approach to speech enhancement tries to overcome the problems of traditional “rule-based” SE algorithms by using machine learning techniques and the framework for feature-based speech segregation proposed by *Computational Auditory Scene Analysis* (CASA, Brown and Cooke, 1994). The algorithms are trained in a supervised manner with exemplar data. When applied to SNR estimation (or T-F mask estimation), a classification (in case of IBM estimation) or regression (in case of IRM estimation) algorithm is used to predict the target mask from features of the noise-corrupted speech signal. The noise-corrupted signal is firstly transformed into a feature space by using one or more acoustic feature extraction algorithms. This is motivated by the human auditory system, that performs a spectral analysis of the sound within the cochlea and maintains this tonotopic representation throughout the auditory pathway, while successively extracting higher-level features from the signal in higher stages of the auditory system (as introduced earlier in this chapter). The features extracted from short-time segments (frames) of the noise-corrupted signal are presented to a machine learning algorithm such as a gaussian mixture model (GMM) for classification or an artificial neural network (NN) for regression. A corresponding target function (e.g. IBM, IRM) is generated for each frame during the training stage, using *a priori* information about the speech and noise components from the training data. The error between the estimated target values from the algorithm and the ideal target function is used to iteratively adapt the parameters of the classification or regression system by using a gradient-based learning scheme to minimize the error. Large amounts of exemplar data are employed for the offline training stage (from several minutes up to hundreds of hours of audio recordings). Once the algorithm has reached a certain level of performance during the training stage (example criteria for stopping the training procedure: error below a certain value, number of training epochs reached or cross-validation), it can be implemented in a real-time algorithm, using the final parameters of the training stage. An exemplified schematic of a machine learning-based speech enhancement framework is shown in Fig. 2.8 and Fig. 2.9, outlining the training stage and application for real-time processing, respectively. This framework can be adapted in a flexible way: the choice of the acoustic features, the estimation algorithm and the target function can be changed to specific use cases (e.g. noise reduction, speech detection, source separation), target user groups (e.g. HA or CI users) or other tasks (for example: VAD, environmental sound detection, music or speaker recognition systems). The main parts of this framework, as shown in Fig. 2.8 and 2.9, are introduced in the following paragraphs.

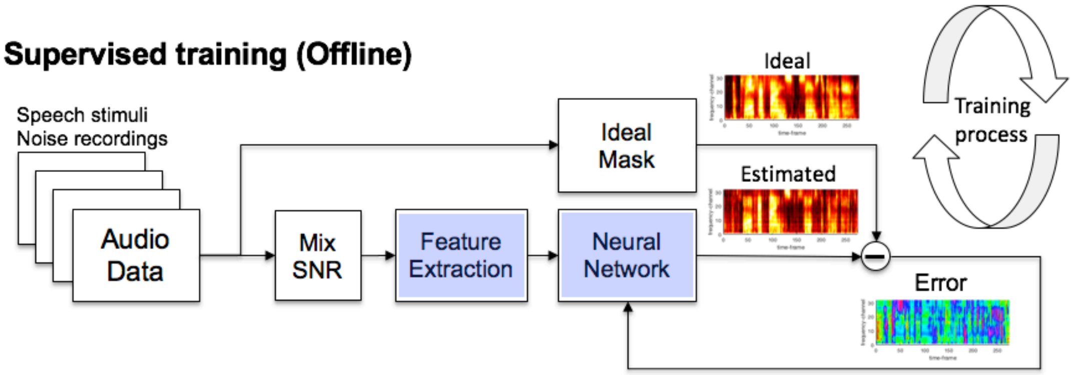


Figure 2.8 Schematic overview of the *offline training stage* for supervised speech segregation. *Audio Data* (speech and noise recordings) are *mixed* at a specific SNR and processed through a *Feature Extraction* stage and the *Neural Network* (or a different machine learning algorithm) and adjusted according to the *Error* between the *Ideal* and *Estimated* masks during the *Training process*. The parts that will be used in the application for Speech Enhancement are indicated in blue (generated by the author).

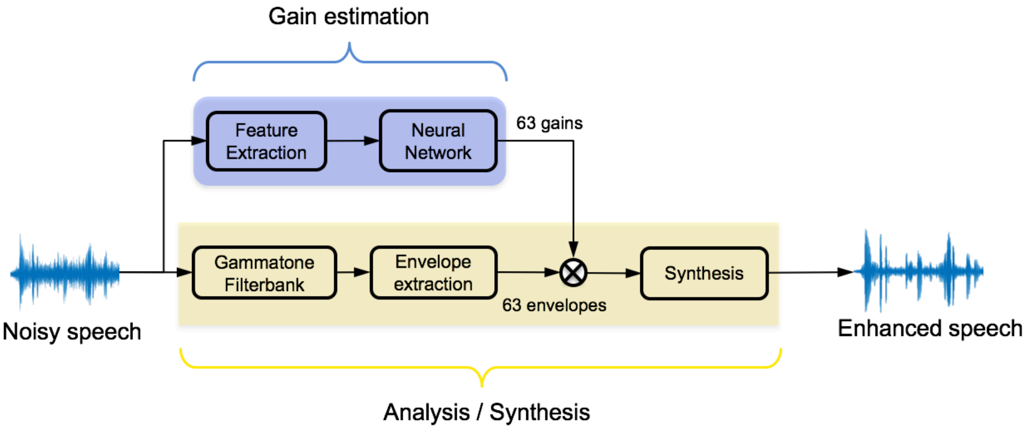


Figure 2.9 Schematic overview of the online application stage for supervised speech segregation. Noise-corrupted speech signals are processed through the *Feature Extraction* stage and the *Neural Network* (or a different machine learning algorithm) and the estimated *gains* are applied to the *envelopes* of the noisy input signal before it is *synthesized* to obtain the *enhanced speech* at the output (generated by the author).

Audio data for training and testing of the algorithm

The first step is to define and generate appropriate training and testing data for the algorithm. The "data-based" nature of the algorithm makes this a crucial step for the final performance of the algorithm and defines the acoustic conditions it can be applied to. The characteristics of the acoustic data are mainly defined by the type and amount of the speech and noise material, together with the choice of mixing ratios (global SNRs) for the noise-corrupted speech.

The speech material is usually taken from one or more research databases for speech that contain well-defined, phonetically-balanced and acoustically-calibrated speech recordings by one or more speakers of different nationalities, genders and ages. There are different types of speech recordings depending on the speech corpus, comprising more easy or difficult material (word rate, articulation, choice of vocabulary, length of utterance), different speech elements such as syllables, words or sentences (with contextual information or without, different number of keywords) and designed for specific target listeners (to account for age and different cognitive and hearing abilities). When evaluating SE algorithms, it is recommended to employ standard speech corpora, that are publicly available to ensure that the obtained results are reproducible, interpretable and usable for comparisons by other researchers (especially when an objective evaluation with prediction algorithms is performed). The most frequently used speech corpora in the field of SE algorithms are the IEEE corpus (*Harvard sentences*, contextual and non-contextual sentences with higher difficulty, containing 5 keywords, spoken by a male speaker, English; IEEE, 1969) and BKB corpus (easier sentences with 3 keywords spoken by a male speaker, developed for children and often used with hearing-impaired listeners, English, Bench et al., 2009), the TIMIT and GRID corpus (TIMIT with more variability, GRID uses a smaller number of words, both with multiple talkers, male and female, English; Garofolo *et al.*, 1993; Cooke *et al.*, 2006) and the LIST corpus (male or female speaker, Dutch, easier sentences with 2-7 keywords and developed for CI users; Jansen et al., 2014).

The same principles apply for the noise material and it is recommended to use well-defined or publicly available recordings. The number of noise types employed for the training stage of the algorithm constitutes to a large extent in which noise conditions the algorithm will perform well. The length and characteristics of the noise recording (spectral shape, stationary or non-stationary with different amounts of spectro-temporal modulations, number of talkers, or how it was generated/recorded) have to be chosen accordingly. Typical noise databases include the NOISEX-92 (Varga and Steeneken, 1992), NOIZEUS (Loizou, 2013) and the recordings from Auditec (Auditec Inc., St. Louis, 1997). Artificial noises that are applied as noise conditions include the speech-weighted noise (SWN or *speech-shaped noise*, SSN; white noise with the spectral shape of the long-term spectrum of the speech material used for training/testing) and multi-talker babble noise (number of both male and female talkers at equal levels, mixed together). Finally, the values of the SNRs employed for the training have to be chosen according to the range of acoustic conditions that the algorithm will be tested in. Generalization performance to novel acoustic conditions can be measured by using testing data different from data used during the training stage of the algorithm in terms of the acoustic "degrees of freedom" of the data (e.g. speech material, sentences, speaker, noise type and recording, SNR).

Feature extraction

Instead of using directly the raw waveform of the input sound, acoustic features are extracted from the input data to obtain a more suitable representation for detecting the speech and noise characteristics within the noisy mixture. The simplest acoustic feature is the output of an FFT analysis that transforms the signal into the spectral domain. Motivated by the processing of the human auditory system, further steps of calculation are performed, that include the compression of amplitudes with a logarithmic function (LOG-FFT) and the mapping of the uniformly spaced spectral channels to a more auditory-inspired frequency scale (such as the *equivalent rectangular bandwith*, ERB, Mel or Bark scales). This yields the so called LOG-MEL features which have been used extensively in the field of speech recognition. A further decorrelation is calculated by applying the discrete cosine transform to the LOG-MEL features to obtain the *mel-frequency cepstral coefficients* (MFCC) - possibly the most popular acoustic feature that has been applied to many tasks (also for music or sound scene classification). A similar approach has been proposed with the *relative spectral transform perceptual linear prediction* (RASTA-PLP) features (Hermansky and Morgan, 1994) that analyze the most important temporal modulations (below 20 Hz) of envelope information for speech intelligibility. The analysis of temporal modulations in individual frequency channels is the main idea behind the auditory inspired *amplitude modulation spectrum* (AMS) features that have been used nearly exclusively in the early studies on machine learning-based speech segregation (Harlander et al., 2012; Hu and Loizou, 2010; Kim et al., 2009; Tchorz and Kollmeier, 2003). One limitation of AMS features used by previous studies was the short temporal window which does not allow for the analysis of very low modulation frequencies (e.g. with 20-ms long frames, only modulations above 20 Hz can be detected) that are most important for speech intelligibility (Hu and Loizou, 2010). More recent studies used a combination of several acoustic features, such as AMS, MFCC and RASTA-PLP with additional delta features (Healy et al., 2014, 2013; Wang and Wang, 2013) or combined with pitch and noise estimation algorithms (May and Dau, 2014a) or simply used LOG-FFT (Huang et al., 2014) or LOG-MEL features directly (Weninger et al., 2014).

The most recent approach tries to mimic the processing inspired by the auditory system by using a Gammatone-based filter bank instead of the more conventional FFT-based analysis scheme (Hohmann, 2002). This yields a so called *cochleagram* representation that was shown to deliver the best performance in a feature comparison study for single-microphone speech enhancement (Chen et al., 2014). As a result of this comparison, Chen et al. propose a *multi-resolution cochleagram* feature (MRCG) that incorporates different degrees of temporal information by using a combination of various frame lengths and temporal overlap between successive frames of Gammatone-based features (e.g. used in Healy et al., 2015).

Machine learning algorithm

A machine learning algorithm is employed to estimate the target function from the extracted set of acoustic features. This has been accomplished using a gaussian mixture model (GMM) or support vector machine (SVM) classification algorithm for estimating the IBM (binary decision; Kim and Loizou, 2009; Hu and Loizou, 2010; Wang and Wang, 2013; May and Dau, 2014). However, one of the first studies of supervised speech separation proposed to use an *artificial neural network* (NN) algorithm for estimating the local SNR in 15 frequency bands (Tchorz and Kollmeier, 2003). Tchorz and Kollmeier reported that the processing with the algorithm provided attenuation of the background noise without the introduction of musical noise, but the system was not evaluated using listening tests or objective predictions (it was tested later by Harlander et al., 2012, but no improvements in speech intelligibility were found). This approach was followed up by Healy et al. (2013) who applied a large *deep neural network* system (DNN, a NN system with multiple hidden layers) to the classification of the IBM and showed remarkable improvements for hearing-impaired listeners. The advancements in the field of machine learning with a strong focus on *deep learning* (a mix of DNNs and machine learning) led to a dominance of DNN-based systems in supervised speech segregation (Harlander et al., 2012; Healy et al., 2013, 2014, 2015; Weninger et al., 2014; Huang et al., 2014) that constitute the state-of-the-art at present. DNNs seem to generalize better to different acoustic conditions, such as different types of noise or SNRs, and are trained efficiently using graphical processing units (GPU) as well as being simpler to implement in the final applications (fast computation of estimation values). During recent years, a variety of software packages that can be used to implement DNNs has been made available within commercial packages such as MATLAB, but also within freely available tools for Python-based platforms (e.g. Tensorflow, Torch, Keras, Caffe etc.). An example of the typical structure of a feedforward deep neural network applied to speech enhancement is shown in Fig. 2.10 comprising the input layer of acoustic features, several hidden layers with nonlinear hidden units and the output layer to estimate a target function such as a time-frequency mask on a frame by frame basis that is used for noise reduction processing.

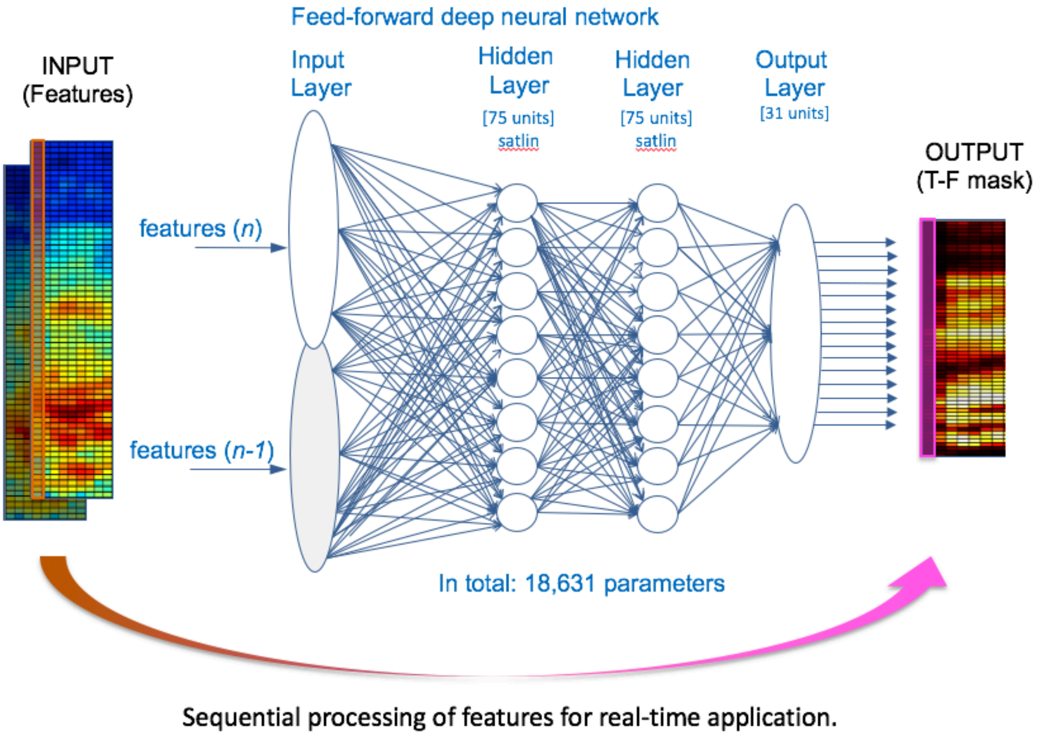


Figure 2.10 Schematic of the *Neural Network* processing block using a feed-forward artificial neural network for regression analysis. A frame of acoustic features is passed to the input, hidden and output layer of the network to obtain an estimate of the target function (T-F mask) at the output layer. This process will be repeated sequentially in a real-time application for *Speech Enhancement* (generated by the author).

Target function

Traditionally, algorithms applied the IBM as target function, mainly because of the assumption that it would be easier for an algorithm to detect speech in a binary fashion (decision between two classes: speech- or noise-dominated), rather than a continuous estimation of speech energy in the form of the SNR or IRM values (Kim and Loizou, 2009; Hu and Loizou, 2010; Healy et al., 2013, 2014). This assumption was proven wrong by the advancement of DNN-based regression systems that have been shown to be able to estimate a continuous value function such as the IRM (Huang et al., 2014; Weninger et al., 2014; Healy et al., 2015; Bolner et al., 2016). The IRM has several advantages over the IBM, mainly the improved perceptual quality of the enhanced speech signal by human listeners and the increased robustness to changes in the global SNR (no need for choosing an appropriate decision threshold, *LC*). Wang et al. (2014) compared several target functions for supervised speech segregation including the IBM, IRM and TBM and more simple approaches such as the FFT amplitudes of the target speech signal or a mask based on the FFT directly (without the mapping to an auditory-inspired frequency scale). In this study, the IRM was found to provide the best results in terms of STOI, SNR-improvement and perceptual scores (PESQ; Rix et al., 2001) for speech segregation at an overall SNR of -5 dB, especially in the case of multi-condition training (when different types of background noise were included in the training data).

Overview of previous studies: Results from listening experiments

Several studies on supervised speech segregation have performed listening tests for evaluating the outcome of the algorithms for speech understanding in noise using different groups of listeners. Kim and Loizou (2009) used a GMM classifier in combination with AMS features and tested normal hearing listeners on their percentage correct scores in different types of background noise. They found large improvements in intelligibility up to 60% in babble noise in low SNR conditions (SNR = -5 dB). A similar system was tested with CI users and again significant improvements in intelligibility were reported, in this case at higher SNR conditions (5 and 10 dB SNR; Hu and Loizou, 2010). These results are promising and show that speech intelligibility in noise can indeed be improved by an algorithm that does not incorporate *a priori* information. However, this statement has to be taken carefully, because the algorithm was trained on very similar data as used for the evaluation, which can be viewed as a different way of *a priori* information. Both studies used the same speaker from the same speech corpus (IEEE), the same SNR conditions and the same recordings of background noise between the training and testing stages. This means that the algorithm may have just learned the specific characteristics of the training data and may not work well in acoustic conditions that are different from the ones in the training data, a behaviour called *overfitting*. May and Dau (2014) reported large decreases in performance for GMM-based systems when different segments of the background noise were used for training and testing. This finding supports the risk of overfitting and reveals a strong limitation of previous GMM-based systems.

Healy et al. (2013) followed up this line of work and replaced the GMM classifier with a large DNN-based system that used a subband-DNN classifier for each frequency channel (as it was done similarly in the GMM-based systems). They used a combination of AMS, RASTA-PLP and MFCC features and tested the performance in speech understanding in noise by normal hearing and hearing impaired listeners. They reported remarkably large improvements in speech intelligibility in low SNR conditions (-2, -5 dB) for both NH and HI listeners up to the level of understanding obtained in quiet acoustic conditions. However, the limitations of the GMM-based system used by previous studies were not addressed as they used the same training and testing framework with strong similarity between the training and testing data (even more eventually by using 10-s long recordings of noise whereas previous studies did not report the length of the noise recordings they used). It can be argued, that also the previously proposed GMM-based system may increase speech intelligibility for HI listeners with similar performance (as demonstrated with NH listeners). In Healy et al. (2015), it was shown that a DNN-based system can generalize to novel recordings of the same noise type, which can be seen as an advantage over the GMM-based approach. Large improvements in speech intelligibility for NH and HI listeners were found, also for these more challenging condition. A noise perturbation approach was used, that incorporated a large variety of the same noise type for training of the system. However, the same speaker, SNR and noise type were used for the evaluation and it remains unclear how supervised speech segregation systems perform in more mismatched conditions that would occur in realistic applications in hearing aids or cochlear implants. Another factor is the computational complexity of the proposed approaches. These algorithms incorporate millions of parameters that would not fit into the digital memory of current hearing devices. Furthermore, the processing power of HAs and CIs is limited due to the restricted battery size that is available in small devices, which does not allow for the implementation of such large-scale algorithms at present. An additional constraint is the processing delay of speech enhancement algorithms that is limited to about 20-30 ms to ensure acceptance by users of hearing devices. This means that SE algorithms for application in hearing devices are restricted to the use of causal processing.

The first results of SE algorithms based on machine learning techniques are very promising and indicate that speech intelligibility benefits may be obtained by hearing impaired listeners and users of cochlear implants if the SE algorithm achieves a high level of estimation accuracy for the target mask. The future task is to address the limitations of previous algorithms, such as improving the generalization performance to unseen or more variable acoustic conditions, while meeting the requirements of limited computational complexity and memory in hearing devices and short processing delay for the application in HAs and CIs. This is the goal that this thesis is aiming for.

3 TOLERANCE OF PROCESSING DELAY

Overview

Processing delay is one of the most important factors that limit the development of novel algorithms for hearing instruments. This is of particular importance for speech enhancement algorithms that extract acoustic features from noise-corrupted speech signals in short time frames to improve speech perception in noise. It is crucial for the acceptance by users of hearing aids that the algorithms perform in real time without introducing degradations to the perceptual quality. The maximum tolerable delay by hearing aid users should be exploited to improve the performance of speech enhancement algorithms by incorporating as much temporal information as possible.

In the *first part* of this chapter, normal-hearing listeners and listeners with hearing loss were tested for their tolerance of processing delay up to 50 ms using a real-time setup for own-voice and external-voice conditions. Participants were asked to speak and listen in five delay and three voice conditions and their perceived subjective annoyance to delay was measured using a 7-point Likert scale. Delay tolerance was significantly greater for the hearing loss group than for the normal-hearing group. The results indicated, even though not statistically significant, that more experienced hearing aid users tolerated longer delays than new hearing aid users, an effect that could be due to long-term acclimatization to processing delay.

In the *second part* of this chapter, normal-hearing listener's long-term acclimatization to processing delay was evaluated. Participants were given audio books and a smartphone app for use at home during a week-long acclimatization period. Pre- and post-tests were performed measuring tolerance of processing delay before and after the acclimatization period. Results indicated potential long-term acclimatization for the group that used a shorter delay. Tolerance to processing delay increased significantly for this group but not for the other group, suggesting that long-term acclimatization to delay is increasing its tolerance if a suitable delay is chosen.

3.1 Tolerable Delay: Effects of Hearing Ability and Experience with Hearing Devices

Declaration of authorship:
Tobias Goehring - Leading the research, developed the technical setup and signal processing, analysis of results and writing the manuscript

Co-authorship:
Josie Chapman - Performed the listening test, study design and analysis
Dr. Jessica Monaghan - Advised the research
Prof. Stefan Bleeck - Advised the research

Parts of this research have been submitted for publication (first authors are underlined):

Goehring, T., Chapman, J., Bleeck, S. Monaghan, J., (2016). Listeners with hearing loss have a greater tolerance of processing delay than listeners with normal hearing. *International Journal of Audiology*. Under revision.

3.1.1 Introduction

State-of-the-art hearing devices such as hearing aids employ digital signal processing to improve perceptual aspects of audio signals for the user. When an audio signal is processed by a hearing aid, the air-conducted signal produced by the device is delayed relative to the direct signal that enters the ear canal. The addition or interaction of these two signals may be perceived by hearing aid users and cause annoyance. The amount of delay depends on the device; in contrast to analog hearing aids that allow for negligible throughput delays between microphone and receiver (under 1 ms; Dillon, Keidser, O'Brien, and Silberstein, 2003), digital technology introduces delays in the range of several milliseconds (typically 2-10 ms). As processing delay increases and as the levels of the direct and processed sound become more equal at the eardrum, sound quality deteriorates and user annoyance ratings increase (Stone and Moore, 1999). In order to avoid this problem, hearing devices are designed to operate with low delay (typically <10 ms, as suggested by Groth and Sondergard, 2003; Stone *et al.*, 2008; Bramslow, 2010). However, this requirement of low processing delay places a strong restriction on the development of novel algorithms for digital hearing devices. If larger delays can, in fact, be tolerated by typical users, this would allow for the use of more sophisticated signal processing techniques and the integration of information over wireless connections (e.g., for binaural algorithms or integration of smartphones), for which a much greater delay is required.

The perceptual effects of processing delay depend on its magnitude (Stone and Moore, 1999). With longer processing delays: the desynchronization of audio-visual information (>80 ms; Summerfield, 1992) and auditory-proprioceptive feedback (>80 ms), perception of distinct echo sounds (>50 ms; Litovsky *et al.*, 1999) or altered speech production rates (>30 ms; Stone and Moore, 2002) have been reported. With shorter delays, a comb-filtering effect may dominate the disruption of speech perception. The comb-filtering effect results from superposition of the direct and delayed sounds. The superposition introduces spectral ripples spaced at the inverse of the delay and leads to a spectral coloration affecting the timbre of the sound that is perceived as unnatural or annoying. A natural situation in which comb-filtering is produced is when a speaker is listening to their own voice. Here, the sound is transmitted via bone conduction and via air conduction to the cochlea. The speech reaches the cochlea via these two paths with different delays (about 0.7 ms difference; Stromsta, 1962) and the two signals sum to result in a comb-filtering effect. Long-term acclimatization to the spectral coloration produced during own-voice listening is thought to be one of the reasons why short delays below about 4 ms are not perceived as bothersome or even noticeable by human listeners (Agnew and Thornton, 2000; Groth and Sondergaard, 2004, Zakis *et al.*, 2012). However, for longer delays in the range of 10-50 ms, the spectral coloration due to the comb-filtering effect is more noticeable and may be perceived as disturbing (Agnew and Thornton, 2000; Stone and Moore, 1999; 2002; 2005).

Stone and Moore performed a series of experiments to investigate subjective tolerance of processing delay. In one study (Stone and Moore, 1999), normal-hearing (NH) participants listened via headphones to recorded stimuli that were obtained by mixing above-ear and in-ear recordings of a person while talking and processed with hearing aid and hearing loss simulations. Four different hearing losses from mild to moderately severe were simulated to account for the effects of threshold elevation and loudness recruitment. Results showed monotonically increasing disturbance with increasing delay. They found the higher the degree of simulated hearing loss, the greater the tolerance to delay (tolerable up to 40 ms for moderately severe losses). They concluded that the delay should not exceed 20 ms for mild to moderate hearing losses. Agnew and Thornton (2000) also tested delay tolerance for expert NH participants listening to their own voice while reading without using hearing loss simulations. Participants controlled the level of the delayed sound and the amount of delay were while being able to compare the delayed condition to the undelayed condition at any point. They found delays about 14 ms to be objectionable. However, their experimental design prevented short-term acclimatisation effects that were later reported to increase tolerance to delay (Stone and Moore, 2005) and they reported that the use of expert listeners can be seen as a worst case in terms of delay tolerance. Stone and Moore (2002) provided NH participants with behind-the-ear devices using real-time processing that used 0 dB insertion gain (at 65 dB SPL) and asked them to rate their disturbance during speech production (reading from a book) for a range of delays (from 7 to 43 ms) without hearing loss simulations. Again, tolerance limits of about 15 to 20 ms depending on the acoustic environment and when defining a rating of 3 out of 7 as the tolerance limit (with 1 labeled as 'not at all disturbing', 4 labeled as 'disturbing' and 7 labeled as 'highly disturbing'). No effects on speech production were reported up to delays of about 30 ms.

A similar experiment was performed using participants with hearing loss (HL) fitted with behind-the-ear devices with real-time processing and using a realistic gain prescription to achieve the same overall loudness as would be perceived by NH participants (Stone and Moore, 2005). The delay tolerance limits were about 14 to 30 ms, with a non-monotonic effect of low frequency hearing loss on delay tolerance. The participants were partly unblinded and trained to hear the effects of changing delay settings. The signal processing implemented a four-channel, fast-acting, wide-dynamic range compression as commonly used in hearing aids. The very low compression threshold of 32 dB SPL and attack and release times of around 9 ms introduced level- and time-dependent variations to the processed signal, but not to the direct signal and thus may have interacted with the perceptual effect of delay. The time-varying dynamics of the processed signal and the direct signal was further increased by the use of a non-linear gain prescription with level compression. Although this processing is similar to what would be used in commercial hearing aids, it is unclear to what degree the subjective ratings of delay tolerance were affected by the introduction of non-linear processing. The authors suggest that this processing may have led to the non-monotonic effect of low frequency hearing loss on delay tolerance and emphasize that the use of other prescription methods and compression parameters may lead to a different pattern of results. Furthermore, perception of external speech was not assessed in this study.

In another study, Stone, Moore, Meisenbacher, and Derleth (2008) investigated delay tolerance for NH participants with open-canal fittings and several types of gain prescription for the delayed signal.

Participants were listening to and asked to rate above-ear and in-ear recordings of a person reading that were mixed and processed with a hearing loss and hearing aid simulation similar to the methodology followed by the experiment in Stone and Moore (1999). In this study participants were much more sensitive and ratings reached tolerance thresholds at delays of just 5 to 6 ms. However, also in this study the authors argued that the use of dynamic range compression in combination with hearing loss simulation for loudness recruitment introduced disturbing dynamic side effects that could have interacted with the disturbing effects of delay and made the interpretation of results difficult. Serveral studies reported no significant effects on delay tolerance or preference for delays up to 10 ms for HL participants (Groth and Sondergaard, 2004; Bramslow, 2010; Zakis *et al.*, 2012) and support the common choice of hearing aid manufacturers to restrict processing delay to values below 10 ms. However, in these studies the greatest tested delay was 10 ms or less, providing no information about tolerance of longer delay.

As a whole, the literature has been taken to support a limit of 10 ms for processing delay. Current hearing devices are constrained to use processing delays below this limit, restricting the development of more complex signal processing methods such as acoustic feedback and noise cancellation algorithms (Dillon *et al.*, 2003). Recommendations in the literature were predominantly based on data obtained using NH participants (Stone and Moore, 1999; 2002; 2008; Agnew and Thornton, 2000) and there are indications that HL participants might tolerate longer delays (Stone and Moore, 1999; Groth and Sondergard, 2004; Bramslow, 2010). This would be beneficial for the development of algorithms that try to improve speech understanding in noisy environments or the suppression of acoustic feedback (Löllmann and Vary, 2009; Chen *et al.*, 2016; Monaghan *et al.*, 2017). Previous studies using HL participants with delays above 10 ms have mainly focused on measuring delay tolerance to own-voice stimuli (Stone and Moore, 2005) or tested across-frequency delay (Stone and Moore, 2003). State-of-the-art hearing devices can actively reduce the occlusion effect by using active feedback cancellation systems (Mejia *et al.*, 2008; Borges *et al.*, 2013) or by detecting when the user speaks to decrease the amplification in lower frequencies for improved own-voice perception (US patents US9094766 B2; US8477973 B2). This means that delay tolerance to external-voice stimuli may become more relevant for future hearing devices. In order to increase the validity of subjective ratings, the use of real-time processing is preferred over pre-recorded stimuli because this provides a more natural perception of own-voice and external-voice sounds by maintaining proprioceptive and audio-visual cues.

We aimed to address some of the limitations of previous studies by testing both NH and HL participants with the same real-time processing setup for both own- and external-voice stimuli over the most relevant delay range of 10 to 50 ms. As previously proposed by Bramslow (2010), we kept the experimental setup simple and did not include any non-linear processing methods (such as dynamic range compression, active noise cancellation etc.) to avoid confounding effects to the perception of delay as reported by Stone and Moore (1999; 2005; 2008). A downside of this approach is that such a simple processing setup is not a realistic and accurate simulation for commercial hearing aids that employ a multitude of adaptive and non-linear processing techniques to improve sound perception. However, the main motivation of this study was to compare the tolerance of processing delay between NH and HL groups that would be informative to different types of hearing devices to some extent (hearing aids, earables etc.) and not the evaluation of a specific setup. Consequently, we chose a linear fitting rationale in contrast to the non-linear prescription gain commonly used in hearing aids. However, the choice of the fitting rationale may not have a strong effect on delay tolerance as reported by Stone and Moore (2002) who found no significant difference in disturbance ratings between linear and non-linear fittings for own-voice perception during speech production with NH participants. Several studies have chosen to use a frequency-independent gain for HL participants regardless of their degree of hearing loss (Groth and Sondergard, 2004; Bramslow, 2010). We aimed to achieve similar audibility between NH and HL groups and therefor provided frequency-dependent gain to compensate for elevated hearing thresholds. In the external-voice condition, two level ratios of 0 and 20 dB between the delayed and direct sounds were investigated to account for the worst-case scenario and a more realistic case that may occur in practice with a hearing aid, respectively. We investigated whether HL participants show increased tolerance to output delay for own-voice and external-voice conditions in comparison to NH participants when tested under the same conditions in a real-time setup. Furthermore, we assessed the dependence of delay tolerance on the degree of hearing loss and the level difference between the direct and delayed sound paths reaching the eardrum. We also aimed to investigate whether there is a difference in tolerable delay between new and experienced users of hearing aids, due to long-term acclimatization effects to delayed sounds that may occur with the regular use of hearing aids.

3.1.2 Method

Participants

Twenty NH participants aged 18 to 45 years (12 female, 8 male, average age of 24 years) and twenty HL participants aged 45 to 81 years (12 female, 8 male, average age of 67 years) were recruited at the University of Southampton and the Royal Berkshire Audiology department. Participants were not paid but reimbursement of travel expenses was offered. The normal-hearing group had hearing levels not exceeding 20 dB HL at octave intervals between 500 and 8000 Hz. The hearing-loss group was tested without hearing aids and had hearing levels higher than 20 dB HL but not exceeding 70 dB HL at octave intervals between 500 and 8000 Hz. Hearing thresholds were confirmed via an audiogram measured by the experimenter before the start of the experiment or were provided by the clinics if a recent measurement was available. Within the HL group there was a significant correlation between age and pure-tone averages (PTA, with a Pearson correlation of r=0.52, p=0.019). Table 1 shows average pure-tone thresholds at 500, 1000 and 2000 Hz and further information for the HL group. Of the 20 HL participants, 10 were regular users of hearing aids and 10 were new to the use of hearing aids (HA, out of whom 7 received their HA on the day of testing and 3 decided against using a HA).

Table 3.1 HL group participant information.

Participant	Gender	Age	HA usage	Mould type	PTA	HL subgroup
HL1	Male	71	Experienced	Closed	44	mid
HL2	Female	66	New	Open	40	mid
HL3	Female	77	Experienced	Closed	64	high
HL4	Male	69	Experienced	Open	54	high
HL5	Male	68	Experienced	Closed	58	high
HL6	Female	77	New	Open	34	low
HL7	Female	76	New	no HA	46	mid
HL8	Female	68	New	Open	42	mid
HL9	Male	50	New	Open	29	low
HL10	Female	83	Experienced	Open	46	mid
HL11	Male	76	New	Open	50	high
HL12	Female	75	New	Open	39	mid
HL13	Female	81	New	Open	41	mid
HL14	Female	79	Experienced	Open	53	high
HL15	Female	53	Experienced	Open	29	low
HL16	Female	75	Experienced	Open	65	high
HL17	Male	48	New	no HA	29	low
HL18	Male	52	New	no HA	24	low
HL19	Female	51	Experienced	Open	38	mid
HL20	Male	45	Experienced	Closed	51	high

Equipment and stimuli

For the NH group, the experiments were performed in a quiet but not sound-proof, carpeted meeting room (5 m x 5 m x 2.5 m, RT60=0.3 s) at the University of Southampton audiology clinics. For the HL group, the experiments took place in a similar room (3 m x 3 m x 2.5 m) at the Royal Berkshire Hospital. The equipment was based on a digital signal processing board (Analog Devices ADSP-BF537-EZLITE) that delivers a minimum throughput delay of 1.5 ms (measured with RME Fireface UC). Stimuli were picked up with a dynamic microphone (JHS GS67), amplified with a pre-amplifier (ART Dual pre USB) and further processed with the DSP board before presentation to the participant via circumaural headphones (Sennheiser HD380pro). The DSP board performed analog-to-digital conversion (48 kHz, 16 bit) and split the signal to produce the two simulated sound paths: the direct air-conducted sound path and the delayed sound path of a hearing device.

Three different stimulus conditions were generated in real time to simulate three scenarios: own-voice perception (OwnV), external-voice perception with no level difference between the delayed and direct sound (Ext0dB) and external-voice perception a level difference of 20 dB between the delayed and direct sound (Ext20dB). Accordingly, the direct sound path was muted (OwnV), kept at the input level (Ext0dB) or attenuated by 20 dB (Ext20dB) (Hétu and Quoc, 1992). In the OwnV condition, the participants spoke into the microphone and listened to their own voice, which they heard directly via bone conduction and via the simulated hearing device delayed by *D* through the headphones. In the two external voice conditions, the experimenter spoke into the microphone and the participant listened to the stimuli through the headphones. The delay *D* was set to 10, 20, 30, 40, or 50 ms. The direct sound signal was then added to the delayed signal. In the case of the HL participants, a hearing-loss dependent gain was applied as final step to compensate for their hearing thresholds via a linear half-gain rule according to the audiogram of each participant using a 10-Band graphic equalizer (MXR M108). It should be noted that the level-ratio between the direct and delayed sound paths was the same for both NH and HL participants. The fitting gain for the HL group was applied after the mixing of the two sound paths in order to compensate for the hearing thresholds, and not to simulate the fitting gain of a hearing aid. The rationale behind this choice was to directly compare the annoyance of delay between NH and HL groups, not confounded by the level-ratio between the signals.

The complete setup was calibrated with a sound level meter (B&K2260) and artificial ear (B&K4153) to give an average sound level of 65 dB(A) when talking at a normal conversational level into the microphone. The overall loudness was assessed by the participant and adjusted if required to make it comfortable.

Experimental procedure

At the beginning of the experiment two practice trials with ratings were performed for the own-voice (OwnV) and one external-voice (Ext0dB) condition. The practice trials used the minimal possible delay produced by the DSP board of 1.5 ms to allow the participant to acclimatise to the sound and usage of the setup. A total of 15 conditions comprising three different stimuli (OwnV, Ext0dB, Ext20dB) and five different delay settings (10, 20, 30, 40, 50 ms) were presented. The presentation order of these conditions was randomized for each participant using a latin square. The participants were asked to rate each condition on a 7-point Likert rating scale for the perceived subjective annoyance (with 1 labeled as 'not at all annoying' to 7 labeled as 'very annoying', and no other labels used).

In line with Stone and Moore (2005) we used a popular narrative book (‘Harry Potter and the Philosopher’s Stone’ by J. K. Rowling) and let the respective speaker (participant or experimenter) read an arbitrarily selected passage of approximately one-minute length for each condition. The participants were instructed to read in a conversational manner not emphasizing or raising their voice for special parts of the text. The total testing time was around 30 minutes and a short break was offered half way through the experiment. All experiments were approved by the local ethics board (ref ID 8978) and the NHS research ethics committee (REC reference number 8978).

3.1.3 Results

The group mean annoyance ratings for all conditions (5 delays per stimulus) are shown in Fig. 1 for the NH and HL groups. A difference in delay tolerance between the two groups occured for the own-voice condition for delays above 20 ms and for the external voice Ext0dB condition over the whole range of delays. The external voice Ext20dB condition led to smaller differences between the two groups and lower ratings overall than for the other conditions. A repeated measures ANOVA (within-subject factors of delay and voice condition and between-subject factor hearing ability) was performed. There were significant main effects of delay [F(4,152) = 85.875, p < 0.001], voice condition [F(2,76) = 13.228, p < 0.001] and hearing ability [F(1,38) = 4.619, p = 0.038]. There was no significant effect of gender when included as the third factor in the repeated measures ANOVA. There was a significant interaction between voice condition and hearing ability [F(2,76) = 6.503, *p* = 0.002] that can be explained with the noticeable effect of hearing ability on delay tolerance for the Ext0dB condition but not for the Ext20dB condition. No other interactions were significant.

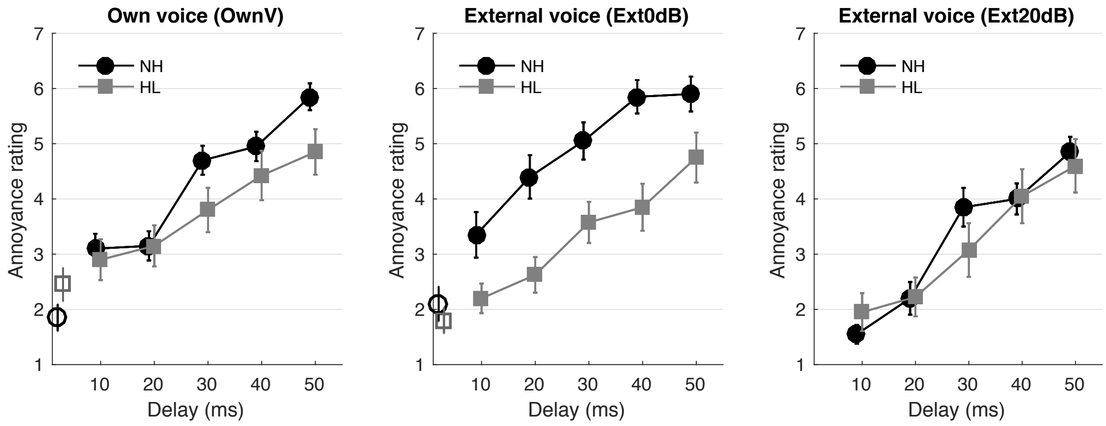


Figure 3.1 Group mean annoyance ratings for the normal-hearing (n=20) and hearing-loss (n=20) groups in conditions (from left to right): own voice, external voice without level difference and external voice with level difference. Error bars show the standard error of the mean. Points plotted at 1.5 ms show average practice trial ratings (OwnV and Ext0dB only).

To assess the effect of degree of hearing loss on delay tolerance, the HL group was split based on their hearing thresholds into low PTA (<35 dB HL, n=5, 3 male, average age of 56 years), mid PTA (35<PTA<50 dB HL, n=8, 1 male, average age of 71 years) and high PTA (>=50 dB HL, n=7, 4 male, average age of 70 years) subgroups. Group mean scores for the three subgroups are shown in Fig. 2. Consistent with differences between the HL and NH groups, a pattern of larger tolerable delay with greater hearing loss occurred in the ratings for OwnV and Ext20dB conditions. The difference between the HL subgroups was most evident for the Ext20dB condition for delays above 20 ms. The difference also occurred for the Ext0dB condition for delays of 30 and 50 ms, but the differences between groups were smaller than for the other two conditions. A repeated measures ANOVA (within-subject factors of delay and voice condition and between-subjects factor of hearing-loss subgroup) was performed. There was a significant main effect of delay [F(4,68) = 31.337, *p* < 0.001], but no effect of voice condition or hearing-loss subgroup. All two-way interactions were non-significant but there was a marginally significant three-way interaction between delay, voice and hearing-loss subgroup [F(16,136) = 1.716, *p* = 0.051] that most likely arises from the differences in the effect of hearing-loss subgroup on delay tolerance between the three voice conditions (smaller or non-existent effect noticeable for the Ext0dB condition than for the other two).

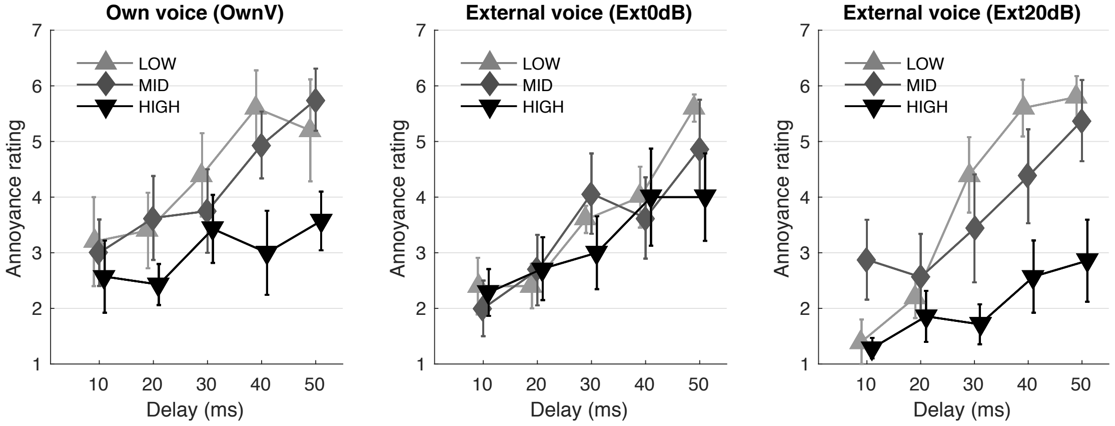


Figure 3.2 Annoyance ratings for the three hearing loss subgroups (low (n=5), mid (n=8), high (n=7)). Otherwise as Fig. 1.

Group mean scores for experienced and new hearing aid users are shown in Fig. 3.3. There was a difference between groups for the OwnV and Ext20dB conditions, for which the experienced users gave lower ratings by about 0.5 to 1 units for all delays (OwnV) or delays above 20 ms (Ext20dB). Ratings of the new hearing aid users in the Ext0dB condition did not monotonically increase with delay and were even lower than for experienced users at delays of 20 and 40 ms. A repeated measures ANOVA (within-subject factors of delay and voice condition and between-subject factor of hearing aid experience) was performed. There was a significant effect of delay [F(4,72) = 27.453, $p < 0.001$], but no significant effects were found for voice condition and hearing aid experience and there were no significant interactions.

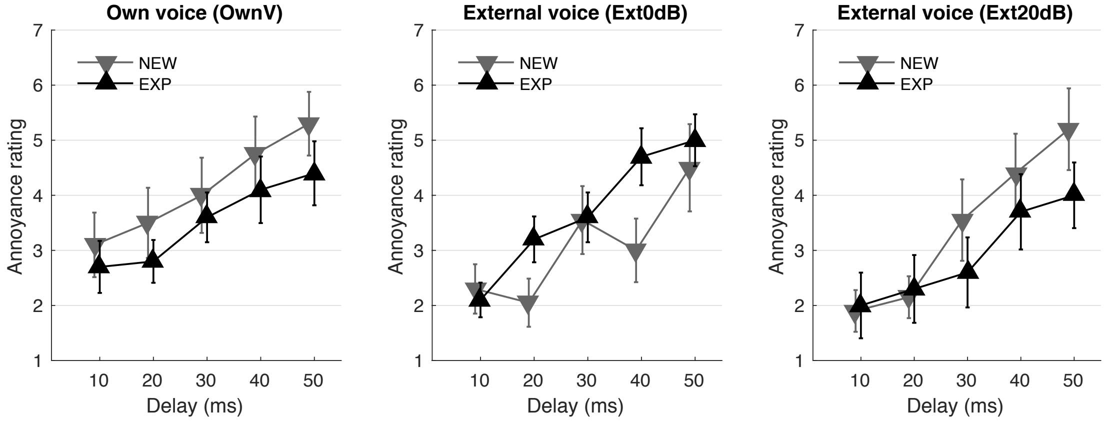


Figure 3.3 As Fig. 1 but for the experienced (n=10) and new (n=10) hearing aid users.

3.1.4 Discussion

Our data demonstrate that participants with a broad range of hearing losses have significantly greater delay tolerance than normal-hearing participants. Generally, the results are consistent with previous work, especially with Stone and Moore (2005). However, we add to this work by comparing the annoyance ratings for normal-hearing and hearing-loss groups using own-voice and external-voice stimuli and a larger range of delays, up to 50 ms. We used real-time processing to allow for a more natural perception than using recorded stimuli and restricted the setup to linear fitting and processing paradigms in an attempt to avoid non-linear side-effects that may interact with the perception of delay. In addition to the perception of delay during speech production (own voice condition, OwnV), the external voice condition with no level difference (Ext0dB) represents a worst-case scenario where the two sound paths (direct and delayed) have equal level over the whole frequency range. This means that the coloration by comb filtering is maximal in the Ext0dB condition. In contrast, the external voice condition with level difference (Ext20dB) was intended to simulate a more typical attenuation of the direct sound by an earmould (in relation to the hearing aid signal). The age difference between the NH and HL group may have confounded the effect of hearing loss and compromises the interpretation of the results. Furthermore, the wide range of hearing losses for the HL group (from 24 up to 65 dB PTA) makes it difficult to formulate quantitative limits for the tolerance of processing delay. Due to the setup used in this study, the level ratio between the direct and delayed sound was similar across participants for both external-voice conditions and for the own-voice condition with the NH group. For the own-voice condition with the HL group, the perceived level ratio between the direct sound (which consisted mainly of bone-conducted sound, that could not be amplified according to the PTA) and the delayed sound was depending on the degree of hearing loss of the participant, with greater level ratios for stronger hearing losses. This may have led to increased tolerance to delay for the own-voice condition by participants with stronger hearing loss. However, this situation would be similar for real hearing devices, because the direct sound would be perceived less with stronger hearing loss.

We apply a rating threshold of 4 to quantify tolerance limits at the midpoint of the 7-point scale. Practice trials showed ratings about 2 when a very small delay was applied. This could represent an edge effect bias and suggest that a threshold of 3, as used by Stone and Moore (2005), might be too strict. NH listeners showed an increase in annoyance rating between 10 and 20 ms delay for Ext0dB and Ext20dB but not for the OwnV condition. This suggests that for NH listeners the delay should not exceed 20 ms. Compared to the NH group, the HL group showed smaller increases in annoyance ratings between the 10 and 20 ms delay conditions. Instead, the HL group showed a stronger increase in annoyance ratings between the 20 and 30 ms delay conditions. This can also be observed for the low and mid HL subgroups for the Ext0dB and Ext20dB conditions and for the low HL subgroup for the OwnV condition. For the HL group, ratings increased more smoothly with delay than for the NH group.

Overall, the annoyance ratings for the HL group were significantly lower than for the NH group and the tolerable delay limit can be estimated to be about 30 ms (passing the threshold of 4). This is consistent with results reported by Stone and Moore (2005) when applying the same threshold value. For the HL subgroups with different degrees of hearing loss, the first group (low) showed a similar trend as the NH group with a tolerance limit of about 20 ms and a strong increase in annoyance between 20 and 30 ms of delay in all three conditions. For the second group (mid) the tolerance limit of delay for which the ratings were below 4 was about 30 ms. The third group (high) only gave ratings above 4 for the Ext0dB condition for delays over 40 ms, and did not reach the tolerance limit in the other conditions. This reflects comments by participants from the third group that they could not hear any difference between the conditions. Thus, also within the HL group, we find a trend towards higher tolerance to delay with higher degree of hearing loss, although this effect was not statistically significant (likely due to the small sample sizes).

Annoyance ratings were higher for the new than for the experienced hearing aid users in two out of three conditions (OwnV, Ext20dB), although this trend was not significant overall. These stimuli represent in a simplified manner the scenarios of own-speech production and listening to external speech with a hearing aid. Potential long-term acclimatization, as suggested by Stone and Moore (2005), could explain this result. In contrast, the Ext0dB condition represents an extreme-case scenario that would normally not occur in a hearing aid (equal level of direct and delayed signals over the whole frequency range). Consequently, the potential long-term acclimatization effect to delay might not apply to the Ext0dB condition. Although, the effect of long-term acclimatization seems likely to account for the higher tolerance in the OwnV and Ext0dB conditions and the two groups were of same average age (67.1 vs 66.9 years), the experienced users had higher PTA values (50.2 vs 37.4 dB HL) than the new users. It is possible that both differences between groups (experience and higher degree of hearing loss) contributed to the increased tolerance to delay shown by the experienced user group. However, this effect did not reach the statistical significance level and has to be treated carefully. Future work that aims to address this question should test a larger sample and try to match PTAs between new and experienced participants.

There are several differences between the technical setup and signal processing implemented in this study (e.g. the use of headphones and linear processing) and commercial hearing aids with earmoulds and a large variety of combinations of non-linear processing techniques. This reduces the potential for predicting tolerable delay for users of a specific type of commercial hearing aid. Given the results presented in this study and earlier results of Stone and Moore (1999), the level ratio between the direct and delayed signals scales the magnitude of the comb-filtering effect and thus directly influences the tolerable delay. This is supported by the finding of lower annoyance ratings for the Ext20dB than for to the Ext0dB condition for both NH and HL groups. Different types of earmoulds and amplification settings of commercial hearing aids will alter the level ratio and thus change the tolerance to delay. However, we considered this by choosing conservative mixing ratios between the direct and delayed sounds to represent worst-case scenarios in terms of comb filtering. In practice, it is likely that users of hearing aids encounter lower magnitudes of comb filtering due to larger level ratios between the direct and delayed sound paths. It should be noted that this study made use of linear processing in an attempt to isolate the perceptual effect of delay by avoiding confounds with dynamic level changes between the direct and delayed sounds. While this approach may give a better estimation of the annoyance caused by processing delay alone, extrapolation of the findings to commercial hearing aids and other hearing devices that make use of non-linear processing is clearly restricted. The setup in this study presented stimuli solely via headphones using a simulation of the summation of direct and delayed sounds which is not a realistic simulation of the sound perception with state-of-the-art hearing aids.

3.2 Tolerable Delay: Effects of Long-term Acclimatisation

Declaration of authorship:
Tobias Goehring - Leading the research, developed the experiment design and technical setup, analysis of results and writing the manuscript

Co-authorship:
Sakeenah Kanji - Performed the listening test
Dr. Nick Clark - Programmed the smartphone app
Dr. Jessica Monaghan - Advised the research
Prof. Stefan Bleeck - Advised the research

I want to thank Nick Clark for providing the smartphone app and tailoring it to our needs.

3.2.1 Introduction

Processing delay is one of the limiting factors for the development of hearing aid algorithms. Users of hearing aids perceive disturbing effects with increasing processing delay leading to non-tolerable sound quality. For normal hearing listeners, a maximum tolerable delay length of 20 ms was found in previous studies and during the listening study described in the previous part of this chapter. Most studies that used normal hearing listeners allowed them to acclimatise shortly to the altered sound quality introduced by processing delay. For example in our study, each delay condition was presented with a duration of 1 minute that did not allow for long-term acclimatisation to the sound quality. In other experiments, similarly short durations of stimuli were used or subjects were allowed to switch between processing delay conditions at any point avoiding long-term acclimatisation to the sound quality. Besides effects of short-term acclimatization that may occur during the course of a listening experiment, there are several indications that human listeners are able to acclimatise to changes in sound quality over longer time periods. Previous studies did not address this question of whether long-term acclimatisation to processing delay increases its tolerance by human listeners.

The simplest example for long-term acclimatisation to the sound quality of processing delay is the perception of ones' own voice when talking. Hereby, a comb-filtering effect occurs due to the delay between the bone-conducted and air-conducted pathways of the own voice (small delay of 0.7 ms), similar to the effect that occurs when listening to two air-borne sounds that are delayed in respect to each other in case of an analog hearing aid (with small processing delays of about 1-2 ms). This shows that human listeners are able to acclimatise to the spectral changes in sound quality due to comb-filtering with small delays over long time periods (in this case their whole life). Human listeners are used to the sound quality of their own voice with comb-filtering and react surprised when hearing their voice for the first time after it has been recorded with a microphone. In the recorded case, only the air-conducted sound reaches the microphone and no comb-filtering effect takes place, leading to an altered sound quality compared to the natural own-voice perception without the spectral changes due to comb-filtering. After being exposed to their recorded voice for several times, listeners become less disturbed by the sound quality of their recorded voice due to long-term acclimatisation.

A key goal for the development of hearing aids is that the user tolerates the sound quality and takes advantage of potential benefits for example in terms of improved speech intelligibility in background noise. The effects of acclimatisation in terms of benefits in speech recognition and perceived sound quality has been investigated in several studies with hearing aid users (Dawes et al., 2014; Munro and Lutman, 2004; Ovegard et al., 1996). These studies reported significant improvements in perceived clarity and total impression (Ovegard et al., 1996), self-reported benefit and satisfaction by the use of HAs (Munro and Lutman, 2004) and subjective benefit reported by new users of HAs (Dawes et al., 2014). These results show that users of hearing aids seem to acclimatize to the sound quality of their devices. It seems likely but remains an open question, whether similar long-term acclimatization effects also apply for the changes in sound quality introduced by processing delay.

Stone and Moore (2005) reported short-term acclimatization effects to processing delay, that increased the subjects' tolerable length of processing delay by about 5 ms over the time-scale of the experiment that lasted 1 hour in total (a rating of 3 out of 7 was considered as tolerable). Disturbance ratings decreased by about 0.4 to 1 units of a 7-point Likert scale. Given the perceptual limit of tolerable processing delay of about 20 ms for normal hearing listeners, the magnitude of this acclimatization effect of 5 ms represents an increase of 25% in tolerable delay length. This short-term acclimatisation effect occured, even though subjects listened to a range of delay conditions between 13 and 40 ms over the course of the experiment. It may be speculated, that an even larger effect of acclimatisation would have occured for a single fixed delay condition in comparison to the use of a range of changing delay conditions.

In this study, we attempt to evaluate the effect of long-term acclimatisation to processing delay using normal hearing listeners who listened to preprocessed audio books and used a smartphone application for several hours to acclimatise to the sound quality of processing delay over a five-day period. We investigate whether the repeated exposure to the sound quality of processing delay over longer time periods increases the subjective tolerance to processing delay by measuring the subjective tolerance to processing delay before and after the acclimatisation period using speech signals that were preprocessed with a simulation of processing delay.

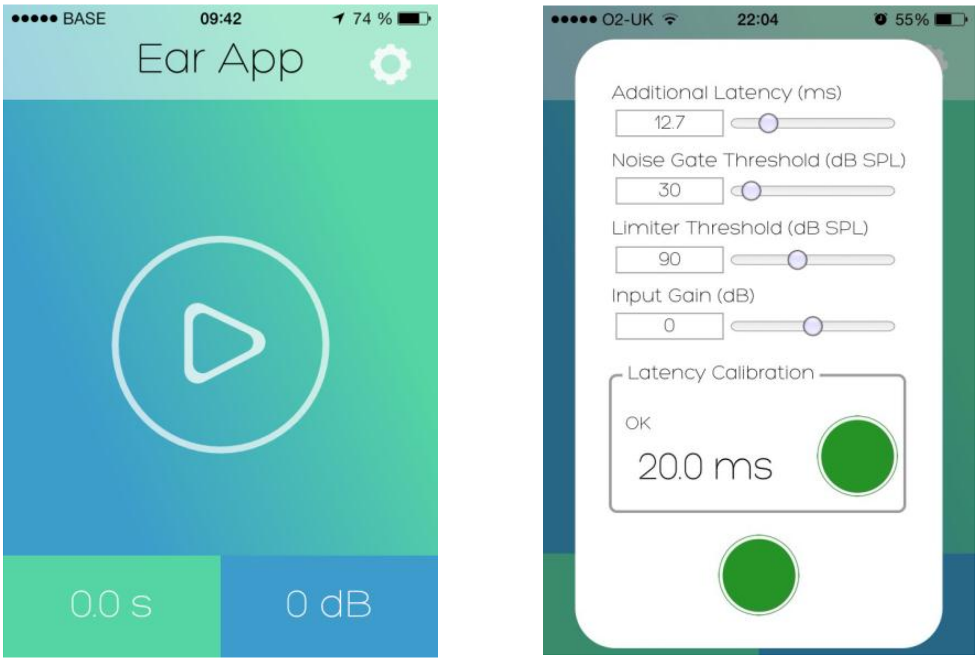
3.2.2 Methods

Subjects

In total, eight native-English speaking adults with an average age of 20.9 years and normal hearing took part in the experiment. Hearing thresholds were confirmed using pure-tone audiometry (PTA) to be within the normal range (less than 20 dB elevation in thresholds at octave intervals between 500 and 6000 Hz). Subjects were required to possess an Apple iPhone (version 4S or 5) to be able to install and use the smartphone application *Earapp*. The subjects were recruited at the University of Southampton and were not paid for participation, but were offered to obtain a pair of earphones after finishing the experiment. Subjects did not have any experience with the use of hearing aids or assistive listening devices and were blinded during the course of the experiment as to which algorithmic technique was applied for the processing of the stimuli, but were informed that a novel hearing aid processing technique was under investigation. The ten subjects were split into two groups according to the iPhone model they possessed. This was done to set up two fixed and different delays in each group depending on the type of iPhone, because the iPhone model 4S allowed for smaller throughput delays (minimum delay < 20 ms) than the iPhone model 5 (minimum delay > 20 ms) due to the internal processing of the devices. There were four subjects in group 1 (using an iPhone 4S) and four subjects in group 2 (using an iPhone 5).

Smartphone application: Earapp

The smartphone application *Earapp* was developed for this experiment by Dr. Nick Clark and allowed the application of a processing delay in combination with the simulation of a simple linear hearing aid for the Apple iPhone (models 4S or 5). The *Earapp* allowed to measure the roundtrip delay of the iPhone using a cross-correlation technique and to add an additional amount of processing delay to achieve a controllable total delay of between 10 and 50 ms (depending on the iPhone model used). The total delay was confirmed with the additional delay applied by again using the in-built function for measuring the total roundtrip delay. The *Earapp* recorded the input sound via the internal microphone of the iPhone, processed the sound with a noise gate function (blocking sounds with levels below 30 dB SPL), applied the additional processing delay and performed a non-linear output limiting function (maximum output level of 90 dB SPL) to avoid very loud sounds from being presented to the subjects and played the output sound via the audio output of the iPhone. There was a password-controlled subsection of the *Earapp* that allowed to change the parameters of the processing (settings menu with control of additional delay, noise gate threshold etc.) that was not accessible to the subjects (see Figure 3.4). The interface used by the subjects consisted of a play button to start the processing, a running timer that showed the time that it was used for and a running average of the RMS level of the input sound (Figure 3.4). For the experiment, the *Earapp* was calibrated to give a comfortable sound level in combination with a set of in-ear earphones (Betron B630) that was given to the subjects, when using the mid-point of the iPhone audio-volume scale (five steps below maximum). This volume setting was recommended to be used by the subjects throughout the experiment, but could be changed if required in specific listening situations.



Additional Latency (ms): 12.7

Noise Gate Threshold (dB SPL): 30

Limiter Threshold (dB SPL): 90

Input Gain (dB): 0

Latency Calibration

OK

20.0 ms

Figure **3.4** The smartphone application Earapp used for the listening experiment. The left panel shows the home-screen and the right panel shows the settings-menu of the app. The latency calibration was used to measure the roundtrip delay of the audio presentation and was used to obtain a total delay 20 ms in this case, using an additional delay of 12.7 ms and the internal delay of 7.3 ms (iPhone 4S).

Audiobooks

In addition to the *Earapp*, a preprocessed audiobook was given to the subjects to listen to during the acclimatisation period of the experiment. We used the audiobook *The Adventures of Tom Sawyer* written by Mark Twain, and spoken by a male, native-English speaker (obtained from Librivox.org). The audiobook was cut into 11-minute long segments and preprocessed to simulate the listening experience with a linear hearing aid. A copy of the audiobook was processed using a high-pass filter to simulate the frequency response of a hearing aid and another copy of the audiobook was processed using a low-pass filter to simulate the frequency response of the direct sound path. The two copies were added together with the simulated hearing aid signal being delayed in respect to the direct signal by a processing delay of either 20 or 40 ms for group 1 or group 2, respectively. The level-ratio of the two signals was set to 10 dB, with the hearing aid signal being 10 dB louder than the direct signal, to simulate the combined amplification of a hearing aid and the attenuation of an earmould in a simplified manner (in practice, hearing aids are likely to lead to larger level differences between the two soundpaths, depending on the frequency range). Four segments of the audiobook were given to the subjects at the beginning of the experiment with a total duration of 44 minutes.

Pre- and post test

The preprocessed audiobook was also used to generate the test stimuli for the pre- and post tests of subjective tolerance to processing delay. A 9-minute long segment of the audiobook, that was not given to the subjects, was used to generate twelve 45-s long test stimuli using 4 delay conditions (10, 20, 30 and 40 ms) and 3 level-ratio conditions (0, 10 and 20 dB) between the simulated direct and delayed soundpaths. This was repeated to generate a second set of 12 test stimuli in a similar manner to obtain a total test set of 24 stimuli that contained each condition twice. Stimuli were chosen to be of moderate length to give enough time for detailed listening to the sound quality, but to avoid acclimatisation effects within the presentation of each stimulus condition (short-term acclimatisation may occur after several minutes of listening).

The pre- and post test of subjective tolerance to processing delay consisted in the presentation of the 24 test stimuli in random order with each stimulus being rated by the subject on a 7-point Likert scale according to: "[...] how disturbing you would find it if you had to listen to it all day." (with 1 - *not disturbing at all*, 4 - *disturbing* and 7 - *highly disturbing* and using ten sub-steps per category yielding a rating-resolution of 0.1). The test part of the experiment took place in a quiet but not sound-isolated room at the University. The stimuli were generated beforehand using MATLAB and played back to the subject using a laptop (Dell Latitude E7440), an external soundcard (RME Babyface) and closed, circumaural headphones (AKG K240 MKII). The setup was calibrated to a presentation level of 65 dB(A) using a sound level meter (Bruel&Kjaer 2660) and artificial ear (Bruel&Kjaer 4153). In total, the pre- and post test lasted about 30 to 40 minutes each including a short break half-way of the tests to avoid listening fatigue.

Experimental procedure

The experiment lasted for 5 days in total. On the first day, PTA was performed and the subjects were instructed how to obtain the audiobooks. The *Earapp* was installed on their iPhones and they were explained how to use it (by starting the application, connecting the earphones and setting the correct volume). Subjects were told to hold the iPhone in one hand or to put it on a table in front of them while using the *Earapp*, and were instructed not to put the iPhone in a pocket or to cover the iPhone microphone in other ways. For Group 1, the processing delay of the iPhone was measured with the *Earapp* and an additional delay was applied to obtain a total delay of 20 ms. For Group 2, the same procedure was followed to obtain a total delay of 40 ms. To complete day one of the experiment, subjects performed the pre-test by listening to and rating their subjective disturbance for each of the 24 test stimuli.

For each of the following three days, the subjects were instructed to use the *Earapp* for a minimum of 30 minutes per day in realistic situations (such as in communicative situations when talking to friends or family, while being in lectures at University or when watching television). They were also required to listen to a 11-minute long episode of the audiobook on each of the three days. Subjects were given paper sheets to document the listening situations in which they used the *Earapp* and to rate their subjective disturbance after using the *Earapp* and listening to the audiobook episodes for each of the following three days.

At the fifth and final day of the experiment, subjects listened to another 11-minute long segment of the audiobook, with the processing settings chosen according to which group they belonged to (either 20 or 40 ms of delay), right before the post-test that was then completed by listening to and rating of the 24 test stimuli again.

Results

When comparing within-group effects of test (pre-/post), there were differences in mean disturbance ratings by about 1-2 rating units for the level-ratios of 0 and 10 dB between the pre- and post tests for group 1, but not for group 2. However, this difference between pre- and post test was not significant (likely because of the small samplesizes). The mean disturbance ratings for group 1 are shown in Fig. 3.5 and for group 2 in Fig. 3.6.

When comparing between-group effects of test (pre-/post), there were differences between groups in mean disturbance ratings for the post test, but not for the pre test. With averaged level-ratio conditions, the ratings for the post test were significantly higher by about 1-1.5 rating units for group 2 in comparison to group 1. The comparison between groups in terms of mean disturbance ratings for the pre- and post tests are shown in Fig. 3.7.

To visualize the effect of delay and level-ratio condition on the subjective disturbance ratings, contour plots for pre- and post tests and all ratings have been calculated and are shown in Fig. 3.8. For group 1, it can be seen that the contour level of tolerable delay (rating of 2.5-3, more green colour) spreads more into higher delay regions (20-25 ms) for the post test than for the pre test. This effect is non-existent for group 2 indicating that group 1 showed acclimatisation effects up to about 20-25 ms of delay, whereas group 2 did not show any acclimatisation effect (no obvious difference in tolerable delay contour levels between pre and post test).

Statistical analysis was performed for both within-group and between-group comparisons. For the within-group comparisons between pre- and post tests (Figures 3.5 and 3.6), a repeated measures ANOVA (within-subject factors of delay and level-ratio and between-subjects factor test) was performed for each group individually. For group 1, Mauchly's Test of Sphericity indicated a violation of sphericity for the factor delay and the Greenhouse-Geisser corrected test was used. There were significant main effects of delay [F(1.249,7.494) = 9.035, p = 0.015], level-ratio [F(2,12) = 16.376, p < 0.001] and a significant interaction between the two [F(6,36) = 5.428, p < 0.001]. There was no significant effect of test [F(1,6) = 1.241, p = 0.308]. For group 2, there were significant main effects of delay [F(3,18) = 10.886, p < 0.001] and level-ratio [F(2,12) = 18.023, p < 0.001] and significant interactions between the two [F(6,36) = 4.351, p = 0.002] and between the factors delay and test [F(3,18) = 3.462, p = 0.038]. There was no significant effect for the factor test [F(1,6) = 0.148, p = 0.714].

For the comparison of pre-test results between the groups (Figure 3.7, left), a repeated measures ANOVA (within-subject factors of delay and level-ratio and between-subjects factor group) was performed. There were significant main effects of delay [F(3,18) = 20.504, p < 0.001] and level-ratio [F(2,12) = 18.605, p < 0.001], but no significant interactions or main effect of group were found [F(1,6) = 0.031, p = 0.867]. For the comparison of post-test results between groups (Figure 3.7, right), a repeated measures ANOVA (within-subject factors of delay and level-ratio and between-subjects factor group) was performed. Mauchly's Test of Sphericity indicated a violation of sphericity for the factor delay and the Greenhouse-Geisser corrected test was used. There were significant main effects of level-ratio [F(2,12) = 15.895, p < 0.001] and group [F(1,6) = 7.665, p = 0.032], but no significant interactions were found.

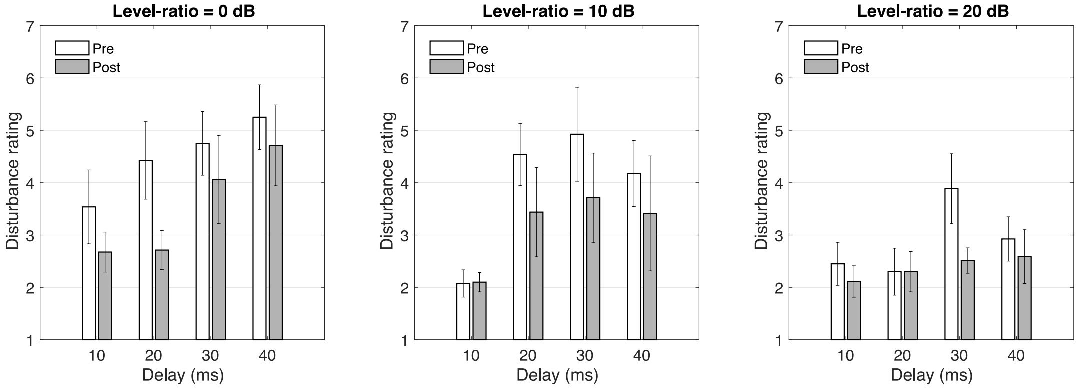


Figure 3.5 Comparison between pre- and post test mean ratings for group 1 for all test conditions (4 delays and 3 level-ratios). Errorbars show standard error of the mean.

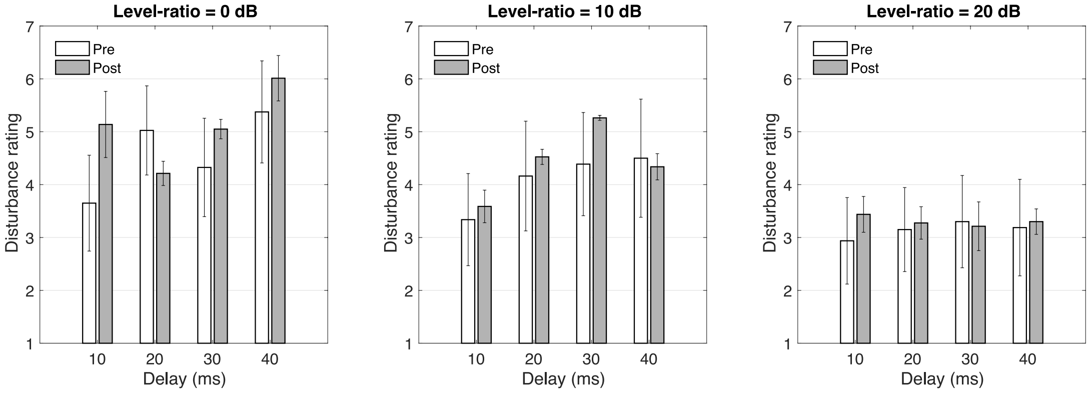


Figure 3.6 As Figure 2 but for group 2.

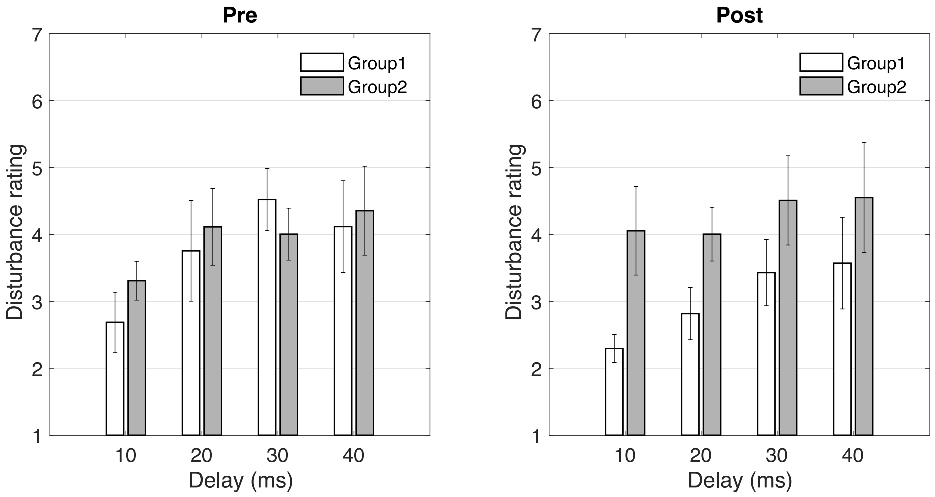


Figure **3.7** Comparison of mean disturbance ratings between groups. The left panel shows pre test results averaged for all level-ratios (0, 10 and 20 dB). The right panel shows post test results averaged for all level-ratios (0, 10 and 20 dB). Errorbars show standard error of the mean.

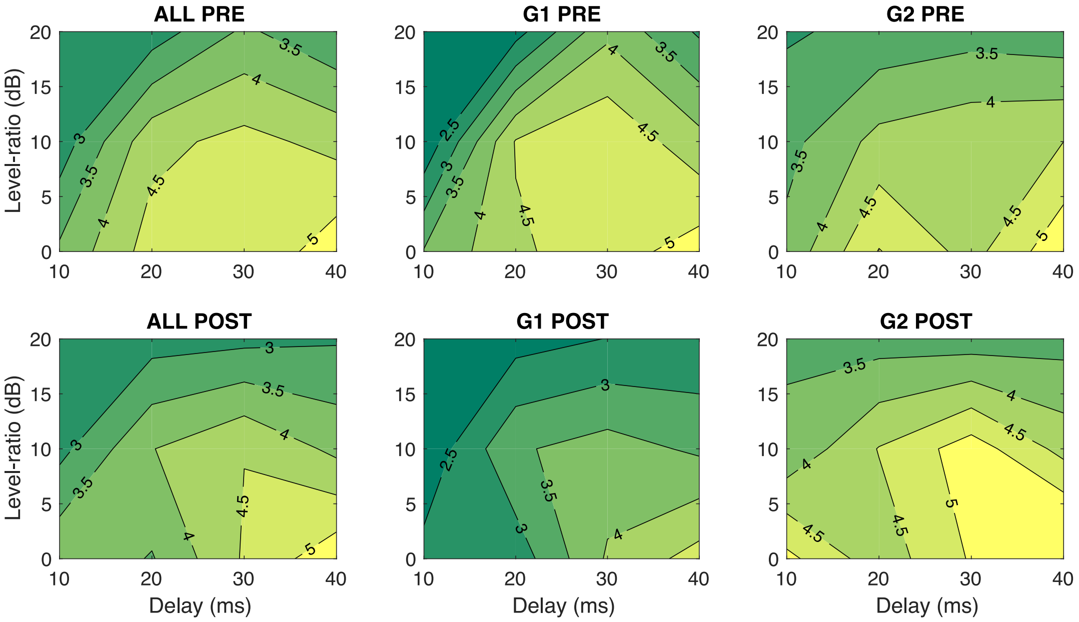


Figure **3.8** Contour plots for pre test (upper panel) and post test (lower panel) mean disturbance ratings. More green indicates lower disturbance ratings and more yellow indicates higher disturbance ratings (numbers on countour lines show mean disturbance ratings of contour level).

3.2.3 Discussion

This study investigated the effect of long-term acclimatisation on the tolerance of processing delay for normal hearing listeners. Two groups of listeners were tested on their subjective tolerance of processing delay before and after an acclimatisation period where they were exposed to the listening experience with a simulated hearing aid using a fixed processing delay for several hours over a five-day long period. Differences in disturbance ratings between groups for the listening test after the acclimatization period indicated a decrease in subjective disturbance to processing delay for the group that was exposed to a delay condition of 20 ms in comparison to the group that was exposed to a processing delay of 40 ms.

There were highly significant effects of level-ratio and delay on the subjective disturbance in all tests ($p < 0.001$). This was expected as both parameters alter the magnitude of comb-filtering effects that change the sound quality of the speech stimuli. The larger the absolute level-ratio and the smaller the delay between the direct and delayed soundpaths, the smaller the change in sound quality due to comb-filtering.

In terms of effects of long-term acclimatisation, the comparison between pre- and post tests for each group yielded no significant differences in subjective disturbance ratings for both groups. However, a larger difference was observed for group 1 than for group 2 at level-ratios of 0 and 10 dB and it can be speculated that this trend may reach statistical significance with a larger sample size (only 4 subjects per group were tested). Interestingly, when comparing subjective ratings between groups, there was a significant difference between group ratings for the post test but not for the pre test. In the post test, group 1 gave significantly lower ratings for subjective disturbance than group 2. This indicates that group 1 showed acclimatisation effects to processing delay but group 2 did not. This result can be explained by the different delay settings between groups that were used during the acclimatisation period. Group 1 listened to audiobooks and used the *Earapp* with a processing delay of 20 ms, that is supposed to be more tolerable than the processing delay used for group 2 of 40 ms. Given these results, we can speculate that acclimatisation to processing delay is possible within certain limits. A delay length of 20 ms that is at the limit of tolerance for normal hearing listeners (as indicated by the experiment conducted in the previous chapter), may enable acclimatisation effects, whereas a delay of 40 ms that is far over the tolerance limit of normal hearing listeners prevents acclimatisation effects. The magnitude of the acclimatisation effect was about 1 rating unit that can be seen as an increase in delay tolerance of about 10 ms. This would yield comparable tolerance of processing delay by normal hearing listeners (group 1) between for example 10 ms before the acclimatisation period and 20 ms after the acclimatisation period. A potential increase in tolerable processing delay by about 100% for normal hearing listeners.

It was a big challenge to find participants that were willing to take part in a 5-day long experiment and that possessed an iPhone model 4S or 5. Consequently, the variability in the data collected in this study was very large due to the differences in subjective perception by the listeners and the small sample sizes in the two groups. In general, conclusions for the population of hearing aid users can not be drawn from these data. Firstly, the results were obtained with normal hearing listeners that are likely to have different perceptual requirements in terms of sound quality. Secondly, the tests of subjective tolerance to delay were performed using preprocessed stimuli and not a real-time setup as in the study in the previous chapter of this thesis, that would allow for more realistic perception of processing delay. This limits the validity of the data to a specific scenario of external voice conditions when listening to a male talker. No evaluation of own-voice perception or other acoustic stimuli was performed and it remains unclear if acclimatisation may have occured also to other stimulus types. Thirdly, exposure to sound quality of processing delay was limited to just about an hour for each of the three days during the experiment. This represents a very small ratio of the total time during the 5-day long experiment and is expected to reduce potential acclimatisation effects. Nevertheless, the findings support the hypothesis that an increase in tolerance to processing delay can be achieved by long-term acclimatisation, even with relatively short exposure to processing delay.

Future studies could improve on the limitations of this study by using hearing impaired listeners and real hearing devices that are worn for the full duration of the acclimatisation period of the experiment. To achieve this, hearing aids could be programmed with an increased delay setting and worn for a week-long period by their users. It seems likely that for hearing impaired listeners, who showed increased tolerance to processing delay compared to normal hearing listeners, further increases in delay tolerance may occur. This would be an interesting study, as it would allow to investigate if the increased tolerance to processing delay by hearing impaired listeners is due to acclimatisation effects by the use of hearing aids (HA experience) or due to different perceptual requirements determined by the amount of hearing loss as it was suggested in the study reported in the previous part of this chapter.

4 IMPROVING SPEECH PERCEPTION IN NOISE FOR USERS OF HEARING AIDS

Overview

Speech understanding in adverse acoustic environments is still a major problem for users of hearing aids. Recent studies on supervised speech segregation show good promise to alleviate this problem by separating speech-dominated from noise-dominated spectro-temporal regions with estimated time-frequency masks. In this chapter a study was performed that evaluated several noise reduction algorithms using 17 hearing impaired listeners by measuring their performance in terms of speech intelligibility in noise and their subjective preference ratings.

Algorithms under test included a conventional Wiener filter algorithm, a novel sparse-coding algorithm inspired by neural coding principles observed in the human brain and two deep *neural network based speech enhancement* (NNSE) algorithms. The two NNSE algorithms used two different acoustic feature sets to evaluate the sub-question of whether auditory-inspired front-end processing using a computational auditory model could provide benefits for speech enhancement processing. The performance of both feature extraction methods was evaluated with objective measurements and a subjective listening test was performed to evaluate speech perception of hearing-impaired listeners. Significant improvements in speech intelligibility and quality ratings were found for the sparse coding and NNSE algorithms and both feature extraction systems. The auditory-model based NNSE algorithm showed good performance overall indicating that auditory-model based processing could provide further improvements for supervised speech segregation systems and their potential applications in hearing aids. The NNSE algorithms used noise-specific neural networks that generalized to novel segments of the same noise type and worked over a range of SNRs. The proposed algorithm has the potential to improve the intelligibility of speech in noise for HA users while meeting the requirements of low computational complexity and processing delay for application in hearing aids.

4.1 Speech Enhancement Based on Neural Networks Using Auditory-Model Based Features for Hearing-Impaired Listeners

Declaration of authorship:

Tobias Goehring - Leading the research, responsible for the neural network based algorithms and wiener filter algorithm, objective prediction analysis and analysis of listening test results, writing the manuscript

Co-authorship:

Dr. Jessica Monaghan - Advised the research, developed the software for the listening experiment, responsible for the development of the sparse coding algorithm, writing the manuscript

Xin Yang - Helped with the development of the feature extraction part of the neural networks

Federico Bolner - Helped with the software development

Shangquo Wang - Performed the listening experiment

Prof. Stefan Bleeck - Advised the research

Parts of this research have been published (first authors are underlined):

Monaghan, J., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. and Bleeck, S. (2017). Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners. *Journal of the Acoustical Society of America*. 141(3), 1985-1998.

4.1.1 Introduction

Individuals with hearing impairment often have difficulty recognizing speech in background noise. In a UK survey of individuals fitted with a hearing aid (HA), a quarter of those who reported never wearing their aids indicated 'lack of benefit in noisy situations' as their reason for not doing so (Kochkin, 2000). Together with the finding that HA users are more tolerant of background noise than are hearing-impaired (HI) individuals who do not choose to use aids (Nabelek et al., 2006), this suggests that solving the problem of background noise could allow more HI people to benefit from HAs. One means of reducing the detrimental effect of noise on recognizing speech is to employ speech-enhancement algorithms (sometimes referred to as noise-reduction algorithms) to improve intelligibility. Single-channel speech-enhancement algorithms operate on the input from a single microphone and are therefore ideally suited to being incorporated into HA processing.

Traditional approaches to single-channel speech enhancement have demonstrated variable degrees of success. For some noise conditions, the "auditory masked threshold noise suppression" technique (Tsoukalas et al., 1997) increased recognition for both HI and NH listeners (Arehart et al., 2003), whilst a sparse-code shrinkage algorithm tested in our laboratory improved speech intelligibility in speech-shaped noise (Sang et al., 2014) and quality in speech-shaped and babble noise (Sang et al., 2015) for HI listeners. Other studies reported no benefit to word recognition for HI listeners with single-channel enhancement algorithms, but did report an increase in listener preference (Bentler et al., 2008; Luts et al., 2010; Zakis et al., 2009), including increased acceptable background noise level for HI listeners (Fredelake et al., 2012; Mueller et al., 2006).

Recently, machine-learning approaches have shown great promise for improving speech intelligibility both for hearing-impaired and normal-hearing listeners (Bolner et al., 2016; Healy et al., 2015, 2013). Rather than calculating a gain function based on estimates of the speech and noise statistics from the incoming signal - the classical approach - machine-learning approaches incorporate prior knowledge of patterns of speech and noise to estimate the optimal gain function to be applied to the incoming signal. Gaussian mixture models have been used to improve speech intelligibility for normal hearing listeners (Kim et al., 2009) and for cochlear implant users (Hu and Loizou, 2010). Healy et al. (2013) demonstrated large improvements in speech intelligibility scores for both NH and HI listeners using a deep neural network (DNN) algorithm. The main limitations to these approaches, however, were the very large classification systems (256 mixtures / 128 sub-band networks), and the specificity of the training set required to achieve this level of performance. These studies used the same noise recordings for both the training and testing stages of the algorithm. A match between training and testing data is likely to overestimate the performance of the algorithm in unseen test conditions. May and Dau (2014) showed that the use of novel noise realizations for testing yielded a substantial decrease in estimation performance with a GMM-based system, such as the one used by Kim et al. (2009). More recently, it has been shown for both NH and HI listeners, that DNN-based algorithms can generalize well to novel realizations of the same noise type (Bolner et al., 2016; Healy et al., 2015) or to completely novel types of noise (Chen et al., 2016).

For potential application in hearing devices such as HAs, algorithms must fulfill the requirements of low computational complexity due to the restricted capacities of HAs in terms of memory and computational power (Löllmann and Vary, 2009) and low processing delay due to the perceptual requirements of HA users (Stone and Moore, 1999). For the class of feedforward neural network algorithms with fully connected layers, the number of units in the consecutive layers define the computational complexity of the whole system (each unit in a given layer is connected to each unit in the next layer via a weight parameter). While it is unclear what the current limits for applications in hearing devices in terms of memory and computational complexity are, large NN algorithms with millions of parameters that require powerful computing units such as graphics processing units (GPU) are unlikely to be implementable on mobile hearing devices with limited battery size in the near future. This motivates for a decrease in the size and complexity of the NNs (Bolner et al., 2016) to allow for potential real-time operation on mobile devices.

Another machine-learning technique, dictionary-based sparse coding, has been used successfully in image denoising (Elad and Aharon, 2006), but applications to speech enhancement have been scarce. This approach is attractive from a biomimetic perspective; evidence suggests a sparse representation of ecologically relevant features in both the auditory (DeWeese et al., 2003; Lewicki, 2002) and visual (Olshausen and Field, 1996) systems. Following the training stage, in which the algorithm learns a ‘dictionary’ of typical speech features from many examples of clean speech segments, an estimate of the clean signal is reconstructed using relatively few of these dictionary components (i.e., the representation is ‘sparse’). In this de-noising stage, if the noise is sufficiently dissimilar to the speech, the reconstructed signal preserves more of the speech information compared to the noise. A neural analogy for the dictionary would be a very large set of neurons, each responding to one specific speech component. If these neurons continued to respond to these particular speech components, even in a noisy background, this would provide robustness to coding speech in noise, since familiar, speech-like elements would be better represented than unfamiliar, noise-like components. In the current study we test a new dictionary based sparse-coding approach to see if it can improve speech intelligibility.

Here, we first assess the performance of neural networks with greatly reduced complexity that have shown promising results in our previous study with NH listeners (Bolner et al., 2016) to determine whether it is still possible to obtain improvements in speech intelligibility for HI listeners with more practically feasible algorithms than other studies used (e.g. Healy et al., 2015). We also assess the performance of a novel machine learning algorithm known as sparse coding, and compare it to both a classical approach, Wiener filtering, and to DNNs. Third, we determine the performance of the DNN approach when it derives its input from an auditory model, comparing its performance to that of an algorithm employing the standard spectrum-based feature vectors of previous studies. Finally, as well as speech recognition scores, we compare the performance of these three approaches in terms of their sound-quality ratings, which have not previously been assessed for algorithms employing neural networks or sparse coding. Each of the four algorithms was assessed in both stationary (speech-shaped) noise and multi-talker babble noise conditions and at signal-to-noise ratios (SNRs) of 0 and +4 dB.

4.1.2 Speech Enhancement algorithms

Wiener Filtering

Wiener filtering was one of the first noise reduction algorithms to be developed (Lim and Oppenheim, 1979), and has been implemented in commercial HAs. In order to obtain the noisy speech spectrum a short-time Fourier transform (STFT) is performed. The clean speech spectrum *X* is then estimated as $\hat{X}_k$ using the following equation:

$$\widehat{X_k} = \sqrt{\frac{\left|\widehat{S_k}\right|^2 - \left|\widehat{N_k}\right|^2}{\left|\widehat{S_k}\right|^2}} Y_k = \sqrt{\frac{\gamma_k - 1}{\gamma_k}} Y_k = \sqrt{\frac{\xi_k}{1 + \xi_k}} Y_k \, , \tag{1}$$

where ξ_k is the *a priori* SNR, γ_k is the *a posteriori* SNR, Y_k is the noisy signal magnitude, $\hat{S}_k$ is the estimated signal magnitude, $\widehat{N}_k$ is the estimated noise magnitude and k indexes the Fourier components. The estimate of the clean signal is derived by minimizing the difference between the clean and enhanced complex speech spectra, taking into account the phase spectra. The Wiener filter is the optimal estimator of the clean speech spectrum, when the speech and noise signals are independent Gaussian processes. Scalart and Filho (1996) reported that using an estimate of the a priori (rather than a posteriori) SNR in (1) would give superior enhancement. Their method was employed in the current study. The noise magnitude spectrum was estimated on a frame-by-frame basis using the algorithm of (Gerkmann and Hendriks, 2011) to estimate the a priori SNR and calculate the gain function for each frame and frequency component.

Hu and Loizou (2007) tested a number of single-channel speech enhancement algorithms and found that the Wiener filtering algorithm described by Scalart and Filho (1996) was the only algorithm that enhanced speech recognition for NH listeners, although this improvement was evident in only one condition (automobile noise at 5 dB SNR). Levitt et al. (1993) found that consonant recognition in non-stationary cafeteria babble noise was significantly increased for half of HI listeners but significantly reduced for half of NH listeners when a Wiener filter was applied. In that study, the gain of the filter was calculated by assuming knowledge of the consonant and noise spectra. This indicated that Wiener filtering can be beneficial for some HI listeners if the filter gain is approximated accurately enough. Luts et al. (2010) tested a Wiener filter algorithm that estimated the noise and speech spectral densities from the signal (in a different way from that employed in the current study) but found no improvements in the recognition of speech in babble noise by NH and HI listeners. Nevertheless, listeners preferred the enhanced speech over the unprocessed condition.

Neural Networks

The next two algorithms to be tested also employed the Wiener gain function to estimate the clean speech signal. The principal difference between these algorithms and the classical Wiener filtering algorithm described in the previous chapter was the use of a more sophisticated approach to estimate the Wiener filter gain, namely the use of an artificial neural network (NN) algorithm.

The NN algorithm consisted of two parts: a front-end that extracted acoustic features from the noisy input signal and a back-end that employed a multi-layer feedforward neural network to estimate the ideal Wiener filter gain (see equation 1) in each frequency channel. The estimated gain was used to enhance the noise-corrupted input signal by applying it to the noisy envelopes after the signal had been passed through a 63 channel gammatone filter bank ranging from 50 to 8000 Hz (Patterson et al., 1987; Hohmann, 2002). A schematic of the NN algorithm is shown in Figure 4.1.

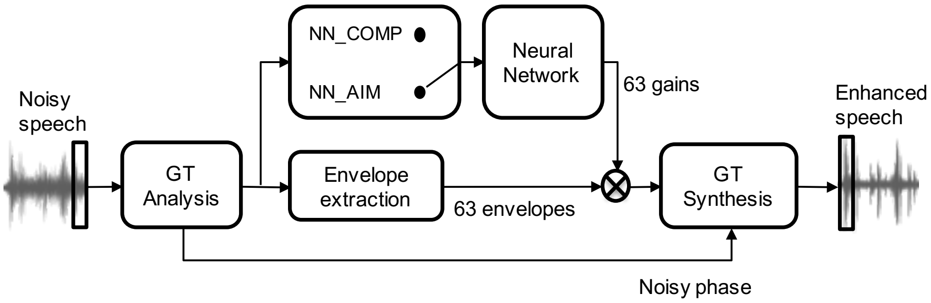


Figure 4.1 Schematic of the neural network based speech enhancement system using either NN_COMP or NN_AIM feature sets as input for the neural network regression algorithm.

The first processing stage for these algorithms was to split the input signal (fs = 16 kHz) into 20-ms long timeframes with 10 ms overlap. Then, two sets of acoustic features were extracted from the broadband signal of each input frame: a comparison feature set (NN_COMP) similar to those used in previous studies (Healy et al., 2013; 2015) and a novel auditory-model based feature set (NN_AIM). Both feature sets comprised several sub-features that were concatenated per timeframe and directly fed to the input layer of the NN. This yielded two distinct NN algorithms: NN_COMP and NN_AIM.

The second processing stage was performed by the NN that consisted of an input layer with a number of units determined by the dimensionality of the feature set, two hidden layers with 100 and 50 units using saturating linear transfer functions and an output layer with linear activations. The output layer had a dimensionality of 63 given by the number of gammatone frequency channels used for calculating the target Wiener filter gain function. The output layer activations of the NN were taken as the estimated Wiener filter gains and applied to the noisy envelopes. The NN was trained using the resilient backpropagation algorithm (Riedmiller and Braun, 1993) to minimize the mean squared error between the estimated and ideal Wiener filter gain in each gammatone frequency channel. The NN was trained in full-batch mode over 500 epochs using weight decay regularization of 0.5 to avoid overfitting. The learning rate was set to 0.01 and weights were updated using increment and decrement factors of 1.2 and 0.5, respectively. These hyper parameters were chosen based on our previous study (Bolner et al., 2016) that yielded improvements in speech perception in noise by NH listeners.

In total 80 sentences (8 lists) from the IEEE database (IEEE, 1969) spoken by a male talker were mixed at 5 SNRs (-2, 0, 2, 4, and 6 dB) to amount to 400 training utterances per noise condition. The training data sets were the same as used for the sparse coding algorithm. A single NN was trained per noise type incorporating all five SNR conditions. As mentioned above, the ideal Wiener filter gain was taken as target signal to be estimated by the NN for each training utterance in each gammatone channel. The target data was calculated using the ground-truth speech and noise signals at the given SNR.

One of the goals of the current study was to assess the performance of more real-time feasible NNs for speech enhancement. Kim et al. (2009) and Healy et al. (2013) used sub-band classifiers that employed two GMMs or NNs for each frequency channel yielding large classification systems. Healy et al. (2014; 2015) and Bolner et al. (2016) used a broadband approach that employed a single NN to estimate the target gains for all frequency channels collectively. This approach yielded a large decrease in NN parameters and computational complexity (a 43-fold increase in processing speed was reported by Healy et al.; 2014). The memory requirements and the number of calculations performed by the NN per timeframe are determined by the number of NN parameters consisting of the weight and bias values of the units in the hidden and output layers. In this study, the auditory-model based NN comprised 39800 parameters which is a 100-fold or 500-fold decrease in parameters in comparison to Healy et al. (2015) or Chen et al. (2016), respectively. Another aspect of real-time processing is the algorithmic processing delay that is limited to a few milliseconds by the perceptual requirements of users of hearing aids (Stone and Moore, 1999). As reported by Healy et al. (2015), the inclusion of future timeframes has to be avoided for real-time processing applications such as HAs. In contrast to two future frames in Healy et al. (2015) and 11 future frames in Chen et al. (2016), the current study used no future frames for the processing to assess a more real-time feasible approach.

Comparison feature set

The comparison feature set NN_COMP was generated based on the same set of features used in Healy et al. (2013; 2014; ‘complementary features’). To generate the feature set, the amplitude modulation spectrum (AMS; Tchorz and Kollmeier, 2003), relative-spectral transform and perceptual linear prediction coefficients (RASTA-PLP; Hermansky and Morgan, 1994), and mel-frequency cepstral coefficients (MFCC) were extracted from each 20-ms long timeframe of the noisy speech mixture (broadband features were computed as described in Healy et al., 2014). The concatenated features had a dimensionality of 445 per timeframe (AMS [25x15] + RASTA-PLP [3x13] + MFCC [31]). NN_COMP was extracted from the current timeframe and concatenated with delta (differences between features in consecutive frames) and delta-delta features for RASTA-PLP only (as described in Healy et al., 2014).

Auditory feature set

The proposed feature set NN_AIM was extracted using the auditory image model (AIM; Patterson et al., 1995; Bleeck et al., 2004). AIM is a time-domain functional model of auditory processing. It generates a stream of two-dimensional sound representations, referred to as 'auditory images', for an acoustic input signal. AIM produces a more stable representation for periodic parts of the input sound, such as for vowels and voiced sounds in speech and tones in music signals, than for non-periodic sounds. The model consists of a cascade of processing stages that simulate peripheral auditory processing, such as pre-cochlear processing, basilar membrane motion (BMM) and the transduction process in the cochlea. Further stages of AIM are intended to model more central auditory processing stages, such as neural activity patterns in the auditory nerve and cochlear nucleus and temporal integration and source size normalization in higher auditory processing stages (finally yielding the size-shape transformed auditory image; SSI). The SSI output of AIM is based on the size covariant processing of the auditory system (Smith et al., 2005) and produces the same pattern for vowels spoken by speakers with different glottal pulse rates or vocal tract lengths. The processing of AIM has been reported to improve the SNR of voiced speech and to yield improved performance in speech recognition experiments (Irino and Patterson, 2002; Monaghan et al., 2008; von Kriegstein et al., 2007).

The NN_AIM feature set combined the output of two processing stages of AIM: the BMM and SSI. The two features were concatenated to obtain a dimensionality of each feature vector of 315, consisting of 63 BMM features and 252 SSI features. The BMM features were obtained by calculating the logarithm of the envelope power of a linear gammatone filterbank with 63 frequency channels (Hohmann, 2002) and represented predominantly spectral information of the current timeframe. The SSI features were obtained by calculating a two-dimensional discrete cosine transform (DCT) of the SSI output of AIM. The DCT was performed for de-correlation and a reduction of the dimensionality of the SSI. Before the DCT was performed, each SSI channel was downsampled to 400 Hz to reduce the temporal resolution of the data. After performing the DCT using the downsampled signal, only the 2nd to 22nd coefficients were used for the NN_AIM feature set. The first coefficient was omitted since it is related to the overall energy of the SSI and more susceptible to noise degradation and the higher order coefficients above the 22nd were found to be numerically close to zero. The SSI represented both spectral and temporal information of the current timeframe in form of enhanced periodicity information and increased SNR for voiced components of speech signals. The NN_AIM feature set was extracted using only the current timeframe.

Objective measures

Two “objective measures” - computationally derived scores intended to predict how well humans will recognize a given sample of noisy or enhanced speech – were used to optimize the performance of the NN and sparse coding algorithms: the Short Time Objective Intelligibility (STOI; Taal et al., 2011) and the Normalized Covariance Metric (NCM; Holube and Kollmeier, 1996). Additionally, for the NN algorithms, hit – false alarms (HIT-FA) and false alarm (FA) rates were determined and used for optimization (Kim et al., 2009). These measures required the estimated gain function to be converted into a binary mask. The hit rate was defined as the percentage of speech-dominated time-frequency bins correctly classified by the binary mask and the false-alarm rate was defined as the percentage of noise-dominated time-frequency bins incorrectly classified as speech-dominated. During optimization of the algorithms, their performance was assessed using objective measure scores from two sentence lists that were not part of the training set or test set (the set used for the human testing). After testing with the human listeners had taken place objective measures were also applied to the sentences in the test set to determine the correlation between these measures and the human performance (see later).

Sparse Coding

The fourth algorithm was a novel speech enhancement algorithm based on dictionary-based sparse coding (Elad and Aharon, 2006). This algorithm was also a machine-learning algorithm but involved a different approach from the NN based algorithms. Rather than estimating a gain function to be applied to the noisy signal (as with the other algorithms tested here), an estimate of the clean filter bank outputs was produced directly.

The algorithm requires a 'dictionary' of typical elements of speech, known as 'atoms'. This dictionary is learned from many frames of clean speech during the training stage. The dictionary is typically over-complete, i.e. the number of atoms in the dictionary is greater than the length of the atoms. Any speech signal can then be approximated by a linear combination of just a few atoms from the dictionary, i.e. it is a 'sparse' representation. Because stationary noise is unstructured, and therefore cannot be predicted, it cannot be sparsely represented. Therefore, for noisy speech, the speech signal can more easily be approximated in the form of a sparse code than can the noise, leading to de-noising.

For a noisy speech frame, **y**, consisting of noise **n**, and clean speech **x**:

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \qquad (2)$$

It is assumed that the clean speech can be represented as:

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} \qquad (3)$$

Where the matrix **D** is a dictionary and **α** is a sparse coefficient vector (i.e., most entries are zero).

The estimate of the clean speech is then given by:

$$\hat{\mathbf{x}} = \mathbf{D}\hat{\boldsymbol{\alpha}}\,, \qquad (4)$$

where

$$\hat{\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \qquad (5)$$

such that

$$\|\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\alpha}}\|_2{}^2 < \varepsilon\,. \qquad (6)$$

The zero norm of $\boldsymbol{\alpha}$, $\|\boldsymbol{\alpha}\|_0$, is the number of non-zero elements in $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ is the desired error, which is chosen to be approximately equal to the estimated noise power.

The signals were processed with the same sampling rate, frame length and using the same gammatone filter as in the NN algorithm except that 30 channels were used rather than 63 channels. Increasing the number of channels for the sparse-coding algorithm to 63 did not improve performance (as assessed using the objective measures described in II.B.1) but greatly increased processing time. Each channel of the filter output was normalized to have a root mean square (RMS) amplitude of 1. In the training stage, the dictionary was trained on eight sentence lists from the same speaker and corpus as used in the testing stage. The K-singular value decomposition (KSVD) algorithm (Aharon et al., 2006) was used to train a dictionary for each channel. The Orthogonal Matching Pursuit (OMP) algorithm (Pati et al., 1993) was modified so that the selection of atoms was optimized across all frequency channels. The first stage of the original OMP finds the atom from the dictionary that gives the highest correlation with the noisy signal. Rather than selecting atoms independently for each frequency channel and dictionary, the atom that gave the highest correlation over each frequency channel and dictionary was selected. The corresponding atom from each dictionary was chosen for the other frequency channels. This was intended to capture across frequency correlations in the speech. Five atoms per frame were chosen as the optimal number for training the dictionary during pilot tests.

For the testing stage the average noise power in each channel was estimated using the approach of Gerkmann and Hendriks (2011) and used to define the desired error in each channel ($\boldsymbol{\varepsilon}$). In the denoising stage the least angle regression algorithm (LARS) algorithm (Efron et al., 2004) was used rather than OMP because it was found to give superior performance in terms of both objective measures. For each frame, atoms were selected until the sum over channels of the RMS difference between the noisy and sparse signals was less than the sum of the desired errors for each channel.

A separate approach was used for the babble noise, as proposed by (Sigg et al., 2012) because in this case the noise and speech are more similar and so the representation of the noise might also be sparse over the speech dictionary. Therefore, in the training stage, in addition to the speech dictionary, a noise dictionary was trained using an example of the babble noise (distinct from the noise segment used in testing). As for the speech dictionary, the KSVD with modified OMP was used to train the noise dictionary. To reduce the similarity of the noise and speech dictionaries (and thus the probability of speech components being misclassified as noise), any atom with a correlation greater than 0.95 was removed from the noise dictionary. In the testing stage the noise and speech dictionaries were concatenated and the LARS algorithm was used to find a fit to the noisy speech.

Atoms selected from the noisy dictionary were discarded and only atoms from the speech dictionary were used to reconstruct the signal. The values of the free parameters were optimized using objective measure scores from a sentence list not used in the training or testing sets.

A similar approach proposed by Sigg et al. (2012) operated instead on the STFT domain. Rather than using the reconstructed speech signal directly (as in the current study) Sigg et al. used it to estimate the speech and noise magnitude to calculate the Wiener filter gain (see equation 1) which was then applied to the original noisy speech. An objective measure of speech quality (the cepstral distance) showed improvement for their approach relative to multiband spectral subtraction (Kamath and Loizou, 2002) and a vector quantization based approach, but tests with human listeners were not performed.

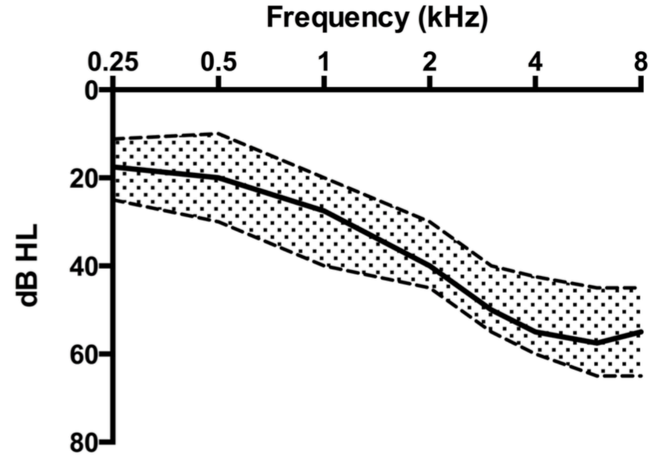


Figure 4.2 Median audiogram of the participants (solid line) and the interquartile range (shaded area).

4.1.3 Methods

Seventeen native speakers of British English (7 female, median age 65 years, IQR 13 years) with mild to moderate sensorineural hearing loss were recruited. Volunteers were recruited using advertisements at the University of Southampton and the Southampton local community, such as churches and libraries. A screening process was performed and participants were excluded from the study if they failed the screening. As part of the screening, otoscopy and tympanometry were performed to check for normal ear-canal anatomy and normal middle ear function. A questionnaire was given to exclude any recent ear surgery, otalgia, tinnitus and hyperacusis and pure tone audiometry was performed. Participants with unilateral, asymmetric or conductive hearing loss were excluded. All participants were experienced hearing aid users (>1 year use). The audiograms of each of the participants are shown in Fig. 4.2.

Speech recognition in each condition was assessed as the percentage of keywords identified correctly in IEEE (Rothauser et al., 1969) sentences spoken by a British male speaker. Two types of noise were tested: speech-shaped noise (SSN) and multi-talker babble noise. In the SSN conditions, a noise generated to have the same long-term average spectrum as the IEEE sentences was used. For the babble noise conditions, the noise was constructed by mixing different sentences from eight speakers (4 male and 4 female) taken from the TIMIT corpus (Garofolo et al., 1993). Both the SSN and the babble noise were 26 s in duration, 18 s of which was used to train the algorithms and the remaining 8 s of which was used in the testing stage. A segment of noise was selected at random from the test noise and added to the sentences at SNRs of 0, and +4 dB. Fixed SNRs were used rather than using an adaptive procedure to find the speech reception threshold (SRT; the SNR for which the speech recognition score is 50%), because the goal was to better compare the performance of the algorithms at a specified SNR. Two sentence lists, each comprising 10 sentences, were used for each condition (each list was used only once per participant). A different, and random, order of conditions was used for each participant, and a Latin square was employed so that the same list would be used in the same condition as seldom as possible. Participants practiced the procedure with a different sentence list from the ones tested, with an SNR of +10 dB and no speech enhancement applied.

Custom Matlab software was used to process and present the stimuli. Pre-processed sentences were loaded using a laptop computer and presented to the participant, who was seated in a quiet room in the clinic, through an RME Babyface soundcard over headphones (HD 380 Pro, Sennheiser, Wedemark, Germany). A finite impulse response headphone filter was designed in Matlab so that the headphones produced a flat frequency response at the ear reference point (measured using a Brüel & Kjær type 4153 artificial ear with the standard cone YJ0304 above the adapter plate for circumaural headphones, type DB 0843). The spectrum and sound level were measured using a Brüel & Kjær type 2250 sound level meter. The stimuli were presented monaurally to the participant's better ear, which was the left ear for nine of the participants. In order to compensate partially for each participant's hearing loss, a linear hearing-loss dependent gain was applied at each audiometric frequency according to the NAL-R prescription formula (Byrne and Dillon, 1986). There was one experimental session lasting approximately two hours. Participants were able to have rest breaks if they felt fatigued.

Noisy sentences were generated by setting the level of the clean speech to 65 dB SPL and adding noise scaled to give an SNR of 0 or +4 dB. Before amplification was applied, the level of the stimulus (the speech and noise mixture) was approximately 68 dB SPL in the unenhanced 0 dB SNR condition and 66 dB SPL in the +4 dB SNR condition. These noisy sentences were processed by each of the four enhancement algorithms and the corresponding enhanced sentences stored. Because the maximum gain that can be applied in enhancement is unity, enhancement will generally result in attenuation of the speech energy as well as the noise. This may reduce the audibility of the speech and render speech enhancement less effective. So that the level of the speech was unchanged between the enhanced and unenhanced conditions, "shadow-filtering" was used as described in Fredelake et al. (2012) for the Wiener filter and Neural Network conditions: the attenuation applied to the speech signal was determined by multiplying the clean speech by the same gain function that was applied to the noisy speech and measuring the corresponding reduction in RMS level relative to the original speech. In the case of the sparse-coding algorithm, there was no gain function applied so instead the reconstructed signal was set to 65 dB SPL, the level of the clean speech.

The experimenter scored each sentence list and condition using a graphical user interface (GUI) without knowledge of which condition was being presented. After each sentence was presented, the participant was asked to repeat what they had heard as accurately as possible. Using the scoring GUI, the experimenter recorded how many of the keywords the participant had identified correctly. After each sentence list, the participant was asked to rate the perceived quality of the speech ("How would you rate the quality of the speech?"). A paper sheet was provided on which the participant indicated the rating on a scale from 0 to 7 (with labels at 0, "bad"; 4, "fair"; and 7, "excellent"). For finer resolution, there were 10 subdivisions for each of its seven values.

4.1.4 Results

Speech recognition

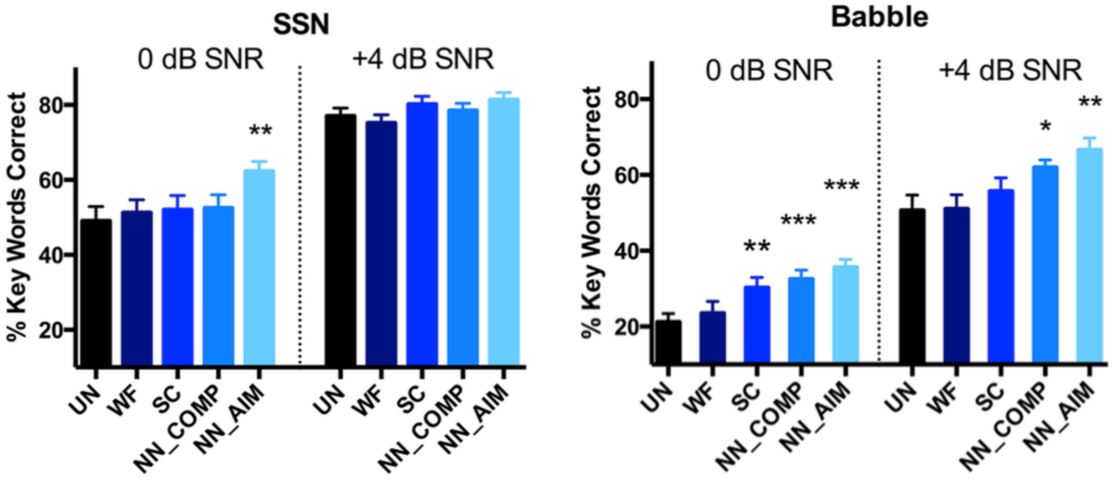


Figure 4.3 Group-mean percentage of key words correctly recognised for each algorithm (WF, SC, NN_COMP and NN_AIM) and the unenhanced condition (UN) in the speech-shaped noise (upper panel) and babble noise (lower panel) conditions. Error bars show standard errors of the mean. Asterisks indicate conditions for which the enhanced scores were significantly different from the unenhanced condition.

Speech intelligibility in speech shaped and multi-talker babble noise at SNRs of 0 and +4 dB was determined for four speech enhancement algorithms and compared to the corresponding unenhanced conditions. Figure 4.3 shows the group-mean percentage of key words correctly recognized in each of the four noise conditions. Across all algorithms, performance improved with increasing SNR, and performance was always lower in babble noise than in SSN. In babble noise, all of the algorithms, other than Wiener filtering, improved performance at least at one SNR, compared to the unprocessed condition. In the SSN conditions, there were significant main effects of algorithm, as determined by a repeated measures two way ANOVA [F(1,16) = 126.88, $p < 0.001$], and SNR [F(4,64) = 5.89, $p < 0.001$], but no significant interaction between the two [F(4,64) = 2.22, $p = 0.077$]. Bonferroni-corrected planned comparisons were performed between the unprocessed and enhanced conditions at each SNR. The only significant improvement in speech recognition for SSN was at 0 dB SNR for the NN_AIM algorithm [F(1,16) = 17.20, p = 0.003]. In the babble noise conditions, there were significant main effects of algorithm [F(1,16) = 17.10, p < 0.001] and SNR [F(4,64) = 323.85, p < 0.001]. The Greenhouse-Geisser correction was applied when testing the interaction between SNR and algorithm since the assumption of sphericity was violated in this case. The interaction was not significant [F(36.27,2.27) = 0.84, p = 0.45]. Bonferroni-corrected planned comparisons were performed between the unprocessed and enhanced conditions at each SNR. The sparse-coding algorithm led to a significant improvement in speech recognition at 0 dB SNR [F(1,16) = 17.37, p = 0.003], as did the NN_COMP [F(1,16) = 47.56, p < 0.001], and NN_AIM [F(1,16) = 114.32, p < 0.001]. At +4 dB SNR there were significant improvements in speech recognition scores for both NN_COMP [F(1,16) = 11.95, p = 0.013], and NN_AIM [F(1,16) = 18.64, p = 0.002].

Speech quality

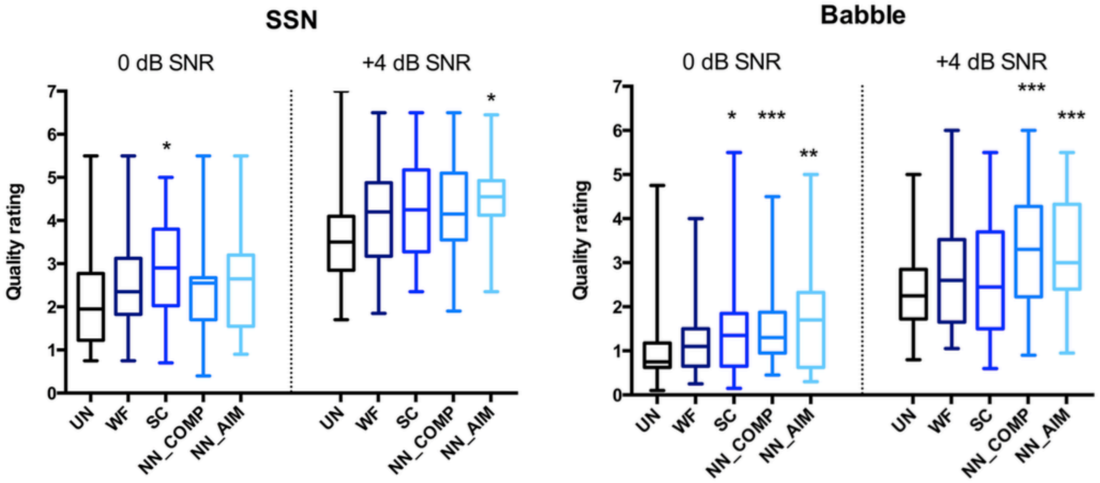


Figure 4.4 Box-and-whisker plots of speech quality ratings for each algorithm in the speech-shaped noise (upper panel) and babble noise (lower panel) conditions. Whiskers indicate the range (1.5 times the interquartile range). Asterisks indicate conditions for which the enhanced scores were significantly different from the unenhanced condition.

Speech-quality ratings were also determined for each algorithm in each noise condition and for both 0 and +4 dB SNR and are plotted in Figure 4.4. The data were not normally distributed in the majority of conditions, so box and whisker plots are shown and non-parametric statistics were used. As for speech intelligibility, speech quality improved at the higher SNR, and the algorithms elicited greater improvements in babble noise compared to SSN. Wiener filtering was ineffectual in improving speech quality. A non-parametric Friedman's ANOVA indicated a significant effect of algorithm for the SSN at +4 dB and the babble noise at 0 and +4 dB SNR. Bonferroni corrected planned comparisons were performed between the unprocessed and enhanced conditions at each SNR. In the SSN conditions there was significant improvement in quality ratings for sparse coding at 0 dB SNR [F(1,16) = 10.25, p = 0.022] and for NN_AIM [F(1,16) = 11.54 , p = 0.015] at +4 dB SNR. In the babble conditions there were significant improvements at 0 dB SNR for sparse coding [F(1,16) =

10.31, p = 0.022], NN_COMP [F(1,16) = 19.98, p = 0.002], and NN_AIM [F(1,16) = 15.07, p = 0.005] and at +4 dB SNR for NN_COMP [F(1,16) = 33.38, p < 0.001], and NN_AIM [F(1,16) = 24.07, p < 0.001].

Objective Measures

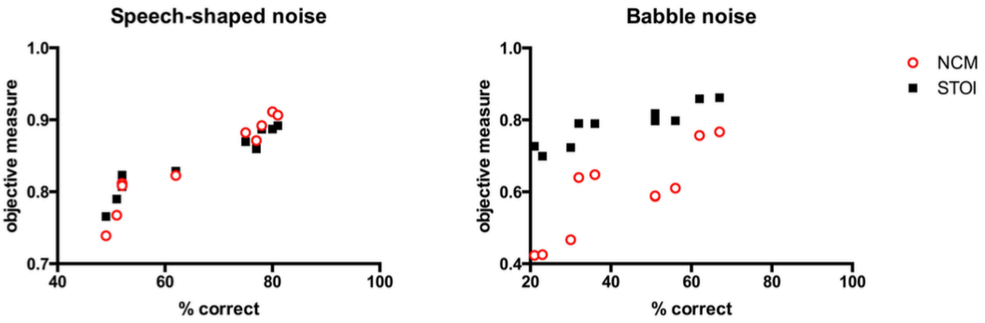


Figure 4.5 Average values of the objective measures NCM and STOI plotted as a function of the mean intelligibility scores obtained from the participants in each of the ten conditions (five algorithm conditions, two SNRs) for each noise type.

Objective measures were used in the optimization stages of the neural networks and sparse coding to help determine those parameters that would produce the best performance. In each of the four noise conditions NCM and STOI scores were calculated for the four algorithms and the unenhanced signals. Figure 4.5 shows the objective measures score plotted as a function of the final intelligibility scores obtained from the participants in the 20 conditions tested. In the SSN conditions, correlations between speech recognition scores and the objective measures were high, with r^2 values of 0.91 for both NCM and STOI. Correlations were lower in the babble noise conditions, with r^2 values of 0.70 and 0.83 for NCM and STOI, respectively. For the two neural-network algorithms HIT-FA scores were also calculated and are shown in Table 4.1.

Table 4.1 HIT-FA scores for NN_COMP and NN_AIM based algorithms.

HIT-FA (FA)	SSN		BABBLE	
SNR	0 dB	4 dB	0 dB	4 dB
NN_COMP	72 (8)	75 (7)	64 (18)	65 (17)
NN_AIM	76 (7)	79 (7)	67 (18)	67 (18)

4.1.5 Discussion

We assessed the performance of four speech enhancement algorithms in improving speech intelligibility and speech quality in two types of interfering noise, speech-shaped noise (i.e. noise with the same long-term spectrum as speech) and 8-talker babble noise. Algorithms based on sparse coding or neural networks improved performance compared to the unprocessed signal, most notably in babble noise and for the lower of the two SNRs (0 dB) we explored. Wiener filtering – commonly applied in hearing technologies – had no effect on speech intelligibility and speech quality compared to the unprocessed signal. This suggests that machine-learning algorithms, particularly those based on neuro-mimetic principles, can improve speech-in-noise performance in challenging listening conditions with fluctuating background noise.

Improvement in speech intelligibility was modest in the SSN conditions, with only significant improvement apparent for NN_AIM at 0 dB SNR, for which an improvement of 13% was evident. Subjective listening indicated that the absence of any improvement in the 0-dB condition may be due to the introduction of fluctuating distortions in the signal counteracting the beneficial effect of noise reduction. Interestingly, greater improvements were seen in babble noise conditions where algorithms typically perform worse than in stationary noise conditions. In this case, significant improvements were seen for all three machine-learning algorithms at 0 dB SNR and for both neural-network-based algorithms at +4 dB SNR.

A similar pattern of results was seen for the speech quality ratings, except that there was a significant improvement in speech quality for the sparse coding relative to the unprocessed speech in SSN at 0 dB SNR, but no significant improvement for the NN_AIM algorithm.

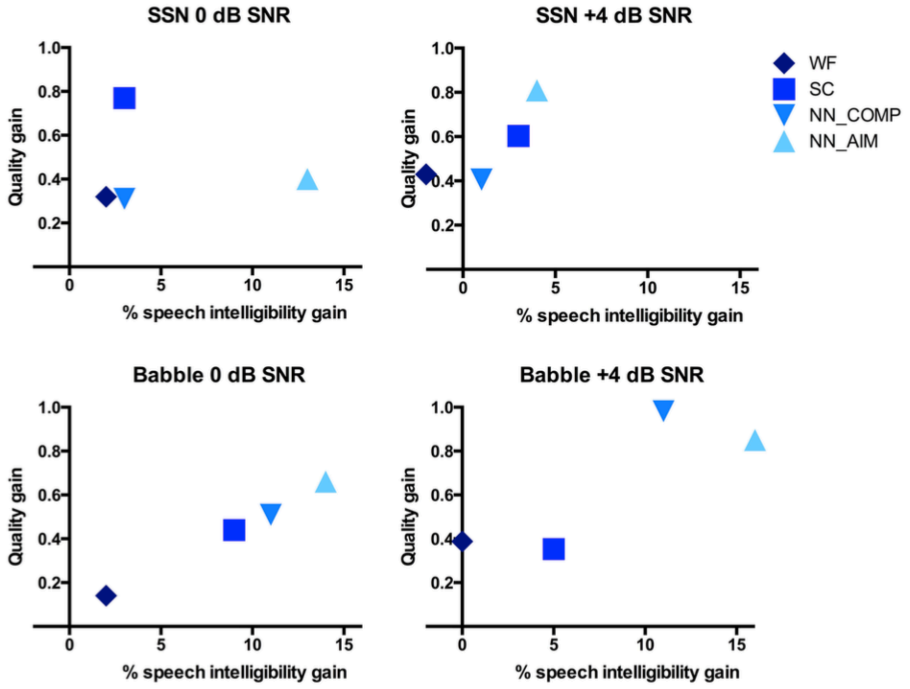


Figure 4.6 Group-mean improvement in speech quality versus improvement in speech intelligibility for the four algorithms in each noise condition.

Figure 4.6 shows the group-mean improvement provided by each algorithm for each of the conditions tested plotted in terms of gain in speech quality as a function of gain in speech intelligibility. Promisingly, almost all algorithms elicited improvements in both quality and intelligibility (albeit not significantly in many cases). Exceptions to this are the Wiener filter at +4 dB SNR for both noises, for which there was a reduction in speech intelligibility despite a small increase in speech quality ratings.

Neural Network based algorithms

Comparison with other studies

Healy et al. (2015) found improvements of 44.4% and 27.8% in babble noise at 0 dB SNR and +5 dB SNR, respectively. In the current study, improvements of 14% and 16% were found for the NN_AIM in babble, at 0 dB SNR and +4 dB SNR, respectively. Although these effects are smaller than those reported by Healy et al. (2015), note that participants in the current study had milder hearing losses, with an average PTA of 31.4 dB compared to 50.5 dB in Healy et al. (2015). This means that algorithm performance across the two populations cannot be completely equated. It has been reported by Healy et al. (2015) that improvements in SI were found to be smaller for NH listeners than for HI listeners. Thus the lower improvement in SI found in the current study may partly be explained by the milder hearing losses of the participants.

To compare algorithms in terms of classification accuracy, Healy et al. reported HIT-FA scores of 78% and 79%, respectively, for babble noise at 0 dB SNR and +5 dB SNR. The neural networks employed in our study achieved lower HIT-FA scores than did those employed by Healy et al., from 64% to 67%, respectively, for babble noise at 0 and +4 dB SNR (see table I). Whereas HIT rates are high for both approaches (over 80%), worse performance in terms of FA rates lead to lower HIT-FA scores achieved by the neural networks employed in the current study. This indicates a lower estimation quality of the masks compromising the removal of background noise. Both studies use ratio masking instead of binary masking, for which predictions drawn from HIT-FA rates will be less accurate. However, consistent decreases in speech intelligibility improvements have been found in respect to Healy et al.'s system as predicted by the HIT-FA metric.

Healy et al. also reported improvements in STOI scores for babble noise of 0.13 and 0.19 for 0 dB SNR and +5 dB SNR, respectively. The NN_COMP and NN_AIM algorithms used in this study both achieved STOI improvements in babble noise of 0.07 and 0.05 in 0 dB SNR and +4 dB SNR, respectively. Thus, STOI scores somewhat predict the difference in improvement of percentage correct scores between the two systems, although the magnitude of this difference is not accurately estimated by STOI.

One major difference between the current approach and that of Healy et al. was the size of the neural networks and the training dataset employed. The networks in the current study had only two hidden layers with 100 and 50 units, whereas Healy et al. (2015) used four hidden layer of 1024 units each. The networks used in this study thus contained 39,863 parameters (degrees of freedom) with the NN_AIM feature set (315 input / 63 output units). Assuming the same input and output dimensionalities, networks sized as in Healy et al. (2015) would comprise 3,536,959 parameters. This difference in network complexity by nearly a factor of 100 explains partly the performance advantage of the networks used in that study. Besides more powerful computational resources, this increase in degrees of freedom requires a much larger number of data points for training. In order to increase the number of training data and to improve generalization performance, Healy et al. (2015)

mixed each of the 560 training sentences with 50 different noise snippets from an 8-min long noise recording. Additionally, Healy et al. (2015) used a noise perturbation technique to further increase the variation of noise characteristics within the training dataset. The fact that the number of training data used by Healy et al. (2015) was greater by a factor of more than 50 compared to the training dataset used in the current study allowed the training of networks of much larger size without the risk of overfitting the training data. Both increases in network and training data size likely resulted in a more powerful regression model and better estimation results. Nevertheless, current hearing devices such as hearing aids have strongly limited capacities in terms of memory and computational power and may not allow for the implementation of algorithms that are computationally too complex.

Another major difference between the current system and that of Healy et al. is the causality of input and output data. The current study uses only the current and past frames as input signals, since this would be the case in real-time implementations. Healy et al. (2015) used five consecutive frames (two past, the current, two future) and Chen et al. (2016) used 23 consecutive frames (11, past, the current and 11 future) for input features as well as target predictions. The inclusion of future frames would introduce large through-put delays (i.e. > 20 ms) that most likely would not be tolerated by users of hearing aids (Stone and Moore, 1999).

To summarize, one of the aims of the current study was to investigate whether a smaller system, which potentially could be implemented in mobile devices in real-time, can still deliver benefits in terms of speech intelligibility and quality improvement. The findings support the results of Healy et al. and Chen et al. (2016) and demonstrate that significant (albeit more modest) improvements in speech intelligibility and quality can be provided by scaled-down neural network approaches that operate in a causal way. However, considering the modification that we made to the network, our data may give a more realistic idea of what performance can be expected in a real-time implementation on a mobile device.

Comparison of comparison and auditory feature sets

In addition to assessing the speech enhancement performance of neural networks with lower complexity, a further goal of the study was to determine whether using feature vectors derived from an auditory model would improve speech enhancement relative to standard feature vectors. Two sets of feature vectors were assessed using the same model architecture: a set derived from an auditory model, NN_AIM, and a standard feature vector set for comparison, NN_COMP.

Although no significant difference was found between intelligibility scores or quality ratings for the two sets of feature vectors, the AIM features gave the highest scores in both dimensions in almost all conditions (see Fig. 5). Indeed, it was the only algorithm tested that generated an improvement in the stationary noise conditions. This consistent overall better performance suggests that the auditory model based features are to be preferred in terms of optimizing speech quality and intelligibility. However, for use in real-time mobile devices other considerations must be taken into account, primarily the amount of computational power required to perform the algorithms, and the ability to generate the features in real-time. Although the generation of AIM features requires considerably more computational complexity than standard spectral features, they can be generated in real time by a modern PC.

A further potential benefit of using AIM feature vectors, not assessed by the current study, is their ability to generalize to different talkers. In the current study, and in previous studies employing neural networks based speech-enhancement techniques (e.g. Healy et al., 2013; 2014; 2015), networks were trained and tested on the same talkers. Performance of the system with a novel talker is likely to be decreased. Although a noise reduction system optimized for a particular talker has practical applications, widespread adoption of NN-based speech enhancement will require generalization to novel talkers and novel listening-situations. Unlike traditional feature vectors, such as MFCCs (Monaghan et al., 2008), the AIM features are robust to changes in speaker-size as well as pitch and so provide a better prospect for good performance with novel speakers.

Speech quality ratings

Most studies of speech-enhancement algorithms have concentrated on improvements in perceived speech-quality in the absence of improvements in speech intelligibility provided by traditional methods. Conversely, the recent studies involving NN-based speech enhancement only assessed intelligibility. Nevertheless, it is important for its general acceptance in hearing devices that an algorithm provides good speech quality (Kochkin, 2010). The results indicate that neural networks produced significant improvements in speech quality in all conditions for which the speech intelligibility was also significantly improved. One factor in the good quality ratings seen here may be the use of the Wiener filter gain which has been shown to produce better speech quality than the binary mask (Madhu et al., 2013) which is often used for neural network speech enhancement. However, speech quality scores in all conditions were similar or higher than those for traditional Wiener filtering. Since the gain function used by the neural network approaches used is identical to that used in WF, this indicates that the greater accuracy of speech and noise estimates provided by the neural network is crucial to the quality of the enhanced speech. These speech quality results support the promise of neural networks as a good candidate for speech enhancement for hearing aids.

Dictionary-based sparse-coding algorithm

An additional goal of this study was to assess the performance of a novel dictionary-based, sparse-coding algorithm. Overall the performance of the sparse-coding algorithm was similar to that of the NN_COMP algorithm except in babble noise at +4 dB SNR and stationary noise at 0 dB SNR. At 0 dB SNR, sparse coding was the only algorithm that generated a significant improvement in speech quality.

A disadvantage of the dictionary-based sparse-coding approach is the relative computational complexity of the de-noising stage. In the neural network approach, after the network is trained, its application in the de-noising stage is straightforward, with the same non-linear formula being applied for each frame to determine the gain. In the case of sparse coding, however, the de-noising stage still requires a sparse approximation to the noisy signal to be found, which is more challenging to optimize. This makes dictionary-based sparse coding a less plausible candidate for a real-time noise-reduction algorithm. In contrast, image de-noising typically takes place offline and so is better suited to a sparse-coding approach. Nevertheless, it remains feasible that the brain employs mechanisms analogous to sparse coding for de-noising speech.

Effects of Audibility

Although the use of the NAL-R gain formula in this study was intended to compensate partially for the hearing loss of the participants, it does not provide equal audibility for all listeners. Since the effect of sensation level on speech quality judgments is not well understood, differences in audibility may have influenced individual differences in speech quality ratings. Therefore, the speech intelligibility index (SII) was calculated for each participant and condition as a measure of the audibility of the speech. In the case of the enhanced conditions, the SII was calculated based on the spectra of the speech and noise after the application of the enhancement gain function and shadow filtering. The sparse coding processing did not make use of a gain function but in order to calculate the SII a gain function was calculated based on the difference in level between the original and enhanced signals in 10-ms frames and one-third octave bands.

Additionally, the SII was calculated for the enhanced speech spectrum and the original noise spectrum, to determine whether any benefit could have been derived by changes in the level of the speech spectrum alone. For most processing and noise conditions, this resulted in a small reduction in the SII values relative to the unenhanced conditions but there were small increases for the Wiener filtering, NN_AIM and NN_COMP in both SSN conditions (mean values of increases in SSI of 0.0157, 0.0131, and 0.0121, respectively, at 0 dB SNR and 0.0081, 0.0043, and 0.0036 at +4 dB SNR). Considering the magnitude of the increases in the SII resulting from the suppression of the noise spectrum, these comparably small increases in SII due to changes to the speech spectrum alone are unlikely to have a strong effect on the ratings of speech quality. The greatest reduction in SSI occurred with sparse coding for the babble noise conditions with a mean reduction in SSI of -0.0859 at 0 dB SNR and -0.0603 at +4 dB SNR. It is possible that for these conditions the performance of the sparse coding algorithm was adversely affected by a lower audibility relative to the other algorithms, although the SII of the sparse coding conditions may have been underestimated by the application of a gain function that was calculated retrospectively.

Overall, there was a significant correlation between audibility and speech quality ratings [r = 0.467, $p < 0.001$]. However, once the mean quality rating and mean SII value were subtracted for each condition there was no significant correlation. This indicates that individual differences in audibility did not influence the rating of speech quality. There was a significant correlation between SII value and intelligibility both overall [r = 0.623, $p < 0.001$] and when the means for each condition were subtracted [r = 0.197, $p < 0.001$], indicating individual differences in audibility accounted for only a small amount of the variance in intelligibility between listeners.

5 IMPROVING SPEECH PERCEPTION IN NOISE FOR USERS OF COCHLEAR IMPLANTS

Overview

Even more so than for hearing aid users, CI users' speech understanding is negatively affected in situations with background noise. Algorithms that attempt to significantly improve speech intelligibility in noise for cochlear implant (CI) users have met with limited success, in particular in the presence of a fluctuating masker. In this chapter, a framework is proposed that integrates a *neural network based speech enhancement* (NNSE) algorithm into the speech processing pipeline of a CI processor for improved speech perception in background noise by CI users.

In the *first part* of this chapter, the proposed algorithm was evaluated in two listening studies with 10 normal-hearing participants listening to CI noise-vocoder simulations. NNSE was compared to a Wiener filter based algorithm, to NNSE algorithms using different target functions and to the unprocessed conditions. Significant improvements in SI in stationary and fluctuating noise were found for the ideal ratio mask (IRM) based NNSE algorithm over unprocessed and Wiener filter processed conditions and over the other target functions that were employed for the training stage.

In the *second part* of this chapter, an optimized version of the algorithm based on the findings in the first experiment using NH listeners was evaluated with 14 CI users by measuring the speech-in-noise performance for three types of background noise. Two NNSE algorithms using the IRM target function were compared: a speaker-dependent algorithm, that was trained on the target speaker used for testing, and a speaker-independent algorithm, that was trained on different speakers. Significant improvements in the intelligibility of speech in stationary and fluctuating noises were found relative to the unprocessed condition for the speaker-dependent algorithm in all noise types and for the speaker-independent algorithm in 2 out of 3 noise types. The NNSE algorithms used noise-specific neural networks that generalized to novel segments of the same noise type and worked over a range of SNRs. The proposed algorithm has the potential to improve the intelligibility of speech in noise for CI users while meeting the requirements of low computational complexity and processing delay for application in CI devices.

5.1 Speech Enhancement Based on Neural Networks Applied to Cochlear Implant Coding Strategies

Declaration of authorship:

Tobias Goehring - Leading the research, developed the algorithm and responsible for the signal processing part, performed the listening experiment and analysis of results, writing the manuscript

Co-authorship:

Federico Bolner - Developed the algorithm and the CI-specific software, writing the manuscript

Dr Jessica Monaghan - Helped with the software development for the listening experiment

Dr Bas van Dijk - Advising the research

Prof Jan Wouters - Advising the research

Prof Stefan Bleeck - Advising the research

Parts of this research have been published (first authors are underlined):

Bolner, F., Goehring, T., Monaghan, J., van Dijk, B., Wouters, J. and Bleeck, S. (2016). Speech enhancement based on neural networks applied to cochlear implant coding strategies. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6520-6524). IEEE.

5.1.1 Introduction

State-of-the-art cochlear implants (CI) allow for near-to-normal speech understanding in quiet acoustic conditions, however environmental noises represent one of the main challenges for CI users' speech understanding in everyday life (Zeng et al., 2008). Several speech enhancement algorithms for cochlear implants have been proposed to alleviate this problem.

Single-channel speech enhancement techniques have been successfully applied to cochlear implant sound coding and demonstrated to improve speech intelligibility (Buechner et al., 2010; Dawson et al., 2011). These algorithms rely on statistical assumptions about the background noise (e.g. stationarity) for the estimation of the SNR in order to modify the spectral content of the signal. Benefits of around 2.5 dB in speech reception threshold (SRT) were demonstrated in stationary noise, but the benefit is much reduced when the interfering noise is non-stationary, as in case of competing talkers.

More recent approaches, such as like supervised speech separation techniques, have been reported to improve speech intelligibility also in fluctuating background noises (Healy et al., 2013; Hu and Loizou, 2010). These algorithms make use of a binary classifier trained on the task of estimating the ideal binary mask (IBM). The concept of the IBM is based on retaining speech-dominant time-frequency (T-F) units while discarding masker-dominant T-F units with lower SNR by the application of an SNR threshold, the local criterion (LC) (typically, the LC is set between -10dB and 0dB) (Brungart et al., 2006; Wang et al., 2009). A demonstration of intelligibility improvement for CI users by a monaural algorithm has been provided by Hu and Loizou. The authors used a Gaussian mixture model classifier to decide whether each CI channel is dominated by speech or by noise. Only speech-dominated channels are then retained for electrical stimulation, resulting in large improvements in speech intelligibility in babble, train and hall noise. More recently, deep neural networks (DNN) have been applied to the task of IBM estimation and have shown significant improvements in SI for NH and hearing-impaired (HI) listeners (Healy et al., 2013). These studies represent a promising direction for improving speech enhancement algorithms, but are yet limited to a specific set of acoustic scenarios utilized during the training stage of the algorithm and depend on the choice of an SNR threshold for the IBM estimation.

In contrast to the IBM, the ideal Wiener filter (also called ideal ratio mask, IRM) applies a gradual weight to each T-F unit according to its local SNR (Lim and Oppenheim, 1979). Recently, listening tests conducted with NH listeners have shown that the IRM is less sensitive to estimation errors and does not depend on the choice of an SNR threshold, and that speech processed with the IRM leads to higher intelligibility scores in low SNR conditions, it is preferred in terms of perceived quality compared to IBM processed speech (Madhu et al., 2013).

The present study aims to investigate the potential improvements in speech intelligibility of a neural network-based speech enhancement algorithm applied to CI coding strategies, hereafter referred as NNSE. The use of a soft-masking function as training target is expected to yield better generalization performance over different SNR conditions. We further extend previous studies by evaluating the performance on unseen noise realizations while reducing the complexity of the algorithm. We evaluated three different target functions for the neural network training: a more aggressive IRM, the less-aggressive square-root IRM (e.g. Healy et al., 2015) and the IBM (e.g. Healy et al., 2013).

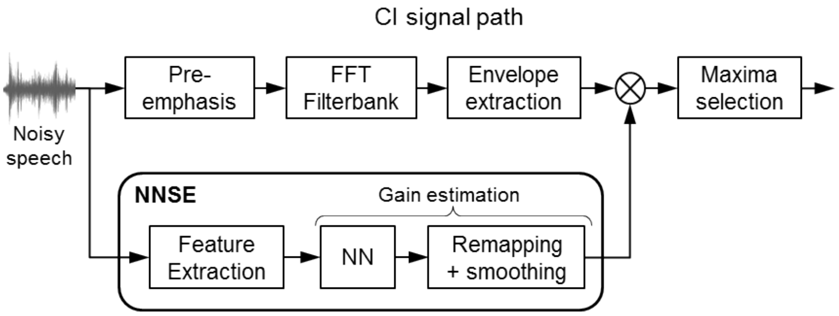


Figure 5.1 System block diagram of the implemented speech enhancement strategy (NNSE).

5.1.2 Algorithm Description

The integration of NNSE into a typical CI signal path is shown in Figure 5.1. We used the Advanced Combination Encoder (ACE), an n-of-m speech coding strategy, where the input signal is decomposed into $m = 22$ frequency channels from which envelope information is extracted through full-wave rectification. Maxima selection then retains only a subset of n channels with the largest amplitudes (maxima) for electrical stimulation. In this study, we chose a typical value of eight maxima.

The proposed algorithm consists of two main components: feature extraction and gain estimation. The integration of the NNSE into CI processing did not require a reconstruction stage, since the energy in frequency channels is directly used to determine the electrode output. Noisy input signals were first downsampled to 16 kHz and divided into 20-ms frames with 10-ms overlap, from which acoustic features are extracted and passed to an artificial neural network trained on the task of estimating the IRM gains over 63 critical bands ranging from 50 to 8000 Hz. The IRM gain in the k-th frequency channel of the n-th frame is defined as:

$$G_{IRM}(k,n) = \left(\frac{\xi(k,n)}{1+\xi(k,n)}\right)^{\beta},$$

where $\xi(k,n)$ is the SNR in the k-th frequency channel of the n-th frame. We compared two different choices for β to obtain the IRM ($\beta = 1$) and the IRMsqr (square-root, $\beta = 0.5$) target masks. We also calculated the IBM target mask by setting an LC threshold of -5 dB SNR. Three different neural networks were trained on either the IRM (NN-IRM), IRMsqr (NN-IRMsqr) or IBM (NN-IBM) target function. The estimated gains are then remapped to 22 CI channels, smoothed (exponential smoothing with a time constant $\tau = 12\ ms$) and applied to the noisy envelopes before the ACE channel selection. This has the main effect of attenuating masker-dominated channels, ultimately affecting maxima selection so that target-dominated channels are more likely to be selected for electrical stimulation.

Feature extraction

In contrast to previous studies that employed sub-band features (Healy et al., 2013; Hu and Loizou, 2010), we use a set of full-band acoustic features extracted from each 20-ms time frame. The set consists of two widely used and robust speech recognition features, the Mel-Frequency Cepstral Coefficients (MFCC) and the Relative Spectral Transform PLP (RASTA-PLP), along with the Gammatone log-energies (GTE). Our experimental results indicated that this combination led to higher estimation accuracy than the individual features alone.

To compute the MFCC and RASTA-PLP features, we applied a Hanning window to the input frame to then derive the power spectrum using a 512-points short-time Fourier transform. For the MFCC feature, the spectrum was converted into Mel scale, followed by log-compression and discrete cosine transform (DCT) to obtain 31 cepstral coefficients. For the RASTA-PLP feature, the power spectrum was instead warped to the Bark scale, log compressed, filtered by the RASTA filter (which emphasizes the modulation frequencies relevant to human speech), and expanded again by an exponential function. Finally, a 12-th order linear prediction model analysis was performed on this filtered spectrum to derive 13 RASTA-PLP features. To extract GTE features, we pass each input signal frame through the same 63-channel gammatone filterbank used to compute the target gains. The energy of each sub-band envelope is computed and log-compressed to obtain the GTE features. Since speech typically exhibits highly structured spectro-temporal patterns, we added contextual temporal-information in the form of the previous frame features, for a total of 214 features for each time-frame.

Artificial neural network training

In this study, we used a feed-forward artificial neural network with two hidden layers of 100 and 50 units. We found that the use of two hidden layers increased estimation accuracy, while more layers provided diminished performance. The number of units in the input and output layers is given by the dimensionality of the input feature set and the output target gains (214 and 63-D, respectively). Both hidden layers use a saturating linear transfer function, whereas the output layer uses a linear transfer function.

We trained the network to estimate the gain mask using the resilient back-propagation algorithm over 500 epochs, with mean squared error performance function and weight decay regularization to avoid overfitting. The network was trained with a total of 80 sentences (eight lists) from the IEEE database spoken by a male talker (IEEE, 1969). The interfering maskers included speech shaped noise (SSN) with the same long-term spectrum as the target speech, and BABBLE noise (4 male and 4 female talkers from the TIMIT corpus). Each noise recording was 26-seconds long. We used 18-seconds long segments of each masker for the training, while the remaining 8 seconds were left for testing. The sentences were mixed with random parts of both noises at 7 SNRs from -6 to 8 dB. The training took around 10-15 minutes on a newer generation personal computer. A different network was trained for each masker tested (but not for each SNR condition).

5.1.3 EVALUATION

Test material

The target speech material for the algorithm included the remaining sentences from the IEEE corpus (male talker) after the neural network training, while the maskers consisted of random parts of 8-seconds long segments of the noise recordings (SSN and BABBLE) reserved for testing.

The noisy mixtures were processed off-line with the Nucleus MATLAB Toolbox ACE implementation in three conditions: unprocessed noisy speech (UN), signal enhanced with the proposed algorithm (NN-IRM, NN-IRMsqr and NN-IBM), and enhanced with a Wiener-filter type algorithm (WFSE). In the WFSE condition, noisy mixtures were pre-processed prior to CI processing with a Wiener filter based on a priori SNR estimation using the unbiased MMSE-based noise power estimator (Gerkmann and Hendriks, 2012). This condition was added in order to compare NN-IRM with a state of the art speech enhancement algorithm.

A broadband correction gain was applied after the enhancement algorithms to restore the level of the speech component to the original level. Finally, after the ACE maxima selection, processed envelopes were passed through a noise band vocoder to simulate CI processing and obtain the test stimuli used in the objective evaluation and listening experiment.

Objective evaluation: procedure and results

To evaluate the accuracy of the neural network estimation analogue to previous studies, we converted the estimated and ideal masks into binary masks by applying a threshold (LC = -6 dB SNR). We then calculated the average correctly classified T-F units (HIT) and type-I-errors (false alarm, FA) percentage rates (Hu and Loizou, 2010).

Additionally, we computed two speech intelligibility measures of the processed vocoded speech (using clean vocoded speech as reference): the short-time objective intelligibility measure (STOI) (Taal et al., 2011), which has been developed for T-F weighted noisy speech, and the normalized covariance metric (NCM), which is closely related to CI processing and was successfully applied to vocoded signals previously (Chen and Loizou, 2011; Ma et al., 2009).

Table 5.1 HIT, FA and HIT-FA rates (expressed in percent) for the four noise conditions for the NN-IRM processing strategy.

	SSN		BABBLE	
SNR	0 dB	5 dB	5 dB	10 dB
HIT	82.15	79.35	78.40	70.58
FA	7.75	3.17	8.75	3.20
HIT-FA	74.40	76.18	69.65	67.38

Table 5.2 STOI and NCM scores for the four noise types and the three processing conditions.

	SSN		BABBLE	
SNR	0 dB	5 dB	5 dB	10 dB
STOI				
UN	0.54	0.63	0.58	0.65
WFSE	0.60	0.69	0.59	0.67
NN-IRM	0.64	0.70	0.64	0.70
NCM				
UN	0.50	0.63	0.42	0.57
WFSE	0.60	0.69	0.44	0.59
NN-IRM	0.64	0.69	0.58	0.66

Accuracy rates and intelligibility scores were computed over 100 sentences and are shown in Table 5.1 and Table 5.2, respectively. The proposed algorithm reaches high performance in terms of HIT-FA rate, a measure that was shown to correlate with SI. Also intelligibility predictions indicate higher scores of the NN-IRM compare to unprocessed speech of NN-IRM over UN in all conditions and over WFSE in all conditions apart in SSN at 5 dB SNR.

Listening experiment: procedure and results

The listening study was split into two sub-studies. In each sub-study, ten normal-hearing native English speakers were recruited (experiment 1: six males and four females, average of 29 years; experiment 2: 1 male and 9 females, average age of 22 years). All the parameters and testing conditions were equal between experiments, apart from the target function used for the neural network processing conditions (NN-IRM, NN-IRMsqr and NN-IBM).

Testing began with a short training to acclimatize the subject to the vocoded stimuli and consisted of one list of clean speech followed by one list at 10 dB SNR for each masker type. The listening test involved a word recognition task of vocoded speech in four noise conditions: SSN at 0 and 5 dB SNR, and BABBLE noise at 5 and 10 dB SNR. Subjects were presented with two lists for each processing strategy, for a total of 12 conditions [Experiment 1: 3 processing strategy (UN, WFSE and NN-IRM) × 2 SNRs × 2 maskers; Experiment 2: 3 processing strategy (UN, NN-IRMsqr and NN-IBM) × 2 SNRs × 2 maskers]. The presentation order of processing strategy and SNR level was randomised for each subject. Stimuli were presented diotically over closed circumaural headphones (Sennheiser HD380 pro) at 65 dB SPL.

Percentage correct word scores are shown in Figure 5.2 and mean improvements over UN for all processing conditions are shown in Figure 5.3.

For SSN, statistical analysis with analysis of variance (ANOVA) with repeated measures indicated significant effects of both SNR level [F(1,9) = 686.2, $p < 0.001$] and processing condition [F(2,18) = 47.3, $p < 0.001$], [F(2,18) = 5.9, $p = 0.011$]. For BABBLE, significant effects of both SNR level [F(1,9) = 265.3, $p < 0.001$] and processing condition [F(2,18) = 70.2, $p < 0.001$] were found.

For experiment 1 (Figure 5.2, left), results indicate significant improvements of the proposed NNSE algorithm NN-IRM over both UN and WFSE in all noise types and SNR levels (Bonferroni corrected post-hoc *p*-values are shown in Fig. 5.2). For experiment 2 (Figure 5.2, right), results indicate significant improvements over UN for both NN-IRMsqr and NN-IBM for SSN at 0 dB SNR and for BABBLE at 5 dB SNR. NN-IRMsqr gave improvements over both UN and NN-IBM for BABBLE at 10 dB SNR, indicating that NN-IRMsqr performed better than NN-IBM.

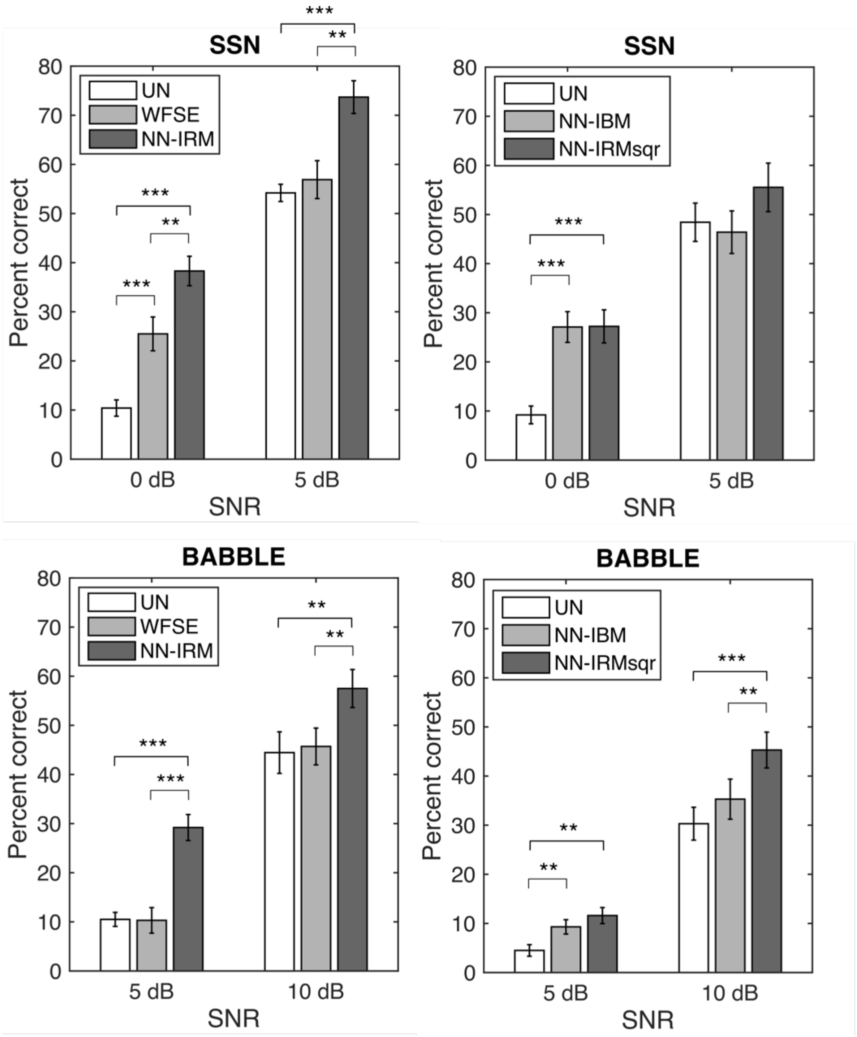


Figure 5.2 Mean speech intelligibility scores in noisy speech (UN), Wiener-filter-based speech enhancement (WFSE), the proposed algorithm (NN) with IRM, IRMsqr and IBM target functions in SSN and BABBLE noise. Error bars represent the standard error of the mean; (**) $p \leq 0.01$; (***) $p \leq 0.001$. Data from experiment 1 are shown on the left and from experiment 2 on the right side.

For the comparison between mean improvements over UN (Figure 5.3), NN-IRM gave largest improvements and significantly better performance relative to NN-IBM for SSN at 5 dB SNR and for BABBLE in both 5 and 10 dB SNR. NN-IRMsqr showed significantly better performance relative to NN-IBM in BABBLE at 5 dB SNR but was significantly worse than NN-IRM for BABBLE in 5 dB SNR. Overall, NN-IRM gave the best performance in speech intelligibility improvements.

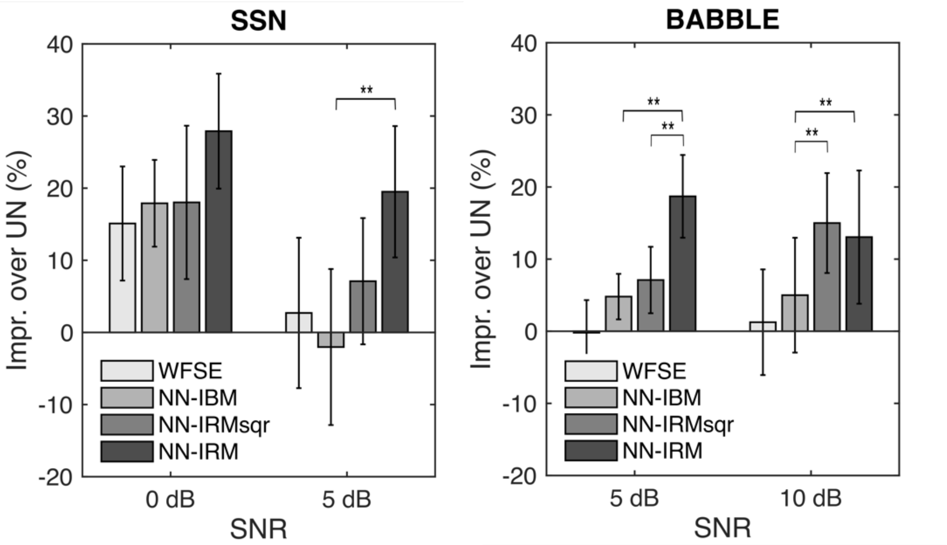


Figure 5.3 Mean improvements over UN for all four processing conditions (WFSE and NN with IBM, IRMsqr and IRM) in both noise types and SNRs (combining the results from experiment 1 and 2). Error bars represent the standard error of the mean; (**) $p \leq 0.01$; (***) $p \leq 0.001$.

5.1.4 Discussion

Significant improvements in intelligibility were observed with the proposed neural-network based speech enhancement strategy for all three target functions. These improvements were consistent for the more aggressive soft-mask NN-IRM in all the conditions tested, both compared with unprocessed noisy (UN) and with the conventional Wiener filter based speech enhancement condition (WFSE). The improvements were more noticeable in the lower SNR level of each masker. For instance, for NN-IRM the improvement in mean scores reached 27% in SSN and 18% in BABBLE noise compared with UN ($p < 0.001$), and 11% in SSN ($p < 0.01$) and 16% in BABBLE noise ($p < 0.001$) compared with the WFSE.

These improvements were predicted by the objective intelligibility scores, even though both STOI and NCM underestimated the increase in performance and did not correctly predict higher scores in SSN at 5 dB SNR for NN-IRM compared with the WFSE. As in the current study we used vocoded stimuli with NH listeners, it does not allow for direct comparison with previous speech separation studies in terms of SI improvements.

The HIT-FA rate is a popular measure found to correlate highly with intelligibility scores obtained with NH listeners (Kim et al., 2009). The NN-IRM condition produced high HIT-FA rates comparable to recent speech separation studies (Healy et al., 2013; Hu and Loizou, 2010) and maintained low FA rates, which according to Kim et al. are necessary to obtain high levels of speech intelligibility. In the same paper, the authors showed how conventional Wiener filter and MMSE noise reduction algorithms reach much lower HIT-FA rates. This is most likely the reason behind the high performance of the proposed algorithm compared with the tested WFSE.

Although previous studies proved supervised speech separation algorithms to be promising strategies for speech enhancement, their implementation poses a major challenge in real-word applications. Previous studies used the same noise realization for both classifier training and testing (Healy et al., 2013; Hu and Loizou, 2010), a situation that is unlikely to occur in practice. May and Dau have shown that the use of unseen noise realizations leads to a substantial decrease in HIT-FA rates with a GMM based system (May and Dau, 2014). In this study, the proposed algorithm was tested with unseen realisations. Although a performance drop can be expected in these more challenging conditions, we found significant SI improvements for the tested maskers and SNR levels.

The generalization of the proposed algorithm to mismatched noises and speakers between the training and testing stage still needs to be investigated. This question could be addressed in several ways, for instance by enlarging the training dataset (Wang and Wang, 2013) or by the integration of a noise-classification system (such as Hazrati et al., 2014; May and Dau, 2013) into the proposed algorithm. Moreover, the computational power required by these algorithms scales with their complexity, reflected by the feature extraction stage and by the employed classifier. Previous studies used sub-band based features and classification systems, while we opted for a smaller architecture. NNSE uses one set of full-band features and one neural network (for each masker type), resulting in reduced algorithm complexity and lower risk of over-fitting the training dataset.

We evaluated different target functions for the NN training and found the IRM to outperform the IRMsqr and IBM target masks. Furthermore, for intelligibility to be maintained with the IBM, the optimal LC value has to be changed with respect to the global SNR. Binary masking is also known to introduce musical noise. Compared with the IBM, the IRM does not depend on the setting of a threshold, is more robust to estimation errors and is preferred in terms of perceived quality (Koning et al., 2015; Madhu et al., 2013). Thus, the IRM is expected to present a better training target for speech enhancement purposes and should be used for experiments using CI users to obtain maximum benefits.

To summarise this study, the results obtained with the proposed algorithm indicate significant improvements in speech intelligibility for NH listeners using CI vocoder simulations. This motivates to investigate the potential benefit of the proposed NNSE strategy NN-IRM with CI users.

5.2 Speech Enhancement Based on Neural Networks Improves Speech Intelligibility in Noise for Cochlear Implant Users

Declaration of authorship:
Tobias Goehring - Leading the research, developed the algorithm, analysis of results, writing the manuscript

Co-authorship:
Federico Bolner - Developed the software for the listening experiment with CI users, helped with the optimization for application to CI technology, writing the manuscript
Dr. Jessica Monaghan - Advising the research
Dr. Bas van Dijk - Advising the research
Prof. Stefan Bleeck - Advising the research

Acknowledgements:
The listening test was performed at Cochlear Technology Centre in Belgium.
I would like to thank:
- the listeners who participated in the study.
- the colleagues from Cochlear for implementing the listening test.
- Prof Brian Moore for his helpful comments and feedback on the manuscript (in the function as editor at Hearing Research).

Parts of this research has been published (first authors are underlined):

Goehring, T., Bolner, F., Monaghan, J., Van Dijk, B., Zarowski, A., Bleeck, S. (2016). Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Research, 344, 183-194.*

5.2.1 INTRODUCTION

A cochlear implant (CI) is an auditory prosthesis that provides a sensation of hearing for listeners with severe to profound sensorineural hearing loss. State-of-the-art CI devices allow many users to achieve near-to-normal speech understanding in quiet acoustic conditions (Fetterman and Domico, 2002; Zeng et al., 2008). However, background noises such as environmental sounds or competing talkers negatively affect CI users' speech understanding. The decrease in performance can be measured with the speech reception threshold (SRT), which is defined as the signal-to-noise ratio (SNR) at which 50% of the speech is intelligible. CI users typically have SRTs that are 10 to 25 dB higher (worse) than those of normal hearing (NH) listeners (Spriet et al., 2007; Wouters and Van den Berghe, 2001). It has been reported that CI recipients can take less advantage of temporal gaps or slow amplitude fluctuations of an otherwise stationary noise masker compared with NH listeners in terms of speech intelligibility (Cullington and Zeng, 2008; Stickney et al., 2004; Zeng et al., 2008, Oxenham and Kreft, 2014). This process is known as release from masking (Miller and Licklider, 1950). Since the spectral information conveyed by a CI is reduced to a small number of effective spectral channels (Friesen et al., 2001), CI users rely strongly on temporal information (in the form of envelope modulations) and thus are more susceptible to modulated masking noise than NH listeners (Cullington and Zeng, 2008; Fu et al., 2013). Most likely, a combination of reduced spectral resolution and increased modulation interference accounts for the decrease in speech understanding performance observed for CI users compared with NH listeners and with NH listeners tested with CI simulations (Cullington and Zeng, 2008; Jin et al., 2013, Oxenham and Kreft, 2014).

Speech enhancement (SE) algorithms have been proposed to alleviate this problem by attenuating the noise component of the noisy mixture to increase the intelligibility and perceived quality of the speech component (Loizou, 2013). SE algorithms can be divided into algorithms that make use of two or more microphones to exploit the spatial properties of target and noise sources, and algorithms that make use of a single microphone (or the output signal of a multi-microphone algorithm). Multi-microphone algorithms have been shown to deliver large benefits in SRT scores when the target signal and the interfering noise source are spatially separated (Mauger and Warren, 2014; Spriet et al., 2007; Wouters and Van den Berghe, 2001). However, in everyday listening situations, these requirements might not always be fulfilled, and single-microphone algorithms are still of interest for numerous applications, such as hearing devices, where the number of microphones is usually limited to two and the two microphones are on the same side of the head.

Single-microphone SE algorithms are based on the assumption that improving the global SNR of noisy speech will lead to improved speech intelligibility (SI) (French and Steinberg, 1947). With such algorithms, the signal is converted into the spectral domain (e.g. by Fourier analysis or filter bank processing) and a filter is applied to retain the signal in frequency channels with high SNR and attenuate the signal in frequency channels with low SNR, leading to an increased global SNR. Numerous algorithms have been proposed to estimate the SNR in each frequency channel (Gerkmann and Hendriks, 2012; Martin, 2001). This estimate is used to calculate a gain function to determine the attenuation of noise-dominated channels. SE algorithms mainly differ in the SNR estimation methods and the gain functions used for noise suppression (e.g. spectral subtraction or parametric Wiener filter, Boll, 1979; Lim and Oppenheim, 1979). In the ideal case (i.e. when the speech and noise components are known), these algorithms can lead to highly increased intelligibility, close to that for noise-free speech for NH listeners (Madhu et al., 2013) and CI users (Koning et al., 2015; Mauger et al., 2012a; Qazi et al., 2013). Similarly, extensive studies on the SI benefits of time-frequency masking with the ideal binary mask (IBM) support the potential of SNR-based suppression criteria for improving the intelligibility of speech in noise (Anzalone et al., 2006; Brungart et al., 2006; Hu and Loizou, 2008; Wang et al., 2009a).

In a real system, where only the mixture of speech and noise is available, SNR estimation errors may lead to speech distortions, introduction of musical noise or insufficient noise suppression. In challenging acoustic environments these artefacts greatly reduce and often completely undo the speech intelligibility benefits observed in the ideal case for NH and hearing-impaired (HI) listeners (Brons et al., 2012; Chen and Loizou, 2012; Loizou, 2013b). For CI users, where a decrease in SI performance is typically observed at higher SNRs than for NH and HI listeners, improvements in SI have been reported with several SE algorithms based on noise-estimation techniques (Dawson et al., 2011; Hu et al., 2007; Mauger et al., 2012b; Ye et al., 2013). This success may be due to the better performance (reduced estimation errors) of the algorithms for higher SNRs. In addition, Mauger et al. (2012a; 2012b) reported that CI users generally preferred a more aggressive gain function than the standard Wiener gain function, suggesting that CI users might be more resistant to speech removal distortions (type-II errors) and less resistant to noise addition errors (type-I) (also reported by Qazi et al., 2013). For CI users, maximum benefits of about 2 dB in SRT were found for speech in stationary noise, but the benefit was much reduced when the interfering noise was non-stationary, as in the case of competing talkers (Dawson et al., 2011; Mauger et al., 2012b).

A recent approach to SE algorithms employs supervised machine learning to estimate the gain function (by using either classification or regression methods), instead of using conventional SNR estimation techniques (Tchorz and Kollmeier, 2003). Using a similar approach, algorithms have been trained on labelled datasets to approximate the IBM. These have been reported to provide remarkably large SI improvements for NH listeners (Kim et al., 2009), HI-listeners (Healy et al., 2013, 2014) and CI users (Hu and Loizou, 2010) for speech in both stationary and non-stationary noise, even at low SNRs. However, these algorithms were trained and tested on datasets using the same speaker, background noise and SNRs. This approach is likely to lead to overfitting of the training data and strongly limits generalization performance to acoustic conditions different from the ones used during training (May and Dau, 2014b). Recently, it has been shown, for both NH and HI listeners, that incorporating more exemplars of the noise recordings in the training stage leads to algorithms that generalize well to novel realizations of the same noise type (Bolner et al., 2016; Healy et al., 2015) or to completely novel types of noise (Chen et al., 2016). These studies indicate that generalization to novel noise conditions is possible when the training datasets incorporate higher degrees of variability. Furthermore, the use of a "soft" gain mask (often called *ideal ratio mask*, IRM) inspired by the Wiener filter gain function (Lim and Oppenheim, 1979) avoids the need to choose an appropriate SNR-dependent classification threshold in IBM-based processing, and can lead to a regression model that worked over a range of SNRs (Bolner et al., 2016) or generalized to untrained SNRs (Chen et al., 2016).

The results from the studies described above are promising. However, generalization to novel, unseen speakers was not tested (Bolner et al., 2016; Chen et al., 2016, Healy et al., 2015). In real-world situations, in the context of SE for hearing devices, an algorithm should work well with any target speaker and meet the requirements of limited computational complexity and short processing delay (Stone and Moore, 2005). The algorithms proposed by Chen et al. (2016) and Healy et al. (2015) include non-causal information (future frames) in the processing and therefor introduce considerable processing delays (>20 ms). As described by Healy et al. (2015), the use of future frames has to be avoided for applications using real-time processing, such as hearing aids and CIs.

In this study, we tested whether an SE algorithm using neural networks (NNSE) can improve the SRTs of CI users for speech in stationary and non-stationary background noises. We address the important aspect of generalization performance to a novel speaker by comparing two identical systems that were trained on either the same or different speakers from the one used during testing. This study used noise-specific networks that were tested on novel segments of the same noise type (similar to Healy et al., 2015). The algorithm complexity and processing delay were chosen to yield a real-time feasible architecture with low latency for potential application in CIs. We employed an aggressive gain function as preferred by CI users (Mauger et al., 2012a, 2012b; Qazi et al., 2013) that indicated the best performance in our study with NH listeners in the previous chapter and integrated the SE algorithm into the coding strategy of a CI to evaluate the performance of the algorithm. The algorithm was designed to work over a range of SNRs (Chen et al., 2016; Bolner et al., 2016) relevant to CI users and to process stimuli adaptively using online processing.

5.2.2 Algorithm Description

The NNSE algorithm, was integrated within an implementation of the Advanced Combination Encoder (ACE™) CI speech processing strategy (Seligman and McDermott, 1995). Figure 5.4 shows a block diagram of the algorithm.

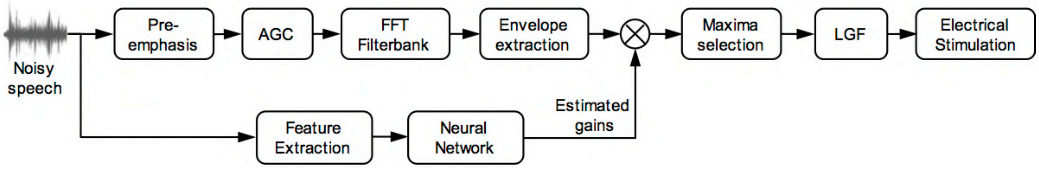


Figure 5.4 Block diagram of the proposed speech enhancement algorithm integrated into the ACE signal path (including an automatic gain control, AGC, and loudness growth function, LGF). The algorithm has two components: Feature Extraction and Neural Network.

2.1 Reference strategy

A research ACE strategy implementation served as the reference strategy. The noisy speech signal was downsampled to 16 kHz, passed through a pre-emphasis filter, and sent through an automatic gain control (AGC). The AGC compressed the acoustic dynamic range such that it could be conveyed into the smaller electrical dynamic range of a CI recipient (with an attack time of 5 ms, a release time of 75 ms, a compression threshold of 73 dB SPL and compression limiting above that level). Next, a filter bank based on a Fast Fourier Transform (FFT) was applied to the compressed signal. The FFT was performed on Hanning-windowed 8-ms long input blocks, with an overlap of 7 ms. The magnitude of the complex FFT output was used to provide an estimate of the envelope for each of the M frequency channels (typically, M=22). Each channel was then allocated to one electrode. Maxima selection was applied to retain the subset of N channels with the largest envelope magnitudes (with N<M set by an audiologist during the fitting of the subject's CI processor). A loudness growth function (LGF) instantaneously mapped the envelope for each channel to the subject's dynamic range between the threshold level (THL) and maximum comfortable loudness level (MCL) for electrical stimulation (using the THL and MCL parameters from the subject's CI processor). Finally, the electrodes corresponding to the selected channels were stimulated sequentially and one cycle of stimulation was completed. The number of cycles per second is called the channel stimulation rate, and the total stimulation rate is N times the channel stimulation rate.

2.2 Speech enhancement algorithm

CI processing directly transforms the envelope of the frequency channels to an electrical output, and it does not require a reconstruction stage. We chose to integrate the NNSE directly into the CI signal path rather than performing preprocessing of the noisy signal. This avoids an unnecessary synthesis stage, which would introduce additional noise and increase the complexity and delay of the system. The NNSE algorithm consisted of two main components: feature extraction and neural network (NN) regression.

After downsampling to 16 kHz, the noisy speech signal was divided into 20-ms long segments with 50% overlap. Feature extraction was performed on each segment of the noisy signal, and the output was fed to the NN. The trained NN (the training is described below) was used to estimate the Wiener gain over 31 frequency channels equally spaced on the equivalent rectangular bandwidth (ERB_N-number, Glasberg and Moore, 1990) scale with centre frequencies ranging from 50 to 8000 Hz. Since the frequency channels assigned to the electrodes varied across subjects, the estimated gains were mapped to each subject's specific filter bank configuration. Exponential smoothing (with a time constant of 12 ms) was performed before applying the gain to the corresponding noisy envelope in the ACE signal path. The main effect of the gain application was the attenuation of noise-dominated channels. This occurred before the ACE channel selection (see Fig. 5.4). Therefore, speech-dominated channels were more likely to be selected for stimulation. Unlike most SE algorithms (Loizou, 2013b), the algorithm does not require a voice activity detector or the estimation of noise statistics. The NNSE was designed so that it could be run in real time, with an algorithmic delay of 10 ms.

An example of an electrodogram of a Dutch sentence ("*Het verhaal is heel spannend*") from the LIST corpus processed by the ACE coding strategy with 11 maxima is shown in Fig. 5.5. An electrodogram represents the stimulation pattern across electrodes (y-axis) over time (x-axes). The height of each vertical bar reflects the normalised amplitude of a single stimulation pulse.

The top panel represents the electrodogram of the clean sentence, in which the boundaries between words are clearly visible. For the second panel, the speech was corrupted by babble noise (SNR = 5 dB). The resulting stimulation sequence changed significantly: periods of silence were filled with noise, envelopes were distorted, and not all of the channels containing speech were selected. The third and fourth panels represent the conditions with NNSE processing using speaker-independent and speaker-dependent training, respectively. The processing steered channel selection to pick the channels containing speech, thus partially restoring information that was masked by the noise (Qazi et al., 2013).

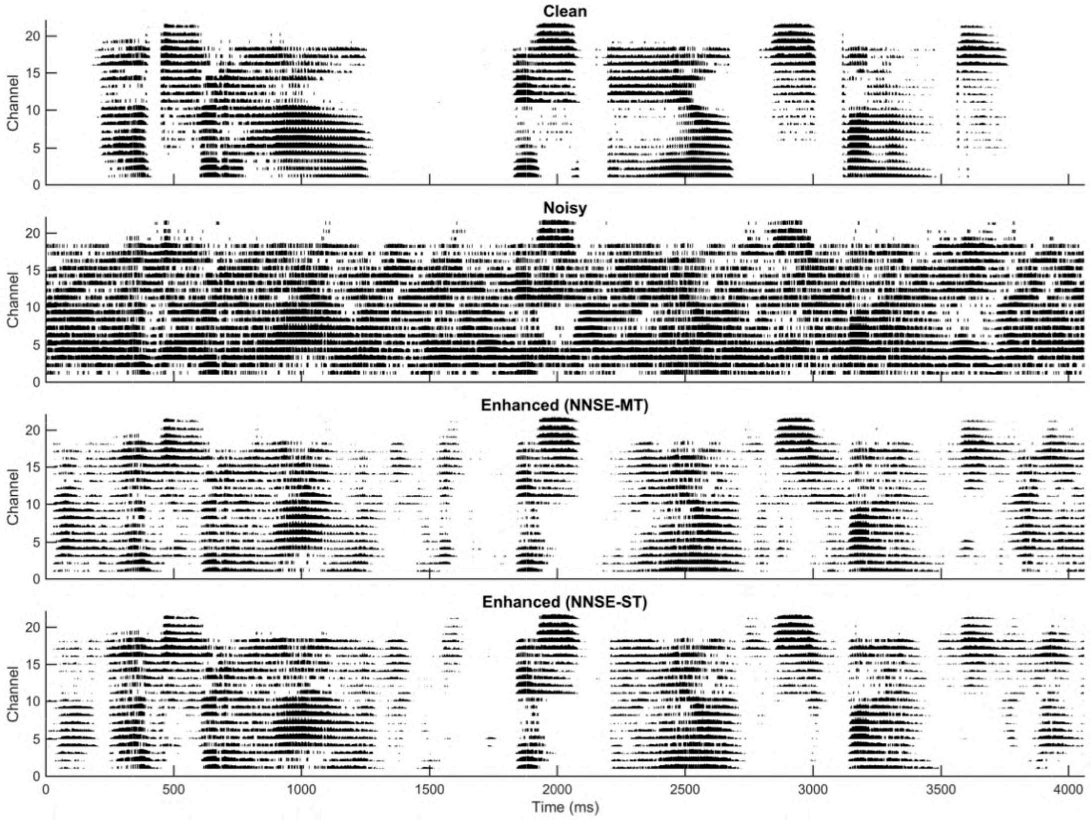


Figure 5.5 Electrodogram of the sentence 'Het verhaal is heel spannend' produced by a male speaker (LISTm) at a level of 65 dB SPL. The top panel is for the noise-free signal. The second panel is for the signal with BABBLE noise (SNR = 5 dB). The third and fourth panels are for the conditions with NNSE-MT and NNSE-ST, respectively.

2.2.1 Feature extraction

Feature extraction was performed on each 20-ms segment, or frame, at a rate of 100 Hz. Each frame was passed through a Gammatone filter bank consisting of 31 channels equally spaced on the $\mathrm{ERB_N}$-number scale with centre frequencies ranging from 50 to 8000 Hz (Hohmann, 2002). Then, the energy of each channel was log-compressed to obtain 31 Gammatone Frequency Energy features ($\mathrm{GFEN_n}$, with *n* denoting the frame number). From the $\mathrm{GFEN_n}$, two additional features were extracted: 26 Gammatone Frequency Cepstral Coefficients ($\mathrm{GFCC_n}$) and 13 Gammatone Frequency Perceptual Linear Prediction Cepstral Coefficients ($\mathrm{GPLP_n}$). The $\mathrm{GFCC_n}$ features were obtained by performing the discrete cosine transform (DCT) on $\mathrm{GFEN_n}$ for frequencies above 200 Hz (and excluding the DC component of the DCT). The $\mathrm{GPLP_n}$ features were obtained by filtering $\mathrm{GFEN_n}$ with the relative spectral transform (RASTA, Hermansky and Morgan, 1994) filter, which emphasises the modulation frequencies relevant to human speech, and performing a 12-th order linear prediction model analysis on the output (perceptual linear prediction, PLP).

The 31 $\mathrm{GFEN_n}$, 26 $\mathrm{GFCC_n}$ and 13 $\mathrm{GFPLP_n}$ features were concatenated to form a 70-dimensional feature vector F_n. Our pilot results (Bolner et al., 2016) indicated that this combination led to higher estimation accuracy than the individual features alone. Note that F_n was derived exclusively from the $\mathrm{ERB_N}$-number spaced spectrum of the signal ($\mathrm{GFEN_n}$). Evaluation with several objective measures (difference between hit and false alarm rates, HIT-FA, Kim et al., 2009; short-time objective intelligibility measure, STOI, Taal et al., 2011; normalized covariance metric, NCM, Holube and Kollmeier, 1996; Ma et al., 2009) indicated that this choice had no detrimental effects on the estimation accuracy of the algorithm compared with the use of the more conventional MFCC (using the Mel-scale) and RASTA-PLP (using the Bark scale), and it avoided two additional filtering stages. Finally, F_n was concatenated with the features extracted from the preceding frame F_{n-1} to provide additional temporal information. The resulting 140-dimensional feature vector [F_n, F_{n-1}] was fed to the NN to estimate the Wiener gain for the current frame n. Note that the NN estimated the Wiener gain using information related to the current and past frames only. This feature set allowed relatively low complexity and low delay making the proposed algorithm suitable for real-time processing, in contrast to most recent speech segregation studies (Chen et al., 2016; Healy et al., 2013, 2015).

2.2.2 Neural network regression: architecture and training procedure

A parametric Wiener gain mask (Lim and Oppenheim, 1979), the IRM, was used as the training target for the supervised training process. The ideal ratio mask is defined as follows:

$$G(f,n) \; = \; \left(\frac{SNR(f,n)}{SNR(f,n)+1}\right)^{\beta},$$

where $SNR(f,n)$ denotes the SNR in frame n and Gammatone frequency channel f. The parameter β controls the slope of the gain function $G(f,n)$. We experimented with different values of β and found $\beta = 1$ to be a good compromise between noise removal and speech distortion when the mask was applied to noisy speech. This choice was also supported by the finding that CI users generally prefer a relatively aggressive gain function (Mauger et al., 2012a, 2012b) as opposed to the square-root Wiener mask ($\beta = 0.5$) used in previous studies with HI listeners (Chen et al., 2016; Healy et al., 2015) and by the listening test results in our pilot study with NH listeners.

The neural network consisted of an input layer, defined by the feature vector, 2 hidden layers of 75 units using a saturating-linear activation function (which resembled a piecewise linearised sigmoidal function) and 31 linear output units. Resilient backpropagation (Riedmiller and Braun, 1993) was used for training the NN in full-batch mode over 500 epochs with a learning rate of 0.01 and weight increment and decrement factors of 1.2 and 0.5, respectively. The cost function was the mean squared error (MSE) between the true and estimated Wiener gain using a weight-decay regularisation of 0.5 to avoid overfitting.

The parameters of the algorithm were chosen based on a previous study of Bolner et al. (2016), who observed significant improvements in speech intelligibility in noise for NH listeners using CI vocoder simulations with a supervised NN-based SE algorithm. The biggest difference between the two algorithm configurations was a reduced number of neural network parameters (node weights and biases), mainly deriving from the use of a Gammatone filter bank with 31 channels both for the feature extraction stage and Wiener gain estimation, as opposed to 63 channels used by Bolner et al. (2016). The Nucleus implants tested in this study maximally use 22 spectral channels, and thus 31 channels seemed a good compromise between algorithm complexity and SE performance for CI application. The 31 estimated Wiener gains were mapped to the 22 CI channels before application to the envelopes. The configuration used in the current study allowed a reduction in the algorithm complexity while maintaining comparable performance in terms of estimation accuracy and with respect to several speech intelligibility objective metrics, such as HIT-FA (between estimated and ideal ratio masks), NCM and STOI (using vocoded simulations of the enhanced and noise-free reference signals, Chen and Loizou, 2011).

The algorithm made use of feed-forward neural networks that were trained using the true Wiener gain along with the features extracted from the noisy speech. Rather than performing large-scale training with thousands of noises (as done by Chen et al., 2016), the networks were noise-specific, i.e. each network was trained for a particular listening situation (similar to Hu and Loizou, 2010). This made it possible to take advantage of the learning of the distinctive spectro-temporal characteristics of each noise while limiting the NN size.

The speech materials used to train the NNSE were LISTm (sentences of equal difficulty with 2-7 keywords, equal number of syllables and key words per list, male Flemish talker, Jansen et al., 2014), LISTf (similar structure to LISTm, but partially different sentences than LISTm, female Flemish talker, Van Wieringen and Wouters, 2008), NVA (lists of 10 bisyllabic words, male Flemish talker, Wouters et al., 1994), and GRID (simple and syntactically identical phrases of 6 words, 18 male and 16 female English talkers, Cooke et al., 2006). Three types of noise were used: steady speech weighted noise (SWN), single-speaker-modulated speech-weighted noise (ICRA), and 20-talker babble (BABBLE). The SWN had the same long-term spectrum as the sentences of the LISTm corpus (Jansen et al., 2014). The modulated speech-weighted noise was the ICRA5-250 (Dreschler et al., 2001) that was generated by sending English male speech through a 3-channel filter bank, randomly reversing the sign of each sample in each channel (with a probability of 0.5), filtering it again with the same filter bank, randomizing the phase in the frequency domain and applying the standard long-term average speech spectral shape of male speech. The ICRA5-250 noise has maximum silent gaps of 250 ms and may contain some intelligible fragments, at least for English native speakers, as reported by Dreschler et al. (2001). The BABBLE signal was recorded at Auditec St. Louis and consisted of a mixture of 20 English competing talkers (8 male, 12 female). The three types of masking noise have different degrees of temporal fluctuation (increasing from SWN to BABBLE to ICRA) and thus introduce varying amounts of modulation masking (Dau et al., 1997).

During training, 4-minute long recordings of the three noises were mixed with two speech material training sets:

- Single talker (ST), containing 10 lists from the LISTm corpus (total of 8 minutes)
- Multiple talker (MT), containing 6 lists from the LISTf corpus, 4 lists from the NVA corpus and 120 sentences from the GRID corpus (total of 15 minutes).

In both cases, the sentences were mixed with random segments of the noise at 7 SNRs, from –6 to +6 dB in steps of 2 dB. This, in turn, produced two networks for each noise type, one trained on a single talker (LISTm) and the other trained on multiple talkers.

5.2.3 MATERIALS AND METHODS

3.1 Software/Hardware

The research ACE strategy and NNSE algorithm were developed in MATLAB (The MathWorks, Natick, Massachusetts). Stimuli were processed through a computer implementing the ACE strategy (with/without NNSE) and directly presented to the implant user. Electrical stimulation was delivered via the Cochlear NIC3 interface connected to an L34 experimental processor. The system delivered radio frequency output to the coil that transmitted stimulus data to the subject's implant.

3.2 Subjects

A group of 14 CI users, all native Dutch speakers and implanted with a Cochlear Nucleus® CI, participated. The study protocol was approved by the Commissie Medische Ethiek GZA Ziekenhuizen (Antwerp) ethics committee, and subjects gave their informed consent to participate in the study. Subjects were not paid, but travel expenses were reimbursed. This study was conducted according to the guidelines for Good Clinical Practice (GCP), ISO14155-2011 (International Standard for Clinical Investigations of medical devices for human subjects) and the Declaration of Helsinki (2013).

Table 5.3 Individual subject demographics: age (years), tested ear (left/right), duration of implant use (years), implant type, origin of hearing loss, etiology, and duration of profound hearing loss (years).

Subject	*Age*	*Tested Ear*	*Implant use*	*Implant type*	*Type of HL*	*Etiology*	*Duration of profound HL*	*Contralateral ear*
01	62	R	12.6	CI24R	Progressive	Unknown	Unknown	-
02	62	L	11.3	CI24R	Progressive	Cholesteatoma	48	-
03	53	L	12.6	CI24R	Progressive	Unknown	47	-
04	68	L	8.1	CI24RE	Progressive	Meniere's Disease	17	HA
05	70	L	13.3	CI24R	Progressive	Otosclerosis	60	HA
06	69	R	10.6	CI24RE	Progressive	Meningitis and Otosclerosis	45	-
07	60	R	5.1	CI512	Sudden	Cholesteatoma	5	HA
08	35	L	11.5	CI24RE	Sudden	Meningitis	3	-

09	81	R	12.6	CI24R	Progressive	Cholesteatoma and Chronic Mastoiditis	Unknown	-
10	69	L	9.6	CI24RE	Sudden	Unknown	53	-
11	72	L	6.6	CI24RE	Progressive	Meniere's Disease	8	-
12	76	R	1.2	CI512	Progressive	Familial	5	HA
13	52	L	8.1	CI24RE	Congenital	Unknown	52	HA
14	23	R	13.6	CI24R	Congenital	Waardenburg Syndrome	1	CI24R

The mean age of the group at the start of the study was 61 years, ranging from 23 to 81 years. Only one ear of each subject was tested. If the subject had a hearing aid (HA) or CI on the contralateral side, it was turned off during the testing. The mean duration of implant use was 9.8 years at the start of the study, with a range from 1.2 to 13.6 years. All subjects were users of the ACE strategy. Demographic data for the subjects can be found in Table 5.3.

Table 5.4 CI parameters for each of the 14 subjects during the study: channel stimulation rate (Hz), number of maxima/number of active electrodes, THL and MCL (threshold and comfort levels in current level, CL), minimum and maximum of the dynamic range (DR, in CL).

Subject	*Channel stimulation rate*	*Pulse Width*	*Maxima / no. active electrodes*	*THL-current level*		*MCL-current level*		*DR*	
UNIT	Hz	µs		Min CL	Max CL	Min CL	Max CL	Min CL	Max CL
01	900	25	14/20	105	130	150	193	39	68
02	900	25	10/19	120	135	174	184	39	60
03	900	25	14/22	108	134	165	194	47	79
04	900	25	14/22	109	176	171	200	24	62
05	900	25	14/20	113	129	159	182	42	66
06	1800	20	10/22	150	180	177	228	27	48
07	900	25	14/22	130	160	153	185	15	28
08	2400	12	10/22	111	125	195	205	75	88
09	900	25	14/20	135	152	157	175	17	28
10	900	25	8/22	78	145	108	168	10	36
11	900	25	11/22	129	171	158	203	28	32
12	900	25	12/22	98	144	132	178	32	34
13	900	25	10/21	109	151	137	190	18	73
14	900	25	14/22	120	145	186	205	50	80

Prior to the speech in noise test, the subjects' existing CI program parameters were transferred from their own sound processor to the control computer. Subjects informally reported that they did not perceive a difference between the daily program on their sound processor and the stimulation delivered via the ACE strategy on the test system. Details of each subject's CI parameters, such as stimulation rate, number of maxima, number of total active channels, THL and MCL, and dynamic range are presented in Table 5.4.

3.3 Stimuli and processing conditions

Sentences from the LISTm corpus (Jansen et al., 2014) were used as the target speech material. The LISTm corpus consists of 38 lists, with 10 sentences for each list, produced by a male Flemish talker. The number of keywords per sentence ranged from 2 to 7, with an average and median of 3. Since 10 lists of the corpus were used during the training stage of the algorithm, only the remaining 28 lists were employed for the listening test.

The maskers were 20-s long novel realizations of SWN, ICRA5-250 and BABBLE, from which a random segment was extracted and mixed with the target speech for each sentence. This was done in order to test the algorithm on sentences and noise segments that were not previously processed by the NNs.

The three processing conditions were:

- UN: unprocessed condition, i.e. ACE.
- NNSE-ST: processed condition with the NNSE algorithm, using the networks trained on the single-talker data. Note that in this case the algorithm was tested on the same speaker as the one used during the training stage (LISTm).
- NNSE-MT: processed condition with the NNSE algorithm, using the networks trained using multiple talkers data, which did not include the target speaker.

The NNSE-MT condition was included to assess the performance of the NNSE in more realistic and challenging conditions when the target speaker was unknown to the system, in contrast to recent SE studies (Bolner et al., 2016; Chen et al., 2016; Healy et al., 2013, 2015; Hu and Loizou, 2010).

3.4 Study protocol

The study used a repeated measures, single-subject design in which each subject served as his/her own control. This approach made it possible to accommodate the heterogeneity that usually characterizes the CI population. At the beginning of the session, each subject was allowed to choose his/her preferred volume. Sentences from one list of the corpus (from the training set) were presented in quiet and in noise (SWN between 0 and 5 dB SNR) until the subject was satisfied with the volume. The chosen volume setting was then fixed for the rest of the testing.

The SRT was measured using an adaptive procedure for 9 conditions [3 maskers (SWN, ICRA, BABBLE) x 3 processing conditions (UN, NNSE-ST, NNSE-MT)] by an audiologist in a sound-treated room. Both subject and audiologist were blind as to which processing condition was being tested.

An SRT was measured using one list (10 sentences) randomly selected from the speech corpus. The speech level was held constant at 65 dB SPL while the noise level was adjusted according to the subject's response to each sentence in steps of 2 dB, in a one-down, one-up procedure to target the 50% correct point. After determining the level of the (hypothetical) 11th item, the SRT was calculated as the mean of the last 6 SNRs. A response was counted as correct when all the keywords in the sentence were correctly identified. Errors for non-keywords were not taken into account, but incomplete keywords or minor variations of verb tenses of keywords were penalised (van Wieringen and Wouters, 2008).

Each of the 9 conditions was tested 3 times, counterbalancing the order in which the conditions were tested for each subject. The order in which the noise and processing conditions were tested was counterbalanced across 12 subjects, and the order for the remaining two subjects was allocated randomly. The final SRT for each condition was obtained by averaging the three SRT values. At the end of the testing, subjects resumed the use of their own sound processor.

5.2.4 Evaluation

Prior to clinical testing, an objective analysis of the performance of each processing condition was performed. Electrodograms were computed at different SNRs, and were compared with a reference electrodogram in terms of type I and type II error rates. Although this method has not been widely used in the literature, it represents a useful way to compare noise reduction performance for CIs (Mauger et al., 2012b).

In an electrodogram, stimuli have normalized values between 0 and 1, representing the electrical perception range between threshold and comfort level in each frame and frequency channel. The reference electrodogram was generated by processing speech in quiet with ACE (without NNSE), and provided the "ideal" outcome of noise reduction.

Error rates were computed as the stimulus amplitude difference of the reference electrodogram (REF-E) and the comparison electrodogram (COM-E), with the method proposed by Mauger *et al.* When the COM-E contained a stimulus (channel-frame) that was lower in amplitude than the corresponding stimulus in the REF-E, a type II error was computed as the stimuli amplitude difference. For example, if the COM-E had a stimulus amplitude of 0.3 and the REF-E had a stimulus of 0.5, this was considered as a type II error of value 0.2. A full type II error (value = 1) occured when no stimulus (amplitude = 0) was present in the COM-E, while the REF-E contained a stimulus with amplitude = 1. In a similar manner, a type I error occurred when the COM-E contained a stimulus of higher amplitude than for the REF-E. The type I error was computed as the difference of the stimulus amplitudes. For example, if the COM-E had a stimulus amplitude of 0.3 and the REF-E had a stimulus amplitude of 0, this was considered as a type I error of value 0.3. A type I error can be viewed as a noise addition error, while a type II error can be viewed as a speech removal error.

Type I and type II errors were summed across all channels and frames and divided by the total number of possible errors to obtain the type I and type II error rates. Error rates for processing condition were computed as the average error rates calculated over 20 sentences at –5, 0, 5, and 10 dB SNR, with 11 selected channels (ACE maxima selection). This was done so as to have the same number of possible errors for both error types and to avoid introducing a bias towards either of the two.

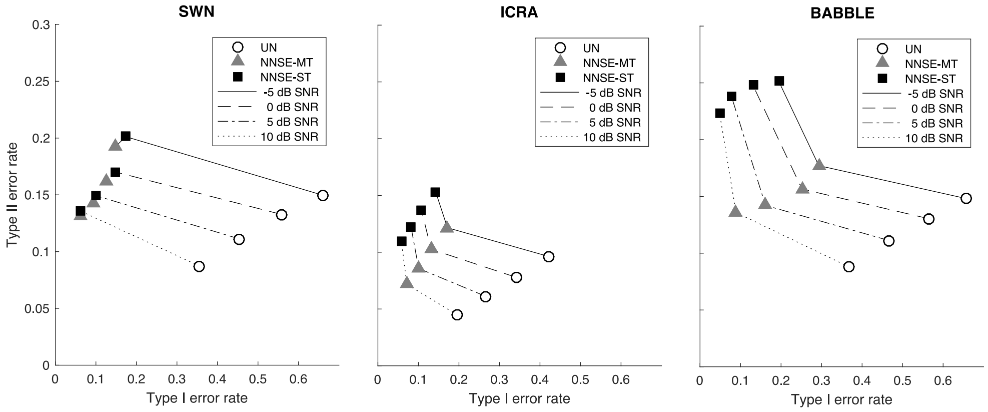


Figure 5.6 Error rate analysis for UN, NNSE-MT and NNSE-ST processing conditions for the three noises, at –5, 0, 5 and 10 dB SNR. Lines join error rates for the same input SNR. The target speech was LISTm sentences (not part of the training database of either of the NNSE algorithms).

Results of the objective analysis are displayed in Figure 5.6. For SWN, UN gave type I error rates from 36% to 66%, and type II error rates ranging from 9% to 15% (SNR = -5 and 10 dB, respectively). The NNSE conditions gave similar error rates, with greatly reduced type I error rates (≤6% and ≤17%, at –5 and 10 dB SNR, respectively), at the expense of slightly higher type II error rates (≤14% and ≤20%, at –5 and 10 dB SNR, respectively).

For ICRA, UN gave type I error rates from 20% to 42%, and type II error rates from 4% to 10% (SNR = -5 and 10 dB, respectively). Again, both NNSE conditions gave greatly reduced type I error rates at the expense of higher type II error rates. Type I errors ranged from 7% to 17% for NNSE-MT, and from 6% to 14% for NNSE-ST, at –5 and 10 dB SNR, respectively, while type II error rates ranged from 7% to 12% for NNSE-MT, and from 11% to 15% for NNSE-ST (at –5 and 10 dB SNR, respectively).

For BABBLE, UN gave type I error rates from 37% to 66%, and type II error rates from 9% to 15% (SNR = -5 and 10 dB, respectively), in line with what was found for SWN. Also for BABBLE, both NNSE conditions gave reduced type I error rates but higher type II error rates compared to the UN condition. Type I errors ranged from 9% to 30% for NNSE-MT, and from 5% to 20% for NNSE-ST, at –5 and 10 dB SNR, respectively. Type II error rates ranged from 14% to 18% for NNSE-MT, and from 22% to 25% for NNSE-ST.

In conclusion, both NNSE algorithms greatly reduced the noise, but also introduced some speech removal distortions. This effect was more pronounced for NNSE-ST than for NNSE-MT for the modulated noises (ICRA and BABBLE), while the performance of the two NNSE strategies was comparable for SWN. Both NNSE-MT and NNSE-ST reduced the total error compared to UN for all noises and SNRs. These results suggested that an improvement in speech perception might be achieved and supported the clinical speech performance testing of CI users.

5.2.5 Results

The group mean SRTs for all processing conditions are shown in Fig. 5.7 and individual SRTs and their changes relative to those for the unprocessed condition (UN) are shown in Fig. 5.8. The data in all conditions were normally distributed, as tested with the Kolmogorov-Smirnov (using Lilliefors significance correction) and the Shapiro-Wilk tests. The SRTs used in statistical analyses were the average of the 3 SRTs obtained for each processing condition and noise type. Performance with UN was poorer (higher SRT) than with the processed conditions for all three noises. Group mean SRTs for speech in UN increased from 2.8 dB in SWN, to 5.1 dB in ICRA, and up to 6.7 dB in BABBLE. For all three noise types, lower mean SRTs were obtained with NNSE-MT and NNSE-ST than with UN. NNSE-ST achieved the lowest SRTs for all three noise conditions with an advantage of about 1 to 1.5 dB SRT over NNSE-MT.

A two-way analysis of variance (ANOVA) with repeated measures was conducted with factors processing condition (UN, NNSE-ST and NNSE-MT) and noise type (SWN, ICRA, and BABBLE). There were significant main effects of processing condition [$F(2,26) = 31.83, p < 0.001$], noise type [$F(2,26) = 37.63, p < 0.001$] and a significant interaction [$F(4,54) = 13.73, p < 0.001$].

Further statistical analysis was conducted separately for each noise type to compare the 3 processing conditions.

For SWN noise, Mauchly's test showed no violation of sphericity and a one-way repeated measures ANOVA indicated a significant effect of processing condition [$F(2,12) = 8.165, p = 0.006$]. *Post hoc* pairwise comparisons using Bonferroni correction revealed significant differences between UN and both NNSE-MT ($p = 0.019$) and NNSE-ST ($p = 0.003$), but not between NNSE-MT and NNSE-ST ($p = 0.10$), with improvements in SRT scores relative to those for UN of 1.4 and 2.3 dB for NNSE-MT and NNSE-ST, respectively. Apart from three subjects for NNSE-MT and one subject for NNSE-ST, subjects benefitted from the processing with both NNSE algorithms for speech in SWN.

For ICRA noise, Mauchly's test showed no violation of sphericity and a one-way repeated measures ANOVA indicated a significant effect of processing condition [$F(2,12) = 28.13, p < 0.001$]. *Post hoc* pairwise comparisons using Bonferroni correction revealed significant differences between UN and both NNSE-MT ($p < 0.001$) and NNSE-ST ($p < 0.001$) but not between NNSE-MT and NNSE-ST ($p = 0.67$), with improvements in SRT scores relative to those for UN of 5.4 and 6.4 dB for NNSE-MT and NNSE-ST, respectively. Apart from subject 14, all subjects benefitted from the processing with both NNSE algorithms for speech in ICRA. For some subjects, there were improvements in SRT scores of more than 10 dB.

For BABBLE noise, Mauchly's test showed a violation of sphericity (p = 0.023) and a one-way repeated measures ANOVA using the Greenhouse-Geisser correction indicated a significant effect of processing condition [$F(1.364,32.727) = 7.45$, p = 0.009]. *Post hoc* pairwise comparisons using Bonferroni correction revealed significant differences between UN and NNSE-ST (p < 0.001) and between NNSE-MT and NNSE-ST (p = 0.035). A significant improvement in SRT scores relative to UN was observed only for NNSE-ST. Apart from subject 4, all subjects benefitted from NNSE-ST for speech in BABBLE. For NNSE-MT, 8 out of the 14 subjects showed SRT improvements relative to UN of 1.5-3 dB. However, the rest of the subjects performed either the same or more poorly with NNSE-MT than with UN.

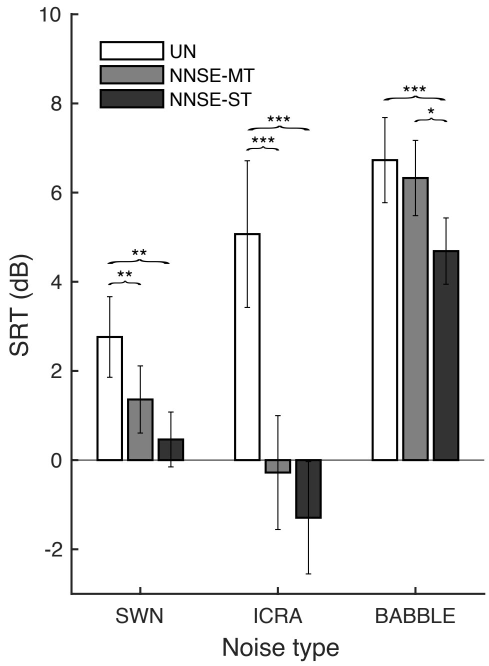


Figure 5.7 Group mean SRTs with UN (ACE), NNSE-MT (multi-talker) and NNSE-ST (single-talker) processing for each noise type (left: SWN, center: ICRA, right: BABBLE). Error bars represent the standard error of the mean; (*) p ≤ 0.05, (**) p ≤ 0.01, (***) p ≤ 0.001.

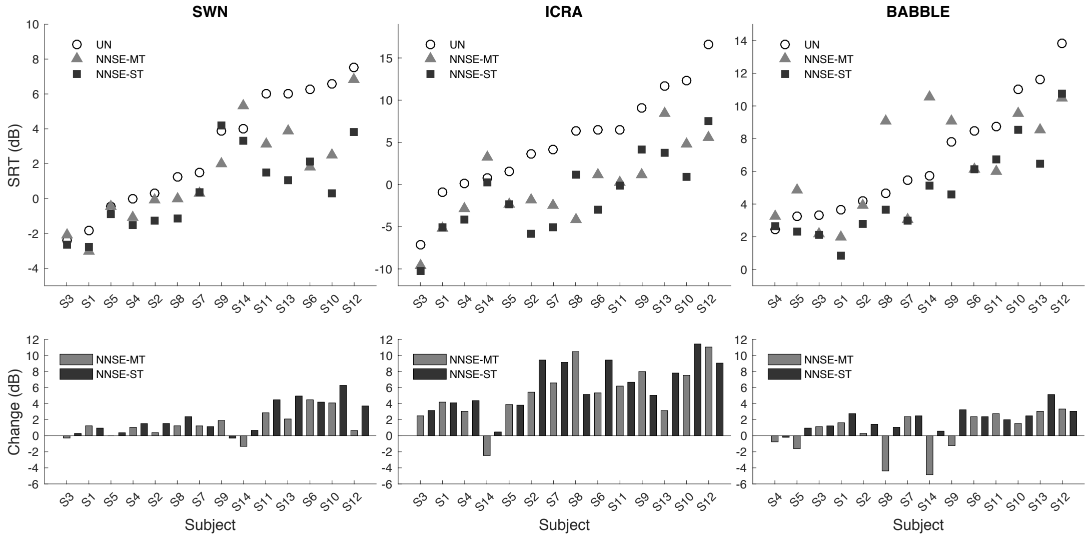


Figure 5.8 Top panel: Individual SRTs for UN (ACE), NNSE-MT (multi-talker) and NNSE-ST (single-talker) processing for each noise type (left: SWN, center: ICRA, right: BABBLE). Bottom panel: individual SRT change (positive is better) relative to the UN condition for NNSE-MT and NNSE-ST, for the three noises. Subjects are ordered by their performance for speech in UN (ascending SRT from left to right).

5.2.6 Discussion

Significant improvements in speech intelligibility for CI subjects were produced by NNSE for the three background noises over a range of SNRs. To accomodate the large variability among CI users, algorithm performance was evaluated using an adaptive procedure measuring SRT scores, in contrast to previous studies that tested at fixed SNRs. The magnitude of the improvements in SRT ranged from 1.4 dB for speech in SWN with NNSE-MT up to 6.4 dB for speech in ICRA with NNSE-ST. Apart from NNSE-MT with BABBLE, significant improvements were found for NNSE relative to UN in all conditions.

For SWN, improvements tended to be larger for NNSE-ST than for NNSE-MT (2.3 / 1.4 dB SRT), but this difference was not statistically significant. There was also a non-significant difference of 1 dB between NNSE-MT and NNSE-ST for ICRA (SRTs of 5.4 and 6.4 dB, respectively) but there was a significant difference of 1.6 dB for BABBLE (SRTs of 0.4 and 2.0 dB, respectively). The advantage of NNSE-ST over NNSE-MT was expected due to the mismatch between training and testing sets for NNSE-MT. Nevertheless, NNSE-MT led to significant improvements relative to UN for speech in SWN and ICRA despite the mismatch in speakers. NNSE-MT failed to give significant improvements relative to UN for BABBLE. For this noise condition, competing speakers might be wrongly detected as the target speaker and not attenuated adequately. Especially for lower SNRs, where the spectral energy of the target speaker was less dominant, NNSE-MT performed worse than NNSE-ST (it should be noted, that the training data were increased by nearly a factor of 2 for NNSE-MT, to increase its robustness to unseen speakers). The latter can use *a priori* information about the target speaker's spectral characteristics.

For ICRA, the improvements produced by NNSE (ST and MT) relative to UN were remarkable (about 5 to 6 dB) and were about 3 times larger than for the other two noise conditions. The average SRT for UN was comparable for ICRA and BABBLE. The processing produced a much larger improvement relative to UN for ICRA than for BABBLE. The ICRA noise employed in this study had much stronger spectro-temporal modulations (obtained from one male talker) than the BABBLE noise (20 talkers), leading to more and larger time-frequency (T-F) regions with a positive SNR. We speculate that the NNSE algorithm exploits these positive-SNR T-F regions in the feature space to predict adjacent or even more distant spectro-temporal patterns of the target speech signal. This would enable the algorithm to extrapolate its prediction over potentially masked T-F regions with lower SNR in the corresponding time frame (similar to the mechanism often called "glimpsing" or listening in the dips by human listeners). The algorithm was presented with numerous examples and variations of potential masking patterns during training and thus learned typical spectral patterns of the speech. This constitutes a potential benefit of machine learning algorithms in conjunction with acoustic broadband features over traditional signal processing schemes that operate independently on separate frequency channels.

The machine learning based algorithm proposed by Hu and Loizou (2010) showed large improvements in percentage correct scores for speech in three different non-stationary noise backgrounds for CI listeners. A direct comparison between the performance of their system and NNSE is difficult because we used an adaptive procedure in contrast to testing at fixed SNRs, and we used different speech materials and background noises. Hu and Loizou showed large improvements with an IBM-based processing scheme, but their system was trained on the same speaker, noise realizations and SNRs as used for testing. May and Dau (2014) showed that the use of novel noise realizations for testing led to a substantial decrease in estimation performance with a Gaussian Mixture Model (GMM) based system, such as the one used by Hu and Loizou. Recently, Healy et al. (2015) and Bolner et al. (2016) have shown that neural network based regression systems can achieve high estimation performance with novel realizations of the same noise type. Both studies tested at fixed SNRs and used acoustic stimuli to test normal hearing and hearing-impaired listeners' speech understanding in noise. Bolner *et al.* tested NH listeners using CI vocoder simulations and reported an improvement of 18% in percentage correct scores for speech in BABBLE at an SNR of 5 dB. This improvement can be compared to the 2-dB improvement in SRT for NNSE-ST, since the two algorithms used the same speaker for training and testing. Jansen et al. (2013) reported that, for CI users, an improvement in SRT scores of about 1 dB corresponds to an improvement in percentage correct scores of 18.7% with the LISTm corpus and SWN. This suggests that CI users benefitted more from NNSE processing than the NH listeners with CI simulations for speech in BABBLE. For SWN at 5 dB SNR, Bolner *et al.* measured an improvement relative to UN of 27%, whereas in this study an improvement of 2.3 dB was achieved by NNSE-ST. Again, this suggests larger benefits for CI users than for NH listeners, but less so than for BABBLE.

Other studies of single-microphone noise reduction for CI users showed consistent improvements in understanding of speech in stationary noise such as SWN (Dawson et al., 2011; Hu et al., 2007; Mauger et al., 2012; Ye et al., 2013). However, the improvements were usually smaller with non-stationary noise and only a few studies achieved significant improvements for both stationary and non-stationary noise (Dawson et al., 2011a). Machine-learning based algorithms like NNSE have the potential to overcome this challenge and achieve consistent improvements in both stationary and non-stationary noises, as indicated by the performance of NNSE with BABBLE and ICRA.

Several architectures for machine learning based noise reduction have been proposed in the last few years. In the studies of Kim et al. (2009) and Hu and Loizou (2010), GMM classifiers were used, which recently have been surpassed by artificial neural networks with several hidden layers (*deep neural network*, DNN) (Chen et al., 2016; Healy et al., 2013, 2015). Similar to the architecture of the previous GMM-based classification systems, where the SNR of each frequency channel is predicted independently, Healy et al. (2013) used two successive stages of multiple-subband DNNs (one DNN for each of the 64 frequency channels) resulting in a very large classification system. Healy et al. (2014) reduced the complexity of the DNN by a factor of 43 by using a single DNN for the prediction of the SNR of all frequency channels simultaneously. They used a DNN with 3 hidden layers, each composed of 1024 rectified linear units, and changed the feature extraction process to broadband features (being extracted across all frequency channels simultaneously) resulting in a greatly reduced number of features (64 times smaller) and an input layer dimensionality of just 259. However, this DNN system still had nearly 2.5 million tunable parameters. In the most recent studies on DNN-based speech separation, the complexity was increased again to DNNs with nearly 4 million (Healy et al., 2015) and more than 20 million tunable parameters (Chen et al., 2016). Recent advances in computational power through the use of supercomputers and graphics processing units (GPUs) made it possible to train and execute such complex algorithms in reasonable amounts of time. However, the application of such complex algorithms to hearing devices with strongly limited computational and memory resources is not feasible at present. In contrast, the NNSE algorithm uses a smaller number of relatively simple features combined with a much smaller NN regression system consisting of 2 hidden layers with 75 units each. This NN system has 18,631 tunable parameters, 2/3 of those used by Bolner et al. (2016). NNSE employs 200 times fewer parameters than the system used by Healy et al. (2015) and has a 1000-fold smaller system complexity than the system used by Chen et al. (2016).

Real-time processing requires a processing delay of less than 20-30 ms to ensure perceived audio-visual synchrony and acceptance by users of hearing devices (Stone and Moore, 2005). Besides the computational complexity aspect, which may become less relevant with the steady increase in computational power, the algorithm architectures used in many studies make use of non-causal processing involving the analysis of "future" frames (e.g. from feature sets using 2 future frames used by Healy et al., 2015, up to 11 future frames used by Chen et al., 2016). Generally, algorithms need to work in a causal way to be implementable in hearing devices that meet the perceptual requirements of potential end-users. The NNSE algorithm proposed in this study satisfies this requirement by using only the past and the current frames.

An important aspect of SE algorithms is their ability to generalize to unseen acoustic conditions. NNSE was designed to satisfy several generalization requirements. Firstly, multiple SNRs were used for training, yielding an algorithm that worked over a range of SNRs. This was assessed by using an error rate analysis where NNSE gave decreased total error rates relative to the unprocessed condition for all noise types and SNRs (and even for an untrained SNR of 10 dB). Secondly, novel realizations of a specific type of background noise were used for evaluation. NNSE performed well in these more challenging conditions (as it was also shown by Bolner et al., 2016, and Healy et al., 2015). Thirdly, NNSE-MT was tested using a novel speaker and substantial improvements were found for two out of three noise types. However, generalization to unseen types of noise was not assessed with the current study that used noise-specific training and testing. A future goal is to design a system that works in completely novel noise conditions, but still meets the constraints on delay and computational power of CI processors.

Kim and Loizou (2010) reported that a GMM classifier using amplitude modulation spectrum (AMS) features for estimating the IBM, that was trained on a large number of noise types, failed to achieve satisfactory performance with unseen noises (low classification rates). This was the case even when a speaker-dependent classifier was used. Instead of employing large-scale training to improve generalization, they proposed incrementally adapting the system to new noises. May and Dau (2014) have shown that a GMM-based classifier trained on AMS features tended to overfit the training data more when they increased the dimensionality of the feature space and the complexity of the classifier. The authors observed a larger decrease in classification performance when the algorithm was tested on novel segments of the same noise type for the more complex classifier and feature combinations than for the less complex ones (no evaluation on unseen noise types was performed). They proposed addressing the problem of overfitting with the use of a less complex classification system and a lower dimensionality of the feature space. Chen et al. (2016) used large-scale training with thousands of background noises in combination with a powerful DNN system and showed that generalization to unseen noises could be achieved when speaker-dependent models were used. This is a promising result and suggests that DNN-based systems might improve generalization to unseen noises compared to the GMM-based systems that were used in previous studies (Kim and Loizou, 2010; May and Dau, 2014).

GMM-based systems have been used mostly in combination with AMS features (Kim et al., 2009; Kim and Loizou, 2010; Hu and Loizou, 2010; May and Dau, 2014). Chen et al. (2014), showed that Gammatone-based features performed better than other features (including AMS) in terms of classification accuracy and HIT-FA rates with a DNN-based system. During the optimization of NNSE, we found similar results, confirming an advantage of Gammatone-based energy features over AMS features. We combined the processing paradigms of Gammatone-based RASTA-PLP features (that incorporate temporal aspects of speech such as modulations), and GFCC features (that perform a de-correlation of the spectral information), with log-compressed Gammatone-energy features in order to increase the robustness to noise and changes in speaker characteristics.

We performed a pilot experiment to evaluate the performance of the NNSE algorithm with unseen types of noise. We used 12 real-world recordings from different noisy environments (various recordings from a stadium, several restaurants and cafeterias, a classroom, a train, city and highway traffic situations; all obtained from freesound.org) and combined 20-s long segments of each recording to form a multi-noise recording with a total length of 4 minutes (the same length as employed for the noise-specific NNSE). The NNSE algorithm was trained on the multi-noise recording using the same procedure as for the listening experiment, and its performance to the noises employed for the training of the noise-specific NNSE was assessed objectively using the NCM speech intelligibility predictor. The NCM scores are shown in Fig. 5.9 for the single- and multi-talker NNSE algorithm for both noise-specific and noise-independent training (the NCM scores were calculated using 20 sentences from the LISTm corpus).

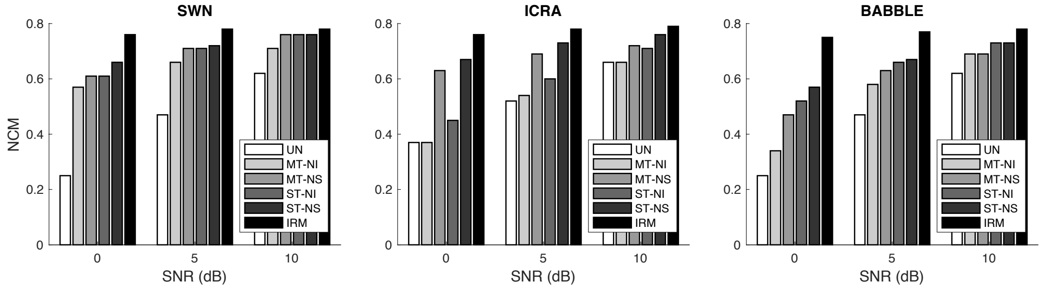


Figure 5.9 NCM intelligibility prediction scores for UN (ACE), MT-NI (NNSE-MT with noise-independent training), MT-NS (NNSE-MT with noise-specific training), ST-NI (NNSE-ST with noise-independent training), ST-NS (NNSE-ST with noise-specific training) and IRM (ideal ratio mask) for each noise type (left: SWN, center: ICRA, right: BABBLE).

For SWN and BABBLE, there was a small decrease in performance with the noise-independent algorithm compared to the noise-specific algorithm for NNSE-ST, and a larger decrease in performance with the noise-independent algorithm compared to the noise-specific algorithm for NNSE-MT. Interestingly, large improvements in NCM scores for both NNSE-ST and NNSE-MT were achieved with the noise-independent algorithms relative to UN. This is promising, because NCM was proven useful for predicting intelligibility outcomes for vocoded stimuli in our pilot study using CI simulations (Bolner et al., 2016) and for CI users (Chen and Loizou, 2011), but it remains unclear if the predicted improvements relative to UN will occur for CI users. For ICRA, the performance of the noise-independent algorithm was much reduced in comparison to that for the noise-specific algorithm for NNSE-ST, and the predicted performance of the noise-independent algorithm equaled that for UN for NNSE-MT (it should be noted that the noise-independent algorithm did not impair intelligibility relative to UN). We speculate that the difference in predicted performance between noise conditions depends on the degree of similarity of the spectro-temporal characteristics between the training and testing noise types. The NCM scores indicate that both the speaker-dependent and the speaker-independent NNSE algorithms generalize better to unseen noise types for cases when the spectro-temporal modulation patterns are somewhat similar between the training and testing noises (as was the case for SWN and BABBLE) than when the training and testing noises contain different spectro-temporal modulation patterns (in the case of ICRA). Instead of using multi-noise training to increase algorithm performance in unseen noise types, a noise-specific algorithm could be combined with an environmental classifier to provide *a priori* knowledge about the noise type (Hazrati et al., 2014; May and Dau, 2013), while retaining the advantages of high SE performance in combination with low processing delay and potentially reduced computational complexity compared to a “one-for-all” large-scale algorithm.

6 User-Controlled Parameters for Speech Enhancement Algorithms

Overview

Hearing aids and cochlear implants employ speech enhancement techniques to improve the perception of speech in difficult listening situations with background noise. The goal is to achieve benefits in speech intelligibility in noise without disturbing the listeners' awareness of the acoustic environment - an aspect that is often not considered in the evaluation of noise reduction algorithms. The ideal ratio mask (IRM) gain function has been demonstrated to provide large improvements in speech intelligibility in background noise up to the level obtained in quiet acoustic conditions. However, little research has been conducted to test listeners' awareness of background sounds when listening to IRM-enhanced speech. For practical use in hearing devices, individual differences between listeners may require tailored parameters for achieving the goal of both improved speech intelligibility and maintained awareness of the acoustic environment.

This study begins to fill this gap in knowledge by investigating user-controlled parameter choices for IRM-processed speech in noise with normal-hearing and hearing-impaired listeners. Using a parameterisation procedure, subjects controlled their indvidually-preferred noise reduction parameters that were evaluated using objective tests of noise awareness, speech intelligibility and perceptual quality ratings in terms of speech distortion, noise intrusiveness and overall speech quality. Results indicate that NH and HI listeners preferred stronger noise reduction settings than the conventional IRM gain function but no significant differences were found for the mean parameter choices between NH and HI groups. There was a significant difference in the variability between NH and HI groups indicating that NH subjects were more robust to changes in NR parameters and tolerate a wider range of parameter values than HI subjects for speech enhancement. For HI subjects, non-optimal parameter choices might affect their speech perception in noise and awareness of background sounds more severely.

6.1 COMPARISON OF USER-CONTROLLED NOISE REDUCTION PARAMETERS BETWEEN NORMAL HEARING AND HEARING IMPAIRED LISTENERS: FINDING THE BALANCE BETWEEN STRENGTH OF NOISE SUPPRESSION AND AWARENESS OF BACKGROUND SOUNDS

Declaration of authorship:
Tobias Goehring - Leading the research, developed the algorithm and software for the listening experiment, study design and analysis of results, writing the manuscript

Co-authorship:
Marina Forbes - Performed the listening experiment, study design
Prof. Stefan Bleeck - Advising the research

6.1.1 INTRODUCTION

One of the most common complaints of hearing aid (HA) users is that they gain little benefit when listening to speech in situations with background noise. This may ultimately lead to a rejection of their HA (Nabelek et al., 1991). An enhancement in both the quality and intelligibility of speech in background noise is desired for users of HAs as it may increase the usage of HAs and reduce the occurrence of social withdrawal (McCay, 1996; Mick et al., 2014). Furthermore, new HA users may be better able to cope and adapt to their HAs if they find that they function in everyday listening situations with background noise.

Speech enhancement techniques have been developed to improve speech perception in background noise such as the Wiener Filter and Spectral Subtraction (Boll, 1979) algorithms. While these have been shown to provide an improvement in speech quality and listening effort, speech intelligibility in background noise remained mostly unaffected (Loizou, 2013). In recent years, algorithms based on time-frequency masking have been proposed for speech enhancement in HA and CI users (Kim and Loizou, 2009; Hu and Loizou, 2010; Healy et al., 2013; Healy et al., 2015; Bolner et al., 2016; Chen et al., 2016). These algorithms were evaluated with HA users in terms of speech intelligibility in background noise and shown to deliver significant improvements.

In practice, SE algorithms are desired to achieve an improvement in speech understanding and perceived quality while maintaining an awareness of the acoustical environment. If too strong noise reduction is applied, background sounds may be attenuated below audibility. This may disturb the listeners' perception of their environment leading to discomfort, feeling of isolation or even dangerous situations if emergency signals or important announcements would be missed out. In order to avoid this problem for speech enhancement algorithms, a compromise between speech enhancement performance and attenuation of background sounds has to be achieved. This compromise may be affected further by individual preferences of the listeners, their ability to understand speech in background noise and their ability to detect and perceive environmental sounds that may also depend on their hearing abilities. While hearing loss negatively affects listeners' performance in understanding speech in background sounds, it is unclear whether hearing loss also affects their ability to detect environmental sounds. This is likely the case because the audibility of environmental sounds is reduced by noise reduction processing making them harder to detect. While hearing ability is assumed to play an important role for determining the optimal balance between background reduction and background awareness, additional individual differences between listeners may require different parameter settings for each listener.

Brons et al. (2012) compared several noise reduction algorithms in their benefits for normal hearing listeners' speech intelligibility (N=10) in background noise and perceptual speech quality. They found the IBM to be most effective for improving speech intelligibility, but listeners did not prefer it over the unprocessed noise-corrupted signal in terms of perceptual speech quality. They reported that listeners preferred the stimuli processed with a tempered binary mask, that applied a more gradual gain function like the IRM and a minimum gain of 0.2 rather than 0 with the IBM and thus avoided the introduction of musical noise. However, no individual differences were reported, the parameter settings (e.g. strength of noise reduction, maximum attenuation) were kept the same for each participant and no evaluation of the awareness of the background sounds was performed.

Brons et al. (2014) investigated the effect of the strength of noise reduction processing on detection thresholds for signal distortion in NH and HI listeners (N=12 per group). HI listeners had higher detection thresholds for signal distortion than NH listeners, indicating that stronger noise reduction may be applied for HI listeners. It was also found that HI listeners preferred noise reduction settings closer to their individual detection threshold for distortion compared to the NH group suggesting that HI listeners tolerate fewer audible distortions than NH listeners. This means that although HI listeners might tolerate higher amounts of noise reduction before they detect distortions compared to NH listeners, they will tolerate less audible distortions above detection threshold compared to NH listeners. The authors suggest that speech distortions should be avoided for HI listeners. Overall, the preferred strength of noise reduction did not differ between NH and HI groups. They reported large inter-individual differences in the preferred strength of noise reduction for both NH and HI groups (with maximum attenuation values ranging from 4.2 up to 27 dB).

The study by Daniel et al. (2013) required normal-hearing listeners (N=15) to adjust a gain function for noise reduction by using a blind-adjustment method with a graphical user interface (GUI). User-controlled adjustments were made to obtain a perceptually optimal gain function using an iterative procedure with stimuli comprised of 3 sentences mixed with 3 types of stationary noise at an SNR of 0 dB. The individually chosen gain functions were not reported but averaged across listeners to obtain a perceptually optimal gain function that was compared to the conventional Wiener Filter gain function (similar to IRM). As a result, listeners preferred stronger noise suppression by about 5 dB than that used in the conventional Wiener Filter and it was found that the experimentally-derived gain function produced less musical noise than the Wiener filter in their setup. However, the resulting gain function was very similar in shape to the Wiener function, mainly shifted towards higher attenuation by about 5 dB, and no objective measures were reported that support the effectiveness of the obtained gain function in terms of speech intelligibility for different SNRs or other types of background noise. The awareness of background sounds or their intrusiveness was not addressed.

For CI users, Mauger et al. (2012) performed a listening study to evaluate perceptually optimized gain functions (N=10). In their first experiment, they found an SNR-threshold of about 5 dB as optimal for improving speech intelligibility in speech weighted noise and 20-talker babble using the IBM approach. This threshold was higher than the standard SNR-threshold used for NH listeners of -5 or 0 dB. However, this difference might have occured because CI users were tested at SNR values where they obtained 50% speech understanding and these SNRs might have been higher than the ones typically used for testing NH or HI listeners. This might have led to a higher SNR-threshold for IBM processing. In a second experiment, CI users chose the threshold and slope values of a parametric Wiener gain function in the ideal case where the *a priori* SNRs were given (similar to

IRM processing). They found large inter-individual differences between the preferred parameter values and a mean gain function threshold that was shifted towards higher *a priori* SNRs by about 6.8 dB in comparison to the standard Wiener filter. This threshold shift of 6.8 dB is somewhat similar to their finding of an increased SNR threshold by about 5 dB in the case of IBM processing and supported CI users' preference for more aggressive noise reduction parameters.

Recently, Neher and Wagener (2016) published a study on the differences in preferred noise reduction strength among hearing aid users (N=27). Listeners were split into two groups depending on their preference for ("NR lovers") or against ("NR haters") noise reduction processing with a binaural-coherence based algorithm, as determined in an earlier study (Neher, 2014). Similar to earlier findings, large inter-individual variability was found for the preferred NR strength and NR strength was chosen to be stronger at higher input SNR (4 dB in comparison to 0 dB). The group of NR lovers were reported to prefer significantly stronger NR settings than the group of NR haters. However, outcome measures in terms of the acceptable noise level (ANL, Nabelek et al., 1991), detection thresholds for speech distortions and self-report measures did not show clear differences between groups. The authors claim that NR strength is an individual trait that could be individually chosen with a self-adjustment procedure during the fitting process of hearing aids.

The ANL determines the highest level of background noise that listeners tolerate and has been used as a predictor for the success of hearing aid use for HI listeners (Nabelek et al., 2006). The procedure to determine the ANL consists in firstly setting a speech stimulus (usually a recording of a story is used) to a comfortable loudness and then adjusting the level of the background noise (usually multi-talker babble is used) to a maximum level at which the listener "can put up with" without becoming tense or tired while listening to and following the words of the story. This procedure yields a maximum tolerated background noise level that may indicate how well a listener can cope with speech in noise, however no comprehension or speech intelligibility test was performed that could confirm the success in finding an apropriate level of background noise, neither can this maximum tolerated level be taken as the perceptually optimal or individually preferred level of background noise by the listener. These drawbacks prevent the ANL from being a suitable measure for self-adjustment of a noise reduction algorithm.

As reported earlier, speech enhancement algorithms have been evaluated with settings that were kept constant for a group of listeners. Only a few studies allowed listeners to choose the parameters of noise reduction processing and evaluated the differences in the preferred choice and the obtained benefits between listeners.

In the present study, the parameters of a single-microphone noise reduction algorithm based on the IRM were adjusted by NH and HI listeners. The listeners adjusted the strength of noise reduction and the attenuation of the background sounds in order to obtain a compromise between improvements in speech intelligibility in background noise and maintaining an awareness of the acoustic environment. These two aspects of noise reduction processing are assumed to be of high importance for the use of hearing aids in practice. We evaluated the user-controlled parameters in terms of their effectiveness for speech enhancement in noisy conditions, by measuring speech intelligibility benefits, the awareness of realistic background sounds by measuring classification and detection accuracy for background sounds and the subjective preference in terms of quality in comparison to the unprocessed condition and a conventional Wiener filter processing condition by using a subjective rating task. We aim to explore the inter-individual differences in listeners' preferred choice of noise reduction parameters and the inter-group differences between NH and HI listeners. The strength of noise reduction as well as the maximum attenuation of background sounds are both assumed to affect the performance in the evaluation tasks and are evaluated in terms of their influence on the outcome of those.

6.1.2 Methods

Subjects

In total, 30 native-English speaking adults took part in this study, of which 16 had normal hearing (NH, 12 female, average age of 28 years) and 14 had a mild-to-moderate sensorineural hearing loss (HI, 7 female, average age of 68 years). Hearing thresholds were confirmed via pure tone audiometry (PTA) to be within normal limits for the NH group (<= 20 dB HL at 250, 500, 1000, 2000, 3000, 4000, 6000, 8000 Hz) and not higher than 80 dB HL for the HI group. The better hearing ear was chosen as the test ear during the study. The HI group's experience with the use of hearing aids ranged from no amount of experience up to 25 years of experience. In addition to PTA, otoscopy and a screening questionnaire were performed to exclude subjects with conductive hearing impairments, tinnitus, hyperacusis or acute consequences of recent surgeries, infections or exposures to loud sounds. The pure-tone hearing thresholds of the HI group are shown in Figure 6.1 and demographic data is provided in Table 6.1. The subjects in the HI group did not wear their hearing aids during the experiment.

Table 6.1 Demographic data for the hearing-impaired group.

Subject	**Age**	**Tested ear**	**Gender**	**HA type**	**HA exper.** in years	**250 Hz PTA**	**500 Hz**	**1000 Hz**	**2000 Hz**	**3000 Hz**	**4000 Hz**	**6000 Hz**	**8000 Hz**
HI1	69	R	F	Starkey ITES	0.8	15	15	25	40	40	50	60	55
HI2	75	R	M	Siemens Octiv S open-fits	0.3	25	20	15	40	70	60	65	65
HI3	50	L	F	Oticon Spirit	20	50	40	40	50	45	45	45	40
HI4	60	L	F	Siemens Impact Pro M	3	25	45	50	40	30	20	15	15
HI5	53	L	F	Siemens Impact L open-fits	3	15	20	25	20	40	35	30	40
HI6	78	R	M	Phonak Ambra open-fits	1	10	10	25	35	30	55	55	50
HI7	80	L	M	Danalogic Resound	20	25	45	45	50	65	70	70	80
HI8	80	L	M	Siemens Octiv M open-fits	0.3	30	30	45	65	60	65	60	60
HI9	50	R	F	Octiv S open-fits	0.5	20	15	20	35	35	55	60	60
HI10	62	L	M	Siemens Impact Pro M	11	10	15	20	25	40	50	55	60
HI11	78	R	M	Specsavers Adv. open-fits	3	35	30	40	50	65	70	70	60
HI12	76	R	F	Bernafon open-fits	25	35	35	40	45	55	65	70	75
HI13	72	R	M	Resound open-fits	2.5	0	5	15	45	60	65	55	60
HI14	65	L	F	Siemens Impact Pro M	15	20	10	15	50	55	65	60	65

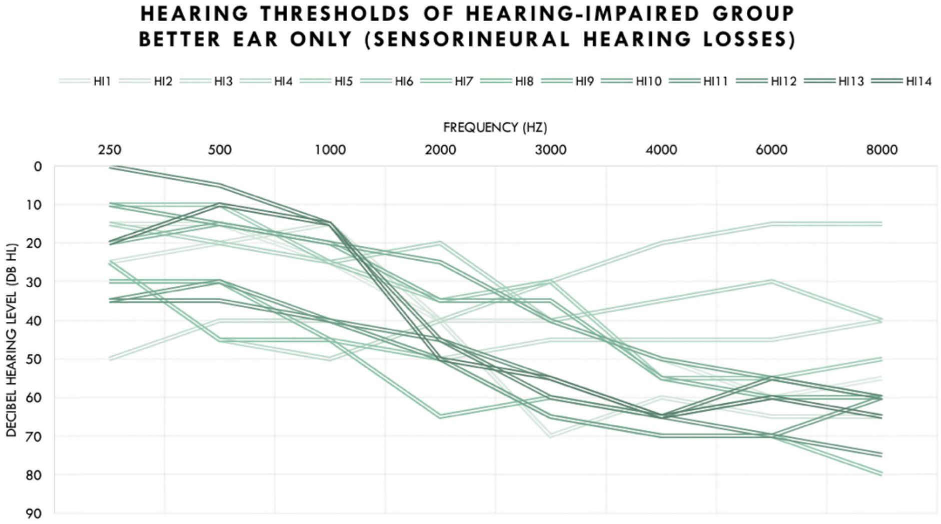


Figure 6.1 Hearing thresholds of hearing-impaired group.

Stimuli and technical equipment

In all parts of the experiment, the target speech stimuli consisted in lists from the IEEE corpus (spoken by a British English, male talker) that contain 5 contextual and 5 non-contextual sentences of moderate difficulty in each list of ten sentences. Each sentence contained five key words. The interfering noises were recordings of realistic, non-stationary background sounds taken from the NOISEX-92 database (100-talker canteen babble, 4-talker babble and factory noise). Both speech and noise stimuli were resampled to a sampling frequency of 16 kHz and mixed together at SNRs ranging from -10 up to 5 dB based on their average RMS level and according to the experimental part.

The stimuli were generated on a laptop (Dell Latitude E7440) that was connected to an external soundcard (RME Babyface Pro) and were presented to the subject via closed, circumaural headphones (Sennheiser HD380pro). The setup was calibrated to a presentation level of 70 dB(A) using the unprocessed noise-corrupted signals and a sound level meter (Bruel&Kjaer 2660) and artificial ear (Bruel&Kjaer 4153). For the HI group, an additional hearing-loss dependent gain was applied in each part of the experiment according to the NAL-R procedure (Byrne and Dillon, 1986). The experiment took place in a quiet but not sound-isolated room at the University of Southampton.

Signal processing for noise reduction

The noise reduction framework used for this experiment consisted in an analysis and synthesis scheme based on a Gammatone filterbank (Hohmann, 2002) that transformed the stimuli into the spectro-temporal domain for the processing. The input signal was segmented into 20-ms long time-frames with an overlap of 10 ms. Each segment was passed through a 32-channel Gammatone filterbank with centre frequencies ranging from 50 up to 8000 Hz (1 channel per equivalent rectangular bandwidth, ERB; Glasberg and Moore, 1990). The enhanced output signal *X* was obtained by applying a gain function *G* to the noise-corrupted envelopes in each Gammatone channel and recombining the enhanced envelopes with the noisy phase information:

with t denoting the time frame, f denoting the frequency channel and φ_Y denoting the phase information of the complex noise-corrupted signal $Y(t,f)$.

Given the challenge of single-microphone speech enhancement, the goal of a noise reduction algorithm is to obtain an estimate *x* of the target speech signal *s* from the noise-corrupted input signal *y*. In the ideal case, when both target and interfering signals are known, uncorrelated and added together to obtain the noise-corrupted signal, the ideal Wiener filter $G_{IWF}(t,f)$ is obtained by computing the Wiener filter gain function in the spectro-temporal domain using the *a priori* signal components:

$$G_{IWF}(t,f) = \frac{S^2(t,f)}{Y^2(t,f)} = \frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)} = \frac{\xi(t,f)}{1+\xi(t,f)},$$

with $\xi(t,f)$ denoting the *a priori* SNR, t denoting the time frame and f denoting the frequency channel in a time-frequency representation of the signals in the Gammatone domain. To obtain the *a priori* SNR for the calculation of the gain function, both speech and noise signals were passed separately through the Gammatone fiterbank and their average envelope power was computed by taking the square of the averaged absolute value of the filterbank output in each time-frequency unit.

For the user-controlled noise reduction parameters, the subjects chose a parametric Wiener filter gain function $G_{IRM\_CH}(t,f)$ (denoted as ideal ratio mask of choice, IRM_CH), where the values of the threshold α (*alpha*) and exponent β (*beta*) controlled the strength of the noise reduction by changing the steepness and threshold of the gain function:

$$G_{IRM\_CH}(t,f) = \left(\frac{\xi(t,f)}{\alpha+\xi(t,f)}\right)^{\beta},$$

Furthermore, a maximum attenuation γ (*gamma*) was used to limit the attenuation of the background sounds by hard-clipping the gain function at this value:

$$G_{IRM\_CH}(t,f) = \begin{cases} G_{IRM\_CH}(t,f) & ,if \;\; G_{IRM\_CH}(t,f) > y \\ \gamma & ,if \;\; G_{IRM\_CH}(t,f) \leq y \end{cases}$$

Self-adjusted noise reduction: parameterization procedure

The user-controlled noise reduction processing was implemented with an iterative parameterization procedure by letting the subjects control the three parameters *alpha*, *beta* and *gamma* of the gain function $G_{CH}(t,f)$. A graphical user interface (GUI) was implemented in MATLAB software that used two continuous sliders (with minimum MIN indicated at the far left and maximum MAX indicated at the far right) to control the strength of noise reduction (*alpha* and *beta*) and the maximum attenuation of background sounds (*gamma*). Note that the parameters *alpha* and *beta* were controlled jointly using the first slider (labelled "Strength of noise reduction"). This was done to simplify the parameterization procedure for the subjects by using only 1 slider for the strength of noise reduction.

The second slider controlled the parameter *gamma* (labelled "Background noise attenuation"). The first five sentences from the IEEE corpus were selected for the parameterization procedure and mixed with two noises at two SNRs (sentence 1: canteen and -5 dB SNR, sentence 2: factory and 0 dB SNR, sentence 3: canteen and 0 dB SNR, sentence 4: factory and -5 dB SNR, sentence 5: canteen and -5 dB SNR). This was done to incorporate different types of background noise and SNR levels within the parameterization procedure.

Before the parameterization procedure started, subjects were instructed that the goal of this part was to choose their individually preferred setting for the speech-in-noise processing of an imaginary hearing aid. The experimenter made clear to the subjects that this part of the experiment was crucial for the following parts. Firstly, the subjects were given the instruction to start adjusting the first slider to control the strength of the noise reduction with the goal to enhance the speech component ("adjust the slider until you can hear the speech perfectly"). Secondly, the subjects were told to use the second slider to adjust the maximum attenuation of the background noise ("adjust the slider until you can just about hear the noise but are still comfortably aware of it").

Figure 6.2 shows how the subjects could control the noise reduction processing using the GUI and choosing different settings of the sliders (from left to right: mild, medium and strong noise reduction settings) and the underlying processing in terms of the chosen gain function (first row), the time-frequency mask that was computed with the chosen gain function (second row) and the waveform of the noise-corrupted input and enhanced output signals (third row). The parameterization procedure started with the playback of sentence 1 processed with the noise reduction once the subject clicked on the slider to adjust the strength of the noise reduction. The subject was allowed to switch between and adjust both sliders as many times as needed until satisfied with the result. Subjects were able to control the strength of NR (*alpha* and *beta*) with a value range of 0.09 up to 11.22 and the maximum attenuation (*gamma*) within a range of 0.95 down to 0.03 (-0.4 down to -31 dB). The GUI allowed to switch processing parameters at any point during the play back. If the subjects were satisfied with their choice, the choice of parameters was saved and the next sentence was used for a new parameterization process. This was repeated until subjects chose their preferred parameters for all five sentences. The initial starting position of the sliders was randomly chosen for each subject. The change in sentence, background noise and SNR condition was found to be sufficient to let the subject explore the parameter space again for each parameterization. The first and last sentence contained the same noise condition and SNR, but different variations were used for the sentences in between (2-4) to let the participant explore several settings of the parameters. It was assumed that subjects were most comfortable with the procedure and their choice of parameters for the last sentence, and this choice of parameters was taken as the final choice used for the following parts of the experiment.

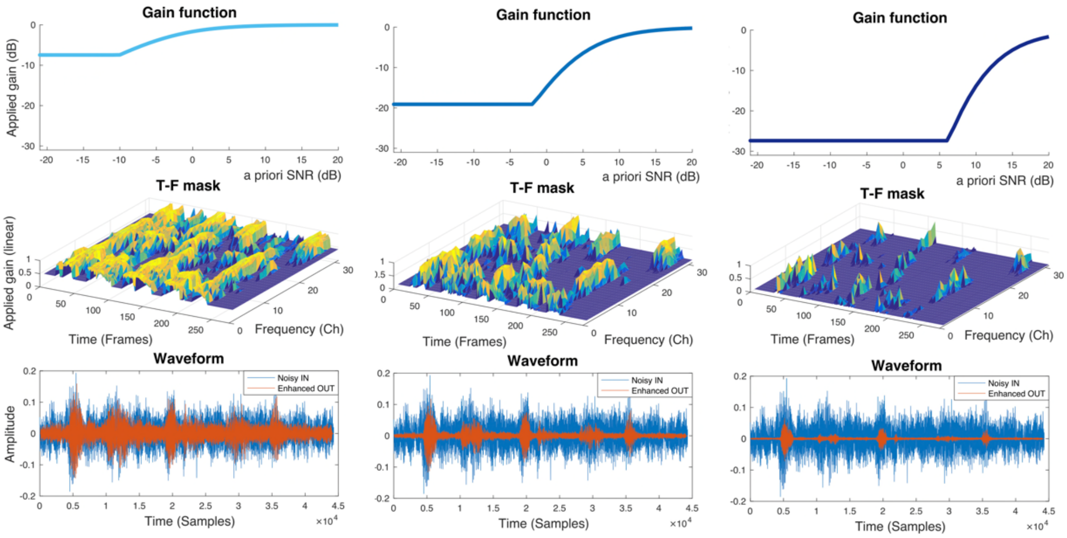


Figure 6.2 From left to right: mild, medium and strong noise reduction settings with gain functions (top), resulting time-frequency mask for a sentence (middle) and comparison between noise-corrupted and enhanced output waveforms (bottom). Subjects controlled the strength of noise reduction by changing the steepness and maximum attenuation of the gain function, as shown in the first row, while listening to the enhanced output sound shown in the bottom row.

Evaluation of awareness of background sounds

This test was designed to evaluate the subjects' awareness of background sounds with their preferred noise reduction parameters in a detection and classification task. To generate the stimuli for this experimental part, sentences from the IEEE corpus (that were not included in the other parts of the experiment) were mixed at an SNR of -5 dB with four recordings of environmental sounds (these included recordings in a garden, coffee shop, by a lawnmower and on a motorway; obtained from musicradar.com). Additionally, one of four recordings of acoustic cue sounds was added to the mixtures at arbitrarily chosen positions (cue sounds: glass crashing, footsteps, siren and horn; obtained from musicradar.com). The four recordings had lengths lasting from 8 to 15 seconds. The subjects were told to ignore the target speech for this task and concentrate on the background sound. Subjects were not allowed to repeat the stimulus presentation. After listening to the stimulus, the subjects chose firstly an acoustic environment (there were four columns, each with three choices of acoustic environments: Forest, Garden, Zoo / Bar, Coffeeshop, Kitchen / Propeller-plan, Lawnmower, Motorboat / Airplane, Motorway, Railway) and secondly an acoustic cue (with the choices: Siren / Glass crash / Footsteps / Horn / None). Correct detection of either acoustic environment or cue was rewarded equally and the final awareness score was calculated as the ratio of correct detections divided by the maximum possible score of 8.

Evaluation of speech intelligibility

This test was performed to evaluate the subjects' performance in terms of speech intelligibility in background noise with the self-adjusted noise reduction in comparison to noise-corrupted speech.

Sentences from the IEEE corpus were mixed with different types of background noise (canteen, factory and 4-talker babble) and at several SNRs (-10, -5 and 0 dB). A short practice session was performed using one list (10 sentences) mixed with canteen noise at an SNR of +8 dB to let the subjects familiarise with the task. For the test, two unprocessed conditions (canteen and factory at an SNR of -5 dB) and five conditions processed with the self-adjusted noise reduction (canteen at SNRs of -10, -5 and 0 dB, factory and 4-talker babble at -5 dB SNR) were generated using seven lists of the IEEE corpus. We also included a new noise type (4-talker babble) and SNR (-10 dB) condition that were not covered by the parameterization procedure to evaluate the self-adjusted noise reduction in both known and unknown acoustic conditions. The order of processing conditions was randomized for each subject. The subject listened to each sentence and repeated back what they heard. The percentage correct score was calculated by counting the correctly recognised key words divided by the total amount of key words.

Evaluation of subjective preference

This test evaluated the subjects' perceptual rating of speech distortion, noise intrusiveness and overall signal quality according to the guidelines of the ITU-T P.835 standard for the unprocessed, processed with the self-adjusted noise reduction and processed with the ideal Wiener filter conditions. The ideal Wiener filter was included as a reference condition that is considered as an optimal choice for noise reduction processing in terms of perceptual speech quality (with parameter values of *alpha* = 1, *beta* = 1 and *gamma* = 0.1, yielding the standard Wiener filter with a maximum attenuation of 20 dB). Three stimuli, one per processing condition, were generated for each of four sentences from the IEEE corpus together with two types of noise (canteen and factory) and three SNRs (-10, -5 and 0 dB): sentence 1 mixed with canteen noise at -5 dB SNR, sentence 2 mixed with factory noise at -5 dB SNR, sentence 3 and 4 mixed with canteen noise at 0 and -10 dB SNR, respectively. A GUI was used that comprised 3 sliders per perceptual rating condition where each slider represented one of the processing conditions. The subjects were instructed to click on a slider for play back and to position the slider at any point within the scale limits to rate the sound according to the condition (with the three perceptual rating conditions: 'Speech Distortion' with labels 'Not Distorted' and 'Very Distorted'; 'Background Noise Intrusiveness' with labels 'Not Noticeable' and 'Very Intrusive'; 'Overall Speech Quality' with labels 'Bad Quality' and 'Excellent Quality'). The middle points of the three scales were explained to the subjects as 'somewhat distorted', 'noticeable but not instrusive' and 'fair' according to the ITU-T P.835 standard. Subjects were allowed to repeat the play back as many times as required and were able to directly compare the different processing conditions by switching between them seamlessly. The starting position of all the sliders was randomized for each of the four sentences. The subjects were blinded as to which processing condition they were listening to and the allocation of the sliders (indicated with numbers 1 to 3) was randomized for each sentence.

Study design

The experiment used a repeated-measures design to evaluate the subjects' individually chosen NR parameters in a set of evaluation tests. The experiment was approved by the Ethics and Research Governance Online (ERGO, Ethics number 17386) procedure of the University of Southampton including the completion of a risk assessment form and noise exposure calculation to ensure the safety of the subjects. All collected experimental data was anonymised and treated confidentially in accordance with the University Data Protection policy. Subjects were not paid for participation but reimbursement of travel expenses was offered. Subjects were given an information sheet to explain the purpose and procedure of the experiment, signed a consent form and filled out a screening questionnaire at the start of the experiment.

The experiment was divided in two main parts: parameterization and evaluation. Subjects firstly completed the parameterization procedure to determine their individually preferred noise reduction parameters. This part lasted about 15-20 minutes. The chosen parameters for noise reduction were then evaluated in the test part of the experiment comprising the three evaluation tests in successive order: evaluation of awareness of background sounds, evaluation of speech intelligibility and evaluation of subjective preference. The three test parts lasted about 30 minutes and short breaks were offered after each test. In total, the experiment lasted about 1.5 hours including the PTA measurement.

6.1.3 RESULTS

Self-adjusted noise reduction parameters

Final gain functions of NH and HI groups are shown in Figure 6.3 and NR parameters are reported in Table 6.2. To compare the choice in user-controlled parameter values between groups, a repeated-measures ANOVA was performed with between-subject factor hearing group (NH/HI) and within-subject factors of chosen NR parameter values (strength and attenuation of background noise). There was a significant difference between the parameter values for strength of noise reduction and the maximum attenuation of the background noise [F(1,28) = 142.795, $p <= 0.001$]. There was no difference between the mean parameter values chosen by the NH and HI groups [F(1,28) = 1.262, $p = 0.828$] for both strength and attenuation of background noise. However, there was a significant effect of unequal variance for the choice of NR strength between the HI and NH groups as determined by Levene's Test of Equality [F(1,28) = 8.866, $p = 0.006$]. For the parameter of attenuation of the background noise, variances were equal between groups [F(1,28) = 0.906, $p = 0.349$].

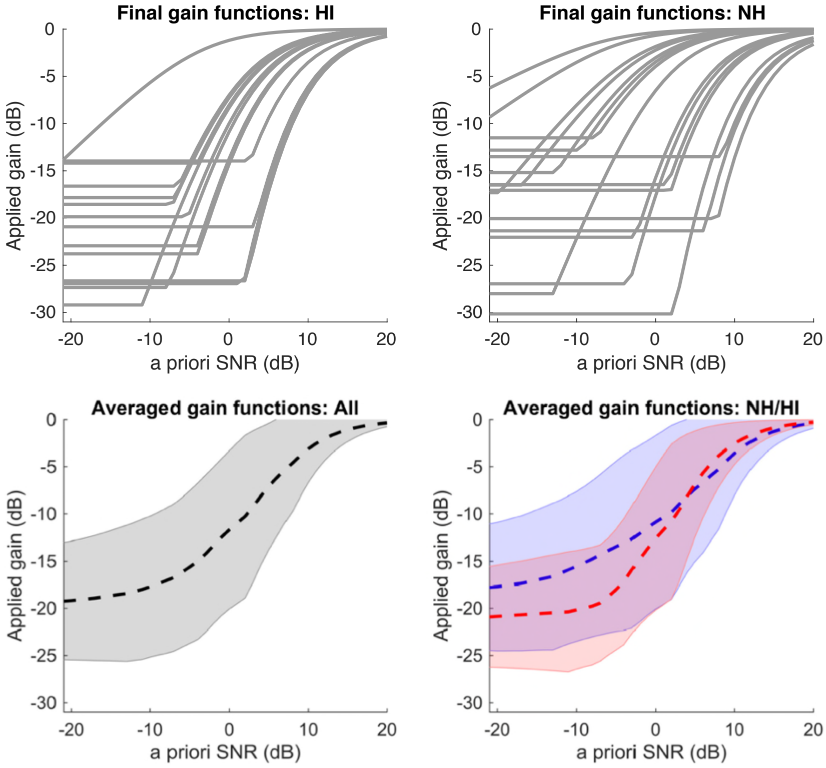


Figure 6.3 Final user-controlled noise reduction gain functions for hearing impaired (HI, top left), normal hearing (NH, top right) and the averages for all (bottom left) and NH and HI groups (bottom right, NH in blue, HI in red).

Table 6.2 Final user-controlled noise reduction parameters for the normal hearing (NH) and hearing impaired (HI) groups.

Subject	Alpha/Beta	Gamma	Gamma (dB)	Subject	Alpha/Beta	Gamma	Gamma (dB)
NH1	0.29	0.05	-26.67	HI1	1.11	0.13	-17.83
NH2	0.79	0.27	-11.51	HI2	1.19	0.20	-14.19
NH3	1.91	0.04	-26.96	HI3	1.09	0.15	-16.62
NH4	0.68	0.17	-15.19	HI4	2.98	0.05	-26.65
NH5	0.22	0.41	-7.74	HI5	2.82	0.09	-20.93
NH6	1.78	0.08	-22.04	HI6	1.42	0.04	-27.35
NH7	3.66	0.21	-13.50	HI7	0.40	0.13	-17.46
NH8	1.05	0.04	-28.01	HI8	1.63	0.07	-22.94
NH9	2.10	0.15	-16.46	HI9	1.21	0.03	-29.20
NH10	3.81	0.09	-21.35	HI10	1.68	0.06	-23.79
NH11	0.51	0.14	-17.31	HI11	2.12	0.20	-13.96
NH12	0.73	0.23	-12.81	HI12	1.12	0.12	-18.55
NH13	4.34	0.10	-20.06	HI13	1.35	0.10	-19.86
NH14	2.23	0.14	-17.07	HI14	2.91	0.04	-26.94
NH15	0.57	0.15	-16.51				
NH16	3.25	0.03	-30.15				
AVG	**1.75**	**0.14**	**-18.96**	**AVG**	**1.65**	**0.10**	**-21.16**
STD	1.37	0.10	6.47	STD	0.78	0.06	5.05

Test-retest variability and mean choice of NR parameters

In order to test the reliability of the parameterization procedure, 10 of the subjects (5 NH and 5 HI) performed a retest experiment under identical conditions about 4 months after the experiment. Subjects completed the parameterization procedure again by choosing their preferred NR settings. The final user-controlled parameters chosen for the experiment and the retest are shown in Table 6.3.

On average, subjects chose similar parameters between both experiments for the strength of NR (1.38 vs. 1.69) and the maximum attenuation of background sounds (0.13 vs. 0.12). One subject, NH12 selected an extreme parameter for NR strength for the retest experiment and can be viewed as an outlier. A paired-samples t-test indicated that there was no difference in the mean parameter choices between the experiment and retest for the strength of NR [t(9) = -1.275, *p* = 0.234; without NH12: t(9) = 0.941, *p* = 0.374] and for the maximum attenuation of background sounds [t(9) = 0.254, *p* = 0.805].

For the test-retest variability, the within-subject standard deviation σ_ω, the *repeatability* $(2.77 * \sigma_\omega)$ and the pearson correlation coefficient *r* were calculated without NH12 for the strength of NR ($\sigma_\omega = 0.67$, with a *repeatability* of 1.86 and *r* = 0.21) and for the maximum attenuation of background sounds ($\sigma_\omega = 0.07$, with a *repeatability* of 0.20 and *r* = 0.28). These estimates of σ_ω were comparable to the within-subject standard deviation $\sigma_{\omega 5}$ for the five iterative steps within the parameterization procedure of the retest experiment ($\sigma_{\omega 5} = 0.62$ and $\sigma_{\omega 5} = 0.1$ for the strength of NR and maximum attenuation, respectively) and the experiment proper ($\sigma_{\omega 5} = 0.70$ and $\sigma_{\omega 5} = 0.06$ for the strength of NR and maximum attenuation, respectively).

Table 6.3 Comparison of final noise reduction parameters between experiment and retest (indicated) for ten subjects (5 NH, 5 HI).

Subject	Alpha/Beta	Alpha/Beta retest	Gamma	Gamma retest	Gamma (dB)	Gamma retest (dB)
NH3	1.91	1.88	0.04	0.09	-26.96	-20.92
NH4	0.68	0.60	0.17	0.21	-15.19	-13.56
NH8	1.05	1.27	0.04	0.04	-28.01	-27.96
NH9	2.10	2.89	0.15	0.07	-16.46	-23.10
NH12*	0.73	9.82*	0.23	0.09	-12.81	-20.92
HI2	1.19	3.16	0.20	0.08	-14.19	-21.94
HI7	0.40	1.85	0.13	0.38	-17.46	-8.40
HI10	1.68	1.57	0.06	0.08	-23.79	-21.94
HI11	2.12	1.01	0.20	0.13	-13.96	-17.72
HI13	1.35	0.94	0.10	0.06	-19.86	-24.44
AVG	**1.32**	**2.50**	**0.13**	**0.12**	**-18.87**	**-20.09**
*w/o NH12	**(1.38)**	**(1.69)**	**(0.12)**	**(0.13)**	**(-19.54)**	**(-20.00)**
STD	0.62	2.70	0.07	0.10	5.56	5.61
*w/o NH12	(0.61)	(0.87)	(0.06)	(0.11)	(5.45)	(5.94)
$\boldsymbol{\sigma_\omega}$	2.13		0.08		4.20	
*w/o NH12	(0.67)		(0.07)		(3.99)	

Awareness of background sounds

Awareness results are shown as average percentage correct recognition scores for both NH and HI groups in Fig. 6.4. For the evaluation of awareness of background sounds, a repeated-measures ANOVA was performed with between-subject factors hearing group (NH/HI) and within-subject factor recognition task (acoustic scene and cue). There was a significant difference between NH and HI groups in the recognition of background sounds [F(1,28) = 24.790, *p* <= 0.001]. Pairwise comparisons indicated highly significant differences (*p* <= 0.001, etasquared = 0.369 and 0.393 for NH and HI, respectively) between groups for both acoustic environment and cue. There was also a significant difference between the recognition of the acoustic scene or cue [F(1,28) = 4.769, *p* = 0.038, etasquared = 0.146] with the acoustic cue being easier to recognize than the acoustic scene by about 11%.

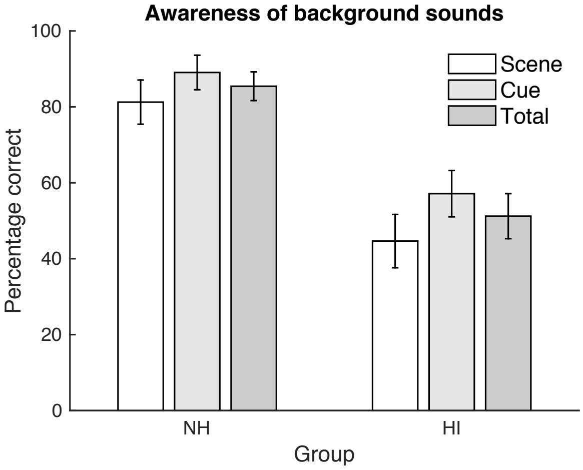


Figure 6.4 Group mean scores for environmental awareness to background sounds for NH and HI groups and the scene, cue and combined detection scores.

Speech intelligibility

Average percentage correct scores of the speech intelligibility test for both NH and HI groups are shown in Fig. 6.5.

For the statistical analysis of speech intelligibility results, the average performance in the unprocessed conditions was compared to the averaged processed conditions with NR processing and to the practice condition. A repeated-measures ANOVA was performed with between-subject factor of hearing group (NH/HI) and within-subject factor processing condition (practice, unprocessed and processed with chosen NR). There was a significant effect for the difference between the NH and HI groups in terms of speech intelligibility [F(1,28) = 4.486, *p* = 0.043]. Pairwise comparisons indicated significantly different SI performance between the NH and HI groups for the practice condition (*p* = 0.007, etasquared = 0.234) and the unprocessed condition (*p* = 0.002, etasquared = 0.292) but not for the processed condition (*p* = 0.567, etasquared = 0.012). There was also a significant main effect of processing condition [F(2,56) = 321.578, *p* <= 0.001]. Bonferroni-corrected pairwise comparisons were calculated and indicated significantly different speech intelligibility scores between all three processing conditions (*p* <= 0.001).

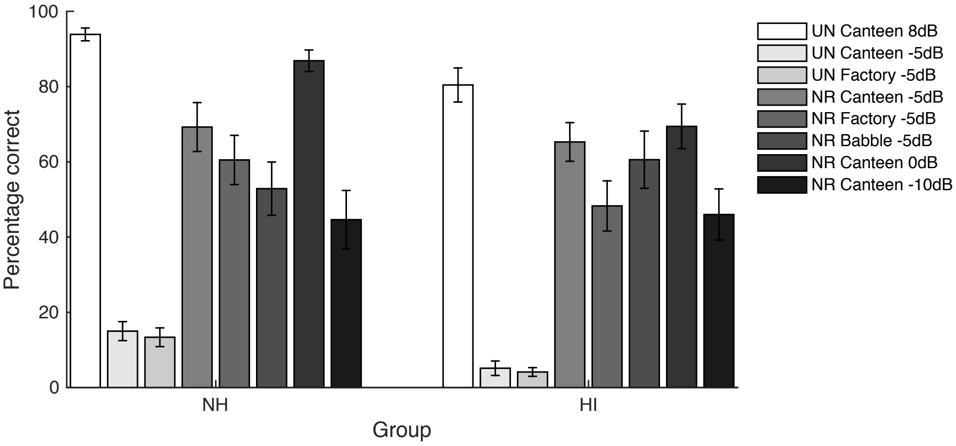


Figure 6.5 Group mean scores for the speech intelligibility test for all test conditions (UN - unprocessed, NR - processed with the individually-chosen noise reduction). Error bars represent the standard error of the mean.

Subjective preference

Mean subjective ratings in terms of speech distortion, noise intrusiveness and overall quality for both NH and HI groups and the three processing conditions (unprocessed, user-controlled NR processing IRM-CH and conventional Wiener filter with a maximum background attenuation of 20 dB) are shown in Figure 6.6.

A repeated-measures one-way ANOVA was performed with the between-subject factor hearing group (NH/HI) and the within-subject factor preference (all three measures were included: speech distortion, noise intrusiveness and overall quality). There was a significant main effect of preference score [F(8,224) = 77.462, *p* <= 0.001] and a significant interaction between hearing group and preference score [F(8,224) = 3.900, *p* <= 0.001]. Bonferroni-corrected pairwise comparisons indicated a significant difference between groups for the speech distortion ratings with IRM-CH processing (*p* = 0.011, etasquared = 0.211) and with the unprocessed condition (*p* = 0.009, etasquared = 0.217).

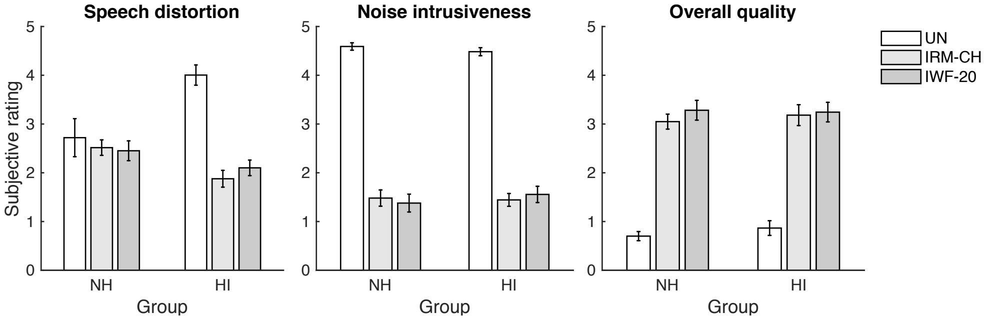


Figure 6.6 Group mean scores for the speech quality test for all three processing conditions (UN - unprocessed, IRM-CH - processed with the individually-chosen noise reduction, IWF-20 - processed with the ideal Wiener filter and 20 dB maximum noise attenuation). Error bars represent the standard error of the mean.

Correlations

Pearson correlations between NR parameter choices, PTA and outcome measures were performed to investigate potential relationships within the data. It can be expected that the choice in NR parameters affected the performance in terms of speech intelligibility and awareness of background sounds and that the parameters itself might be affected by the amount of hearing loss.

For this analysis, the parameter controlling the strength of noise reduction was transformed to represent a monotonic performance function which is described hereafter. A deviation from the standard Wiener filter parameter in terms of the strength of noise reduction (parameters *alpha* and *beta*) yields a more mild (values < 1) or more aggressive (values > 1) noise reduction. Both deviations may lead to a decrease in speech intelligibility performance in comparison to the IWF because of either more noise addition (values < 1) or more speech distortion (values > 1). The parameter for the strength of noise reduction was corrected as to represent both deviations on a monotonic scale by converting the parameter value into the dB-scale and taking the absolute values. When using this correction, there was a significant correlation between the parameter for the strength of noise reduction and the improvement in speech intelligibility for the NH group (the stronger the deviation from IWF, the lower the improvement in SI, with r = -0.83) but this relationship was weaker and not significant for the HI group (r = -0.32).

For the HI group, there was a significant correlation between the maximum attenuation of the background noise and the improvement in speech intelligibility performance (the lower the attenuation of the background noise the lower the improvement in SI, with r = -0.60). This relationship was much weaker for the NH group and not significant (r = -0.24). There was also a significant trend within the HI group, to choose a milder attenuation of the background noise with stronger hearing loss (r = 0.54).

6.1.4 Discussion

This study investigated user-controlled parameters for noise reduction processing chosen by normal hearing and hearing impaired subjects during a parameterization procedure and evaluated their choice in terms of performance in speech-in-noise understanding, awareness of background noise and subjective perception of speech quality. Subjects were able to control the strength and maximum background attenuation of a noise reduction algorithm that was based on *a priori* information about the speech and noise components. For the parameterization procedure, the goal was to control the NR parameters to obtain a balance between enhancing the intelligibility and quality of speech and maintaining an awareness of the background sounds. Potential differences in perceptual requirements between NH and HI groups and within each group are of interest for the optimization of speech enhancement algorithms for hearing devices.

For the comparison of noise reduction parameters between NH and HI groups, there was no significant difference in the mean values of user-controlled NR parameters for both strength of noise reduction and maximum attenuation of background sounds. This is an interesting finding, as it suggests that hearing ability does not seem to affect the user-controlled NR parameters on average and it might suggest the existence of a set of optimum values that could be used for both NH and HI subjects. The average user-controlled value for the strength of noise reduction of both groups was significantly larger than the conventional Wiener filter and resulted in an average gain function for all subjects that is shifted by about 5 dB (see Fig. 6.7 for a comparison between the average user-controlled gain function in comparison to the conventional Wiener filter with 20 dB maximum attenuation). This finding is consistent with Daniel et al. (2013) who reported a shift in gain function by about 5 dB for normal hearing listeners and with Mauger et al. (2012a) who observed a shift in the gain function towards stronger noise suppression for CI users. Together, these results suggest that there is a consistent deviation of user-controlled NR parameters from the conventional Wiener filter that is often used for speech enhancement algorithms in hearing devices. The retest of NR parameter values used a subset of 10 subjects and showed that the parameterization procedure used in this experiment produced repeatable NR parameter choices. In the retest experiment performed 4 months later, the 10 subjects chose similar NR parameters on average as in the first experiment.

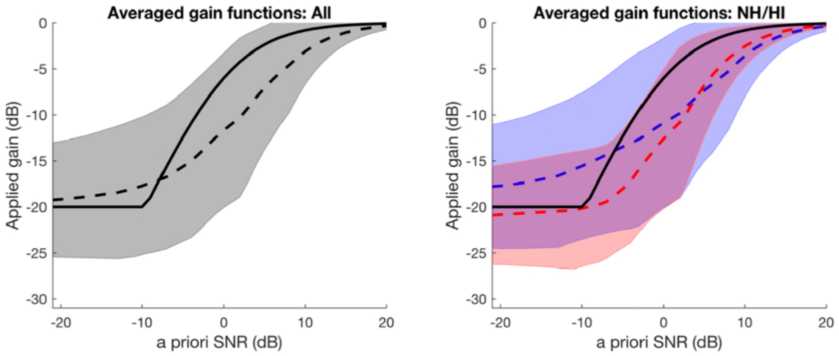


Figure 6.7 Final gain functions (averages between all, left, and for NH and HI groups, right as in Figure 7.3) in comparison to IWF gain function (black).

There was a significant difference between NH and HI groups in the variability of user-controlled parameters for the strength of NR processing. NH subjects varied more in their choice of the strength of NR processing than HI subjects (with standard deviations of 1.37 and 0.78, respectively). However, NH and HI groups chose very similar values for the strength of NR processing on average (mean values of 1.75 and 1.65, respectively). This may indicate that NH subjects are more robust to changes in NR parameters and tolerate a wider range of possible parameter values than HI subjects. For HI subjects, non-optimal parameter choices in terms of NR strength might affect their perception more strongly and lead to less variability in their final choices. This can be explained with the occurence of sensorineural hearing loss in the HI group that leads to an impaired ability in terms of speech understanding in noise from the presence of artefacts such as noise addition and speech distortion. The parameter choice of strength of NR processing has the effect to balance out artefacts in terms of noise addition and speech distortion and HI subjects might have a more narrow window of parameter choices that allows them to obtain benefits in terms of speech understanding.

The maximum attenuation of background noise was user-controlled to about 20 dB by both NH and HI groups. There was a small difference in mean values between groups by about 2 dB but this difference was not significant. This suggests that a value of about 20 dB of background noise attenuation represents an optimum choice for the balance between improving speech intelligibility in background noise and maintaining an awareness of background sounds. For the attenuation of background noise, there was no difference in the variability between NH and HI groups (with standard deviations of 6.47 and 5.05, respectively).

The results of the test of awareness to background sounds showed that NH subjects performed better than the HI group in recognizing both the acoustic scene and cue sounds. NH subjects performed at a high level with a percentage of correct recognition of about 85% compared to HI subjects with 52% recognition performance. This indicates that NH subjects were successful in choosing a parameter set for maintaining a good awareness of background sounds. HI subjects struggled more to achieve this goal, and might have been less successful because of their hearing loss. Background sounds were attenuated by about 20 dB for the HI group and this might have resulted in some parts of the background sounds falling below their audibility thresholds and led to decreased performance in the recognition task.

For the test of speech intelligibility, both NH and HI groups were successful in obtaining large improvements with self-adjusted NR parameters over the unprocessed conditions. There were differences between NH and HI groups in terms of their performance in the practice and unprocessed conditions (with lower performance for the HI group), but there was no difference in performance between groups for the processed condition. This suggests that the HI group was as good as the NH group in obtaining improved speech understanding with their choice of NR parameters. The HI group benefitted more from their choice of NR parameters than the NH group, as their improvements over the unprocessed condition was larger by about 11% and their performance with NR processing was closer to their performance in the practice condition. This indicates that the HI group was slightly more successful in making use of the potential in improved speech understanding in noise by user-controlled NR parameters. This is a promising finding as it supports the possibility of integrating a user-controlled process into the fitting process or control software of hearing aids - to allow the users to control the NR parameters themselves.

For the test of subjective preference, there was no significant difference between the rating of the user-controlled NR processing and the conventional Wiener filter. This suggests that in terms of perceived quality, subjects were able to achieve good performance but did not improve the quality in comparison to the conventional Wiener filter. This might have resulted from the goal of the parameterization procedure to improve in terms of speech intelligibility and awareness of background sounds and not to improve the perceptual quality of the processed speech. Nevertheless, results show that the user-controlled NR parameters still achieved high performance similar to IRM processing. There was a difference in the ratings of speech distortions between NH and HI groups, with the HI group rating the speech distortions of the unprocessed condition much worse in comparison with the NH group. This is an interesting finding and might indicate that for HI subjects the background noise might affect the perceived speech quality more than for NH subjects. It can be speculated that a decrease in spectral resolution, as it occurs with sensorineural hearing loss, may lead to more speech distortions in background noise because of overlapping auditory filters. The spectral information of speech might be more affected by the noise energy in adjacent auditory filters for HI subjects than for NH subjects and lead to more speech distortions perceived by HI subjects.

The data collected within this study indicates that NH and HI subjects choose similar parameters for noise reduction processing in terms of strength and maximum attenuation of background noise for NR processing. The values obtained in this study are limited to noise reduction algorithms that use *a priori* information about the signal components. In practice, when these components are not available and have to be estimated from the noise-corrupted speech signal, the choice of NR parameters might be affected by the type and amount of estimation errors introduced by the specific algorithm. For example, if a SE algorithm introduces more speech distortions and is not able to identify the background noise accurately, different values for the parametric gain function are likely to be necessary to achieve optimum performance (eventually less aggressive parameters in this case). Another constraint of this study is that the HI group had relatively mild hearing losses and it might be the case that subjects with more severe hearing loss would have chosen different parameters for NR processing.

7 General Discussion

This chapter aims to discuss the findings of the individual chapters in a more general format by connecting the findings across chapters and by including a significant amount of speculative arguments that may lead to interesting follow-up questions to be investigated in future studies.

Discussion of studies on processing delay and their relevance for speech enhancement

The findings reported in chapter 3 indicate greater tolerance of processing delay due to hearing loss and long-term acclimatisation effects by the listeners. Given that current hearing devices are restricted to extremely short processing delays (often smaller than 10 ms), the integration of speech enhancement algorithms that make use of more complex processing techniques is compromised by the restricted computational power and temporal window that can be used for the processing. Such complex algorithms, for example the approach followed in chapter 4 and 5 of this thesis, indicate benefits for users of hearing devices since they show good promise to overcome the current limitations of speech enhancement performance in challenging noise conditions with fluctuating background sounds. The algorithms developed in this thesis have been optimized for application to real-time processing by using short temporal windows for the processing that would fit into the tolerable range of delays for hearing impaired listeners (<20-30 ms) and by scaling down the computational complexity of the neural network algorithm to yield a system that is computationally less heavy than previous approaches. For future hearing devices, the processing power is expected to increase with time but there may even be a quicker route to make more complex algorithms available for users of hearing devices. The wireless connectivity to external devices such as smartphones may serve as an alternative route to overcome restrictions in terms of computational complexity. Smartphone devices' development in terms of computational power is proceeding at a much faster pace than for hearing devices and already approached or even surpassed the computational power of modern personal computers. This could make it possible to shift parts of the processing over to the external device - for example the detection of and adaptation to novel acoustic environments, or even the complete estimation of speech signals in background noise with the approach followed in chapter 4 and 5 of this thesis. If the wireless data transmission would be fast enough to transmit information within the tolerable range of delay, then the whole speech enhancement processing could eventually be shifted to the smartphone device. This could be a promising way to improve the performance of hearing devices in challenging acoustic situations and would be an interesting topic for future research studies.

In chapter 3, the tolerance of processing delay was only measured using normal hearing and hearing impaired listeners. There is a lack of literature on the perceptual range of delays tolerable for users of cochlear implants. For external stimuli, there should be even more tolerance of processing delay because the detrimental effects due to comb-filtering would not occur for cochlear implant users that have no residual acoustic hearing (there are no acoustic paths or interaction between them). For the case of own-voice perception, it can be assumed that similar perceptual requirements in terms of processing delay would be measured between users of cochlear implants and normal hearing persons due to the perception of proprioceptive cues (for a delay of 43 ms, small effects of processing delay have been measured for speech production rates of normal hearing listeners in literature). Still, a delay of about 40 ms is much larger than currently used for hearing devices and may allow for the wireless streaming of information between external devices. However, these assumptions are completely hypothetical and would be an interesting topic for further research.

Another topic of potential research on processing delay would be the long-term acclimatisation of hearing aid users to specific delay lengths. It would be interesting to measure the individual processing delay of a specific hearing aid that users of hearing aids are exposed to in their daily life and to perform a blinded listening study to explore if and to what degree a long-term acclimatisation effect to processing delay is existent. If the acclimatisation effect to processing delay would be similar for delays larger than 10 ms and delays smaller than 10 ms, then it could be assumed that the absolute length of delay is not a crucial factor as long as it stays within the tolerable range of delays for that user group (e.g. <20-30 ms).

The effect of processing delay was only measured in terms of subjective disturbance or annoyance of speech stimuli using mean opinion scores during the tests with listeners in chapter 3. The effects of processing delay on other aspects of speech perception, especially on speech intelligibility in background noise have not been investigated in literature and could be an interesting field of study. For short delays and quiet acoustic conditions without reverberation, no significant effect on speech intelligibility would be expected but this situation might change when background noise and reverberation are added and longer delays are used. Such potential detrimental effects of processing delay would counteract the benefits of speech enhancement processing and should be avoided. Before longer delays would be allowed for hearing devices to accommodate more complex speech enhancement algorithms this potential conflict should be investigated to avoid that longer delays would diminish the benefits from speech enhancement for the user.

Discussion of studies on neural network based speech enhancement

The neural network based speech enhancement system that was employed in chapters 4 and 5 represents a promising example for machine learning based algorithms that may provide benefits for users of hearing devices in terms of improved speech perception in noise. Both hearing impaired listeners, users of hearing aids, and listeners using a cochlear implant obtained significant improvements in several types of background noise, including fluctuating types. This is motivating and shows that indeed improvements in intelligibility may be obtained when more powerful estimation techniques, such as artificial neural networks, are adopted for speech enhancement algorithms.

During the listening study performed in chapter 4, an auditory-inspired set of acoustic features was compared to a more conventional feature set and indicated superior performance in the objective prediction scores and with higher group scores in the listening experiment with hearing impaired listeners. However, this effect was not significant and it can only be speculated that auditory-inspired features may provide improvements in performance for speech enhancement purposes. For the listening studies reported in this thesis, the low-level auditory features in form of the Gammatone filter bank energies provided the best performance and were the core feature that enabled the neural network regression to perform well. A possible explanation for the good performance of the Gammatone features are the fine resolution at lower frequencies that provide important information about the fundamental frequency and first formants of speech signals (<1000 Hz). In order to reach a similar spectral resolution with the conventional approach, the FFT, a higher order of FFT has to be computed (minimum of 512 frequency bins) than is usually done in speech processing applications (about 64 to 256 frequency bins usually). This would introduce longer processing delays and more computational load and a large amount of redundant information in the higher frequency channels that may be less important for speech detection purposes than the lower frequency regions.

It may be of interest to include more sophisticated auditory-modelling techniques that simulate higher stages of the auditory system, such as the cochlear nucleus or inferior colliculus parts in the auditory brainstem. These stages are believed to extract higher level features of speech that may provide better robustness to background noise. The inclusion of these more robust speech features may improve the performance of the neural network based speech enhancement in background noise and may be an interesting venue for further improvements of the approach in future.

Another important aspect concerns the amount and type of temporal information that is processed by the neural network. This thesis and previous literature made use of concatenated frames of the input data that are presented to the neural network simultaneously (in this study only 2 frames were used: the current and past frame). The NN has to learn the temporal information based on this simplified presentation of temporal information and it can be speculated that a neural network with a more sophisticated internal processing of temporal information may perform better than the simple feed-forward architecture used in this thesis. Recently, recurrent neural network architectures, such as the long-short term memory (LSTM) neural network, have been shown to provide stronger performance in speech detection and recognition tasks than the feed-forward architecture. Future studies on neural network based speech enhancement should evaluate these more sophisticated types of NNs that are recently becoming more available and easier to train successfully due to improved implementations and freely available software packages.

The third main aspect of the neural network based algorithm is the target gain function that is used for the training of the algorithm and the application of the noise reduction filter. Based on traditional speech enhancement studies, SNR-based target masks such as the ideal ratio mask or ideal binary mask were employed in this thesis and the ideal ratio mask using a stronger attenuation of the noise was found to provide benefits for users of cochlear implants. This concept is based on the principal proposed by the articulation index that suggests that retaining the frequency channels with high SNR while discarding or attenuating the frequency channels with low SNR yields an improved global, long-term SNR that provides benefits in terms of speech intelligibility. This processing paradigm gives large improvements in speech intelligibility in both stationary and fluctuating and spectrally broadband types of noise. For more spectrally narrow and stationary types of masking noise this paradigm might not present the best option, since speech components will be attenuated solely based on the SNR in time-frequency regions and without incorporating knowledge about the natural pattern of the speech signal. With those kinds of narrowband noises or even more in general, a speech model based mask function may provide better results by preserving the temporal envelope patterns of the underlying speech signal instead of constantly attenuating the frequency channels with low SNR even though that may destroy the natural pattern of speech and lead to distortions harmful to speech intelligibility.

In practice, the neural network based speech enhancement needs to work in unseen acoustic environments that were not included in the training data to obtain benefits for the user in the real world. This challenge could be tackled in different ways. When the acoustic environment changes, an automatic acoustic environment detection algorithm could be used to trigger an adaptation procedure of the neural network to the environment of interest. Given the low computational complexity of the networks employed in this thesis, this approach would be feasible by using a smartphone processor to compute the updated weights for the network. The adapted network parameters could then be transmitted from the smartphone to the hearing device using a wireless connection such as Bluetooth. Another way would be to include a much larger variety of acoustic material in the training data to obtain a system that generalizes to a wide range of background noise types (as done by Chen et al., 2016 and in the pilot experiment in chapter 5). This approach would require a more powerful neural network algorithm that is able to learn a large amount of characteristic patterns in the training data and is less likely to perform similarly well as an environment-optimized algorithm in every possible situation or acoustic environment. Therefore this approach seems less feasible for applications in hearing devices, on the one hand because the computational complexity is too big for current hearing devices and for the foreseeable future, on the other hand because the variability of acoustic environments that exist in the real world is likely to be too large to be fully included in the training data. A related challenge for single-microphone noise reduction techniques such as the one used in this thesis is to detect or decide the speaker of interest in a multi-talker situation (cocktail party effect). In such a challenging acoustic environment, the background noise is very similar to the target signal (both are speech signals) and it remains a challenging task for the neural network to decide which speaker is the one that should be retained and which ones to attenuate. In chapter 5 of this thesis, the multi-talker NNSE algorithm had to perform in this condition and showed better performance at higher SNRs where the target speaker is more dominant in respect to the other talkers. At lower SNRs approaching 0 dB, where the target speaker is more similar in level to the other competing speakers, the algorithm was not able to successfully detect the speaker of interest and this confusion of speakers resulted in a lack of significant improvement in intelligibility for this condition (NNSE-MT for BABBLE). When a priori knowledge on the target speaker is available, as with the speaker-dependent algorithm that was trained beforehand on the speaker of interest, significant improvements in intelligibility were found for this noise type. This shows that for the type of background noise and the speaker of interest, *a priori* information is certainly helpful but may not be available in the real world. This requires clever adaptation techniques to be developed that enable the neural network to adapt to certain speakers or environments of interest during the final application in the real world.

Another aspect of successful real-world application of the proposed speech enhancement algorithm is the implementation in real time. The algorithm was implemented for demonstration purposes using the MATLAB/Simulink environment and the Speedgoat real-time machine that is able to perform ultra-low latency processing of audio signals. A simplified feature set was used to further reduce the computational complexity of the algorithm, but this was done due to time constraints for the development and it was not evaluated whether this was necessary to enable real-time operation on the Speedgoat system. The algorithm performed with short processing delay (about 15 ms) and provided a strong attenuation of background sounds even in acoustic environments that were not included in the training data. This indicates that the algorithm can indeed generalize to novel situations, but decreases in performance should be expected for mismatched conditions (eventually down to the level of intelligibility obtained without speech enhancement processing). Another project in our lab dealt with the implementation of the speech enhancement algorithm using the Raspberry Pi low-cost computing platform. This device is not optimized for ultra-low latency processing but processing delays of about 30 ms were still achieved. The neural network based speech enhancement employed a similarly sized network as the one used in the listening studies of this thesis (chapters 4 and 5) and was able to perform in real time on the Raspberry Pi. This indicates that indeed the optimization of the neural network architecture that was intended to provide a more realistically applicable algorithm for hearing devices was successful. However, practical challenges that involved the input normalization of acoustic data, a problem that only occurs in realistic applications where the input sound can change strongly in level, complicated the development and may have led to decreased performance of the algorithm. The non-linear processing of the neural network by using non-linear transfer functions within the hidden units may lead to saturation effects if the input data is not scaled appropriately. This requires an automatic gain control that ensures that a constant level of input sound is provided. A simple standardization procedure was applied in the pilot experiments for this experimental application but did not yield satisfactory results (the mean and standard deviation for the standardization should be estimated continuously which represents a challenging task for future studies). Nevertheless, the efforts made to implement the NNSE algorithm in real time were successful to constitute a proof-of-concept that could be optimized in future studies.

Discussion of studies on user-controlled speech enhancement

The study on user-controlled parameters in chapter 6 of this thesis indicates that potential users of hearing aids are able to tune the noise reduction parameters to obtain benefits in terms of speech intelligibility in background noise. Both listener groups, with normal hearing and hearing loss, were successful to find suitable parameters for the algorithm using the parameterization procedure. This setup could be used to obtain more individually acceptable noise reduction parameters for the NNSE approach in chapter 4 and 5 and may lead to higher acceptance by users of hearing aids or cochlear implants. Furthermore, giving more control to the user would allow for situation-specific fine-tuning by the user to access the best performance in each situation. This could be beneficial since in different acoustic situations the optimal choice for noise reduction parameters is expected to change depending on the level of background noise or the type of interfering sounds and would allow for both improvements in intelligibility and maintained awareness to background sounds. Noise reduction parameters of hearing devices are usually determined by the manufacturer and kept the same for each user leading to a lack of individualization that may be one of the reasons why listeners do not make use of these algorithms because they find them to be not effective enough or not matching their personal requirements. In the listening study in chapter 6 of this thesis, similar parameter choices have been found for NH and HI listeners on average, however with a significantly smaller variability for the HI group. This result indicates that HI listeners were less robust to changes in parameters and may require a more specific tuning of a speech enhancement algorithm to match their individual hearing capabilities allowing to access the benefits of the processing.

Overall, the work described in this thesis has obtained several interesting findings that may motivate for further investigations and could potentially represent a step towards application of user-optimized neural network based speech enhancement techniques in hearing devices such as hearing aids and cochlear implants that operate with short processing delays. This would be of interest to the users if the indicated benefits in speech perception in noise hold also for more realistic acoustic scenarios. To achieve this goal, further work is needed that takes up the challenges of generalization to unseen acoustic conditions and real-time applications in hearing devices, based on the findings presented in this thesis.

8 CONCLUSION

This thesis is concerned with the investigation of several aspects related to improving the perception of speech in background noise by listeners with hearing loss.

Processing delay is one of the main factors that limits the development of novel and more complex speech enhancement algorithms for hearing devices such as hearing aids. The first study performed in chapter 3 of this thesis found a significantly greater tolerance of processing delay for listeners with hearing loss than for normal-hearing listeners. Accordingly, there was a trend of increased tolerable delay with higher degree of hearing loss within the hearing-loss group, but this effect did not reach statistical significance due to small sample sizes. Quantitatively, delays of up to 20 ms are not expected to exceed tolerance limits for the average normal-hearing listener and the average listener with mild hearing loss. Higher degrees of hearing loss may allow an increase in delay up to 30 ms without causing excessive annoyance. These findings are in line with previous recommendations for listeners with hearing loss and extend the testing conditions to the scenario of external-voice stimuli with linearly processed signals. It should be noted, that the testing setup used in this study differed from commercial hearing aids in several aspects and therefor limits the extrapolation of the findings. The second study that was performed within this line of work evaluated the potential of long-term acclimatisation to processing delay by normal-hearing listeners. Participants in the group listening to a processing delay of 20 ms during the acclimatisation period showed a significant increase in tolerance of processing delay in the post-test relative to the group listening to a processing delay of 40 ms during the acclimatisation period. This indicates that long-term acclimatisation to delay increases its tolerance when a suitable delay is chosen that lies within the tolerable range of delays (<20-30 ms). This finding is promising since it occurred after a relatively short period of acclimatization with only a few hours of total listening time to stimuli with processing delay and it can be speculated that further acclimatisation may occur for longer acclimatisation periods.

The results reported in chapter 3 indicate that delays above 10 ms, the limit imposed for current hearing aids, may well be tolerated when a hearing loss is present and when potential acclimatisation effects are taken into account. Still, long-term acclimatisation to processing delay for hearing-impaired listeners remains an interesting question for future studies.

In chapter 4 of this thesis, three *speech enhancement algorithms based on machine learning* were evaluated against a traditional Wiener filtering approach using noise estimation and the unprocessed condition in terms of speech recognition and speech quality ratings in hearing-impaired listeners. The three machine learning based algorithms were noise-specific techniques that were evaluated on unseen segments of the same noise type. Significant increases in speech-recognition scores and quality ratings were seen for all three machine-learning approaches in at least one of the four noise conditions. In contrast, the Wiener filtering algorithm produced no significant improvement in either speech recognition or quality rating in any noise condition. The three machine learning approaches included two neural-network based approaches, comparing a standard feature set to one using a novel set of features derived from the auditory image model, a computational model of the human auditory system. The results obtained with the auditory-based feature set indicated better performance than the standard feature set in terms of both objective measures and speech recognition and quality ratings by hearing-impaired listeners in all conditions (except speech quality ratings in babble noise at 4 dB SNR), although none of these differences reached statistical significance for the listening experiment results. Neural networks seemed preferable over the sparse coding approach, both because of their better performance (even for small networks like those used here) and because they were more computationally efficient in the testing stage.

In the first part of chapter 5 of this thesis, significant improvements in speech intelligibility were found with a neural-network based speech enhancement algorithm for normal hearing subjects listening to noise-vocoded stimuli to simulate speech perception with a cochlear implant. Consistent improvements were found in both stationary and non-stationary noise types and two SNRs. In most conditions, a conventional speech enhancement algorithm based on Wiener filtering was not able to show benefits for speech perception in background noise (apart from the lower SNR condition in the stationary noise). The NN-based algorithm used a simplified feature set in comparison to the one used in chapter 4 to yield low computational complexity for potential application in real time devices such as cochlear implants. Several target functions were compared and the ideal ratio mask with stronger noise reduction relative to the one used by previous studies was found to give the best results for vocoded stimuli. This finding is in line with the literature reporting preference for stronger noise reduction by cochlear implant users. This study indicated the potential *benefit for CI users' speech perception in noise* and motivated the evaluation of the neural network based algorithm with CI users in the second part of chapter 5. The algorithm architecture was further optimized to yield a faster and less complex algorithm than in the previous experiment that was able to run in real time. Speech intelligibility in noise by CI users was significantly increased and ranged from 1.4 to 6.4 dB in SRT with the neural network based speech enhancement algorithm for several types of background noise and SNRs. An adaptive procedure to target the SRT was used due to the large variability among CI users' performance in speech understanding requiring an algorithm that worked successfully over a range of SNRs. Furthermore, generalization to an unseen speaker was evaluated by comparing two NN-based algorithms: one that used speaker-dependent training and testing and one that used speaker-independent training and testing. Even though improvements in SRT scores were about 1 to 1.5 dB lower than for the speaker-dependent algorithm, substantial and statistically significant improvements were found for 2 out of 3 noise conditions for the speaker-independent NN-based algorithm. The benefits in CI users' speech in noise understanding are promising and provide motivation for further investigations of this approach.

This thesis employed real-time feasible architectures for the NN-based algorithms in chapters 4 and 5 to ensure that the perceptual requirements of target users are being met in terms of processing delay. The results regarding tolerance to processing delay reported in chapter 3 of this thesis have been an important source of information for defining the temporal window that was used for the feature extraction part within the NN-based algorithms employed in chapter 4 and 5. In order to obtain a real-time feasible algorithm with low computational complexity that may be able to run on future hearing devices, the neural network based approach was further optimized by reducing the computational complexity in comparison to previous studies.

It should be noted that noise-specific neural networks were used for all studies in this thesis. This constitutes a limitation for the potential application of the algorithms in hearing devices such as hearing aids and cochlear implants that are designed to work in any acoustic environment. This challenge needs to be addressed in future studies by either including an intelligent adaptation procedure or by using large-scale training data that incorporates more variability and thus may be able to improve generalization performance to unseen noise types. Future development in the rapidly growing field of machine learning can be expected to improve the estimation accuracy and generalization performance to unseen acoustic conditions by incorporating methods from unsupervised training techniques or reinforcement learning.

In chapter 6, a listening study was performed to investigate the differences in user-controlled parameters for speech enhancement among and between normal hearing and hearing impaired subjects. Firstly, subjects completed a parameterization procedure to choose their preferred strength of noise reduction and maximum attenuation of background sounds with the goal to find a balance between improved speech understanding and maintained awareness of background sounds. Evaluation was performed in terms of detection and classification of background sounds, intelligibility of speech in background noise and subjective rating of speech quality. Most subjects were successful in choosing parameters leading to large increases in speech intelligibility in noise, with similar performances between NH and HI groups. Normal hearing subjects had better awareness of background sounds which may have resulted from better hearing abilities and higher robustness to the attenuation of background sounds relative to hearing impaired subjects. This conclusion was supported by the finding of a significantly larger variability in noise reduction parameter choices for the NH group than for the HI group indicating that HI subjects had a narrower window of optimal noise reduction parameters and were less successful in finding a balance in terms of intelligibility improvement and awareness of background sounds. Interestingly, group mean scores for the noise reduction parameters were similar between NH and HI groups indicating that a noise reduction strength of about 5 dB stronger than the conventional ideal ratio mask (ideal Wiener filter) in combination with a maximum background attenuation of 20 dB constitutes a sweet spot that may be optimal for both NH and HI listeners.

To conclude, the work described in this thesis represents a step towards the application of speech enhancement based on neural networks in hearing aids and cochlear implants to obtain benefits in speech perception in noise for the users of those devices and to overcome the limitations of current technology.

9 REFERENCES

Agnew, J., Thornton, J.M., 2000. Just noticeable and objectionable group delays in digital hearing aids. J. Am. Acad. Audiol. 11, 330–6.

Aharon, M., Elad, M., Bruckstein, A.M., 2006. The {K-SVD}: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representations. IEEE Trans. Signal Process. 54, 4311–4322.

Anzalone, M.C., Calandruccio, L., Doherty, K.A., Carney, L.H., 2006. Determination of the potential benefit of time-frequency gain manipulation. Ear Hear. 27, 480–492.

Arehart, K.H., Hansen, J.H.L., Gallant, S., Kalstein, L., 2003. Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners. Speech Commun. 40, 575–592.

Bench, J., Kowal, Å., Bamford, J., 2009. The Bkb (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. Br. J. Audiol. 13, 108–112.

Bentler, R., Wu, Y.H., Kettel, J., Hurtig, R., 2008. Digital noise reduction: outcomes from laboratory and field studies. Int. J. Audiol. 47, 447–460.

Boll, S.F., 1979. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Trans. Acoust. 27, 113–120.

Bolner, F., Goehring, T., Monaghan, J., van Dijk, B., Wouters, J., Bleeck, S., 2016. Speech enhancement based on neural networks applied to cochlear implant coding strategies, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6520–6524.

Borges, R. C., Costa, M. H., Cordioli, J. A., and Assuiti, L. F., 2013. An adaptive occlusion canceller for hearing aids. In Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European (pp. 1-5). IEEE.

Boymans, M., Dreschler, W., 2000. Field Trials Using a Digital Hearing Aid with Active Noise Reduction and Dual-Microphone Directionality: Estudios de campo utilizando un audifono digital con. Int. J. … 6091.

Bland, J. M., Altman, D. G., 1996. Statistics notes: measurement error. Bmj, 313(7059), 744.

Bramslow, L., 2010. Preferred signal path delay and high-pass cut-off in open fittings. Int. J. Audiol. 49, 634–44.

Brons, I., Dreschler, W.A., Houben, R., 2014a. Detection threshold for sound distortion resulting from noise reduction in normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 136, 1375.

Brons, I., Houben, R., Dreschler, W. A., 2014b. Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners. Trends Hear. 18.

Brons, I., Houben, R., Dreschler, W. A., 2012. Perceptual effects of noise reduction by time-frequency masking of noisy speech. J. Acoust. Soc. Am. 132, 2690–9.

Brown, G., Cooke, M., 1994. Computational auditory scene analysis. Comput. Speech Lang.

Brungart, D.S., Chang, P.S., Simpson, B.D., Wang, D., 2006. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. 120, 4007.

Buechner, A., Brendel, M., Saalfeld, H., Litvak, L., Frohne-Buechner, C., Lenarz, T., 2010. Results of a pilot study with a signal enhancement algorithm for HiRes 120 cochlear implant users. Otol. Neurotol. 31, 1386–1390.

Chen, F., Loizou, P.C., 2012. Impact of SNR and gain-function over- and under-estimation on speech intelligibility. Speech Commun. 54, 272–281.

Chen, F., Loizou, P.C., 2011. Predicting the intelligibility of vocoded speech. Ear Hear. 32, 331–338.

Chen, J., Baer, T., Moore, B.C.J., 2013. Effect of spectral change enhancement for the hearing impaired using parameter values selected with a genetic algorithm. J. Acoust. Soc. … 133, 2910–20.

Chen, J., Wang, Y., Wang, D., 2014. A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Trans. Speech Lang. Process. 22, 1993–2002.

Chen, J., Wang, Y., Yoho, S.E., Wang, D., Healy, E.W., 2016. Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. J. Acoust. Soc. Am. 139, 2604–2612.

Cohen, I., Berdugo, B., 2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. 9, 12–15.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120, 2421–2424.

Cullington, H.E., Zeng, F.G., 2008. Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. J. Acoust. Soc. Am. 123, 450–61.

Dahlquist, M., Lutman, M.E., Wood, S., Leijon, A., 2005. Methodology for quantifying perceptual effects from noise suppression systems. Int. J. Audiol. 44, 721–732.

Daniel, A., Lepauloux, L., Yemdji, C., Evans, N., Beaugeant, C., 2013. An experimental framework for the derivation of perceptually-optimal noise suppression functions, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 7800–7804.

Dau, T., Kollmeier, B., Kohlrausch, A., 1997. Modeling auditory processing of amplitude modulation .1. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102, 2892–2905.

Dawes, P., Munro, K., Kalluri, S., Edwards, B., 2014. Acclimatization to hearing aids. Ear Hear. 35, 203–212.

Dawson, P.W., Mauger, S.J., Hersbach, A.A., 2011. Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients. Ear Hear. 32, 382–390.

DeWeese, M.R., Wehr, M.S., Zador, A.M., 2003. Binary spiking in auditory cortex. J. Neurosci. 23, 7940–9.

Dillon, H., 2001. Hearing aids. Boomerang press, Sydney.

Dillon, H., Keidser, G., O'Brien, A., Silberstein, H., 2003. Sound quality comparisons of advanced hearing aids. Hear. J.

Dreschler, W.A., Verschuure, H., Ludvigsen, C., Westermann, S., 2001. ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment. Audiology 40, 148–157.

Edwards, B., 2007. The future of hearing aid technology. Trends Amplif. 11, 31–45.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., Loubes, J.M., Massart, P., Madigan, D., Ridgeway, G., Rosset, S., Zhu, J.I., Stine, R.A., Turlach, B.A., Weisberg, S., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407–499.

Elad, M., Aharon, M., 2006. Image denoising via sparse and redundant representation over learned dictionaries. IEEE Transations Image Process. 15, 3736–3745.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoust. Speech Signal ….

Festen, J.M., Plomp, R., 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J. Acoust. Soc. Am. 88, 1725–1736.

Fetterman, B.L., Domico, E.H., 2002. Speech recognition in background noise of cochlear implant patients. Otolaryngol. Head. Neck Surg. 126, 257–63.

Fredelake, S., Holube, I., Schlueter, A., Hansen, M., 2012. Measurement and prediction of the acceptable noise level for single-microphone noise reduction algorithms. Int. J. Audiol. 51, 299–308.

French, N.R., Steinberg, J.C., 1947. Factors Governing the Intelligibility of Speech Sounds. J. Acoust. Soc. Am. 19(1), 90–119.

Friesen, L.M., Shannon, R. V, Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am. 110, 1150–1163.

Fu, Q.J., Shannon, R. V, Wang, X., 2013. Effects of noise and spectral resolution on vowel and consonant recognition : Acoustic and electric hearing. J. Acoust. Soc. Am. 104, 3586–3596.

Furman, A.C., Kujawa, S.G., Liberman, M.C., 2013. Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. J. Neurophysiol. 110, 577–586.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio, Speech Lang. Process. 20, 1383–1393.

Gibak Kim, Loizou, P.C., 2010. Improving Speech Intelligibility in Noise Using Environment-Optimized Algorithms. IEEE Trans. Audio. Speech. Lang. Processing 18, 2080–2090.

Glasberg, B.R., Moore, B.C.., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47, 103–138.

Goldsworthy, R.L., Greenberg, J.E., 2004. Analysis of speech-based Speech Transmission Index methods with implications for nonlinear operations. J. Acoust. Soc. Am. 116, 3679–3689.

Groth, J., Sondergaard, M., 2004. Disturbance caused by varying propagation delay in non-occluding hearing aid fittings. Int. J. Audiol.

Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., Rass, U., 2005. Signal processing in high-end hearing aids: state of the art, challenges, and future trends. EURASIP J. Appl. Signal Processing 2005, 2915–2929.

Harlander, N., Rosenkranz, T., Hohmann, V., 2012. Evaluation of model-based versus non-parametric monaural noise-reduction approaches for hearing aids. Int. J. Audiol. 51, 627–39.

Hazrati, O., Sadjadi, S.O., Hansen, J.H.L., 2014. Robust and efficient environment detection for adaptive speech enhancement in cochlear implants. 2014 IEEE Int. Conf. Acoust. Speech Signal Process. 900–904.

Healy, E.W., Yoho, S.E., Chen, J., Wang, Y., Wang, D., 2015. An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. J. Acoust. Soc. Am. 138, 1660–1669.

Healy, E.W., Yoho, S.E., Wang, Y., Apoux, F., Wang, D., 2014. Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners. J. Acoust. Soc. Am. 136, 3325.

Healy, E.W., Yoho, S.E., Wang, Y., Wang, D., 2013. An algorithm to improve speech recognition in noise for hearing-impaired listeners. J. Acoust. Soc. Am. 134, 3029–38.

Healy, E.W., Youngdahl, C.L., Apoux, F., 2014b. Evidence for independent time-unit processing of speech using noise promoting or suppressing masking release (L). J. Acoust. Soc. Am. 135, 581–4.

Henshaw, H., Sharkey, L., Crowe, D., Ferguson, M., 2015. Research priorities for mild-to-moderate hearing loss in adults. Lancet 386, 2140–2141.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2, 587–589.

Hohmann, V., 2002. Frequency analysis and synthesis using a Gammatone filterbank. Acta Acust. united with Acust. 88, 433–442.

Holube, I., Kollmeier, B., 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 100, 1703–1716.

Hopkins, K., Moore, B.C.J., Stone, M., 2008. Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. J. Acoust. Soc. Am. 123, 1140–1153.

Hu, Y., Loizou, P.C., 2010. Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. J. Acoust. Soc. Am. 127, 3689–95.

Hu, Y., Loizou, P.C., 2008. A new sound coding strategy for suppressing noise in cochlear implants. J. Acoust. Soc. Am. 124, 498–509.

Hu, Y., Loizou, P.C., 2007a. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. 122, 1777.

Hu, Y., Loizou, P.C., 2007b. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49, 588–601.

Hu, Y., Loizou, P.C., Li, N., Kasturi, K., 2007. Use of a sigmoidal-shaped function for noise attenuation in cochlear implants. J. Acoust. Soc. Am. 122, EL128-34.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2014. Deep learning for monaural speech separation, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1562–1566.

IEEE, 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. AU-17, 225–246.

Irino, T., Patterson, R.D., 2002. Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. Speech Commun. 36, 181–203.

Jansen, S., Koning, R., Wouters, J., van Wieringen, A., 2014. Development and validation of the Leuven intelligibility sentence test with male speaker (LIST-m). Int. J. Audiol. 53, 55–9.

Jin, S.H., Nie, Y., Nelson, P., 2013. Masking release and modulation interference in cochlear implant and simulation listeners. Am. J. Audiol. 22, 135–146.

Kamath, S., Loizou, P., 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. 2002 IEEE Int. Conf. Acoust. Speech, Signal Process. 4, IV-4164-IV-4164.

Killion, M.C., 1997. Hearing aids: Past, present, future: Moving toward normal conversation in noise. Br. J. Audiol. 31, 141–148.

Kim, G., Lu, Y., Hu, Y., Loizou, P.C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. 126, 1486–94.

Kjems, U., Boldt, J.B., Pedersen, M.S., Lunner, T., Wang, D., 2009. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. J. Acoust. Soc. Am. 126, 1415–26.

Kochkin, S., 2000. MarkeTrak V: "Why my hearing aids are in the drawer": The consumers' perspective. Hear. J. 53, 34–41.

Koning, R., Madhu, N., Wouters, J., 2015. Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners. IEEE Trans. Biomed. Eng. 62, 331–341.

Koning, R., Wouters, J., 2012. The potential of onset enhancement for increased speech intelligibility in auditory prostheses. J. Acoust. Soc. Am. 132, 2569–81.

Krawczyk, M., Gerkmann, T., 2014. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE/ACM Trans. Speech Lang. Process. 22, 1931–1940.

Kujawa, S.G., Liberman, M.C., 2015. Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss. Hear. Res. 330, 191–199.

Kuk, F., Keenan, D., Korhonen, P., Lau, C.-C., 2009. Efficacy of linear frequency transposition on consonant identification in quiet and in noise. J. Am. Acad. Audiol. 20, 465–479.

Levitt, H., 2007. A historical perspective on digital hearing AIDS: how digital technology has changed modern hearing AIDS. Trends Amplif. 11, 7–24.

Levitt, H., Bakke, M., Kates, J., Neuman, A., Schwander, T., Weiss, M., 1993. Signal processing for hearing impairment. Scand. Audiol. Suppl. 38, 7–19.

Lewicki, M.S., 2002. Efficient coding of natural sounds. Nat. Neurosci. 5, 356–363.

Li, N., Loizou, P.C., 2008. Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction. J. Acoust. Soc. Am. 123, 1673–1682.

Lim, J., Oppenheim, A., 1979. Enhancement and band width compression of noisy speech. Proc. IEEE 67, 1586–1604.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J., 1999. The precedence effect. The Journal of the Acoustical Society of America, 106(4), 1633-1654.

Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC Press.

Loizou, P.C., Dorman, M., Tu, Z., 1999. On the number of channels needed to understand speech. J. Acoust. Soc. Am. 106, 2097–2103.

Loizou, P.C., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio. Speech. Lang. Processing 19, 47–56.

Löllmann, H.W., Vary, P., 2009. Low Delay Noise Reduction and Dereverberation for Hearing Aids. EURASIP J. Adv. Signal Process. 2009, 437807.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., Moore, B.C.J., 2006. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc. Natl. Acad. Sci. U. S. A. 103, 18866–9.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Dillier, N., Houben, R., Dreschler, W.A., Froehlich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., Spriet, A., 2010. Multicenter evaluation of signal enhancement algorithms for hearing aids. J. Acoust. Soc. Am. 127, 1491–1505.

Ma, J., Hu, Y., Loizou, P.C., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125, 3387–3405.

Madhu, N., Spriet, A., Jansen, S., Koning, R., Wouters, J., 2013. The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses. IEEE Trans. Audio, Speech Lang. Process. 21, 61–70.

Magnusson, L., Claesson, A., Persson, M., Tengstrand, T., 2012. Speech recognition in noise using bilateral open-fit hearing aids: The limited benefit of directional microphones and noise reduction. Int. J. Audiol. 2027, 1–8.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504–512.

Mauger, S., Warren, C., 2014. Clinical evaluation of the Nucleus® 6 cochlear implant system: Performance improvements with SmartSound iQ. Int. J. Audiol. 53, 564–76.

Mauger, S.J., Arora, K., Dawson, P.W., 2012a. Cochlear implant optimized noise reduction. J. Neural Eng. 9, 65007.

Mauger, S.J., Dawson, P.W., Hersbach, A.A., 2012. Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction. J. Acoust. Soc. Am. 131, 327–36.

Mauger, S.J., Dawson, P.W., Hersbach, A. a., 2012b. Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction. J. Acoust. Soc. Am. 131, 327.

May, T., Dau, T., 2014a. Computational speech segregation based on an auditory-inspired modulation analysis. J. Acoust. Soc. Am. 136, 3350.

May, T., Dau, T., 2014b. Requirements for the evaluation of computational speech segregation systems. J. Acoust. Soc. Am. 136, EL398.

May, T., Dau, T., 2013. Environment-aware ideal binary mask estimation using monaural cues, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

McCreery, R.W., Venediktov, R.A., Coleman, J.J., Leech, H.M., 2012. An Evidence-Based Systematic Review of Directional Microphones and Digital Noise Reduction Hearing Aids in School-Age Children With Hearing Loss. Am. J. Audiol. 21, 295–312.

Monaghan, J.J., Feldbauer, C., Walters, T.C., Patterson, R.D., 2008. Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition. J. Acoust. Soc. Am. 123, 3066.

Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C., and Bleeck, S., 2017. Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners. The

Journal of the Acoustical Society of America, 141(3), 1985-1998.

Moore, B.C.J., 2014. Auditory Processing of Temporal Fine Structure. World Scientific.

Moore, B.C.J., 2012. An Introduction to the Psychology of Hearing. Emerald Group Publishing Limited.

Moore, B.C.J., 2007. Cochlear Hearing Loss: Physiological, Psychological and Technical Issues, Wiley series in human communication science. Wiley.

Moore, B.C.J., 1995. Perceptual Consequences of Cochlear Hearing Loss and their Implications for the Design of Hearing. Ear Hear. 17, 133–160.

Mueller, H.G., Weber, J., Hornsby, B.W.Y., 2006. The effects of digital noise reduction on the acceptance of background noise. Trends Amplif. 10, 83–93.

Munro, K.J., Lutman, M.E., 2004. Self-reported outcome in new hearing aid users over a 24-week post-fitting period. Int J Audiol 43, 555–562.

Nabelek, A.K., Freyaldenhoven, M.C., Tampas, J.W., Burchfiel, S.B., Muenchen, R.A., 2006. Acceptable noise level as a predictor of hearing aid use. J. Am. Acad. Audiol. 17, 626–639.

Nabelek, A.K., Tucker, F.M., Letowski, T.R., 1991. Toleration of Background Noises - Relationship With Patterns of Hearing-Aid Use By Elderly Persons. J. Speech Hear. Res. 34, 679–685.

Neher, T., 2014. Relating hearing loss and executive functions to hearing aid users' preference for, and speech recognition with, different combinations of binaural noise reduction and microphone directionality. Front. Neurosci. 8, 1–14.

Neher, T., Wagener, K.C., 2016. Directional Processing and Noise Reduction in Hearing Aids: Individual and Situational Influences on Preferred Setting. J. Am. Acad. Audiol. 19, 1–19.

Nelson, P.B., Jin, S.H., Carney, A.E., Nelson, D.A., 2003. Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. J. Acoust. Soc. Am. 113, 961–968.

Ng, L., Kelley, M.W., Forrest, D., 2013. Making sense with thyroid hormone--the role of T(3) in auditory development. Nat. Rev. Endocrinol. 9, 296–307.

Nordrum, S., Erler, S., Garstecki, D., Dhar, S., 2006. Comparison of Performance on the Hearing in Noise Test Using Directional Microphones and noise Reduction Filters. Heal. (San Fr. 15, 81–91.

Olshausen, B.A., Field, D.J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature.

Ovegard, A., Lundberg, G., Hagerman, B., Gabrielsson, A., Bengtsson, M., Brandstrom, U., 1996. Sound quality judgement during acclimatization of hearing aid. Scandanavian Audiol. 26, 43–51.

Pati, Y.C.C., Rezaiifar, R., Krishnaprasad, P.S.S., 1993. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. Proc. 27th Asilomar Conf. Signals, Syst. Comput. 1–5.

Pickles, J.O., 2008. An Introduction to the Physiology of Hearing. Emerald Group Publishing Limited, p. 410.

Plack, C.J., 2013. The sense of hearing. Psychology press.

Qazi, O.U.R., van Dijk, B., Moonen, M., Wouters, J., 2013. Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility. Hear. Res. 299, 79–87.

Qin, M.K., Oxenham, A.J., 2003. Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J. Acoust. Soc. Am. 114, 446–454.

Rangachari, S., Loizou, P.C., 2006. A noise-estimation algorithm for highly non-stationary environments. Speech Commun. 48, 220–231.

Rhebergen, K.S., Versfeld, N.J., 2005. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. J. Acoust. Soc. Am. 117, 2181–2192.

Ricketts, T.A., Hornsby, B.W.Y., 2005. Sound quality measures for speech in noise through a commercial hearing aid implementing digital noise reduction. J. Am. Acad. Audiol. 16, 270–277.

Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. IEEE Int. Conf. Neural Networks - Conf. Proc. January, 586–591.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 336, 367–73.

Ryan, A.F., 2000. Protection of auditory receptors and neurons: Evidence for interactive damage. Proc. Natl. Acad. Sci. 97, 6939–6940.

Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M.E., Bleeck, S., 2015. Speech quality evaluation of a sparse coding shrinkage noise reduction algorithm with normal hearing and hearing impaired listeners. Hear. Res. 327, 175–185.

Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M.E., Bleeck, S., 2014. Evaluation of a sparse coding shrinkage algorithm in normal hearing and hearing impaired listeners. Eur. Signal Process. Conf. 310, 1074–1078.

Santos, J.F., Cosentino, S., Hazrati, O., Loizou, P.C., Falk, T.H., 2013. Objective speech intelligibility measurement for cochlear implant users in complex listening environments. Speech Commun. 55, 815–824.

Schmidt, R., 1986. Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. 34, 276–280.

Schnupp, J., Nelken, I., King, A., 2012. Auditory Neuroscience. MIT Press.

Schweitzer, C., 1997. Development of Digital Hearing Aids 2.

Seligman P. M. and McDermott H. J., 1995. Architecture of the spectra 22 speech processor. Ann. Otol. Rhinol. Laryngol. 104, 139–141.

Shannon, R. V, Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–4.

Sigg, C.D., Dikk, T., Buhmann, J.M., 2012. Speech enhancement using generative dictionary learning. IEEE Trans. Audio, Speech Lang. Process. 20, 1698–1712.

Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., van Dijk, B., Van Wieringen, A., Wouters, J., 2007. Speech understanding in background noise with the two-microphone adaptive beamformer BEAM in the Nucleus Freedom Cochlear Implant System. Ear Hear. 28, 62–72.

Steeneken, H.J.M., Houtgast, T., 1980. A physical method for measuring speech-transmission quality 67, 318–326.

Stickney, G.S., Zeng, F.G., Litovsky, R., Assmann, P., 2004. Cochlear implant speech recognition with speech maskers. J. Acoust. Soc. Am. 116, 1081.

Stone, M.A., Moore, B.C., 1999. Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses. Ear Hear. 20, 182–92.

Stone, M.A., Moore, B.C.J., 2002. Tolerable hearing aid delays. II. Estimation of limits imposed during speech production. Ear Hear. 23, 325–38.

Stone, M.A., Moore, B.C.J., 2005. Tolerable hearing-aid delays: IV. effects on subjective disturbance during speech production by hearing-impaired subjects. Ear Hear. 26, 225–35.

Stone, M.A., Moore, B.C.J., Meisenbacher, K., Derleth, R.P., 2008. Tolerable hearing aid delays. V. Estimation of limits for open canal fittings. Ear Hear. 29, 601–17.

Stromsta, C., 1962. Delays associated with certain sidetone pathways. The Journal of the Acoustical Society of America, 34(4), 392-396.

Summerfield, Q., 1992 Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 335(1273), 71-78.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. IEEE Trans. Audio. Speech. Lang. Processing 19, 2125–2136.

Tchorz, J., Kollmeier, B., 2003. SNR estimation based on amplitude modulation analysis with applications to noise suppression. IEEE Trans. Speech Audio Process. 11, 184–192.

Tsoukalas, D.E., Mourjopoulos, J.N., Kokkinakis, G., 1997. Speech Enhancement based on Audible Noise Suppression. IEEE Trans. Speech Audio Process. 7, 497–513.

Turner, C.W., Gantz, B.J., Vidal, C., Behrens, A., Henry, B.A., 2004. Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing. J. Acoust. Soc. Am. 115, 1729–1735.

Van Wieringen, A., Wouters, J., 2008. LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands. Int. J. Audiol. 47, 348–355.

Verschuur, C., Lutman, M., Wahat, N.H., 2006. Evaluation of a non-linear spectral subtraction noise suppression scheme in cochlear implant users. Cochlear Implant. Int 7, 193–196.

Von Kriegstein, K., Smith, D.R.R., Patterson, R.D., Ives, D.T., Griffiths, T.D., 2007. Neural Representation of Auditory Size in the Human Voice and in Sounds from Other Resonant Sources. Curr. Biol. 17, 1123–1128.

Wang, D., Kjems, U., Pedersen, M.S., Boldt, J.B., Lunner, T., 2009. Speech intelligibility in background noise with ideal binary time-frequency masking. J. Acoust. Soc. Am. 125, 2336–47.

Wang, Y., Wang, D., 2013. Towards scaling up classification-based speech separation. IEEE Trans. Audio, Speech Lang. Process. 21, 1381–1390.

Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., Rigoll, G., 2014. Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. Comput. Speech Lang. 1–15.

Wood, S.A., Lutman, M.E., 2004. Relative benefits of linear analogue and advanced digital hearing aids. Int. J. Audiol. 43, 144–55.

Wouters, J., Damman, W., Bosman, A.J., 1994. Vlaamse opname van woordenlijsten voor spraakaudiometrie. Logop. informatiemedium van Vlaam. Ver. voor Logop.

Wouters, J., Vanden Berghe, J., 2001. Speech recognition in noise for cochlear implantees with a two microphone monaural adaptive noise reduction system. Ear Hear. 22, 420–430.

Yang, L.-P., Fu, Q.-J., 2005. Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. J. Acoust. Soc. Am. 117, 1001–1004.

Ye, H., Deng, G., Mauger, S.J., Hersbach, A.A., Dawson, P.W., Heasman, J.M., 2013. A Wavelet-Based Noise Reduction Algorithm and Its Clinical Evaluation in Cochlear Implants. PLoS One 8.

Zakis, J.A., Hau, J., Blamey, P.J., 2009. Environmental noise reduction configuration: Effects on preferences, satisfaction, and speech understanding. Int. J. Audiol. 48, 853–867.

Zakis, J. A., Fulton, B., and Steele, B. R., 2012. Preferred delay and phase-frequency response of open-canal hearing aids with music at low insertion gain. International journal of audiology, 51(12), 906-913.

Zeng, F.G., Rebscher, S., Harrison, W., Sun, X., Feng, H., 2008. Cochlear implants: system design, integration, and evaluation. IEEE Rev. Biomed. Eng.