

University of Southampton
Faculty of Engineering and Applied Science
Institute of Sound and Vibration Research
Signal Processing and Control Group

Enhancement of Body-Conducted Speech from an Ear-Microphone

Kyriakos Papanagiotou

A thesis submitted for the degree of
Doctor of Philosophy

October 2003

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE

INSTITUTE OF SOUND AND VIBRATION RESEARCH

Doctor of Philosophy

ENHANCEMENT OF BODY-CONDUCTED SPEECH FROM AN
EAR-MICROPHONE

by Kyriakos Papanagiotou

This thesis is concerned with the use of optimal filtering to enhance the intelligibility and quality of body-conducted speech. In the context of a mobile communication system, body-conducted speech picked up by an ear-microphone has the advantage of being immune to background noise compared to air-conducted speech, which can be easily degraded by external noise sources. However, their intelligibility and quality differ significantly, as air-conducted speech is subjectively superior to body-conducted speech. This study investigates the design of a filter to enhance the overall quality of body-conducted speech. This is pursued through the application of optimal filtering which uses the body-conducted speech as the input, and the air-conducted speech as the desired signal. The optimal filter, which is generated for specific phonetic material, attempts to match the two signals by minimising their difference in a least-squares sense.

Analysis of simultaneous speech recordings with acoustic and vibration transducers in quiet and noisy conditions demonstrated the noise-immunity characteristic of body-conducted speech, and its low-pass filtering effect due to body-conduction attenuation. Experiments have also shown that the temporo-mandibular joint of a speaker is a good position at which to detect speech vibrations. Optimal filtering has been compared subjectively to high-pass filtering, a technique that is currently used in communication systems. The statistical analysis of standard intelligibility and quality tests showed that, overall, significantly better performance may be obtained with the optimal filter derived for different words/speakers, under both quiet and noisy conditions.

Furthermore, the analysis of the optimal filters suggested that the filter's performance is affected by the electronic noise-floor of the accelerometer, and is not strongly dependent on the frequency content of the speech input. Instead, the performance depends mainly on the anatomical characteristics of a speaker, and the positioning of the transducers. This led to the design of several types of fixed optimal filters, which were generated as averages across a number of speakers and/or speech materials. The results of formal listening tests revealed that an optimal filter that is generated specifically for one speaker outperforms the optimal filters designed for a number of speakers of mixed, or single gender.

Acknowledgements

My deepest gratitude and appreciation to Boaz and Christine for the help and support, both as supervisors as well as friends. Their patience, understanding and guidance is unique.

Many thanks to Professor Mark Lutman and Dr Anna Barney for their constructive comments on this thesis. I would also like to thank EPSRC for funding this project the first 18 months (grant GR/M86026). Much obliged to everybody else that helped me, especially Scott Notley and Nick Tzavides, and to all the subjects that participated in my experiments

A big smile to all the fine people who have partied, danced, laughed, talked, rocked with me, loved and listened to me for the last years. A big frown to all the people who have not done the above. I love you all.

Contents

Contents	iii
1 Introduction	1
1.1 Project Aims	1
1.2 Thesis Structure	3
1.3 Main Contributions	5
1.4 Publications	6
2 The Speech and Hearing Mechanisms	7
2.1 Introduction	7
2.2 The Speech Mechanism	8
2.2.1 Lungs, Larynx and Vocal Tract	8
2.2.2 Modes of Phonation and their Categorisation	10
2.3 The Hearing Mechanism	17

2.3.1	Physiology of the Ear	17
2.3.2	Response of the Ear to Sound	18
2.4	Main Factors Influencing the Reception of one's own Speech .	19
2.4.1	The Transfer Function of the Open Ear (TFOE)	21
2.4.2	Speech Directivity	22
2.4.3	Body-Conduction and Speech Vibrations	24
2.4.4	Jaw Dynamics	26
2.5	Conclusions	29
3	Detection and Enhancement of Body-Conducted Speech	31
3.1	Introduction	31
3.2	History of Speech Vibration Detection and Assessment	32
3.2.1	Devices for Speech Vibrations Detection and Enhancement	36
3.3	Present Communication Systems Using Speech Vibrations . .	41
3.3.1	Laryngeal Microphones	42
3.3.2	Forehead Microphones	42
3.3.3	Mask Microphones	44
3.3.4	Body-conduction Microphones	44
3.3.5	Sound-Pressure-type Ear Microphones	45

3.4	Conclusions	46
4	Review of Speech Intelligibility and Quality Tests	48
4.1	Introduction	48
4.2	Diagnostic Rhyme Test (DRT)	49
4.3	Diagnostic Acceptability Measure (DAM)	51
4.4	Conclusions	54
5	An Optimal Filtering Approach for Speech Vibration Analysis	56
5.1	Introduction	56
5.2	Optimal Filters in the Time Domain	57
5.2.1	The Wiener Filter	60
5.3	Modelling of the Optimal Filter	61
5.4	Conclusions	65
6	Development of the Experimental System	67
6.1	Introduction	67
6.2	Initial Experiment: Aims for the Development of the Experimental Setup	68
6.2.1	Place of Experiments	68
6.2.2	Experimental Equipment	69

6.2.3	Noise-Floor Measurements	72
6.2.4	Speech Recordings	76
6.2.5	Investigation of the Best Conduction Point	78
6.2.6	Preliminary Speech Enhancement and Subjective Eval- uation	79
6.3	Re-design of the Experimental Setup	87
6.3.1	Recording Location	88
6.3.2	Changes in the Experimental Equipment	89
6.3.3	New Experimental Setup	91
6.4	Conclusions	92
7	Subjective Assessment of Optimal Filtering Performance	93
7.1	Introduction	93
7.2	Experiment Aims	94
7.3	Speech Recordings	94
7.3.1	Acquisition and Randomisation Programs	95
7.3.2	Recording Session	96
7.4	Processing of the Signals	98
7.5	Subjective Listening Tests	100

7.6	Results from Listening Tests	102
7.6.1	Results from DRT Test	103
7.6.2	The Variables in the Statistical Analysis	104
7.6.3	Effect of Transducer and Condition on Specific Features	105
7.6.4	Effect of Filter and Condition on Specific Features . . .	110
7.6.5	Effect of Speaker	112
7.6.6	Results from DAM Test	114
7.6.7	Signal Qualities	115
7.6.8	Background Qualities	119
7.6.9	Total Qualities	122
7.7	Conclusions	124
8	Quantitative Assessment of Optimal Filtering Performance	126
8.1	Introduction	126
8.2	Variability of Speech Spectra and Optimal Filter Responses . .	127
8.2.1	Reference Microphone Signals	127
8.2.2	Accelerometer Signals	129
8.2.3	Optimal Filter Responses	131
8.3	Factors Influencing the Performance of Optimal Filtering . . .	135

8.3.1	Effect of Electronic Noise	135
8.3.2	Effect of Non-Causality	138
8.4	Summary and Conclusions	142
9	Design and Assessment of a Fixed Optimal Filter	143
9.1	Introduction	143
9.2	Motivation for the Design of a Fixed Filter	144
9.3	Speech Recordings	145
9.4	Design of the Optimal Filters	146
9.5	Subjective Assessment of the Optimal Filters	151
9.6	Results from Listening Tests	152
9.6.1	Results from DRT Test	152
9.6.2	Effect of Transducer on Specific Features	154
9.6.3	Effect of Filter on Specific Features	155
9.6.4	Results from DAM Test	158
9.6.5	Signal Qualities	159
9.6.6	Background Qualities	161
9.6.7	Total Qualities	163
9.7	Conclusions	164

10 Conclusions and Suggested Further Work	166
10.1 Summary	166
10.2 General Conclusions	167
10.3 Suggestions for Further Work	168
 A Linear Frequency Scale Plots of Chapter 8	 175
A.0.1 Reference Microphone Signals	176
A.0.2 Accelerometer Signals	177
A.0.3 Optimal Filter Responses	178
 B Acoustic and Anatomical Path Delay Calculations	 179

List of Figures

2.1	The main components of the vocal tract.	9
2.2	Main places of constriction for place categorisation.	14
2.3	The three principal parts of the ear.	17
2.4	The auditory area used for intelligible speech.	19
2.5	Components for the reception of one's own voice.	20
2.6	Ear canal transfer function.	21
2.7	Long-term average speech spectra for males and females.	22
2.8	Horizontal and vertical directivity patterns of speech.	23
2.9	The three components of hearing by body-conduction.	26
2.10	Anatomical view of body-conduction.	27
2.11	Frequency spectrum of a typical TMJ click.	28
3.1	The decline of vibrations amplitude produced by the vocal folds as they travel over the surface of the body.	32

3.2	Anatomical locations investigated by Moser and Oyer (1958).	33
3.3	Attenuation characteristic of the path through the head between the mouth and the outer ear canal	37
3.4	The headband microphone and amplifier components.	38
3.5	Configuration of the in-ear transducer by Shigeaki et al. (1999).	40
3.6	The new NTT mobile phone.	41
3.7	Example of a laryngeal microphone.	43
3.8	Example of a forehead microphone.	43
3.9	Example of a mask microphone.	44
3.10	Example of a body-conduction microphone.	45
3.11	Two examples of sound-pressure ear microphones used for mobile phones.	46
4.1	The DAM response sheet given to the subjects of the listening experiment.	53
5.1	The main blocks of an FIR optimal filter.	58
5.2	Physical model of the optimal filtering problem.	62
5.3	All-digital physical model of the optimal filtering problem. . .	64
5.4	Magnitude spectrum of $ \frac{G_{an.}}{G_{ac.}} ^2$	65

6.1	Background noise level spectrum in the lab where the experiments were conducted.	69
6.2	Amplitude and phase response of the accelerometer's amplifier at 40 dB setting.	72
6.3	Experimental setup for the measurement of equipment noise-floors.	73
6.4	Noise-floor of the Data-Acquisition system.	74
6.5	Electronic noise-floor versus speech spectra from the three transducers.	75
6.6	Experimental setup for preliminary experiments.	77
6.7	Speech spectra from 3 anatomical positions.	78
6.8	Frequency and phase responses of the high-pass and band-pass filters.	80
6.9	Block diagram showing the high-pass or band-pass filtering of the accelerometer signal.	80
6.10	Setup for the design of our optimal filter.	81
6.11	Block diagram showing the optimal filtering of the accelerometer signal.	81
6.12	Frequency response and output error spectrum of a characteristic W_{opt}	81
6.13	Desired, accelerometer input and optimal output spectra from four phonemes.	83

6.14	Reference microphone spectrum vs. output spectra from all filters for the sentence “The sunlight strikes raindrops in the air”.	84
6.15	The response scale for ease of understanding and naturalness of presented signals.	85
6.16	The results from the first subjective test.	85
6.17	The results from the second subjective test.	86
6.18	Spectra of speech and noise recorded from the acoustic and vibration transducers.	90
6.19	Experimental setup for the second experiment.	91
7.1	Spectra from recorded speech and noise from the reference microphone and the accelerometer.	98
7.2	Schematic diagram of the Matlab program for optimal filtering.	99
7.3	Schematic diagram of the Matlab program for high-pass filtering with/without noise.	99
7.4	Results from DRT subjective test	103
7.5	Results from DAM subjective test.	114
7.6	Results from DAM subjective test (Signal Qualities)	116
7.7	Results from DAM subjective test (Background Qualities)	119
7.8	Results from DAM subjective test (Total Qualities)	122

8.1	Reference microphone signal spectra from male and female speakers for the DRT and DAM tests.	128
8.2	Average quiet accelerometer DAM spectra for male and female speakers.	130
8.3	Average noisy accelerometer DAM spectra for male and female speakers.	131
8.4	Average quiet W_{opt} magnitude and phase responses for the DAM test.	132
8.5	Average quiet error spectra for the DAM test.	133
8.6	Average noisy optimal filter magnitude and phase responses for the DAM test.	133
8.7	Average noisy error spectra for the DAM test.	134
8.8	Physical model of the optimal filtering problem with electronic noise.	136
8.9	Average coherence functions for the DRT and DAM tests. . .	137
8.10	Impulse responses of an original and an artificially-delayed optimal filter.	139
8.11	R_{xd} 's from the sentence ' <i>Dirt was blown in my face</i> ' for male and female.	140
9.1	Schematic diagram of the Matlab program for optimal filtering with different types of W_{opt}	146

9.2	Magnitude and phase responses of a W_{opt} for a male and a female subject, for the same sentence.	147
9.3	Magnitude and phase response of a $W_{opt1sub}$ for a male and a female subject.	148
9.4	Magnitude and phase response of $W_{opt8sub}$ for eight subjects. .	149
9.5	Magnitude and phase responses of W_{optmal} and W_{optfem}	150
9.6	Results from the DRT subjective test.	153
9.7	DRT results from presentation of the cross-filtered signals. . .	157
9.8	Results from the last DAM quality test.	159
9.9	Results from the last DAM subjective test (Signal Qualities) .	160
9.10	Results from the last DAM subjective test (Background Qualities)	162
9.11	Results from the last DAM subjective test (Total Qualities) .	163
A.1	Reference microphone signal spectra from male and female speakers for the DRT and DAM tests.	176
A.2	Average quiet accelerometer DAM spectra for male and female speakers.	177
A.3	Average noisy accelerometer DAM spectra for male and female speakers.	177
A.4	Average quiet W_{opt} magnitude and phase responses for the DAM test.	178

A.5 Average noisy optimal filter magnitude and phase responses for the DAM test.	178
---	-----

List of Tables

2.1	English phonemes and corresponding features	16
4.1	The DRT words list.	51
4.2	The four DAM sentence groups.	54
7.1	Intelligibility scores for each feature and every transducer con- figuration.	104
7.2	6×2×2 level ANOVA matrix for effect of transducer vs. condi- tion for every feature	106
7.3	Effect of transducer and condition on specific features	107
7.4	6×2×2 level ANOVA matrix for filter vs. condition	111
7.5	Effect of filter and condition on specific features	112
7.6	2×2 level ANOVA matrix for speaker vs. transducer for quiet condition	112
7.7	2×2 level ANOVA matrix for speaker vs. transducer for noisy condition	113

7.8	DAM scores for each Signal sub-category and every system configuration	116
7.9	DAM scores for each Background sub-category and every system configuration	120
7.10	DAM scores for each Total sub-category and every system configuration	122
7.11	2×2 level ANOVA results for filter vs. condition (Individual sub-categories of Total Effect)	124
8.1	Speed of sound (m/s) in different materials	141
8.2	Total delays from all measurements approaches.	141
8.3	Main conclusions from the study of W_{opt} 's.	142
9.1	Main hypotheses for the design of a fixed W_{opt}	145
9.2	Intelligibility scores for each feature and every transducer/filter implementation.	153
9.3	Intelligibility scores for each feature and every transducer configuration.	154
9.4	Intelligibility scores for each feature and every filter implementation.	155
9.5	Effect of filter for specific features. (NS indicates non-significance with $p > 0.05$).	156

9.6	DAM scores for each Signal sub-category and every system configuration	161
9.7	DAM scores for each Background sub-category and every system configuration	161
9.8	DAM scores for each Total sub-category and every system configuration	163
B.1	Speed of sound (m/s) in different materials.	179

Chapter 1

Introduction

1.1 Project Aims

The number of mobile and personal telecommunication systems in use is continuing to grow steadily in the form of dedicated devices for specialised applications, indicating a rising demand for mobility and robustness in this field. However, a frequent problem with such systems is poor speech intelligibility and quality during reception and transmission. The most common factors contributing to this are: (1) non-ergonomic design of the speech pick-up element, (2) obstruction by an apparatus (e.g. breathing mask), and (3) background noise, which degrades the transduced speech signal and has to be minimised so that the signal is as clear as possible. The requirements are, therefore, **ergonomy**, **compactness** and **noise-immunity**.

Ordinary close-talk microphones are widely used and can satisfy the compactness and ergonomy requirements. However, because the microphone in such units is air-conductive, they are not immune to noise. Ear microphones constitute another approach that is sometimes adopted. This satisfies the er-

gonomy requirements, but still presents the problem of noise-susceptibility of the signal.

The other solution that has been recently adopted is the use of a body-conduction, or contact, microphone. This type of system consists of a miniature accelerometer that picks up the speech vibrations near the ear of a speaker, which are caused by the vocal folds and the sounds internal to the vocal tract. This has the advantage that the signal is less influenced by ambient noise. However, it is degraded by the effect of body-conduction, so it is not as clear as an air-conducted signal transduced by a microphone.

Hence, the question is whether the filtering of the body-conducted speech signal by a digital filter could be effective concerning the enhancement of its intelligibility and quality, and if so, how well it could perform compared to the currently used high-pass or band-pass filters. The main aims of this project are therefore:

1. The investigation of body-conducted speech. This is approached with the recording of speech with microphones and a miniature accelerometer in order to better understand the characteristics of both air- and body-conducted speech signals.
2. The study of optimal filtering and its application to the accelerometer signal, to compensate for the loss of quality due to body-conduction attenuation. The filter under investigation uses the microphone as the desired signal, and the accelerometer as the input to the filter. It then tries to match the input to the desired signal in a minimum mean-square error sense, improving the intelligibility and/or quality of the accelerometer signal.
3. The design of a fixed filter that could potentially correspond to a broad range of speech inputs and speakers, and could therefore be more suitable

for a practical implementation.

1.2 Thesis Structure

The first part of this thesis (Chapters 2-4) is focused on the literature relevant to this study. Chapter 5 presents the theory of the optimal filtering approach. Chapters 6-9 discuss all the experimental work, while the last part (Chapter 10) discusses the general conclusions of this project and proposes ideas for future work.

Chapter 2 starts with a description of basic speech physiology, and moves onto the classification of speech according to various phonetic features. It then presents the basic physiology of hearing and discusses its dynamic and frequency response. Finally it gives information on the factors that affect the detection of one's own speech, such as the transfer function of the open ear (TFOE), speech directivity, body conduction, speech vibrations, and the dynamics of the jaw.

Chapter 3 discusses past attempts at speech vibration detection, enhancement and assessment. The topic of mobile communication using speech vibrations is not recent, so in this section a historical background dating back to the 1950's and reaching contemporary communication systems is presented.

Chapter 4 presents the theory behind the design and application of the Diagnostic Rhyme Test (DRT), and the Diagnostic Acceptability Measure (DAM), the two listening tests used in our experiments, for assessing, respectively, the intelligibility and quality of processed speech.

Chapter 5 comprises a review section on optimal (Wiener) filtering theory, the main enhancement strategy that was adopted throughout this study. The

chapter also includes the development of a model for our optimal filtering problem, that takes into account the acoustic and anatomical influences on the filter's design.

Chapter 6 presents the development of the experimental setup. The chapter includes a presentation of the experimental equipment, the electronic noise-floor measurements, and speech recordings at different anatomical positions in order to choose a good speech vibration conduction point. The outputs from optimal, high-pass and band-pass filters are compared, and initial conclusions are derived for the potential of optimal filtering in enhancing speech vibration signals.

Chapter 7 acquaints the reader with the formal experiments that investigate the performance of word- and sentence-specific optimal filters compared to fixed high-pass filters. It contains a detailed description of the speech recording sessions, the processing of the results, a description of the programs that were written in Matlab for the randomisation and analysis of the signals, and finally the subjective listening tests, using the DRT and DAM procedures in order to assess the processed speech in quiet and noisy conditions. The chapter then concludes with the results from the statistical analysis of the two tests.

Chapter 8 comprises an objective analysis of the optimal filters, in order to understand the factors that influence their performance. This is approached initially by a presentation of the spectra of the microphone and accelerometer signals, and the optimal filters' magnitude and phase responses and output error spectra. The chapter proceeds to discuss the effects of electronic noise and non-causality on the coherence of the system with the help of the model from Chapter 5.

Chapter 9 presents the design and assessment of a 'general' optimal filter that corresponds to a broader range of inputs and speakers. The chapter presents an

optimal filter which is an average of optimal filters across speech materials and speakers, and its subjective comparison to optimal filters designed specifically for words or sentences.

Finally, **Chapter 10** discusses the main conclusions of this project and proposes some ideas for a potential future expansion of this study.

1.3 Main Contributions

The main contributions of this study, which also satisfy the main objectives of the project, are the following:

- The investigation of the potential performance of a more ‘general’ fixed filter that corresponds to a broader range of phonetic inputs and speakers, and is easier to implement in practice.
- The experimental demonstration of the optimal filter’s superior performance compared to the currently used high-pass filtering.
- The study and interpretation of the nature of body-conducted speech picked up by an accelerometer, compared to air-conducted speech picked up by a microphone.

Other contributions are:

- A solid understanding of the most efficient position for the detection of speech vibrations in the context of a mobile communication system.
- The development of an efficient system for the detection of speech vibration signals. This was accomplished with the design of a low-noise experimental setup incorporating simultaneous recordings of air- and body-conducted speech.

- The development of an analysis system and an enhancement method based on optimal (Wiener) filtering for speech vibration signals.
- The development of a model, and the derivation of an optimal filtering expression for this specific application.

1.4 Publications

The work from Chapter 6 was presented in a poster at the 141st meeting of the Acoustical Society of America (Chicago, IL, June 2001).

The work from both Chapters 6 and 7 was presented in a conference paper at the 1st meeting of the Hellenic Institute of Acoustics (Patra, Greece, September 2002).

Chapter 2

The Speech and Hearing Mechanisms

2.1 Introduction

Speech communication can be considered to be the passage of information from one speaker to a listener via an air-conducted acoustic speech signal and includes two processes: production, which is the generation of the speech signal, and perception, which interprets the speech signal in the ear and the brain. The human speech production and hearing mechanisms are believed to have evolved in parallel, since they work in a complementary way. The human ear is especially responsive to those frequencies in the speech signal that contain most information. The aim of this chapter, therefore, is to present the basic concepts of the speech and hearing mechanisms that are relevant to this study. The motivation is that a basic understanding of the physiology, the dynamic characteristics, as well as the factors that influence the function of those two mechanisms is vital in order to comprehend the concept of the reception of a speaker's own speech, whether this is an air-conducted signal detected by a

microphone, or a body-conducted signal detected by an accelerometer.

The first part presents an introduction to the speech mechanism's physiology. This is followed by an analysis of the various modes of phonation and their manner and place categorisations. The chapter then proceeds with a basic coverage of the ear's anatomy, while its dynamic and frequency responses are discussed. The last part presents the main factors that influence the detection of one's own speech. The points that are covered here include the transfer function of the open ear and the effect of speech directivity as influencing factors for the detection of the air-conducted signal. The analysis continues by presenting the concept of body-conduction in relation to speech vibrations and the dynamics of the jaw, as an influencing factor for the detection of the body-conducted signal.

2.2 The Speech Mechanism

The speech production system can be divided into three main sections: the lungs, the larynx and the vocal tract, which includes the various articulators that create the speech signal. Of particular importance to our study is the larynx and the vocal tract. We will present these two parts briefly, since an appreciation of their functions and physiology is important in order to discuss the manner and place categorisations of speech, as well as the theory of intelligibility and quality tests in Chapter 4.

2.2.1 Lungs, Larynx and Vocal Tract

The larynx connects the lungs and trachea to the vocal tract. (see Figure 2.1). Of particular importance are the vocal folds, two small muscular cushions

forming a diaphragm with a slit-like opening that modulates the airstream as it vibrates open and shut. The opening between them is called the glottis and the term glottal has come to be used as a general term for laryngeal function (length of the opening is about 2.5cm in men and 1.5cm in women). The tension applied to the vocal folds, the lung pressure and their mass determine the fundamental frequency (F0) of their modulation. The shorter and lighter vocal folds of most women vibrate almost twice as rapidly as those of men. This rate of vibration is closely related to the perception of the speaker's vocal pitch. The range of F0 for an adult male is from about 80 Hz to about 160 Hz with an average value of about 132 Hz. For an adult female, typical values give a lower limit at about 150 Hz while the upper limit of the range is at approximately 500 Hz, with an average value of about 223 Hz (O'Shaughnessy, 2000).

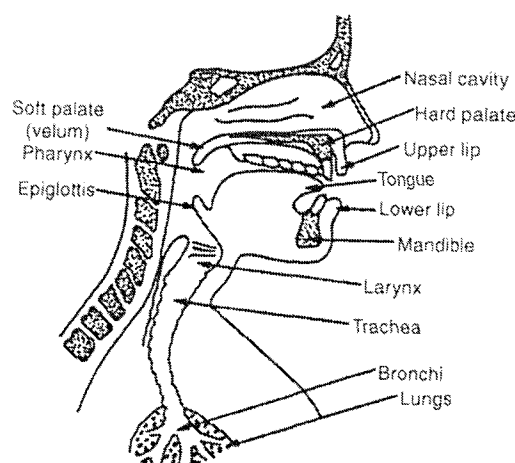


Figure 2.1: The main components of the vocal tract (after Owens (1993)).

The vocal tract is a tubular passageway surrounded by muscular and bony tissues. It provides the means to produce various sounds that characterize spoken language. The vocal tract has two functions: (1) it can modify the spectral distribution of energy in glottal sound waves, and (2) it can contribute to the generation of sound for stops and fricatives. It consists essentially of the pharynx and the oral and nasal tracts (see Figure 2.1). The vocal tract is a non-uniform acoustic tube, approximately 17cm long in an adult male, terminated

at the front by the lips and at the back by the vocal folds or larynx. Its cross-sectional area can be varied by muscular control of the speech articulators (lips, tongue, jaw and velum). The nasal tract is also a non-uniform tube of fixed area and length (about 12cm in an adult male). The vocal tract can be modelled as an acoustic tube with resonances, called formants, as well as antiresonances. The formants are abbreviated F_i , where F_1 is the formant with the lowest frequency. By moving the various articulators, the shape of this acoustic tube is altered, which in turn changes its frequency response; the simple presence of the vocal tract boosts output speech intensity by 10-15 dB as an impedance match between the glottis and free space beyond the lips. The vocal tract's filter amplifies energy at the formant frequencies while it attenuates energy at the antiresonances.

2.2.2 Modes of Phonation and their Categorisation

The speech sounds are categorized in different schemes, however, voicing, manner and place is one set concerned with the way in which they are produced (articulatory description). *Voicing* describes whether or not the vocal folds are vibrating. The *manner* of articulation is concerned with airflow, the path that it takes and the degree to which the vocal tract impedes it. The *place* of articulation describes the most constricted section of the tract during the phoneme production. The classification of a phoneme, therefore, includes elements from the above set.

Voicing Categorisation

- (a) **Voiced** sounds include all vowels, glides, liquids, nasals, and some fricatives, affricates and stops. Those are produced when the vocal folds are tensed together and they vibrate in a relaxation mode as the air pressure

builds up, forcing the glottis open, and then subside as the air passes through. This vibration of the folds produces an airflow waveform, which is approximately triangular. Being periodic, or at least quasi-periodic, it has a frequency spectrum of rich harmonics at integral multiples of the fundamental frequency of vibration. The vocal tract acts as a resonant cavity, which amplifies some of these harmonics and attenuates others to produce voiced sounds.

- (b) In the production of **unvoiced** sounds the vocal folds do not vibrate. For instance, a steady forced exhalation will produce a hissing sound caused by turbulence set up in the flow of air through the numerous irregularities along the vocal tract. Spectrum analysis of the unvoiced sounds reveals a band of practically continuous frequency coverage largely confined to the upper portion of speech range (Kinsler et al., 1982). There are three basic types of unvoiced speech sounds: fricatives, stops and affricates, which are described below.

Manner Categorisation

- *Vowels* are produced by voiced excitation of the vocal tract with the position of the articulator remaining static. There is no nasal coupling and sound radiation is from the mouth. Examples are /i/ (e.g. *beat*), /ɛ/ (e.g. *bet*), /æ/ (e.g. *bat*), /ʌ/ (e.g. *but*), /o/ (e.g. *coat*). Vowel energy is primarily concentrated below 1 kHz and falls off at about -6 dB/oct with frequency. At the acoustic level, vowels are characterised primarily by the locations of their first three formant frequencies. Usually F_1 decreases as tongue height increases during articulation, and F_2 decreases as the tongue is shifted backward, while lip rounding lowers the first two formants by reducing the size of the mouth opening.
- *Diphthongs* are a combination of two vowel sounds. They are therefore

similar to vowels except that the gesture is created when the articulators move slowly from one static vowel position to another. Examples are /ɛ i/ (e.g. *bay*), /aɪ/ (e.g. *eye*), /aʊ/ (e.g. *how*).

- *Semivowels* are also vowel-like gestures. They include the *glides* and *liquids*.
- *Glides*, (/w/ as in *wow*, /j/ as in *you*), employ narrow vocal tract constrictions that under conditions of unusually strong airflow may cause friction; the glides are close in articulation to high vowels (i.e. /j/-/i/).
- *Liquids*, (/r/ as in *row*, /l/ as in *lull*), are similar to vowels but use the tongue as an obstruction in the oral tract, causing air to deflect around the tip.
- The *nasals* /m/ (e.g. *mother*), /n/ (e.g. *no*), /ŋ/ (e.g. *doing*) are produced by vocal fold excitation with the vocal tract totally constricted at some point along the oral passageway. The velum is lowered, which allows airflow through the nostrils. Thus the oral cavity acts as a side-branch resonator that traps acoustic energy at certain frequencies. In the sound radiated from the nostrils, these resonant frequencies of the oral cavity appear as antiresonances.
- *Stop* (plosive) phonemes involve the complete closure and subsequent release of a vocal tract obstruction. The velum is raised to prevent nasal airflow during the closure. After a rapid closure, pressure builds up behind the occlusion and is suddenly released with a rush of air that creates a short burst. This transient excitation may occur with or without vocal cord vibration to produce voiced (/b/ as in *boy*, /d/ as in *dad*, /g/ as in *gig*), or unvoiced plosive sounds (/p/ as in *play*, /t/ as in *toy*, /k/ as in *keg*). Plosives are characterized by transient bursts of energy and as a result their properties are highly influenced by the sounds which precede or succeed them.

- When turbulent airflow occurs at a constriction in the vocal tract, the category of sounds produced are called *fricatives*. The point of this constriction occurs near the front of the mouth in English and its exact location characterizes the particular fricative sound that is produced. The unvoiced fricatives /f/ (e.g. *fire*), /θ/ (e.g. *thin*), /s/ (e.g. *super*), /ʃ/ (e.g. *show*) are produced without vocal cord excitation whereas in their voiced counterparts /v/ (e.g. *vote*), /ð/ (e.g. *that*), /z/ (e.g. *zoo*), /ʒ/ (e.g. *treasure*), the vocal folds are vibrating.
- *Affricates* consist of the unvoiced /tʃ/ as in *church*, and the voiced /dʒ/ as in *judge*. They are produced when a stop and fricative consonant are both shortened and combined. They may therefore be interpreted as the consonant equivalent of a diphthong (Owens, 1993).

Place Categorisation

The place categorisation of phonemes (see Figure 2.2) involves the following groups:

- *Labials*: these sounds are produced when the main constriction is at the lips (/m, p, b/).
- *Dental*: produced when the tongue tip (apex) touches the edge or back of the upper incisor teeth (/θ/).
- *Labio-dental*: produced when lower lip forms a constriction with the upper teeth (/f, v/).
- *Alveolar*: produced when the tongue blade approaches or touches the alveolar ridge (/n, t, d, r, l, s, z/).
- *Alveo-palatal*: produced when the tongue blade is against the place which overlaps the alveolar ridge and palate.

- *Palatal*: when the main constriction is formed between the back of the tongue (dorsum) and the hard palate, (/j/).
- *Velar*: when the the main constriction is formed between the tongue dorsum and the soft palate, (/ŋ, k, g/).
- *Uvular*: when the dorsum approaches the uvula.
- *Pharyngeal*: when the main constriction is formed at the pharynx.
- *Glottal*: when the main constriction or closure is formed at the vocal folds (/h/).

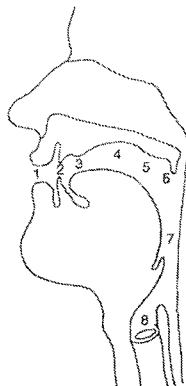


Figure 2.2: Places of articulation: (1) labial, (2) dental, (3) alveolar, (4) palatal, (5) velar, (6) uvular, (7) pharyngeal, (8) glottal (after O'Shaughnessy (2000)).

Feature Categorisation

There is one more categorisation of phonemes according to their articulatory and acoustic characteristics. Those features are more commonly used in listening tests involving speech intelligibility and quality. It is important to present them, as they will be used later during the discussion of our listening tests in Chapter 4. The features are the following (Crystal, 1991):

- *Continuant* (sustained): Continuant sounds have been defined as those produced with an incomplete closure of the vocal tract. All vowels and fricatives are continuant.

- *Compact*: Compact sounds are defined as those which involve a stricture relatively far forward in the mouth and a relatively high concentration of acoustic energy in a narrow central part of the sound spectrum (e.g. high or mid. vowels, 2-3 kHz).
- *Grave*: Grave sounds are defined as those involving a peripheral articulation in the vocal tract and a concentration of acoustic energy in the lower frequencies. Back vowels and labial or velar consonants are grave.
- *Sibilant*: Refers to a fricative sound made by producing a narrow, groove-like stricture between the blade of the tongue and the back part of the alveolar ridge. These sounds such as /s/ or /ʃ/ have a high frequency hiss characteristic.
- *Nasal*: Refers to sounds produced while the soft palate is lowered to allow an audible escape of air through the nose. Both consonants and vowels may be articulated in this way. Nasal consonants occur when there is a complete closure in the mouth and all the air thus escapes through the nose, e.g. /m, n, ŋ/. English has no nasal vowels, but nasalisation is often heard on English vowels, when they display the articulatory influence of an adjacent nasal consonant as in *mat* or *hand*.
- *Voicing*: A fundamental term used in the phonetic classification of phonemes, referring to the auditory result of the vibration of the vocal folds. Sounds produced while the vocal folds are vibrating are voiced sounds, e.g. /b, z, a, i/ as opposed to those produced with no such vibrations which are called voiceless or unvoiced, e.g. voiceless stops /p, t, k/.

Table 2.1 (O'Shaughnessy, 2000) shows all the phoneme classifications (voicing, manner, place, and feature), each accompanied by an example word.

Phoneme	Manner	Place	Voiced	Nas.	Sust.	Sib.	Grav.	Comp.	Ex.
/ i /	vowel	high front	+	-	+	-	-	-	beat
ɪ	vowel	high front	+	-	+	-	-	-	bit
ɛ	vowel	mid front	+	-	+	-	-	-	bet
e	vowel	mid front	+	-	+	-	-	-	pay
æ	vowel	mid front	+	-	+	-	-	-	bat
ɐ	vowel	low back	+	-	+	-	+	+	sofa
ɑ	vowel	low back	+	-	+	-	+	+	hard
ɔ	vowel	mid back	+	-	+	-	+	-	caught
o	vowel	mid back	+	-	+	-	+	-	coat
ʊ	vowel	high back	+	-	+	-	+	-	book
u	vowel	high back	+	-	+	-	+	-	boot
ʌ	vowel	mid back	+	-	+	-	+	-	but
ɜ	vowel	mid central	+	-	+	-	-	-	bird
ə	vowel	mid central	+	-	+	-	-	-	about
ɔɪ	diphthong	mid back→high front	+	-	+	-	+	-	boy
eɪ	diphthong	mid front→high front	+	-	+	-	-	-	hey
oʊ	diphthong	mid back→high back	+	-	+	-	+	-	hoe
aʊ	diphthong	low back→high back	+	-	+	-	+	-	how
aɪ	diphthong	low back→high front	+	-	+	-	-	-	eye
ju	diphthong	front→high back	+	-	+	-	-	-	you
j	glide	front	+	-	-	-	-	+	you
w	glide	back	+	-	-	-	-	-	wow
l	liquid	alveolar	+	-	-	-	-	-	lull
r	liquid	retroflex	+	-	-	-	-	-	roar
m	nasal	labial	+	+	-	-	+	-	aim
n	nasal	alveolar	+	+	-	-	-	-	none
ŋ	nasal	velar	+	+	-	-	+	-	bang
f	fricative	labiodental	-	-	+	+	+	+	fin
v	fricative	labiodental	+	-	+	+	+	+	valve
θ	fricative	dental	-	-	+	+	-	+	thin
ð	fricative	dental	+	-	+	+	-	+	then
s	fricative	alveolar	-	-	+	+	-	+	sin
z	fricative	alveolar	+	-	+	+	-	+	zoo
ʃ	fricative	alveopalatal	-	-	+	+	-	-	shoe
ʒ	fricative	alveopalatal	+	-	+	+	-	+	measure
h	fricative	glottal	-	-	+	+	-	-	how
p	stop	labial	-	-	-	-	+	+	pop
b	stop	labial	+	-	-	-	+	+	bob
t	stop	alveolar	-	-	-	-	-	+	tot
d	stop	alveolar	+	-	-	-	-	+	did
k	stop	velar	-	-	-	-	+	-	kick
g	stop	velar	+	-	-	-	+	-	gig
tʃ	affricate	alveopalatal	-	-	-	-	-	+	church
dʒ	affricate	alveopalatal	+	-	-	-	-	+	judge

Table 2.1: English phonemes and corresponding features (after O'Shaughnessy (2000)).

2.3 The Hearing Mechanism

This section presents the principal aspects of the human ear's physiology and response in order to complement the information concerning the speech mechanism. The aim is to aid the understanding of the various factors that influence the reception of a person's own speech, as will be described in section 2.4.

2.3.1 Physiology of the Ear

The three principal parts of the human auditory system as shown in Figure 2.3 are the *outer ear*, the *middle ear* and the *inner ear*.

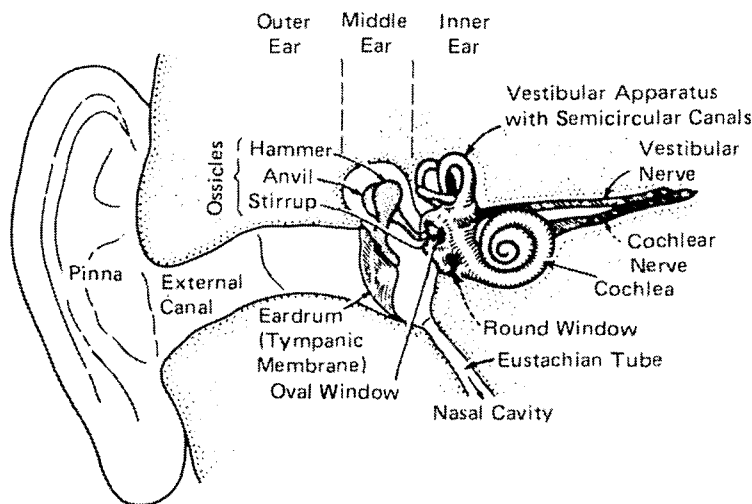


Figure 2.3: The three principal parts of the ear. Reproduced with permission from Roederer (1995).

The **outer ear** is composed of the pinna and the auditory canal (or auditory meatus). The auditory canal is terminated by the tympanic membrane or the eardrum. The outer part has a cartilage skeleton and the deep part is bony. The **middle ear** is an air-filled cavity spanned by the three tiny bones, the ossicles, called the *malleus*, the *incus* and the *stapes*. The malleus is attached to the eardrum and the stapes is attached to the oval window of the inner ear. Together these three bones form a mechanical, lever-action connection

between the air-actuated eardrum and the fluid-filled cochlea of the inner ear. The middle ear is vented to the upper throat behind the nasal cavity by the Eustachian tube which is used to equalise the static air pressure of the middle ear with the outside atmospheric pressure. The middle ear also contains two small muscles and is traversed by the facial nerve before it exits the skull (Westermann, 1987). The **inner ear** comprises a dense bony capsule containing a membranous labyrinth, which forms the cochlea, vestibule and semicircular canals. The membranous part is sealed from the middle ear by the stapes footplate and round window membrane. The cochlea contains the organ of hearing, which is connected by the auditory nerve to the brain stem.

2.3.2 Response of the Ear to Sound

The sensitivity of the ear to an external excitation of sound decreases considerably towards low (and also towards very high) frequencies. Maximum sensitivity is achieved around 3 kHz. The shape of this threshold curve is influenced by the acoustical properties of the auditory canal and by the mechanical properties of the bone chain in the middle ear (Roederer, 1995).

Curve A of Figure 2.4, the threshold of hearing, tells us that human ears are most sensitive around 3 kHz. At this most responsive region, a sound pressure level of 0 dB can just barely be heard by a person with normal hearing. Curve B represents the level at each frequency at which a tickling impression is felt in the ears. This occurs at a sound pressure level of about 120-130 dB at which a sensation of pain is produced. In between the thresholds of hearing and feeling is the area of audibility.

However, speech uses only a portion of the auditory area. This region is located centrally in the auditory area; neither the extremely soft or extremely loud sounds, nor sounds of very low or very high frequency, are used in common

speech sounds. With frequencies ranging from 50 Hz up to approximately 10 kHz or more, speech has amplitudes between 30 and 90 dBA SPL (measured at a distance 1m from the lips). Therefore, the speech area should have fuzzy boundaries to represent these excursions in level and frequency. The auditory region used for intelligible speech has an average dynamic range of about 42 dB and a frequency range of about 5 octaves (150 Hz-5.5 kHz) (Everest, 1994), and is shown in Figure 2.4.

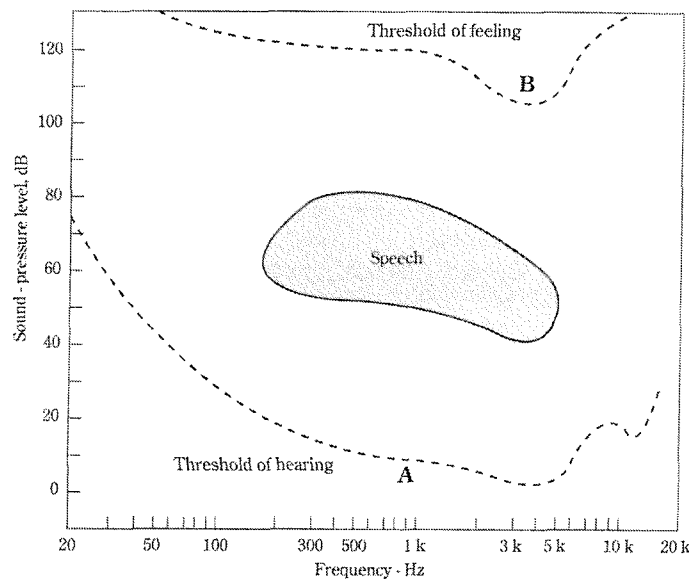


Figure 2.4: The portion of the auditory region used for intelligible speech communication (after Everest (1994)).

2.4 Main Factors Influencing the Reception of one's own Speech

The analysis of speech picked up by a microphone or an accelerometer near a speaker's ear, as it is done in this study, is a difficult task, since the recorded signal is influenced by a number of acoustic and anatomical factors. The aim of this section, therefore, is to advance the knowledge from the two previous parts by giving some insight on those multi-disciplinary factors that influence the reception of one's own air- or body-conducted speech. This section will

help us understand the experimental results of Chapters 6 to 9, since it conveys the necessary information to interpret the spectra of speech recorded by both transducers.

The reception of one's own voice is influenced by three main pathways, as shown in Figure 2.5.

- (a) Direct sound transmission from the mouth through the air around the speaker's head to the ear drums.
- (b) Internal sound transmission inside the head via bones and skull to the cochlea.
- (c) Reflections from one's own voice from surfaces in the surrounding environment.

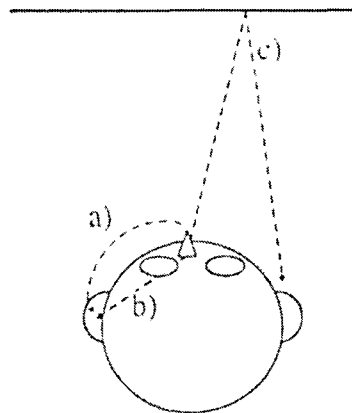


Figure 2.5: Components for the reception of one's own voice (Porchmann, 2000). Reproduced with permission from Acustica-Acta Acustica.

Some examples may help to illustrate these different components. When wearing an in-ear microphone for communication or an earmuff, one's own voice is perceived as significantly changed. The sound transmission around the speaker's head (path a) is attenuated, and the body-conducted component is enhanced due to the occlusion imposed by the device. If a speaker listens to a recording of his/her own voice, the absence of internal sound transmission

due to body-conduction (path b) becomes apparent. For our study, path (c) representing reflections from an enclosed space is not so important as we are more concerned with path (b) compared to path (a) (see Chapters 6 and 8).

2.4.1 The Transfer Function of the Open Ear (TFOE)

An important parameter that characterises the behaviour of the external ear is the transformation of sound pressure level from the free field to the eardrum. It is affected by the structure of the external ear, the diffraction from the head and the pinna, together with resonances in the ear canal. These diffraction and resonance effects influence the Transfer Function of the Open Ear (TFOE). This can be defined as the ratio of the increase in sound pressure at the eardrum, due to the presence of the head, to the relative free-field pressure without the head present (Wiener and Ross, 1946). Therefore it should not be unexpected that the sound at the eardrum is found to be greater than the free-field pressure. TFOE is important for our study as it influences the recorded speech and noise signals from the miniature transducers (see Chapters 7 and 8).

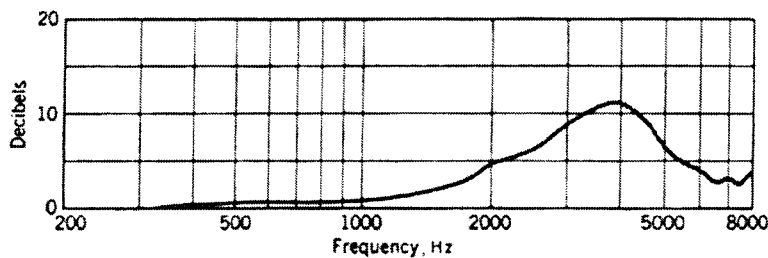


Figure 2.6: Ear canal transfer functions at 0° azimuth (after Wiener and Ross (1946)).

Wiener and Ross (1946) concluded that the above ratio is a function of frequency and can reach values of about 20 dB in the region of 3 kHz (see Figure 2.6), where we have the ear canal's first major resonance, i.e. approximately 2.3cm corresponding to a quarter-wavelength resonance. The amplification is greater if the source is at an azimuth angle approaching 90° .

2.4.2 Speech Directivity

Most of the energy in speech is found in the lower frequencies, particularly below 1 kHz, whereas intensity falls off as frequency increases above this range. Although the long-term average speech spectrum tends to be similar for male and female speech in the range 250-5 kHz, male levels are considerably higher at frequencies below 160 Hz, and female levels are slightly higher in the range above 5 kHz (see Figure 2.7).

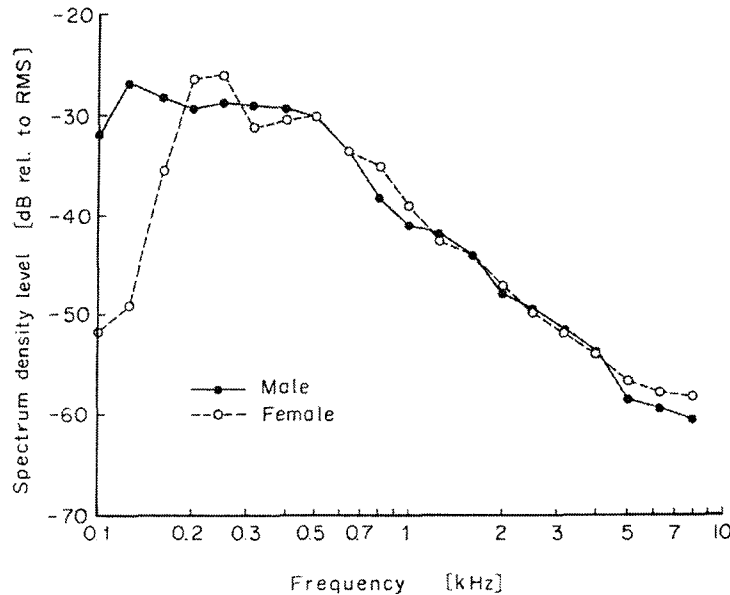


Figure 2.7: Long-time averaged spectrum calculated for utterances made by 80 speakers (after Sadaoki (1989)).

However, speech sounds do not have the same strength in all directions. This is due primarily to the sound shadow cast by the head. Previous studies (Flanagan, 1960; Kuttruff, 1979; McKendree, 1986) concerning speech directivity concluded three main points. First, the free-field radiated sound spectrum from the mouth varies as a function of distance and angle. Second, the distance attenuation is caused by the medium in which sound is transmitted, whereas the human head and torso cause the directional attenuation. Third, the angular dependence affects the perceived speech spectrum and intelligibility. The first point is the one directly related to our study, and becomes

apparent with the simultaneous recording of speech with a reference microphone 60cm from the speaker's mouth, and a miniature in-ear microphone (see main experiments in Chapters 7 and 8).

One of the first and most basic studies concerning speech directivity is the one conducted by Dunn and Farnsworth (1939), which constitutes the starting point for many recent studies on this topic. A single speaker in a seated position repeated a fifteen-second sample of a phonetically balanced text, while r.m.s pressure measurements were made in thirteen frequency bands (62.5 Hz-12 kHz), and at seventy-six positions, in different directions and distances. A reference microphone in front of the mouth was also used so that the various speech levels in different positions could be compared with it. In this study, the radiation from surfaces such as the throat, the chest and the back of the neck, are mentioned for the first time as possible factors that could influence the spatial distribution of speech at a particular frequency. The results appear in Figure 2.8 for the horizontal and vertical planes.

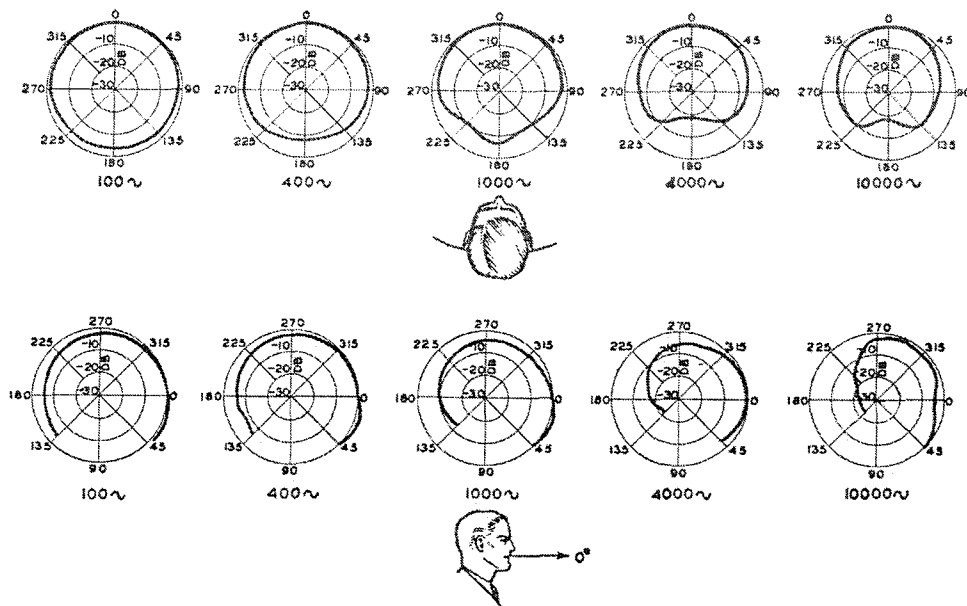


Figure 2.8: The directional characteristics of the human voice in the horizontal (top) and vertical (bottom) planes, for five frequency bands (Dunn and Farnsworth, 1939), (Note: \sim = Hz). Reproduced with permission from Olson (1967).

One can see that in the horizontal and vertical planes, there is a directional effect of about 5-10 dB in the 100 Hz to 400 Hz band. This is to be expected because the head is small compared to the wavelengths associated with this frequency band. There are significant directional effects, however, for the 1-4 kHz band. For this band, which contains important speech frequencies, the front-to-back difference is about 20 dB, except for the torso. Those results also agree with those of Kuttruff (1979), Flanagan (1960) and Huopaniemi et al. (1999). Generally, for head directivity in the horizontal plane, as the angle of incidence increases (the angle between the mouth and the receiving transducer), it results in low-pass filtering. The attenuation of high frequencies (greater than 1.5 kHz) becomes especially significant as the wavelength approaches, and decreases beyond, the dimensions of the human head, at about 2 kHz.

2.4.3 Body-Conduction and Speech Vibrations

It is important to differentiate between “artificial” body-conducted sounds such as the ones induced by bone-conduction hearing aids, and “self-generated” body-conducted sounds that are generated during speaking. The second type of body-conducted signal is the one that is of interest to us, and is essentially the reciprocal of the first, since the stimulation happens internally rather than externally, but the mechanism of vibration transmission is similar. The motivation for this section is the need to comprehend this common mechanism and the factors that affect the transmission of the speech vibrations. The term body-conduction, however, refers only to the conduction of speech sounds through the bones. In our case, the transmission path from the various speech sources in the head to the pick-up point near the ear is much more complex, involving tissues, muscles, tendons as well as bones. ‘Body-conduction’, therefore, is a more appropriate term to use, and will be adopted throughout this

study.

Regarding the excitation of body-conducted speech, Howell et al. (1988) identified various types of vibrations. He distinguished between vibrations associated with the movement of the vocal folds, which force the bones to vibrate, and vibrations caused by the resonances of the cavities in the vocal tract. Some common ground has been found in terms of the conduction pathways to the ear. This can be seen in the work by Tonndorf (1972), as well as in the paper by Khanna et al. (1976), in accordance with Schroeter and Poesselt (1986). Those pathways are:

- The external ear component, in which the air in the ear canal is excited by vibrations of the ear canal walls, as well as the soft-tissue conduction of sound originating from vibrations of circumaural skin and pinna.
- The middle-ear component, which is caused by relative vibrations of the ossicles and the skull due to middle-ear inertia.
- The inner-ear component, which excites the cochlea directly, causing the so-called distortional compressions.

Bekesy (1960) concluded that the mechanism for reception of body-conducted sounds was due to three phenomena. For the low frequencies, the relative movement of the ossicles controlled the body-conduction response. Above 1.5 kHz, compressional waves become apparent, and hence the response is attributed to compression of the labyrinth. He also suggested a third possibility as a source of body-conducted stimuli: relative movement between the jaw and the skull. All three cases appear in Figure 2.9. According to Khanna et al. (1976), vibrations on the walls of the ear canal produce most of the signal that can be detected in or around the ear canal. In this respect, the effects of the middle and inner-ear components are minimal.

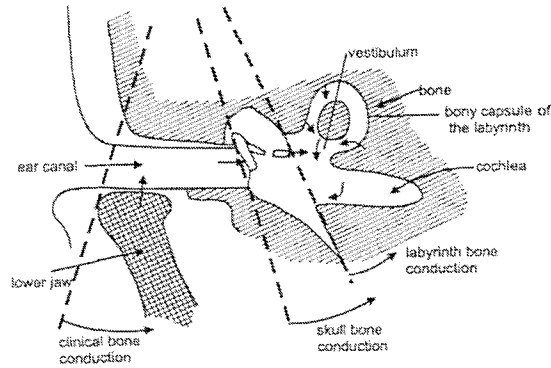


Figure 2.9: An illustration of the three components of hearing by body-conduction. The left part, clinical body-conduction, includes the movement of the lower jaw. Reproduced with permission from Stenfelt (1999).

One of the most detailed studies on body-conducted sound is that of Schroeter and Poesselt (1986). They modelled the contributions from the middle ear and the ear canal walls to the total sound pressure in the occluded ear canal. However, they considered only the sound produced by a bone vibrator placed on the skull and not by one's own voice. The literature provides little data on body-conducted sound produced by a speaker's own speech. As a general comment inferred from the very few studies, vibrations from different speech sources in the head are transmitted, not only to the ear canal, but to the middle and probably to the inner ear as well. All three pathways, therefore, contribute to the detection of the body-conducted sound. The area of speech vibration detection, assessment and enhancement will be presented more thoroughly in Chapter 3.

2.4.4 Jaw Dynamics

A factor that influences the reception of the body-conducted part of speech is the mechanism of the lower mandible (jaw). The lower jaw is attached by its condyle to the skull at the Temporo-Mandibular Joint (TMJ), close to the outer ear canal. When the skull is set in motion, either when moving or speaking, the jaw oscillates with the same frequency as the rest of the skull,

however with a different phase and amplitude, since it is not firmly attached to the other skull bones. This difference in phase and amplitude yields a relative motion due to which, this part is compressed and expanded, something that induces sound in the outer-ear canal and acts on the eardrum (see Figure 2.10).

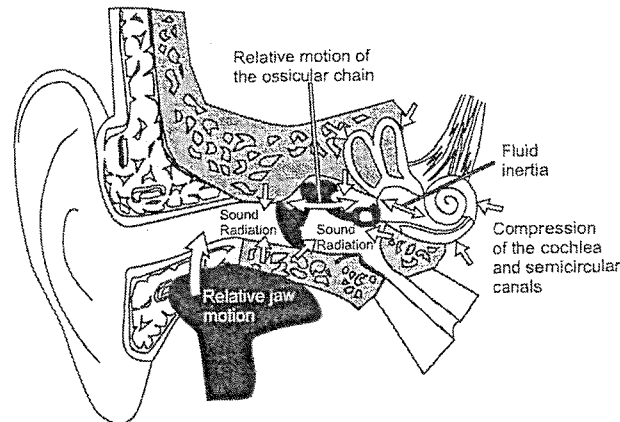


Figure 2.10: An anatomical view of the body-conduction route during jaw movement. Reproduced with permission from Stenfelt (1999).

Experiments by Franke et al. (1952) and Howell et al. (1988) indicated that opening and closing of the mouth increases the sound pressure produced by body-conduction in the closed auditory canal by as much as 6-10 dB in the frequency range between 40 and 700 Hz. This action was also investigated by Prinz (1998) and Shiga and Kobayashi (1995), who found that sounds in the TMJ were created during mandibular movement (opening of the jaw as far as possible and then closing it). Prinz (1998) recorded the vibration signals from 216 subjects with in-ear microphones and inferred that TMJ clicks are produced by an impact of the condyle with the fossa, and that any energy release in the joint excites resonances in adjacent structures. This point can be supported by the findings of Wee and Ashley (1990) who observed that it is impossible to isolate the recording of sounds transmitted through a single muscle without interference by other muscles, when a group of agonist muscles are contracting simultaneously. The averaged spectra of clicks all showed a similar spectrum with a major peak at approximately 180 Hz (see Figure 2.11). The function of the jaw is an important detail for our study, as the signals

detected by the accelerometer on the subjects' TMJ's as well as from the in-ear microphone during our speech recordings were influenced by its movement (see Chapters 6 and 7).

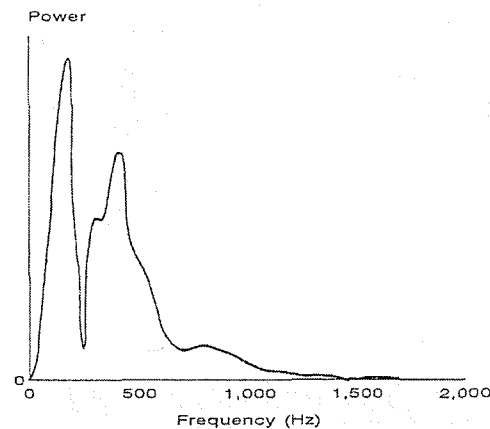


Figure 2.11: Frequency spectrum of a typical TMJ click (after Prinz (1998)).

As will be seen in Chapters 6 and 7, the main experiments involve body-conducted speech recordings with a miniature accelerometer placed on the TMJ of one or more subjects. A reasonable question, therefore, concerns the useful information we can extract from these body-conducted signals. From the theory of this chapter, we saw that jaw movement is likely to influence the recorded speech spectra. The answer is difficult, since specific experiments focused on the dynamic characteristics of each speaker's jaw need to be planned, something that is beyond the scope of this study. The only solution that was adopted was to listen to the recorded body-conducted speech, observe the time-domain signals and discard any data contaminated with 'mechanical' noise, a frequent point during the recording/playback of the Temco signals (see Chapter 7).

2.5 Conclusions

The first point that we can infer from the study of the speech mechanism's physiology concerns the generation of speech vibrations from different anatomical sources. Therefore, the larynx, being the main source of voiced speech, is expected to produce a vibration signal with a relatively high amplitude. This is shown in the literature presented in Chapter 3, and also in Chapter 6 with the recording of body-conducted speech using an accelerometer from different anatomical locations (see section 6.2.5). On the same assumption, both the nose and the larynx are expected to be the main source of vibrations for nasal speech. For fricative sounds, however, it is difficult to assume a point that could carry the main weight of speech vibration generation, as frication takes place along the vocal tract for different fricative phonemes. Therefore, we can support the fact that the identification and analysis of the speech vibration sources and transmission pathways is a complex task, since speaking generates vibrations on the whole body (Bekesy, 1960; Tomatis, 1987).

Another point that can be inferred from this chapter concerns the effect of the TFOE on the spectra of recorded speech and acoustic noise with the accelerometer. Since the accelerometer in our speech recordings was placed on the skin surface over the speaker's TMJ, the amplification of sound (speech and noise) due to the TFOE, is expected to influence the spectrum of the body-conducted speech radiated from the subject's ear canal to the circum-aural area. This is also shown in Chapter 8 with the presentation of acoustic noise spectrum, recorded by the accelerometer on the skin surface over the TMJ, which is amplified in a similar fashion to the TFOE's main amplification response.

Speech directivity is another point that, as will be seen in Chapter 6, influences the spectrum of the in-ear microphone signal. Therefore, the miniature

microphone signal, which for our experiments was placed at the opening of the subject's ear-canal, gave a spectrum which, at the higher frequencies, was lower than the on-axis reference microphone spectrum.

The last point concerns the TMJ as a possible vibration pick-up point. As was seen from section 2.4.4, the TMJ plays an important role as a mechanism of vibration transmission from the ear canal during speaking, to the circumaural area. Its role, therefore, as a means to convey speech vibrations is important. This point is supported with the literature of Chapter 3 where the TMJ's vibration conduction properties are discussed, as well as from Chapter 6, where it is chosen as the most appropriate position for speech vibration pick-up.

To summarise, this chapter has contributed towards an understanding of the main principles of the speech and hearing mechanisms while, at the same time, it has given a broader picture of the factors influencing the reception of a speaker's own air- and body-conducted speech.

Chapter 3

Detection and Enhancement of Body-Conducted Speech

3.1 Introduction

Not many publications on the detection and assessment of speech vibrations were found in the open literature. The aim of this chapter, therefore, is to discuss the topic of speech-induced vibrations, their detection, measurement and enhancement from the few studies on this field. The first section presents selected papers in the history of speech vibration detection and assessment. We then describe previous attempts to design communication devices that use speech vibrations, while the last section presents the main classes of present systems that exploit this family of signals. The motivation behind this chapter is the study of the experimental procedures and results discussed in the literature. This will not only give a starting point for the development of our own experimental setup, but will also reveal some possible weak points that could be challenged by our approach, as far as the speech recording procedure and the enhancement strategies are concerned.

3.2 History of Speech Vibration Detection and Assessment

In the 1920's, it was first noticed that when a stethoscope was laid over any part of the known resonance area of the speech mechanism, i.e. larynx, throat, cheeks, nose, etc., a distinct vibration was communicated to the ear of the observer when the underlying area was functioning as a resonating chamber. Wise (1932) rank-ordered the conductive efficiency of different types of tissue. From most efficient to least efficient, they are as follows: (1) bony tissue, (2) tendonous tissue, (3) tense muscle tissue, (4) relaxed muscle tissue and (5) soft non-muscular tissue. In 1939, Bekesy reported an investigation on areas of the body from the abdomen to the mastoid process with reference to the relative amplitude of skin surface vibration during phonation. According to his findings, the amplitude of vibration decreases with progressively greater distance from the larynx (see Figure 3.1). No details are given for this study, apart from the fact that the measurements were conducted with a stethoscope.

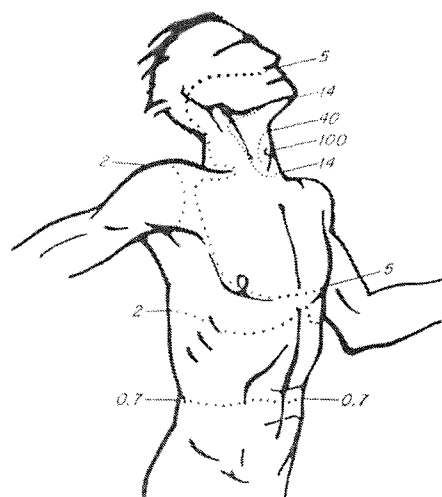


Figure 3.1: The decline in relative amplitude of vibrations produced by the vocal folds as they travel over the surface of the body (Bekesy, 1960).

Figure 3.1 is the best representation of speech vibration contours in the literature, and can be used as a 'map' for our study in order to see how speech

vibrations travel through the body and especially through the head, which is our primary interest. As one can notice, the main vibratory pathway on the head passes from the TMJ, a location that is exploited in the experimental phase of the project.

In a related study, Mullendore (1949) reported relative amplitudes of vowel sounds at various locations of the body. He ranked 10 locations in decreasing order as follows: (1) larynx, (2) mandible, (3) nose, (4) top of head, (5) clavicle, (6) vertebra, (7) sternum-superior end, (8) sternum-inferior end, (9) mastoid and (10) fifth rib. However, the most fundamental study, on which many contemporary papers were based, is the one by Moser and Oyer (1958). During their investigation on speech vibration detection, they used 3 subjects with a reference microphone placed 30cm from their lips and an accelerometer on 16 different locations on their heads and necks (see Figure 3.2). The subjects sustained different vowel sounds at a constant level (about 76 dB) for 5s. The way the accelerometer was mounted on the head was examined, and it was found that with a medium pressure applied to the accelerometer, the measurements were constant.

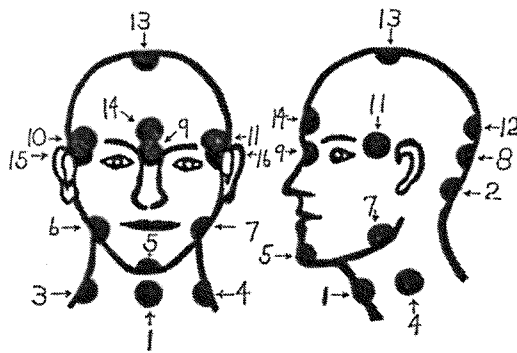


Figure 3.2: Anatomical locations investigated by Moser and Oyer (1958).

The resultant 576 recordings (12 vowels at 16 locations from three subjects) were analysed for relative intensities by spectral power level measurements. The anatomical locations were divided into four groups according to the relative intensity of the transduced vowels. The composite intensities of the test

vowels were highly similar at the larynx, the sides of the neck and at the angles of the mandible. Those were the three positions where the highest intensity of the transduced vowels was recorded. The difference between the strongest point (larynx and side of mandible) and the weakest point (sphenoid bone) was found to be 21 dB. The recorded signals were also compared subjectively. There was unanimous agreement that the most faithful reproduction of vowels occurred at the forehead but the loudest signals were obtained from the mandible and laryngeal areas. In respect to the larynx, it was noted that positioning of the transducer was extremely critical. When it was positioned slightly below the thyroid notch, it became impossible to make a phonetic distinction even between /i/ and /a/. The above findings were a starting point for the selection of the accelerometer positions in our experiments (see Chapter 6). As will be seen, we only tested a certain number of speech vibration pick-up points (larynx, mastoid, TMJ) as most of the other locations, as seen above, have poor vibration conduction properties.

This topic was also studied experimentally by Maurer and Landis (1990) from a psychophysical angle. In their tests they used 20 subjects with normal hearing. The voices of the subjects were recorded with two different transducers separated in two channels of a tape recorder with an integrated mixing desk. On one channel, the voice was recorded by a commercially available microphone placed 10cm in front of the mouth. On the other channel, the voice was recorded from the skull by an accelerometer with a suitable preamplifier. The accelerometer was pressed by the subject onto the mastoid bone behind the right ear by the subject while speaking. The subjects were trained prior to the tests, to press the accelerometer until distortions occurred in the recording and then to release the pressure just enough for the distortions to disappear. Hence, the voice was simultaneously recorded by means of an air- and body-conduction pathway. The recordings could then be played back by a mixing desk, in such a way that the subject could adjust the settings (loudness) of

the two separate recordings until the perceived voice matched the subject's own voice when speaking. The results were revealing in the sense that all subjects added body-conducted speech when searching for their own familiar voice. Most subjects showed a constant mixture of body/air-conducted speech within the limits of 10 dB.

Similar work has been conducted by Giua (1998), who performed a study of "accelerometric" speech. He recorded 7 seconds of continuous speech through a reference microphone and an in-ear microphone (Temco 'voiceducer') which has also been used in this project (see Chapter 7). The main difference that he found between the two spectra is that above 600-800 Hz, the signal picked up by the accelerometer was low-pass filtered due to the effect of body-conduction. Other possible parameters that could affect the accelerometric speech, according to him, are head and body movements that can introduce low frequency noise.

An important finding was the reduced intelligibility of the accelerometric voice, especially for back vowels. Generally, voiced stops (/b, d, g/), nasals (m, n, ŋ), liquids (/l, r/) and voiced fricatives (/v, z, ð, ʒ/) were enhanced, while unvoiced consonants (/p, t, k, f, s/) were attenuated. A simple intelligibility test was performed, and it was found that the nasals and semivowels were the most intelligible (89%) followed by the stops (83%), then the voiced fricatives (81%) and finally the unvoiced fricatives (56%). It is interesting to note that the range of the above results is similar to our results from the intelligibility test of Chapter 7. In his paper, he also proposes the simultaneous recording of body- and air-conducted speech in order to overcome the noise problem.

Analogous work was also performed by Ishimitsu and Kitazake (2001), who recorded three sets of Japanese words simultaneously with a reference microphone placed in front of the mouth of three male subjects, in combination with an accelerometer tested on three positions on the subjects' head: on the cheek

(possibly the TMJ), on the larynx and on the upper jaw. The results showed that the upper jaw was the best pick-up point for speech vibrations. With a basic intelligibility test, he found that the accelerometric speech obtained a high score in both high (17 dB) and low (-6 dB) signal-to-noise ratio cases (94% and 91% respectively).

3.2.1 Devices for Speech Vibrations Detection and Enhancement

Probably the first attempt to test the idea of speech vibration detection for communication systems dates back to 1957 with Black's study on an in-ear microphone. In this paper, the disadvantages of having a microphone in front of the mouth (e.g. moisture on the capsule, impediment to user's movements) are discussed for the first time, while the usefulness of having a compact, noise-immune system in the ear, is noted. The speech material consisted of the phrase "*Joe took father's shoe bench out*", which the subjects had to cite at a constant level. The material was picked up by a reference microphone in front of the subject's mouths, and by the in-ear microphone which picked-up the subjects' speech vibrations radiated in their ear canals. Both signals from 6 subjects were eventually recorded on tape. The attenuation characteristic of the head path, through the different resonating cavities of the head and body-conduction is then approximated as the difference between the recorded voice spectrum at the mouth and the voice spectrum in the ear (see Figure 3.3). He also comments on the difficulty of detecting consonants through the body-conduction path and proposes the complementary use of both a conventional and an in-ear microphone.

Another attempt to approach the topic of accelerometric speech was conducted by Sebesta et al. (1970). In this study an accelerometer referred to as an iner-

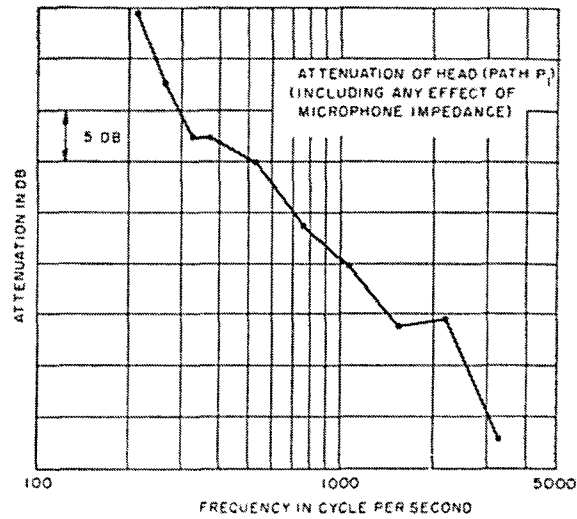


Figure 3.3: Attenuation characteristic of the path through the head between the mouth and the outer ear canal (after Black (1957)).

tial microphone, mounted on a headstrap at the back of the head, was tested. The system was completed with the addition of an amplifier and a loudspeaker for the purpose of listening to transmitted speech for two-way communication (see Figure 3.4). Laboratory speech tests were performed in an ambient sound field of 110 dB SPL. The speech samples that were used were characteristic of the various phoneme categories. Therefore the syllables *ah* and *oh* were used as being representative of vowel sounds, and *sixth* and *fifth* were chosen as being rich in sibilants. The material was recorded with a reference microphone in front of the subject's mouth and the accelerometer. The results in this study came in close agreement with the findings of Giua (1998); that is, after about 800 Hz, there is a steep roll-off of the body-conducted sound, and vowels are detected more clearly than consonants.

Ono (1977) also performed a comparison for speech vibration signals, radiated into the ear canal of a subject, and then picked up by a sound-pressure type in-ear microphone. The shortcomings that he noticed were that:

- (a) The output signal was very dependent upon how well the ear microphone was inserted in the ear-canal.

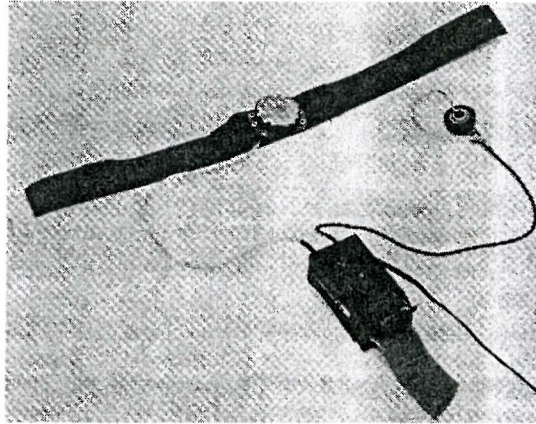


Figure 3.4: The headband microphone and amplifier components, (Sebesta et al., 1970).

- (b) Listeners were very concerned about their safety and security caused by their inability to hear outside sounds since the ear microphone obstructed their ear-canal.
- (c) It was necessary to wear earmuffs to reduce high ambient noise, in order to increase the signal-to-noise ratio.

Consequently, he designed a speech vibration transducer that picked-up speech vibrations directly. This utilised a piezoelectric element, which ran parallel to the temporal bone in the ear-canal. This position was chosen instead of other places on the skull since it was assumed that the collection of speech vibrations conducted to bones deeper than the ear canal cartilage, would be spectrally richer and could be unconstrained from the effect of the lower jaw movement. He then compared the spectra of sustained vowels, obtained from three different positions in the ear canal of a subject, with the spectra obtained simultaneously from a reference microphone fixed close to the subject's lips. The three positions were the following:

1. In contact with the inlet of the external auditory meatus.
2. In contact with the cartilagenous part of the external auditory canal and,
3. In contact with the bony part of the ear-canal.

In a simple quality test he compared the signals recorded at the positions (1)-(3), and observed that the signal picked up from points (1) and (2) was enhanced in the lower frequency region (300-600 Hz) and showed an attenuation of 10 dB/oct after 1 kHz. However, the signal from point (3) showed a smaller attenuation, even for high frequencies (> 1.5 kHz) and hence chosen as the optimal speech vibration pick-up point. For its enhancement, he used various high-pass filters with cut-off frequencies of 200, 500, 1000 and 1500 Hz, with a roll-off rate of 12 dB/oct in quiet conditions. In an intelligibility test using one hundred vibration-recorded syllables, with four subjects (2 female, 2 male), the scores were 96% for 200, 500 and 1000 Hz cut-off frequencies, and 94% for 1500 Hz (unfiltered signals were not used in his intelligibility test). The above procedure was also helpful for the development of our experimental setup (see Chapters 6 and 7), as it provided an enhancement strategy that could be challenged and compared with optimal filtering. We, therefore, applied the above high-pass filters to our speech vibration signals, and performed an intelligibility test to see their effectiveness compared to optimal filtering, not only in quiet but in noisy conditions as well.

The use of a combination of acoustic and vibration transducers, was also explored by Viswanathan et al. (1985). In this study a shaping filter for the enhancement of accelerometric speech was used (sharp high-pass at about 800 Hz, 5 dB/oct boost over the range 800-2800 Hz, flat response over 2.8-4.7 kHz, sharp low-pass at 4.7 kHz). The accelerometer was placed on the side of the lower jaw during speech recordings. The recordings employed the use of the DRT and DAM test for intelligibility and quality testing respectively (see Chapter 4). A selected set of two-sensor systems (acoustic and vibration) and individual sensors including the frequency-shaped accelerometer, were tested in 95 dB and 115 dB SPL of simulated fighter aircraft cockpit noise. The two-sensor systems produced essentially the same DRT and DAM scores in the 95 dB SPL case, and even higher scores in the 115 dB SPL case, as compared

to individual microphones, indicating the noise-immunity characteristic of the body-conducted signal. The shaping filter from this study was also applied to our accelerometer speech signals (see Chapter 6), and its outputs were compared subjectively to the ones of optimal filtering with an intelligibility test.

In the study of Shigeaki et al. (1999), a microphone/receiver unit using an accelerometer has been developed for use in noisy environments. This unit consisted of an acoustic transducing part that picked-up and reproduced voice signals, and an equalising circuit that controlled the optimum sound quality for various ambient noise levels. The transducers involved a small loudspeaker referred to as a receiver, enabling two-way communication, a microphone and an accelerometer. All three components were included within a single earplug-shaped case, which fitted easily into the user's ear (see Figure 3.5). This device was adopted by the Japanese company Temco, which designed the Temco-voiceducer that was tested in our experiments (see Chapter 6).

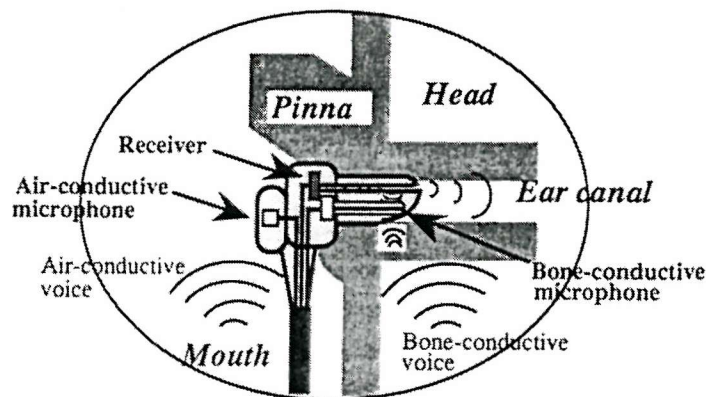


Figure 3.5: Configuration of the in-ear transducer (after Shigeaki et al. (1999)).

In the proposed design, the accelerometer and receiver units are acoustically isolated so that there is no interaction leading to acoustic feedback. The output signal of the equalising circuit consists of three band-restricted signals, i.e. low-pass and high-pass filtered air-conducted voice and low-pass filtered body-conducted voice with an optimum weight which depends on the environmental noise level. The system estimates the noise level automatically by

integrating the air-conducted sound as the noise level. Most of the components of this unit were embedded in CMOS. When the noise level is low, the frequency characteristic is almost the same as the air-conducted sound. When the noise increases, the frequency characteristic includes few high-frequency components (air-conducted) but more, enhanced lower-frequency components (body-conducted).

The latest development in speech vibration detection comes from the Japanese firm NTT DoCoMo, which has come up with a mobile phone that is worn on the wrist. Its receiver sends vibrations through the hands to the end of the fingers. Hence, in order to hear the caller, the user presses the finger on the ear's pinna and listens (see Figure 3.6).



Figure 3.6: The new mobile phone, created by NTT DoCoMo, Japan, (IEE Telecommunications magazine).

3.3 Present Communication Systems Using Speech Vibrations

Hands-free interaction between mobile communication devices is an important feature, since it can increase the flexibility of present applications where the user cannot be impeded by a hand-held or a head-mounted transducer (e.g.

a boom microphone). In this context, one could adopt the solution of a directional microphone, when the position of the talker is fixed, e.g. a fixed microphone in a car compartment (Omologo et al., 1998). However, a small movement of the speaker or the effect of environmental noise can degrade the speech signal. Hence, for mobile and robust speech transduction, a small and noise-immune transducer, like an accelerometer, is an attractive solution. The aim of this section is to present the different classes of commercial products which are currently available as communication devices, using speech vibrations as the signal to be detected. Sound-pressure type in-ear microphones are also presented as complementary information. The motivation is to provide information concerning their operation, the enhancement techniques applied to the body-conducted speech signal and the positioning of their vibration pick-up element. The section could, therefore, trigger some ideas that could be tested in our experiments concerning the positioning of the accelerometer and/or microphone, and the enhancement of the speech vibration signal.

3.3.1 Laryngeal Microphones

The laryngeal microphone is fixed on the anterior wall of the trachea to pick up the voice signal during phonation (see Figure 3.7). Since speech at this area involves an enhanced low-frequency content, the clarity is generally low even without noise. When it is worn, it presses on the neck from the outside and is fairly uncomfortable. For this reason, it is not usually the first choice for specialised applications.

3.3.2 Forehead Microphones

The forehead microphone is essentially an accelerometer-based system that picks-up the voice conducted to the skull. The voice quality is theoretically

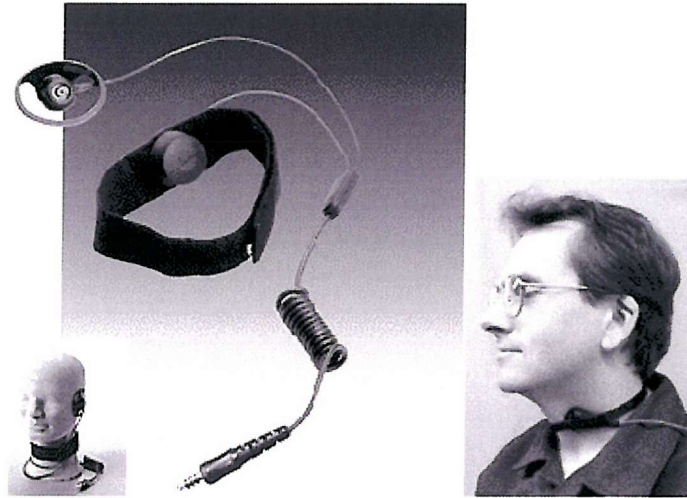


Figure 3.7: An example of a laryngeal (throat) microphone combined with a receiver (left) and the same unit in use (right) by SavoxTM.

almost the same (see section 3.2) as that obtained with the body-conduction microphone. The greatest disadvantages of the forehead microphone is the low gain because of long distance from the vocal cords and the difficulty of placement. If the forehead microphone is fixed with a helmet (see Figure

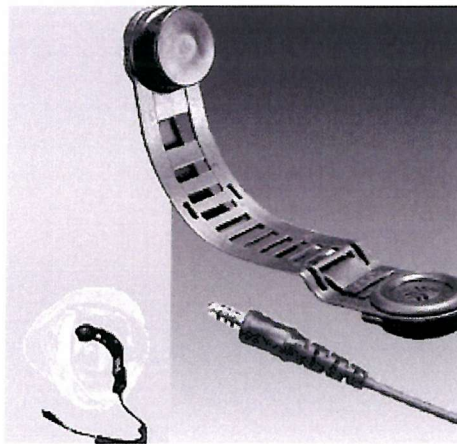


Figure 3.8: Example of a forehead microphone used inside a helmet.

3.8), the helmet functions as if it were a sounding board and transmits the noise to the forehead microphone, thereby increasing the noise output of the microphone. The helmet also introduces resonant peaks in its response.

3.3.3 Mask Microphones

Those are accelerometers, which are placed on a mask for fire-fighting and for radioactivity protection, as well as for underwater applications (see Figure 3.9). The drawbacks of this design are that it responds poorly to a deep voice, has low clarity, and is sensitive to moisture. The mask has a built-in microphone which would be unnecessary on some contingencies, since it is sensitive to rough handling.



Figure 3.9: Example of a mask microphone used for underwater and firefighters' communication.

3.3.4 Body-conduction Microphones

One type of transducer that has been used in communication systems where suppression of background noise is desirable, is the *body-conduction microphone*. Usually, this is nothing more than a piezoelectric accelerometer, built in a small device, which can be placed on different points around the head. The manufacturing companies often propose a different location around the head for their products. The motivation behind this application is that vibrations caused by speech are immune to acoustic noise, in contrast to an air-conducted signal transduced e.g. from a microphone, that will be degraded significantly

by the presence of acoustic noise. Different commercial systems are available, where the accelerometer is often housed in an in-ear type device that is placed in the user's ear-canal.



Figure 3.10: Example of a body-conduction microphone placed in the ear-canal of the user.

However, because the vibrations are transmitted through bone or skin, the high frequency speech components are easily attenuated. Hence, the transduced speech is unnatural and there is always the need to consider where and how to put the microphone in order to get the best results. The speech enhancement strategies adopted in those systems are fairly simplistic (high-pass or band-pass filtering). A typical in-ear body-conduction microphone appears in Figure 3.10.

3.3.5 Sound-Pressure-type Ear Microphones

The basic concept of sound-pressure-type ear microphones is to pick up the sound pressure produced in the external auditory canal (see Figure 3.11). A drawback of the device is that the external auditory canal must be sealed off. Hence, there is an abnormal sensation in speaking because of obstruction of the ear canal (also known as the occlusion effect). The sound-pressure type does not provide noise reduction, unless the ear canal is sealed. It has however, an advantage that it can be bilaterally used as a transmitter/receiver with the microphone and loudspeaker (receiver) in the same casing. In terms

of voice quality, since the low frequency is increased and high frequency is decreased when the external auditory canal is blocked, the voice sounds deep, but the use of an adequate high-pass filter provides sufficient clarity (Ono, 1977). The development of the body-conduction microphone, later on in the history of speech vibration detection, originated as an improvement to the sound-pressure-type ear microphone, since it proved to be superior regarding noise reduction properties.



Figure 3.11: Two examples of sound-pressure ear microphones used for mobile phones (EarmarkTM left, and JabraTM, right).

3.4 Conclusions

This chapter contributed towards a broader understanding of the area of speech vibration detection and its use by commercial systems. The limiting factor when reviewing this area was the small number of research papers available; this topic is still not studied in depth, since the literature does not provide any information on the identification of the transmission pathway of speech vibrations. A general remark for all the studies listed in this chapter was the use of a reference microphone in front of the speaker's mouth in order to compare air- and body-conducted speech, a point that was adopted throughout our experiments. Another point is the use of a miniature in-ear microphone

to pick-up the radiated speech vibrations from the speaker's ear-canal, a point that was also adopted in the experiments of Chapter 6.

As an overall comment stemming from this chapter, it can be inferred that body-conducted speech is most important at lower frequencies. For frequencies between approximately 600 Hz and 1200 Hz, it dominates the perception of one's own voice, meaning that the body-conducted signal has a relatively high amplitude. This is demonstrated in our speech recordings of Chapters 6 and 7 where, as will be seen, sounds like vowels or nasals with strong low-frequency energy content are detected more clearly than other sounds like fricatives.

As far as the enhancement of the body-conducted speech is concerned, most of the studies made use of a linear high-pass or band-pass filter applied on the body-conducted speech signal. This approach has the advantage of low complexity and ease of implementation. However, only Shigeaki et al. (1999) took advantage of both air- and body-conducted signals by mixing them with an equalising circuit. In this case, optimal filtering appears as a better solution, since it improves the quality of the body-conducted signal, by matching it to the air-conducted signal in a minimum mean-square error sense. This is investigated in Chapters 6 to 8, where optimal filtering is compared subjectively to the high-pass and band-pass filters discussed above.

Chapter 4

Review of Speech Intelligibility and Quality Tests

4.1 Introduction

Speech intelligibility and quality are two concepts that must be based on subjective human responses when they are tested. Speech intelligibility tests are only appropriate for communication systems that produce moderate to severely degraded speech, since, by their very nature they are unable to distinguish among signals that are all highly intelligible. These systems, however, may still differ in other perceivable characteristics such as pleasantness, naturalness, or the recognizability of the talker. Subjective tests that point out those fine subjective preference details are more appropriately termed speech quality tests.

The main aim of this chapter, therefore, is to present the theory behind two standardised procedures for testing speech intelligibility and quality, the Diagnostic Rhyme Test (DRT) and the Diagnostic Acceptability Measure (DAM),

respectively. The motivation is the need to understand their theory and function since they are used as the main subjective tests in Chapters 7 and 9 to study the performance of optimal filters in enhancing speech picked up by a vibration sensor. The chapter presents the theory of the DRT test and continues with the theory of the DAM test. The speech materials and the rules when administering both are also presented and analysed as far as the phonetic content and the recording-listening procedures are concerned.

4.2 Diagnostic Rhyme Test (DRT)

Voiers (1977b) advanced the concept of the already existing rhyme tests for intelligibility in his Diagnostic Rhyme Test (DRT). The DRT has been extensively used to test the intelligibility of communication systems. It uses 192 monosyllabic English words arranged in 96 word pairs. The words are constructed from a consonant-vowel-consonant sound sequence and differ only in their initial consonants. Listeners are shown a word pair, then asked to identify which word is presented by the talker. The test results reveal errors in the discrimination of initial consonant sounds. The distinctive features considered in the DRT are *voicing*, *nasality*, *sustention*, *sibilation*, *graveness* and *compactness* (see section 2.2.2 for definitions). The advantage of the DRT is that it gives not only an overall intelligibility rating for a speech system, but also a “diagnostic” rating consisting of intelligibility scores in each of the distinctive feature categories (see feature categorisation in section 2.2.2).

When administering the DRT, it is best if more than one individual is used in the recording, while the stimulus words should be cited at approximately 1.5s intervals. In each listening session, at least eight to ten listeners should be used with all listener responses pooled together (Quackenbush et al., 1988). The listener’s task is simply to strike out the member of the word pair she/he

judges to have been uttered, whether the material is played back processed or not. The most straightforward scoring of the DRT yields seven scores, one for each of the six consonant distinctive feature categories and one mean score. All scores are corrected for random guessing effects as follows:

$$S_j = \frac{100(R_j - W_j)}{T_j} \quad (4.1)$$

where S is the “true” percent-correct responses, R is the number of correct responses, W is the number of incorrect responses, T is the total number of responses, and subscript j denotes the distinctive feature class. The mean score is an average of the diagnostic scores, which is equivalent to grouping all responses into a single class. In essence, what equation 4.1 does, is to convert a chance level (say 50%) in R to, respectively, 0% in S .

In recognition of evidence that the state of a feature may not be equally intelligible in every manifestation, various constraints were observed in assembling the corpus of test words which can be seen in Table 4.1:

1. For half of the word pairs designed to test for the intelligibility of voicing, both critical phonemes involve friction (i.e. are affricates or fricatives); for half the critical phonemes are stops.
2. Half of the nasality word pairs in each vowel context involve a grave phoneme-pair (i.e. /m-b/); half involve an acute pair (i.e. /n-d/).
3. Half of the word pairs designed to test for the intelligibility of sustention involve a voiced phoneme-pair; half, an unvoiced pair.
4. Half of the word pairs concerned with sibilant involve voiced phonemes; half, unvoiced phonemes.
5. In the case of graveness, word pairs were constructed such that for each vowel context, one pair is voiced and the other is unvoiced.

6. Compactness word pairs were constructed such that states of vowel-likeness, sibilant, voicing, and sustention were given equal representation in the test.

Voicing		Nasality		Sustention		Sibilant		Graveness		Compactness	
veal	feel	meat	beat	vee	bee	zee	thee	weed	reed	yield	wield
bean	pean	need	deed	sheet	cheat	cheep	keep	peak	teak	key	tea
gin	chin	mitt	bit	vill	bill	jilt	gilt	bid	did	hit	fit
dint	tint	nip	dip	thick	tick	sing	thing	fin	thin	gill	dill
zoo	sue	moot	boot	foo	pooh	juice	goose	moon	noon	coop	poop
dune	tune	news	dues	shoes	choose	chew	coo	pool	tool	you	rue
vole	foal	moan	bone	those	doze	joe	go	bowl	dole	ghost	boast
goat	coat	note	dote	though	dough	sole	thole	fore	thor	show	so
zed	said	mend	bend	then	den	jest	guest	met	net	keg	peg
dense	tense	neck	deck	fence	pence	chair	care	pent	tent	yen	wren
vast	fast	mad	bad	than	dan	jab	gab	bank	dank	gat	bat
gaff	calf	nab	dab	shad	chad	sank	thank	fad	thad	shag	sag
vault	fault	moss	boss	thong	tong	jaws	gauze	fought	thought	yawl	wall
daunt	taunt	gnaw	daw	shaw	chaw	saw	thaw	bong	dong	caught	thought
jock	chock	mom	bomb	von	bon	jot	got	wad	rod	hop	fop
bond	pond	knock	dock	vox	box	chop	cop	pot	tot	got	dot

Table 4.1: The DRT words list.

4.3 Diagnostic Acceptability Measure (DAM)

The Diagnostic Acceptability Measure or DAM is a relatively new quality measure, first described by Voiers (1977a). DAM is able to gauge a wide range of speech degradations, and hence is useful for a wide range of speech qualities. It is different from other subjective measures of speech quality in that it incorporates a multidimensional approach to its realisation. It assesses a speech signal on 19 separate scales, all of which have a range of 0 to 100 points. These scales can be divided into three groups: *Signal Quality*, *Background Quality* and *Total Quality*. The three total quality sub-scales I, P and A represent principally the standard direct approach of evaluating speech quality in which the speech is assessed relative to some general probes such as “Intelligibility”,

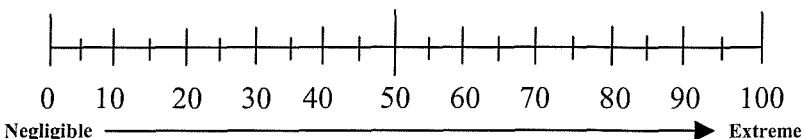
“Pleasantness” or “Acceptability” respectively.

The speech materials used with the DAM consist of four groups of twelve sentences each (see Table 4.2). Each sentence is six syllables in length and has its phonetic makeup controlled in the following manner: it is constrained to have at least one vowel from each of the three categories (Front, Mid, Back) and at least one consonant from each of the three categories (Sibilant, Stop, Fricative). The DAM is typically used to assess speech from six to twenty four communication systems in a single testing session. For each processing system, twelve sentences of speech with total duration of approximately one minute are required. Sentences are presented to a number of listeners at a rate of one sentence every four seconds and systems are presented in random ordering.

The DAM combines a direct (isometric) and an indirect (parametric) approach to the overall speech sample evaluation.

- Isometric: **I**, (Intelligibility), **P** (Pleasantness) and **A** (Acceptability).
- Parametric: Signal and Background quality scales.

The isometric approach requires the listener to provide a direct, subjective assessment of the speech sample in terms of intelligibility, pleasantness and overall acceptability. The parametric approach requires him/her to evaluate the sample with respect to its signal and background perceived qualities. Based on experimental evaluation of a large amount of test speech using a large team of quality gauges, Voiers (1977a) developed a set of response scales, shown on the following DAM listener response sheet (Figure 4.1). Of the response scales, nine are used to rate the speech signal, seven to rate the background, and three to rate the overall quality.



0 10 20 30 40 50 60 70 80 90 100
Negligible → Extreme

Signal Qualities **Score (0-100)**

- Fluttering (e.g. amplitude modulated speech).....
- Muffled (e.g. low-pass speech).....
- Distant (e.g. transmitted speech).....
- Rasping (e.g. peak-clipped speech).....
- Thin (e.g. nasal speech).....
- Unnatural (e.g. mechanical, lifeless).....
- Babbling (e.g. nonsense).....
- Irregular (e.g. fluctuating, imbalanced).....
- Interrupted (e.g. internet-packetised speech).....

Background Qualities

- Hissing (e.g. noise-masked speech).....
- Chirping (e.g. clicking sounds).....
- Roaring (e.g. loud, deafening).....
- Crackling (e.g. scratching sounds).....
- Buzzing (e.g. vibrant, lively).....
- Rumbling (e.g. low-frequency noise).....
- Bubbling (e.g. narrow-band filtered speech).....

Total Effect

- Intelligible (understandable, meaningful).....
- Pleasant (rich, mellow).....
- Acceptable.....

Figure 4.1: The DAM response sheet given to the subjects of the listening experiment.

Group 1	Group 2
I read the news today Wrap the bones in tin foil. He stole all her hats. I hated the poems. You are the biggest man. We all ate fresh mushrooms. The cold pond made me sneeze. Don't permit that cheap talk. The boat sailed off the edge. You kids get off the fence. Dirt was blown in my face. Doubtless he was too thin.	This road forks up ahead Jack was lost in the woods. Do not present the prize. Don't feed that vicious hawk. The trees grew ten feet tall. Snow fell all last winter. Let's sit in the cool bar. This box is much too flat. You should be there on time. The cook can boil beef stew. Look at those peacocks strut. The wind blows hard at sea.
Group 3	Group 4
That hose can wash her feet. The goose laid an odd egg. That quiz was much too hard. Those are pudgy old men. We cracked all those pecans. He had the bluest eyes. Don't meet me first today. These shoes were black and brown. They are too loud in church. We saw a bad movie. I suggest you leave now. The rabbits and dogs drowned.	Go and sit on the bed. The book is about trash. Moths still turned yellow. Fire consumed the paper. The bowl dropped from his hands. Don't thrash around that way. Those children are dirty. Dress sleeves are much too warm. Tractor plowed the fields. We made some fine brownies. They broke out of prison. He drank our lusty brew.

Table 4.2: The four DAM sentence groups.

4.4 Conclusions

When performing a subjective speech quality evaluation it is difficult to decide exactly what quantities to measure when making the quality assessments. Speech intelligibility is indisputably a significant factor in shaping the overall acceptability of voice communication systems, but it is clearly not sufficient in all cases. For example, in high quality systems where easy and efficient communication might be the goal, it is also desired that speech sound natural and that the user be able to recognize the speaker.

This chapter, therefore, contributed towards the understanding of DRT and

DAM, two standardised procedures that test speech intelligibility and quality, respectively. Both of them are used by the U.S Department of Defense for the evaluation of communication systems. This is the reason that the reader may notice in the corpuses of the two tests that some words are not British English, and also that some word pairs that rhyme in American English, are not rhymes in British English. The DRT is also the NATO standard for evaluating speech intelligibility in communication systems. In addition to government agencies, DAM is widely used by many companies involved in speech communication research.

To summarise, the chapter presented the theory behind two standardised subjective listening test, DRT and DAM, used for evaluating the intelligibility and quality of processed speech in communication systems. Details concerning the scoring and the design of the two tests were also set forth, since they will be extensively used in Chapters 7 and 9.

Chapter 5

An Optimal Filtering Approach for Speech Vibration Analysis

5.1 Introduction

Added noise and imperfect transmission process can degrade the intelligibility or quality of speech. The enhancement of degraded speech is therefore an important problem with numerous applications ranging from suppression of environmental noise for communications systems and hearing aids, to preprocessing for speech recognition systems. In our study, the accelerometer signal is immune to noise but it has poor speech quality due to body-conduction losses, in contrast to the microphone which gives a good quality signal, but is susceptible to acoustic noise degradation. The objective is, therefore, the investigation and application of an optimal filter which will compensate for the loss of higher frequencies due to body-conduction attenuation, while maintaining the small amount of acoustic noise. The choice of the Wiener filter, as it is also known, can be justified from the fact that Wiener estimate is optimal. Therefore, no other enhancement method can yield better results in

a mean-square error sense. It uses the microphone as the desired signal, and the accelerometer as the input to the filter. The filter then tries to match the input to the desired signal in a minimum mean-square error approach. This helps in improving the intelligibility and/or quality of the accelerometer signal, something that is verified with listening tests in Chapters 6 and 7.

The aim of this chapter, therefore, is to present the theory of optimal filtering, which is the main signal processing technique that is investigated in Chapters 6 to 9, and describe its application to speech vibration signals. The chapter starts with the theory of optimal filtering. It then progresses to the modelling and derivation of an optimal filter for speech vibrations, by using a block diagram which encompasses the various transfer functions that affect the reception of self-generated speech, air- and body-conducted. The motivation for this is the need to have a general model that will help in the interpretation of the experimental results in subsequent chapters.

5.2 Optimal Filters in the Time Domain

Filters are commonly used to extract a desired signal from a background of random noise or deterministic interference. The aim of this section is the presentation of the time-domain optimal filter theory, that will be used as our proposed speech enhancement strategy to accelerometric speech, in Chapters 6 to 9.

The optimal filters are designed to minimise the mean-square error between their output and a desired signal. Thus, they are said to be optimum in a mean-square error (MSE) sense. This particular definition of optimality is convenient because it leads to closed form solutions for the filter coefficients in terms of the autocorrelation of the signal at the input to the filter, and

the cross-correlation between the input and the desired signal. Figure 5.1 illustrates the Wiener filter configuration.

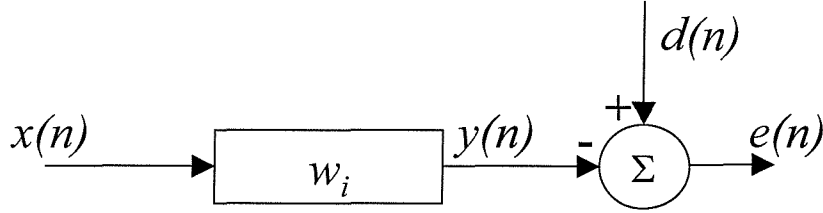


Figure 5.1: The main blocks of an FIR optimal filter.

In this figure the error signal $e(n)$ is given by the difference between the desired signal $d(n)$ and the input signal $x(n)$ filtered by an FIR filter with coefficients w_i so that

$$e(n) = d(n) - \sum_{i=0}^{I-1} w_i x(n-i) \quad (5.1)$$

The summation over $w_i x(n-i)$ in equation 5.1 can be conveniently represented as a vector inner product, such that

$$e(n) = d(n) - \mathbf{w}^T \mathbf{x}(n) = d(n) - \mathbf{x}^T(n) \mathbf{w} \quad (5.2)$$

where

$$\mathbf{w} = [w_0 \ w_1 \ \dots \ w_{I-1}]^T, \quad (5.3)$$

$$\mathbf{x}(n) = [x(n) \ x(n-1) \ \dots \ x(n-I+1)]^T \quad (5.4)$$

and the superscript T denotes the transpose of the vectors, which are assumed to be column vectors.

The objective is to find the values of each of the filter coefficients $w_0 \dots w_{I-1}$ that minimise the quadratic *cost function* given by the mean square error (MSE),

$$J = E[e^2(n)], \quad (5.5)$$

where E denotes the *expectation operator*. If $x(n)$ and $d(n)$ are not stationary, then J and the optimal filter coefficients will be functions of time. We assume

here that all the signals are stationary, so that the expectation is time invariant, and can be calculated by averaging over time. The cost function given by equation 5.5 is thus equal to the average mean-square value of the error signal.

Using equation 5.2 the cost function can be written as (Elliott, 2001):

$$J = \mathbf{w}^T \mathbf{A} \mathbf{w} - 2\mathbf{w}^T \mathbf{b} + c, \quad (5.6)$$

where

$$\mathbf{A} = E[\mathbf{x}(n)\mathbf{x}^T(n)], \quad (5.7)$$

$$\mathbf{b} = E[\mathbf{x}(n)d(n)], \quad (5.8)$$

$$c = E[d^2(n)] \quad (5.9)$$

In a quadratic equation having the general form of equation 5.6, the matrix \mathbf{A} is known as the *Hessian matrix*, and in this case its elements are equal to the values of the autocorrelation function of the input signal,

$$\mathbf{A} = \begin{bmatrix} R_{xx}(0) & R_{xx}(1) & \dots & R_{xx}(I-1) \\ R_{xx}(1) & R_{xx}(0) & & \\ \vdots & & \ddots & \\ \vdots & & & \ddots \\ R_{xx}(I-1) & & & R_{xx}(0) \end{bmatrix} \quad (5.10)$$

where $\mathbf{R}_{xx}(m)$ is the symmetric *autocorrelation* function of $x(n)$, defined for the entirely real time sequences assumed here to be

$$\mathbf{R}_{xx}(m) = E[x(n)x(n+m)] = \mathbf{R}_{xx}(-m) \quad (5.11)$$

The Hessian matrix is not necessarily equal to this matrix of autocorrelation functions, and in order to be consistent with the discussion below, the Hessian matrix is written here as \mathbf{A} rather than the more specific form \mathbf{R} , which is widely used for FIR filters in the signal processing literature. An example of this more general form for \mathbf{A} is when the cost function to be minimised

includes a term proportional to the sum of the squared filter weights, $\mathbf{w}^T \mathbf{w}$, so that

$$J = E[e^2(n)] + \beta \mathbf{w}^T \mathbf{w} \quad (5.12)$$

where β is a positive, real *coefficient-weighting* parameter. This cost function can also be written as a quadratic equation of the form of equation 5.6, but now the Hessian matrix has the form

$$\mathbf{A} = \mathbf{R} + \beta \mathbf{I} \quad (5.13)$$

where \mathbf{R} is the autocorrelation matrix given by the right-hand side of equation 5.10 and \mathbf{I} is the identity matrix.

The vector \mathbf{b} in equation 5.8 has elements that are equal to the values of the *cross-correlation function* between the input and the desired signals so that

$$\mathbf{b} = [R_{xd}(0) \ R_{xd}(1) \ \dots \ R_{xd}(I-1)]^T \quad (5.14)$$

where for the entirely real and stationary time sequences assumed here

$$R_{xd}(m) = E[x(n)d(n+m)] = E[x(n-m)d(n)] \quad (5.15)$$

Finally, c is a scalar constant equal to the mean-square value of the desired signal.

5.2.1 The Wiener Filter

The value of the coefficients of the FIR filter that reduces the mean-square error to a minimum can be found by differentiating the cost function with respect to each coefficient and setting all of the resulting derivatives to zero. It is convenient to express this differentiation using vector notation, and so we define the vector of derivatives to be

$$\frac{\partial J}{\partial \mathbf{w}} = \left[\frac{\partial J}{\partial w_0} \quad \frac{\partial J}{\partial w_1} \quad \dots \quad \frac{\partial J}{\partial w_{I-1}} \right]^T \quad (5.16)$$

Using the definition of J in equation 5.6 and the properties of derivatives of vectors, equation 5.16 can be expressed as

$$\frac{\partial J}{\partial \mathbf{w}} = 2[\mathbf{A}\mathbf{w} - \mathbf{b}] \quad (5.17)$$

The vector of optimal filter coefficients can be obtained by setting each element of equation 5.17 to zero, giving

$$\mathbf{w}_{opt} = \mathbf{A}^{-1}\mathbf{b} \quad (5.18)$$

The filter that has these optimal coefficients is often called the Wiener filter.

5.3 Modelling of the Optimal Filter

As was seen in Chapter 2, a speaker produces a speech signal in the form of acoustic and vibration waves that travel around the speaker's body as well as from the speaker's mouth to his ears. The air-conducted signal consists of variations in pressure as a function of time and is usually measured directly in front of the mouth, the primary sound source (although sound also comes from the nostrils, cheeks and throat). In addition to that, the body-conducted signal consists of vibrations around the speaker's body, which are difficult to quantify, and can only be described by making assumptions about the medium in which they travel (Tonndorf, 1972). The speech signal is non-stationary, with changing characteristics as the vocal tract changes shape and the source characteristics change (O'Shaughnessy, 2000). However, the long-term average speech spectrum can be assumed stationary, something that supports the design and modelling of the optimal filter for the enhancement of speech vibrations that is used in this study.

Hence, the optimal filtering problem that we try to fit to this case can be modelled as the mixed analogue-digital block diagram shown in Figure 5.2

where the two types of speech signal are picked up by an equivalent transducer, i.e. a microphone for the air-conducted signal and an accelerometer for the body-conducted signal. The microphone is placed at a close distance from the speaker's mouth, while the accelerometer is attached on a position on the speaker's head where speech vibrations are conducted sufficiently well. For this section, the blocks will be assumed linear and time-invariant.

The motivation for the development of this model is the need to have a tool that will aid the analysis of the optimal filtering performance, and the interpretation of the filters' responses (see Chapter 8). What we try to achieve, therefore, is to derive a final optimal filter expression that will potentially show the dependence of the optimal filter on certain factors. This, not only will help in explaining certain discrepancies in the filters' performance, but will also aid the proposition of a speech enhancement solution for the front-end user of a communication system.

The model links the electroacoustic with the physiological aspects of our experimental setup, as it is presented in Chapter 6. Therefore, Figure 5.2 is an expanded version of Figure 5.1 to include the effects of the body-conduction and the air-conduction path from the speech source $s(t)$ to the accelerometer and microphone, respectively. The signal from the microphone is used as the desired signal as it is of higher quality than that of the accelerometer, which is the input to the filter.

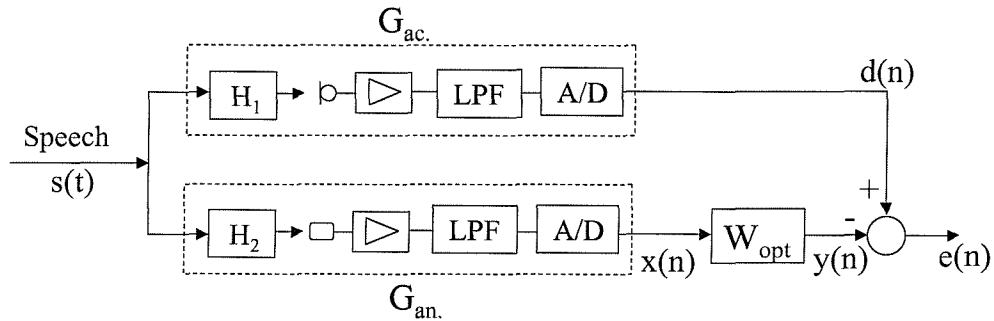


Figure 5.2: Mixed analogue-digital physical model of the optimal filtering problem.

The multi-source speech input $s(t)$ is influenced by the acoustic and anatomical transfer functions $G_{ac.}$ and $G_{an.}$, respectively. The air-conducted signal is filtered by $G_{ac.}$, forming the desired signal $d(n)$. On the other hand, the body-conducted signal is filtered by $G_{an.}$ and constitutes the input to the optimal filter W_{opt} , which tries to match $x(n)$ to $d(n)$.

However, $G_{ac.}$ includes various individual transfer functions. H_1 is the acoustic transfer function from the various speech sources to the microphone, and includes the vocal tract. $G_{ac.}$ also includes the acoustic transfer function from the speaker's lips and nose to the microphone, the electroacoustic transfer function of the microphone, the electrical transfer function of the microphone's amplifier, the electrical transfer function of the anti-aliasing filter (appearing as LPF), and that of the A/D conversion process.

$G_{an.}$, on the other hand, includes H_2 , the body-conduction transfer function from any possible speech vibration generation point on the speaker's head to the accelerometer. $G_{an.}$ also includes the electroacoustic transfer function of the miniature accelerometer, the electrical transfer function of its amplifier, the electrical transfer function of this channel's anti-aliasing filter (also appearing as LPF), and that of the A/D conversion.

The optimal filter W_{opt} can be written in terms of the power spectra of $d(n)$ and $x(n)$ (Elliott, 2001) as:

$$W_{opt} = \frac{S_{xd}}{S_{xx}} \quad (5.19)$$

where

$$S_{xd} = G_{ac.} S_{xs} = G_{ac.} G_{an.}^* S_{ss} \quad (5.20)$$

$$S_{xx} = |G_{an.}|^2 S_{ss} \quad (5.21)$$

Substituting 5.20 and 5.21 in 5.19, we see that the optimal filter becomes equal to:

$$W_{opt} = \frac{G_{ac.}}{G_{an.}} \quad (5.22)$$

An equivalent all-digital approach to Figure 5.2 is the one in Figure 5.3.

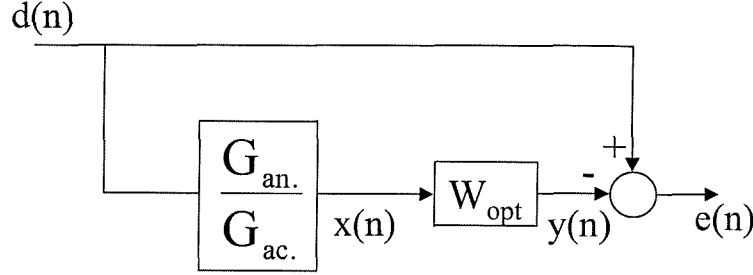


Figure 5.3: All-digital physical model of the optimal filtering problem.

We treat the air-conducted signal $d(n)$ as the only reliably measured quantity, since we cannot actually measure $s(t)$, and working backwards we end with the accelerometer signal $x(n)$ which is related to the air-conducted signal filtered by the transfer functions of the two paths combined together. This approach makes the derivation of the optimal filter more straightforward, since:

$$S_{xd} = \left(\frac{G_{an.}}{G_{ac.}} \right)^* S_{dd} \quad (5.23)$$

$$S_{xx} = \left| \frac{G_{an.}}{G_{ac.}} \right|^2 S_{dd} \quad (5.24)$$

Therefore, the optimal filter W_{opt} is equal to:

$$W_{opt} = \frac{S_{xd}}{S_{xx}} = \frac{1}{\frac{G_{an.}}{G_{ac.}}} = \frac{G_{ac.}}{G_{an.}} \quad (5.25)$$

i.e. W_{opt} is the same as the one derived in equation 5.22, indicating that Figure 5.3 is indeed equivalent to Figure 5.2. Equation 5.25 reveals that the optimal filter that we try to design for our application, is dependent on the two transfer paths. From that we can infer that the response of the filter will not depend significantly on the frequency characteristics of the speech input, but instead, on the anatomical characteristics of the speaker that define $G_{an.}$, and the placement of the reference microphone that defines $G_{ac.}$.

An approximate estimation of the filtering effect imposed by $\left| \frac{G_{an.}}{G_{ac.}} \right|^2$ was calculated as the ratio of the magnitudes of $\frac{S_{dd}}{S_{xx}}$ based on Figure 5.3. The calculation was performed for the male and the female subjects, using average spectra of

the DAM sentences list (see Chapter 6 for the recording procedure). The resulting filtering magnitude appears in Figure 5.4.

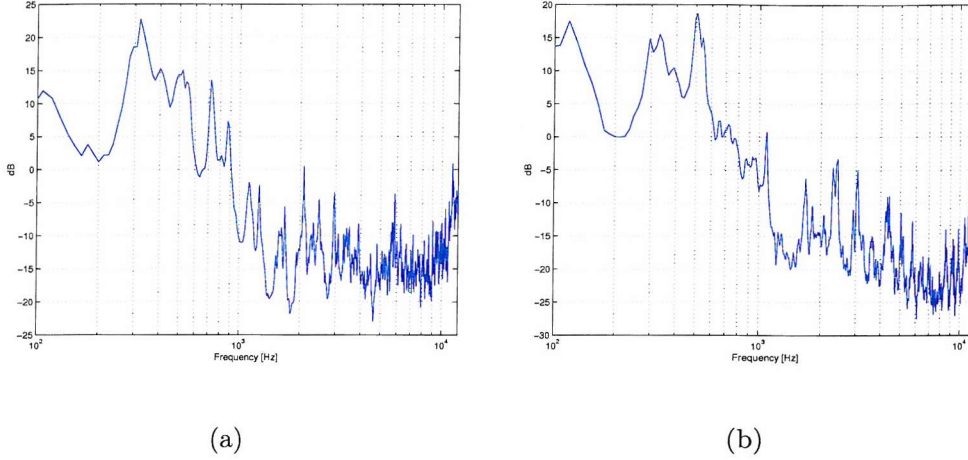


Figure 5.4: Magnitude spectrum of $|\frac{G_{an.}}{G_{ac.}}|^2$ for (a) male, and (b) female speakers.

This filtering effect, which also includes the low-pass filtering $G_{an.}$ due to body-conduction, appears to have a main cut-off at about 500-600 Hz for both male and female speaker, while the roll-off rate is approximately 18-20 dB/oct. Since $W_{opt} = \frac{1}{|\frac{G_{an.}}{G_{ac.}}|}$, it is expected to have the form of a high-pass filter in this frequency range.

5.4 Conclusions

This chapter has contributed to a better understanding of the theory of the FIR Wiener filter that has been adopted as our main enhancement strategy for the intelligibility and quality improvement of the recorded accelerometric signals.

Optimal filter theory was adapted to our case of detecting a speech vibration signal, which the filter tries to match to an air-conducted speech signal. This diagramatic approach proved to be useful as it facilitated the derivation of the optimal filter by lumping together the two main speech transfer pathways that

we are investigating, and which can be evaluated with power-spectra estimations. This approach helped the interpretation of the optimal filter responses, presented in Chapter 8, where the diagram is expanded to include the effect of electronic noise of the accelerometer. However, the main distinction that has to be made is that section 5.3 was used to understand the modelling of the system, since $W_{opt} = \frac{G_{ac.}}{G_{an.}}$ is not restricted to be causal, thus is not always implementable. In practice, W_{opt} was computed as in section 5.2.1 to ensure causality.

The main conclusion, therefore, is the derivation of an optimal filter expression which showed that the filter that we try to design for our application is not significantly dependent on the frequency content of the speech input, but instead, on the anatomical characteristics of a speaker and the positioning of the transducers, that define $G_{ac.}$ and $G_{an.}$. This poses the question whether an optimal filter which is designed for specific phonetic material performs better or equally well to a fixed optimal filter that is designed for a speaker, something that is answered in Chapter 9.

This chapter concludes the theory sections of this thesis. We have presented the theory of optimal (Wiener) filtering that has been adopted as our main enhancement strategy. The chapter also presented a block diagram which linked the physiological principles of speech vibration generation, transmission and detection, with the design of an optimal filter focused on the problem of enhancing them.

Chapter 6

Development of the Experimental System

6.1 Introduction

The aim of this chapter is to present the development of a vibration-based recording system that was used throughout the experimental sessions. The chapter starts with a description of the setup used in the initial experiment. This includes a presentation of the acoustic and vibration transducers, the amplification and filtering devices, as well as the signal acquisition and generation devices. It is then followed by a discussion on the electronic noise-floor measurements of the above equipment. The next section presents the first attempt of speech vibration recordings on three anatomical locations in order to see which one is the most appropriate for detecting speech. The chapter then proceeds to portray the initial speech enhancement strategies that were adopted (high-pass, band-pass and optimal filtering), some indicative results, and their assessment through an informal listening test. The chapter ends with a presentation of the experimental setup used in the later experiments,

the changes in the equipment, and the motivation behind its re-design.

(The work included in this chapter was presented as a poster at the 141st meeting of the Acoustical Society of America, June 2001).

6.2 Initial Experiment: Aims for the Development of the Experimental Setup

The experimental setup described in this chapter was aimed at guaranteeing a viable, practical, and stable measuring system for speech vibration recordings. This was accomplished first by investigating the electronic noise-floor of the equipment to ensure that there would be no significant interference with the speech recordings. It was further approached by investigating the best contact point for speech vibration detection. For this reason, selected speech material that was representative of different classes of speech was recorded from three anatomical locations in order to decide which one had the best speech vibration conduction properties. The subjective comparison among various high-pass filters and a band-pass filter with optimal filtering applied on the speech vibration signals was the last step in order to see which type of enhancement gave promising results for its use in a larger-scale experiment. The above steps contributed towards a better understanding of the equipment, and the design of a reliable and functional recording system that could be easily set up and used in the future.

6.2.1 Place of Experiments

The first experiments were conducted in the laboratory of the Signal Processing and Control Group at the Institute of Sound and Vibration Research.

This laboratory is dedicated to applications of signal processing on sound or vibration. It is a space equipped with various measuring equipment, filters, data-acquisition devices and computers, making it an all-purpose lab for sound and vibration applications. It is a sufficiently quiet space for speech recording, although it cannot be considered perfect, as there is a background noise level from the different pieces of equipment when they are switched on. A typical background noise level spectrum of the lab, recorded with our reference microphone, which is described below, can be seen in Figure 6.1. The spike at about 1.5 kHz can be attributed to the cooling fan from PC's.

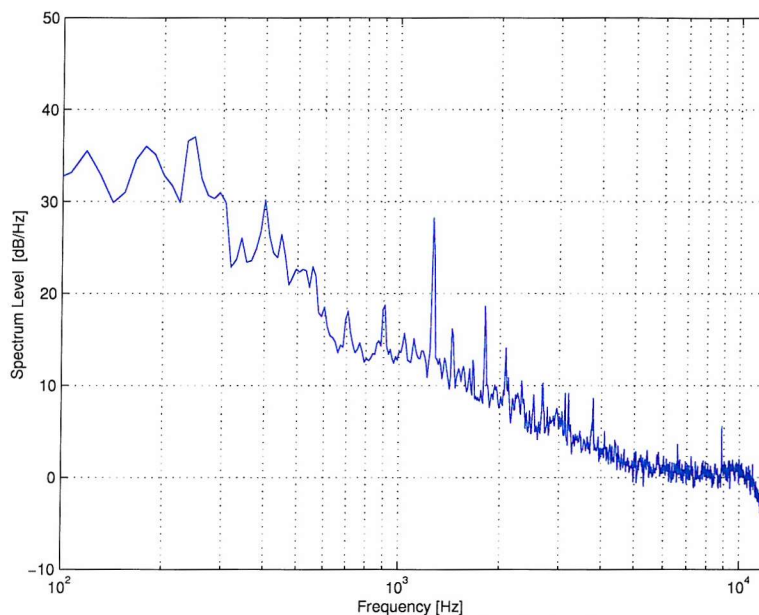


Figure 6.1: Background noise level spectrum in the lab where the experiments were conducted.

6.2.2 Experimental Equipment

The equipment that was used throughout our experimental sessions can be divided into five main categories: acoustic transducers, vibration transducers, amplification equipment, acquisition devices and finally filtering and analysis equipment.

Acoustic Transducers

The ideal size of microphone for our application would be a miniature one, since it would be possible to place it at the aural area easily. For our experiments we used the standard EM-3046, manufactured by Knowles, a model that is broadly used in hearing aid applications. It is a millimetre-scale, condenser microphone with a robust construction to withstand severe environmental conditions. In addition, we also used a 1/2" condenser microphone, B&K 4133, as reference microphone. This served later as the desired signal of the optimal filter.

Vibration Transducers

The miniature accelerometer (model BU-1771) was also obtained from Knowles. This is a low-mass accelerometer containing a ceramic vibration transducer with high sensitivity and a flat response up to 3 kHz, above which it rises sharply. The in-built FET amplifier and high mechanical shock resistance also make it suitable for severe environmental conditions. The presence of this FET amplifier makes the use of a simple voltage amplifier possible. Hence, the use of a charge amplifier, which is more commonly used with piezoelectric accelerometers, is not essential.

Filtering Equipment

KEMO anti-aliasing filters (model VBF8) were used prior to the A/D conversion for every channel, with cut-off frequencies set at 11 kHz. Each filter rack has two channels with a switch to adjust for parallel, series or separate modular use. In our case, since we had three independent channels, we used the modules separately. The filter chosen was an 8-pole, 6-zero Butterworth filter (approx. 48 dB/oct). These modules are also provided with low-noise in-

built amplifiers with 0, 10 or 20 dB additional amplification for every channel, which could be selected with a switch.

Signal Acquisition and Generation Devices

For the acquisition of signals, we used a National Instruments data-acquisition card (type PCI-6035E), fitted in a standard desktop PC (PII, 800 MHz, 128MB RAM) along with the appropriate interface (National Instruments BNC-2090) where all the transducers could be attached with BNC connectors. This specific card has 16 analog inputs, 2 analog outputs and 16-bit resolution. Its maximum sampling rate for one channel is 200 kHz and is compatible with the Matlab software, hence no hardware-mismatch problems were encountered. During all measurements, a sampling frequency of $f_s=24$ kHz was used, since it was assumed that speech contains negligible energy above $f_s/2=12$ kHz (Meakawa and Lord, 1994). For the measurement of the amplifiers' responses, we used a dual-channel FFT spectrum analyser (Advantest R9211C).

Amplification Equipment

During all measurements, the amplification for each transducer was set to the appropriate level so that the acquired signals did not distort, while we could get the best signal/noise ratio. For the miniature microphone and accelerometer signals, two individual amplifiers were used, built in ISVR. These have 3 amplification levels at 20 dB steps (20, 40 and 60 dB) and each one accepts a single input. The response of the amplifiers was measured with the FFT spectrum analyser, which served as a signal generator, by using a broad-band input at 0.5V peak. During all our experiments, only the 40 dB setting was used. The amplifier's amplitude and phase response appears in Figure 6.2.

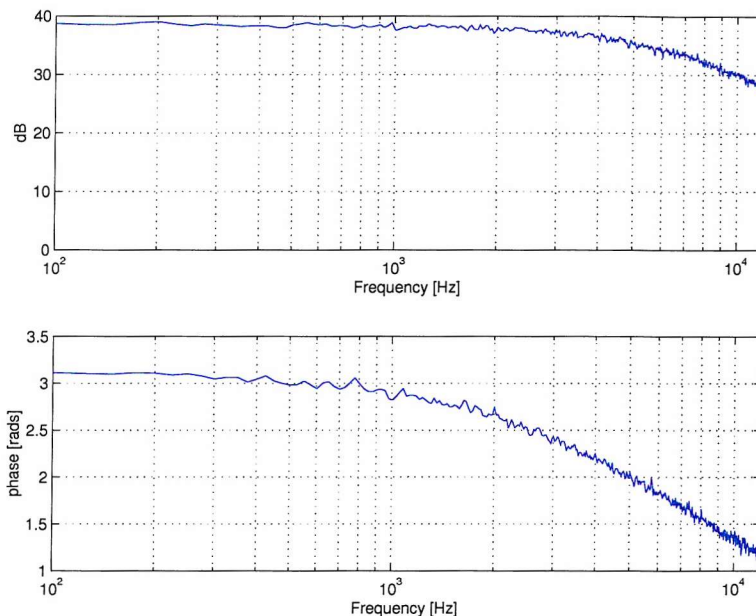


Figure 6.2: Amplitude and phase response of the accelerometer's amplifier at 40 dB setting.

For the B&K 4133 reference microphone, we used the standard B&K measuring amplifier type 2609, with the capsule pre-amplifier, type 2619.

6.2.3 Noise-Floor Measurements

After setting up our equipment in the laboratory, it was decided to check whether there would be any electrical noise interference with the acquired speech signals. Several tests were done involving the acquisition of 5s signals with all transducers in the laboratory, while switching electrical equipment on and off, but none seemed to interfere drastically with the acquired signals. The custom-made amplifiers for the miniature transducers needed to be switched on for about 10min prior to the experiment so that consistent measurements could be taken. A suggested reason for this is that charges in the circuitry may take some time to settle. A schematic diagram for the noise-floor measurements can be seen in Figure 6.3.

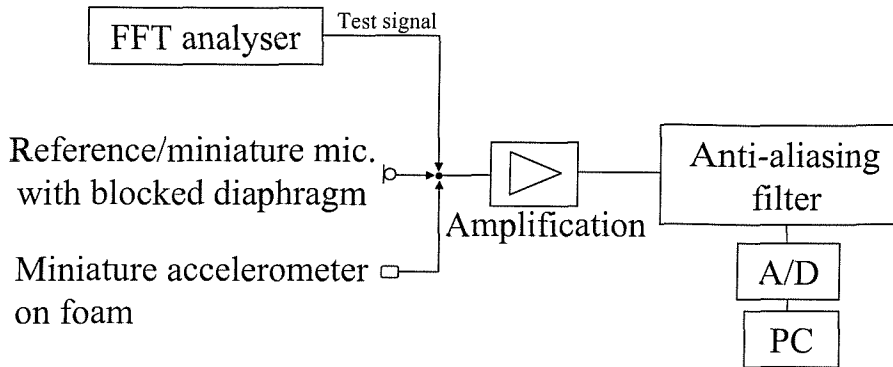


Figure 6.3: Experimental setup for the measurement of equipment noise-floors.

Noise-Floor of the Data-Acquisition (DAQ) System

A starting requirement was to check the electronic noise-floor of the DAQ system on its own, and in conjunction with the other pieces of equipment. A way to check this is to feed it with a 0V signal in order to see the inherent noise disturbances. The reason for using such a signal is that if the DAQ system is fully disconnected from the other equipment, the system may suffer from impedance mismatches giving unreasonable results. Hence, a 0V input signal was fed to the DAQ system from the FFT analyser source and a 5s sample was acquired at a sampling frequency of 24 kHz. Immediately after, a pure tone at 1 kHz and 2.5V amplitude was acquired for the same duration at the same sampling frequency.

The spectra of the 0V signal, corresponding to noise, and the 2.5V sine wave, corresponding to the signal, appear in Figure 6.4, where the SNR was calculated theoretically according to Watkinson (1995) as equal to 98.1 dB, based on the approximate 6 dB/bit-rule. However, this result is applicable for a full-scale input signal, so in order to compare the measured value to the theoretical value, we have to subtract 6 dB from the theoretical value, as we are using half of the maximum peak-peak voltage that the DAQ card can handle ($\pm 5V$ p-p). The fluctuations of both the signal and noise signals, however, indicate that there is additional electrical interference, most probably from the

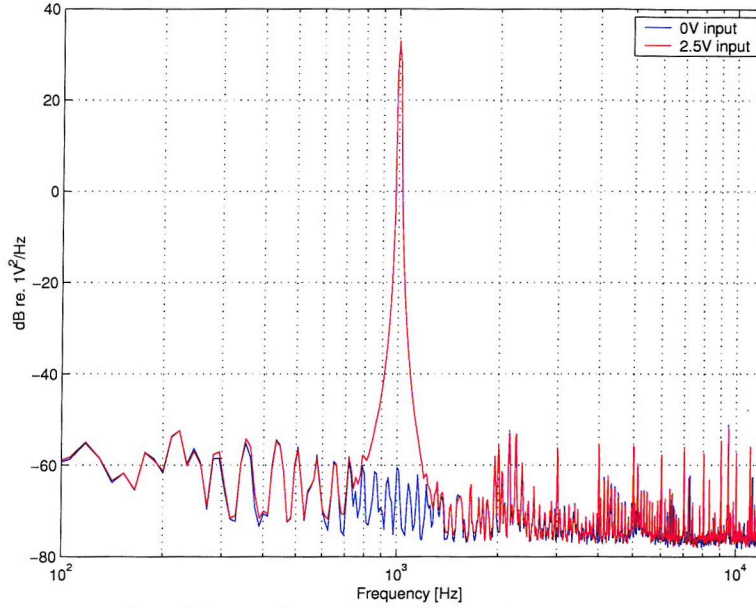


Figure 6.4: Noise-floor of the Data-Acquisition system.

PC-amplifier-filter interaction, giving another 20 dB approximately of noise, which have to be subtracted from the theoretical calculation. Therefore, the measured SNR was calculated in Matlab as follows:

$$SNR = \frac{\text{Power of Signal}}{\text{Power of Noise}} - 6dB - 20dB \quad (6.1)$$

This calculation gave a SNR value of 72.2 dB. This coincides with the theoretical calculation if we subtract the 26 dB as in equation 6.1 and, nevertheless, shows that our system has low inherent noise-floor and that any contributions other than Analogue-to-Digital Conversion (ADC) noise, are not significant.

Noise-Floors of the Microphones and Accelerometer

No intentional noise sources were present except background noise. The amplification settings were the same as in the experimental recording session. The reference and miniature microphones were isolated from any acoustic interference by blocking their diaphragms firmly with compressed acoustic foam, and 5s of data were acquired. The accelerometer was placed on a piece of foam so that any possible structure-borne vibrations from the laboratory would be

minimised, and again 5s of data were recorded (see Figure 6.3). A typical spectrum of speech recorded with the same procedure as described in section 6.2.4, (the sentence “*Sunlight strikes raindrops in the air*”), is plotted along with the electronic noise-floor spectra from all three transducers in Figure 6.5.

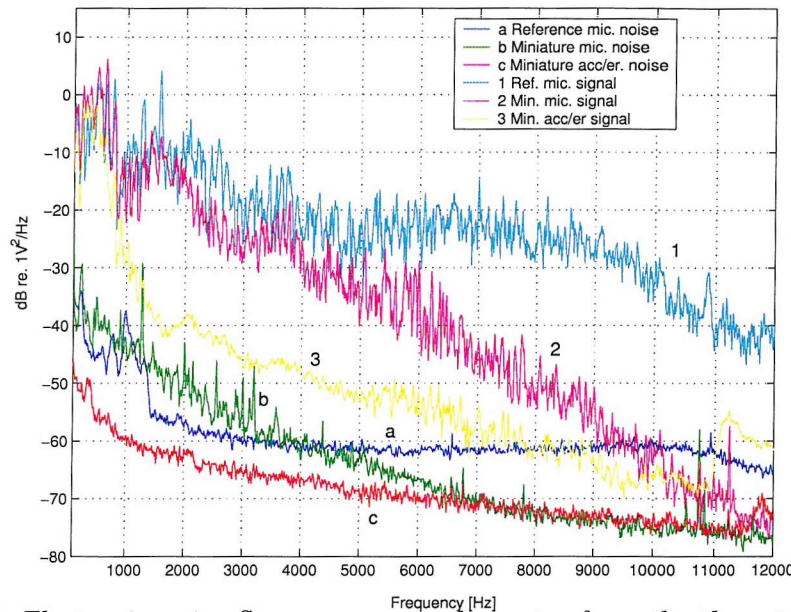


Figure 6.5: Electronic noise-floor versus speech spectra from the three transducers.

It can be seen that the noise-floor curves converge at high frequencies. This can be explained by the attenuation imposed by the anti-aliasing filter cut-off at 11 kHz. Overall, the electronic noise-floor levels of the transducers were lower than the recorded signals, thus giving us a satisfactory signal-to-noise ratio (36.3 dB for the reference microphone; 26.4 dB for the miniature microphone; 10.2 dB for the miniature accelerometer, calculated as a ratio between the power of the speech over the power of the noise signal). Figure 6.5 also shows the low-pass filtering effect of speech directivity (see section 2.4.2), with the miniature microphone spectrum being about 5-10 dB lower than the reference microphone spectrum in the range 1-4 kHz, while above 4 kHz this difference increases to more than 25 dB.

6.2.4 Speech Recordings

For the first set of speech recordings, the material that was used consisted of single, sustained phonemes (sustained for 2s) and phonetically balanced sentences. The phonemes were vowels (/a, o, i/), nasals (/m/), and voiced and unvoiced fricatives (/s, f, v/). The sentences were the following:

- “*High jets whiz past screaming*”, which contains a mixture of vowels, and consonant clusters.
- “*The sunlight strikes raindrops in the air*”, a phrase from the “Rainbow passage” which is often used for psychoacoustic tests. It contains a balanced quantity of phonemes, i.e. stops, nasals, voiced and unvoiced fricatives, vowels, etc.
- “*We were away a year ago*”, which contains entirely voiced phonemes, no nasals or fricatives, and,
- “*Should we chase those young outlaw cowboys?*”, which contains a mixture of fricatives (voiced and unvoiced), diphthongs, and affricates.

A data-acquisition program was written, using the Data Acquisition Toolbox for Matlab. It employed three channels (as many as our transducers), each acquiring data at 24 kHz. For the acquisition of phonemes, a duration of 2s was used, while for the sentences the duration was increased to 3s. Those times were considered large enough to capture all the phonemic information of phonemes or phrases respectively.

One Greek, male subject was used. The recording session began with the subject getting accustomed to the positioning of the miniature accelerometer on his face. For this reason, adhesive stickers, used to attach electrodes on the skin, were used for the miniature accelerometer and the miniature microphone.

It was made sure that the contact for the miniature accelerometer was the best possible, without much force exerted on the surface of the skin, since that would distort the results. Three contact points were investigated: the larynx, the skin surface over the TMJ, and similarly over the mastoid bone. Our motivation for using only those locations was based on the results of Moser and Oyer (1958), who showed that these three locations were the most conductive for speech vibrations. However, in the context of our study, we had to decide on a location that would be vibrationally conductive, close to the ear (as the miniature microphone is placed there as well), and applicable for an ergonomic design. The miniature microphone was attached at the opening of the ear-canal, making sure that it was not blocking it. The experimental setup which follows the structure of the model in Figure 5.2 can be seen in Figure 6.6 below.

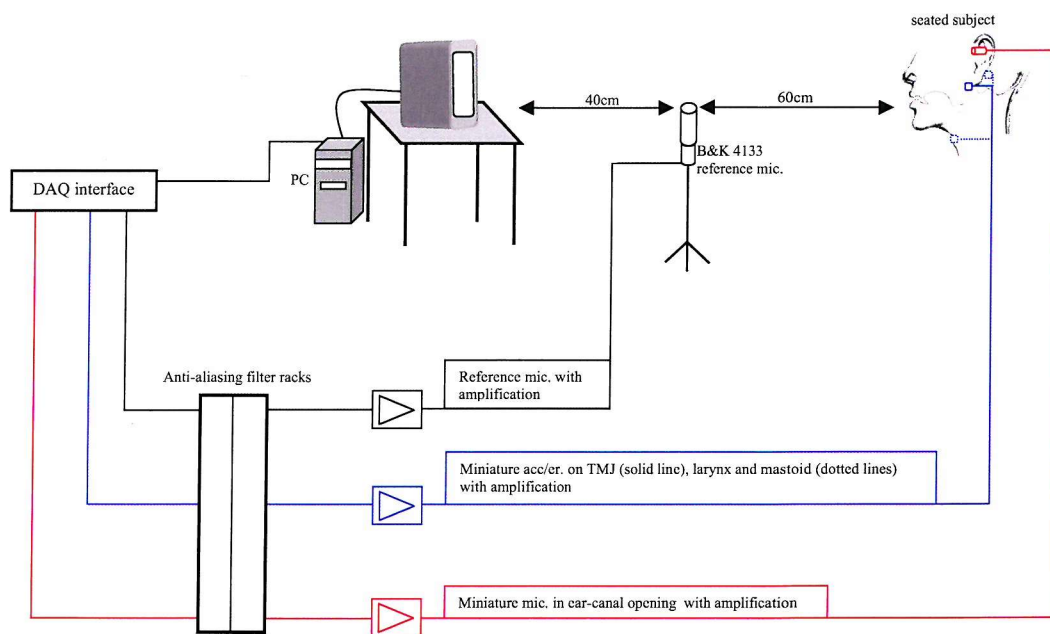


Figure 6.6: *Experimental setup for preliminary experiments.*

The subject had control over the program, so when he was ready to cite the material, he would press a button on the keyboard resting on his lap, and the acquisition would begin for all three channels simultaneously. The recorded data were stored in the computer as Matlab files ready for processing.

6.2.5 Investigation of the Best Conduction Point

The spectra from four characteristic sustained phonemes recorded at the three positions stated above can be seen in Figure 6.7.

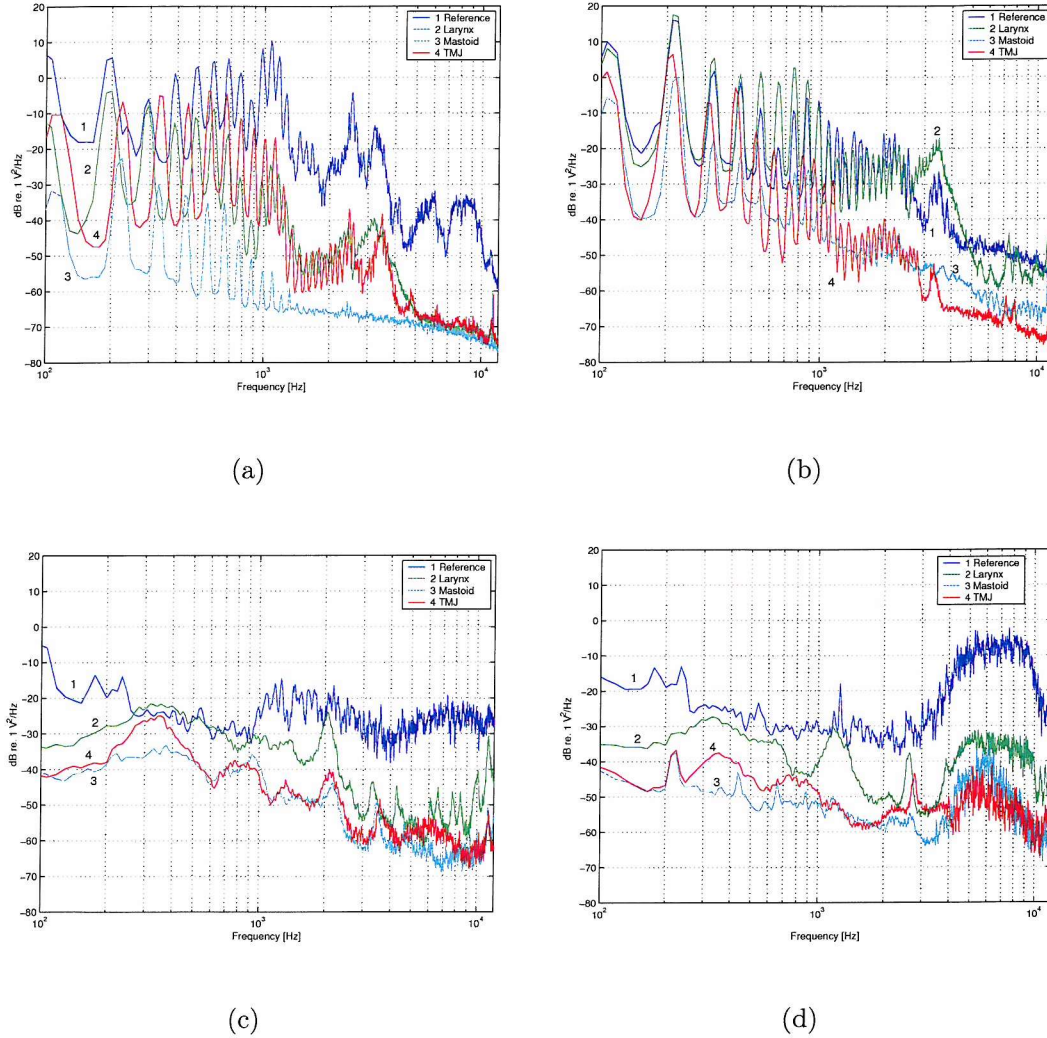


Figure 6.7: Speech spectra recorded by the reference microphone and the accelerometer located at 3 anatomical positions. (a) phoneme /a/, (b) phoneme /m/, (c) phoneme /f/, (d) phoneme /s/.

We note that for vowels (only /a/ shown here) and nasals (similarly, only /m/) the peaks for the reference microphone and larynx positions do not line up with those for the mastoid and TMJ positions because they were not recorded simultaneously, so the actual speech signal was different. Also, spectral amplitude decreases noticeably above 1 kHz for all accelerometer positions. Moreover,

the accelerometer located on the mastoid delivers the lowest amplitude signal in most cases, except for the /a/ where at low frequencies (up to 800 Hz), it is a better transmission position than the TMJ. The accelerometer signal from the larynx also delivers a strong signal, but is not as practical for the user as the mastoid or the TMJ. This can be understood in the context of a final design, in which the user would most likely prefer a small device that could be positioned easily around the aural area, without any encumbrance in other parts of his/her head. Hence, the temporo-mandibular joint (TMJ) appears to be the best accelerometer position for this system. The recording of the sentences, therefore, that were recorded in subsequent recording sessions were done with all the transducers in the same place, except the accelerometer which was placed only on the TMJ.

6.2.6 Preliminary Speech Enhancement and Subjective Evaluation

The aim of this section is to compare the effects of high-pass, band-pass and optimal filtering on speech vibration signals. This is done through the design of the filters, the recording and processing of speech, and the conduction of a listening test. Results are then used in the design of an improved experimental setup and analysis methods described in the next section.

High-pass and Shaping Filters

In his study, Ono (1977) investigated the use of a high-pass filter with different cut-off frequencies (200, 500, 1000 and 1500 Hz) for the enhancement of speech picked up by an accelerometer. In his paper, he did not give any specific details about the construction of the filter. We therefore built a 2nd order Butterworth high-pass filter with Matlab, in which we incorporated one more

cut-off frequency (800 Hz), based on the results of Giua (1998), who found that the characteristic low-pass filter of body-conduction has an approximate cut-off frequency at 800 Hz. The other filter we built was a shaping filter (band-pass), based on the study of Viswanathan et al. (1985). This has a sharp high-pass at about 800 Hz, a 5 dB/oct boost over the range 800-2800 Hz, a flat response over 2.8-4.7 kHz and a sharp low-pass at 4.7 kHz. The reason we designed the above filters was to use a reference point from the literature (see section 3.2.1), with which the optimal filter could be compared as far as its enhancement properties is concerned. The two filter responses appear in Figure 6.8. Figure 6.9 shows the block diagram for the high-pass and band-pass filtering of the accelerometer signals.

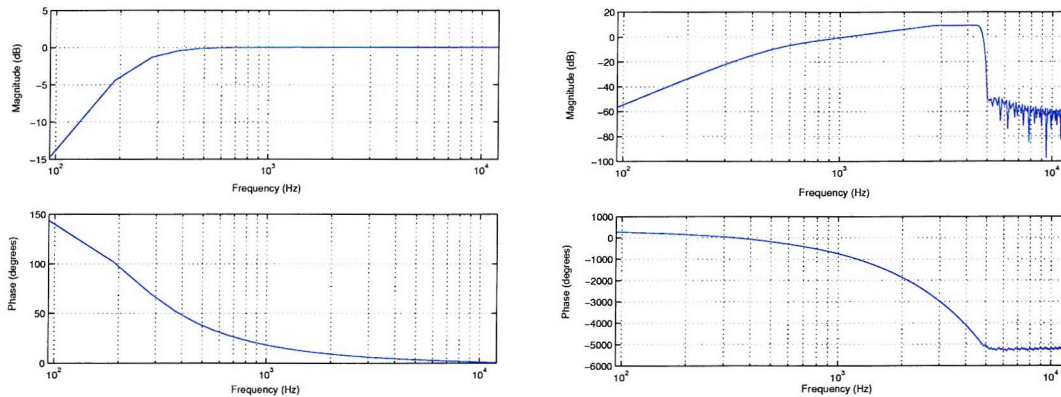


Figure 6.8: Frequency and phase responses of the high pass filter (left) and band-pass filter (right).



Figure 6.9: Block diagram showing the high-pass or band-pass filtering of the accelerometer signal.

Optimal (Wiener) Filter

Based on the theory of section 5.2.1, a Matlab program was written which used the reference B&K microphone as the desired signal, and the accelerometer signal as the input. Six hundred taps were used and a different W_{opt}

was derived for every phoneme and every sentence, in order to investigate its enhancement characteristics for inputs with specific phonetic properties. The filter design stage can be understood with the aid of Figure 6.10. Figure 6.11 shows the block diagram for the filtering of the accelerometer signals.

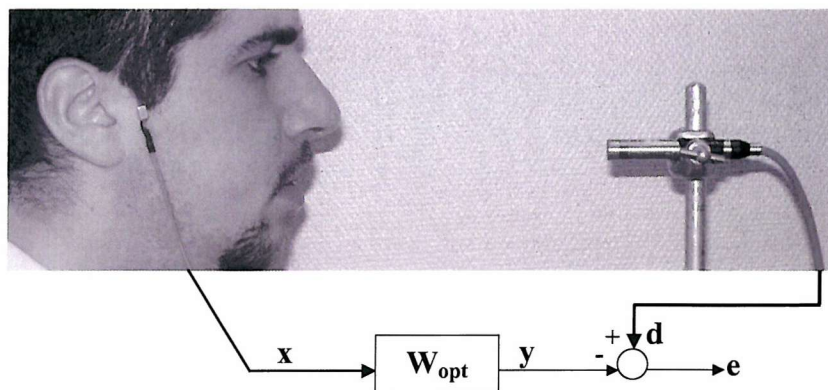


Figure 6.10: The setup for the design of our optimal filter: subject with accelerometer on his TMJ as the input signal, and reference microphone on the right as the desired signal.

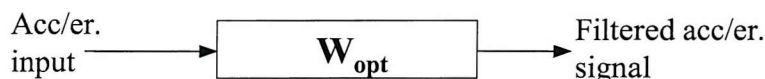


Figure 6.11: Block diagram showing the optimal filtering of the accelerometer signal.

A typical frequency response and output error spectrum of a W_{opt} filter derived for the sentence “The sunlight strikes raindrops in the air” can be seen in Figure 6.12.

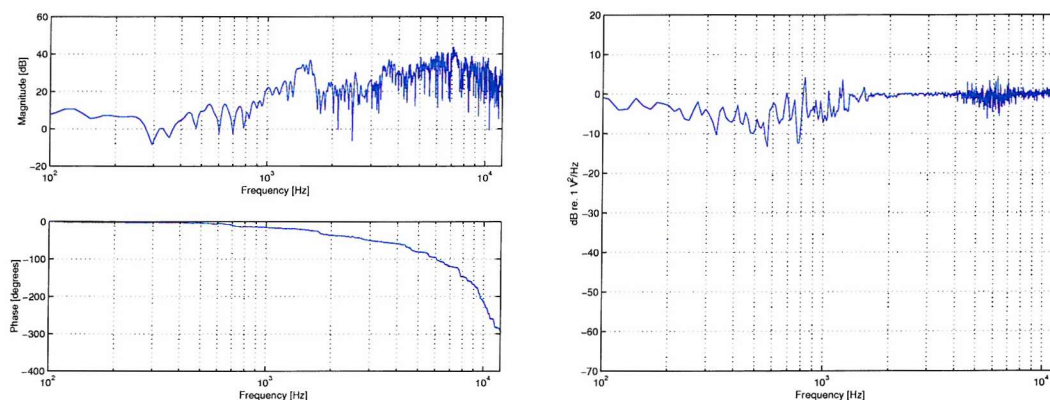


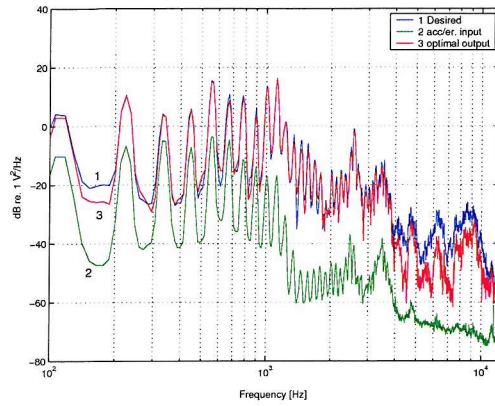
Figure 6.12: Magnitude $|W_{opt}|$ and phase response (top left) and output error spectrum (top right) S_{ee} of W_{opt} derived for the sentence “The sunlight strikes raindrops in the air”.

Here, the W_{opt} , which is derived for every phoneme or sentence individually, includes all of the impulse responses for all of the air- and body-conduction pathways. The optimising process attempts to match the TMJ accelerometer signal to the reference signal. Hence, as can be seen from its frequency response in Figure 6.12, the optimal filter compensates for the attenuation of the speech signal due to body-conduction losses (above about 1 kHz). This can also be justified by the error response, which is the difference of the desired signal d (reference microphone signal) and the output of the optimal filter y . A negative error response (in dB) tells us that there is a good match at the lower frequencies (up to approximately 1.5 kHz). At higher frequencies however, (above 4 kHz), some fluctuations in the error response indicate that the filter may need more data points, or that the coherence of the system is degraded by noise. These fluctuations that appear in the frequency response as well, could also be due to the fact that the filter tries to process two kinds of different, uncorrelated noise: acoustic noise picked up from the reference microphone, and mechanical noise due to the movement of the jaw, picked up by the accelerometer.

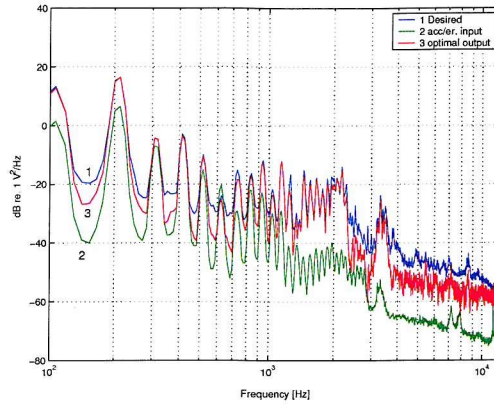
Results from Speech Processing

The miniature microphone signals were not processed at all at this stage, as our main interest was focused on the processing of the body-conducted speech. Characteristic input-output spectra from the optimal filtering operation on sustained phonemes can be seen in Figure 6.13. An example of all filtered output spectra for the sentence *“The sunlight strikes raindrops in the air”* can be seen in Figure 6.14.

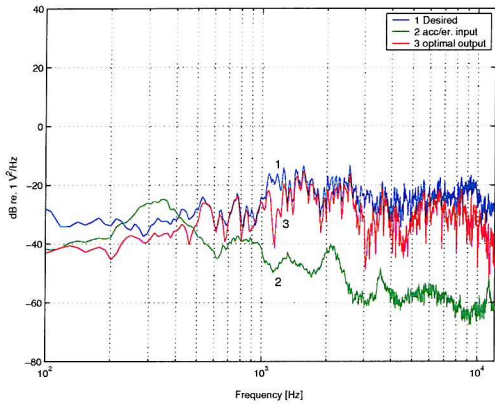
Optimal filtering seems to have the best performance of all filters, judged by the match between accelerometer input and reference spectra for most of the frequency range. For nearly all speech test material (sustained phonemes and



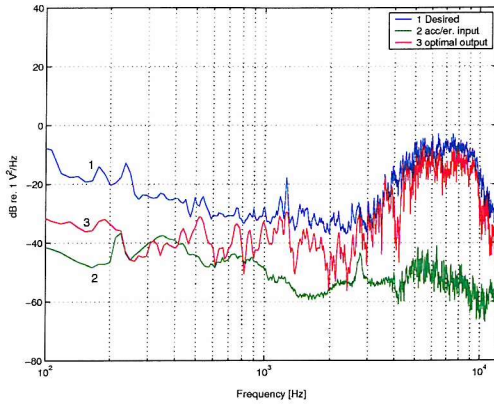
(a)



(b)



(c)



(d)

Figure 6.13: Desired, acc/er. input and optimal output spectra from four phonemes: (a) phoneme /a/, (b) phoneme /m/, (c) phoneme /f/, (d) phoneme /s/.

phrases), the error signals (not presented here) indicated an enhancement for frequencies up to 3 kHz or higher. More specifically, for vowels and nasals (see Figure 6.13(a,b)), the match was good up to about 3.5 kHz. For the /f/ case (see Figure 6.13(c)), the enhancement of the signal was better between 3.5 kHz-8 kHz, while for /s/ (see Figure 6.13(d)) the filter performed well above about 3 kHz.

The above observations are encouraging as far as the performance of the optimal filter is concerned. Thus, the signal enhancement for vowels and nasals

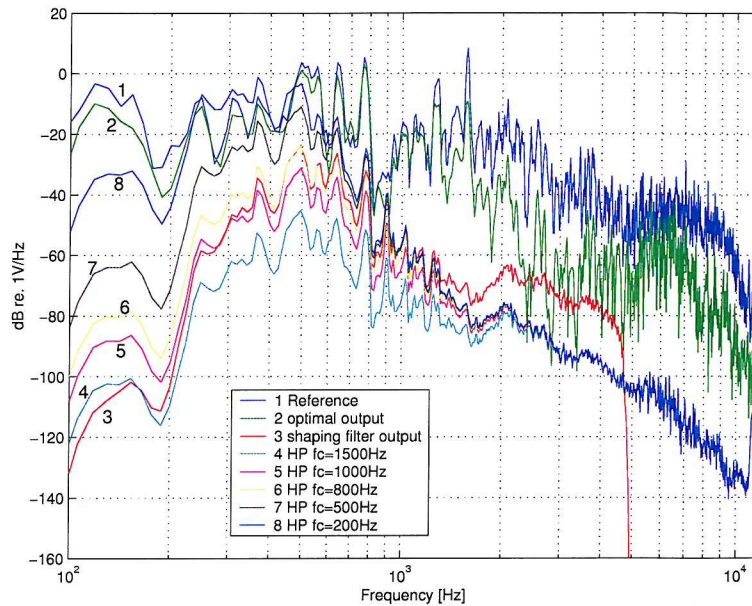


Figure 6.14: Reference microphone spectrum vs. output spectra from all filters for the sentence “The sunlight strikes raindrops in the air”.

occurred at lower frequencies (less than 3.5 kHz), where the input had most of its energy. Equivalently, for fricative sounds, the filter performed better at higher frequencies (above 3.5 kHz), where most of the fricative sound’s energy is found. The high cut-off frequency (4.7 kHz) of the shaping filter makes it inappropriate for processing of fricative speech sounds, as most of the useful information is excluded (see Figure 6.14). The same applies for the high-pass filter with high cut-off frequencies (1 and 1.5 kHz).

Listening Tests

The aim of this informal listening test was to investigate the subjective responses for the processed speech. For this reason, four subjects were used (2 female and 2 male). Only the sentences were tested as there was no point in assessing single sustained phonemes. In the first test, the subjects were asked to rate each processed sentence for naturalness and ease of understanding, each on a scale from 1-5 (see Figure 6.15).

The speech test material consisted of the reference microphone signal, the original unfiltered accelerometer signal, and all the outputs from the various filters discussed in section 6.2.6, thus giving us 9 different versions of each sentence. The results can be seen in Figure 6.16 after averaging the results across subjects.

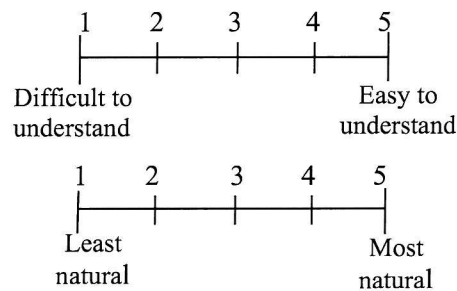


Figure 6.15: The response scale for ease of understanding (top) and naturalness (bottom) of presented signals.

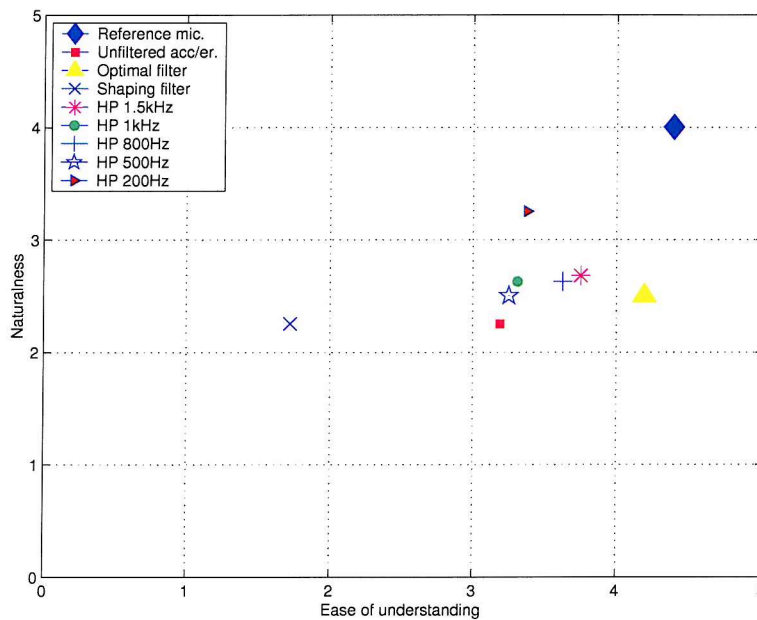


Figure 6.16: The results from the first subjective test.

In the second test, four versions of each sentence were presented to the subjects in order to test intelligibility. The versions were:

- Original unfiltered accelerometer signal.
- Optimal output.

- Shaping (band-pass) filter output.
- High-pass filter ($f_c=800$ Hz) output.

All four sentences were presented in a randomised order, and the subjects were asked to write down what they thought they had heard. At that time, a simple scoring was devised based on the number of syllables of every sentence. A score of 0 was given if a syllable was missing or if the sentence for the specific version was poorly reproduced. A score of 1 was given to an incorrect sentence which did not deviate too much from the original. Finally, a score of 2 was given to a fully reproduced sentence. The results, again, were found after averaging across subjects, and can be seen in Figure 6.17.

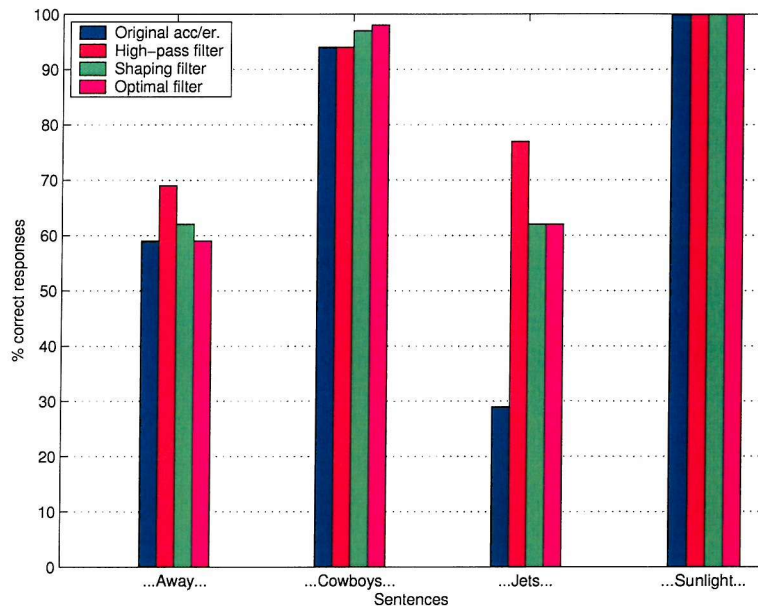


Figure 6.17: The results from the second subjective test.

Overall, as can be seen from Figure 6.16, the optimal filter outputs were judged easiest to understand on average, in comparison to the shaping filter outputs which were judged the hardest to understand. Naturalness did not have a big variability since most of the results were grouped between marks 2 and 3, in contrast to the ease of understanding which presents a higher variability. As far as the results of the second test are concerned (see Figure 6.17), it can be

seen that in the first sentence containing mostly vowels and glides (“*We were away a year ago*”), the unfiltered accelerometer signal gives the same results as the optimal filter, while the highest score is obtained with the high-pass filter. For the second sentence containing mostly diphthongs (“*Should we chase those young outlaw cowboys?*”), the unfiltered signal gives the same results as the high-pass filter, while the optimal and shaping filter give a slightly higher score. The interesting case however, is the third sentence containing mostly affricates, glides and diphthongs (“*High jets whiz past screaming*”). Here, the high-pass filter gives the highest score of all (about 78%), with the optimal and shaping filters following at a lower level (about 62%), and finally with the unfiltered accelerometer giving the lowest score (about 28%). For the last sentence, containing a good proportion of all classes of speech sounds, all filter outputs and the unfiltered accelerometer signal resulted in 100% correct responses, due to the fact that this specific sentence was already familiar to the subjects. These results indicate for the first time that different classes of speech may need different filters.

6.3 Re-design of the Experimental Setup

The initial experiment served as preparation for a larger-scale experiment which would test the effectiveness of optimal filtering compared only to high-pass filtering, since we saw that band-pass filtering was not really appropriate for speech vibration enhancement. The motivation for the re-design of the setup was, initially, the acoustic noise in the laboratory where the speech recordings took place. This pointed out the need for a quieter and more comfortable space, where the equipment could be easily setup without the obstruction from other experimental rigs. The other detail that surfaced after the completion of the first experimental period was the need to test a commercial body-conduction microphone, and compare its signal quality to that

of the miniature accelerometer. Another point that became important, was the use of a speech corpus which could comply with a standardised procedure for testing the intelligibility and quality of processed speech in communication devices. The last motivation was the subjective comparison of high-pass to optimal filtering both for quiet and noisy conditions, since, in the literature, high-pass filtering had not been compared to any other type of filtering for this application in either condition.

The second version of our experimental setup, therefore, approached the above problems in an improved way. The aim of this section is to present the main features of the new setup, i.e. the new place where speech recordings were to be made, the new equipment, and the new rig.

6.3.1 Recording Location

All the recording experiments of the revised experiment were conducted in a paediatric clinic of the Audiology department at the Institute of Sound and Vibration Research, used for audiological testing of children. The clinic is divided into a control room and a test room, which communicate via a door and a triple-glazed window of dimensions $1.5\text{m} \times 1\text{m}$. In order to comply with International Standards (ISO 8253-1, aimed at standardizing audiometric test methods), all the walls of the test room are twice the thickness (102mm) of a normal wall, and covered with sound-absorbing material, approximating the conditions to semi-anechoic. The test room's reverberation time as given by the building contractor does not exceed approximately 0.3s for the 1/3-octave bands 31.5 Hz-8 kHz, and the average temperature with one person in the room and all our equipment switched on, was measured with a digital thermometer at around 20°C.

The control room communicates visually with the test room through the win-

dow, and electrically via dedicated connector blocks on both sides of the separating wall. Hence, all equipment can be monitored from the control room, while the subject participating in an experiment can stay unaccompanied in the test room. An important piece of equipment that proved very advantageous for our setup was a monitor in the test room, which could communicate with the monitor of the PC in the control room via an embedded connection on the wall. Therefore, a subject could actually see on the screen whatever appeared in the experimenter's screen, something that is quite useful if automated programs are used as in our case.

6.3.2 Changes in the Experimental Equipment

All the equipment used for this experiment was exactly the same as the equipment presented in section 6.2.2, except the reference B&K microphone and an additional in-ear transducer. As was seen in section 6.2.2, we used a condenser 1/2" B&K 4133 type microphone, which gave us a high level of electric noise that appeared as a hissing sound in the preliminary recordings. Therefore, it was decided to use the 1" B&K 4145 which has a 10 dB lower inherent noise-floor than its predecessor. This is a microphone for general laboratory use and can handle a broad dynamic range of sound levels.

Another transducer was kindly sent to us from Temco, a Japanese company specialised in radio communication equipment. This "voiceducer" (model EP-2) is a body-conduction microphone (see section 3.3.4). The device contains both a subminiature electromagnetic receiver (loudspeaker) as well as a high performance vibration pick-up microphone. Temco also sent us a custom-made amplifier with an adjustable amplification knob for use with the in-ear microphone.

During some preliminary recordings, it was found that by using some addi-

tional amplification from the Kemo filter racks, we could achieve a satisfactory suppression of the electronic noise level for the miniature accelerometer and the Temco microphone, while the speech-to-electrical noise ratio improved to the following values: 28 dB for the miniature accelerometer, and 44.5 dB for the Temco voiceducer, measured with a recorded sentence from the DAM list serving as the speech signal (see section 7.3), and electrical noise recorded as described in section 6.2.3 as the noise signal. The SNR was calculated as a ratio between the power of the speech over the power of the noise signal. No extra amplification was used for the miniature microphone except its dedicated ISVR-made amplifier.

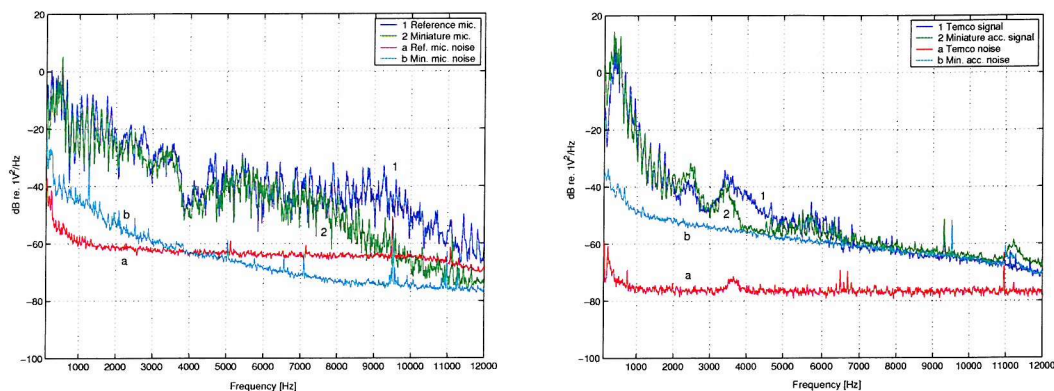


Figure 6.18: Average power spectrum of the sentence “The boat sailed off the edge”, compared to spectra of electronic noise-floors from the acoustic (left) and vibration transducers (right).

The noise-floor of the Temco in-ear microphone and of the other transducers was measured in exactly the same way as described in section 6.2.3. The total electronic noise-floor of all the transducers, along with the spectrum of the sentence “The boat sailed off the edge” recorded from the male speaker, appears in Figure 6.18. As seen, the spectral magnitude of the recorded signals is generally higher than the noise level. For the miniature accelerometer case however, the signal and noise converge at about 7 kHz which, as will be seen in Chapter 8, is the main reason of the poor performance of the optimal filter at higher frequencies.

6.3.3 New Experimental Setup

Two native British-speaking subjects were chosen for the recording of the speech samples so that any variations in expression due to different native language would be eliminated. One was male (MJ, 24 years old) and the other was female (EP, 29 years old). Both subjects were from the area of acoustic research, so they had a certain level of experience in speech recording. More specifically, the female subject came from the Audiology group of ISVR, hence she was quite acquainted with the way she should cite the speech samples.

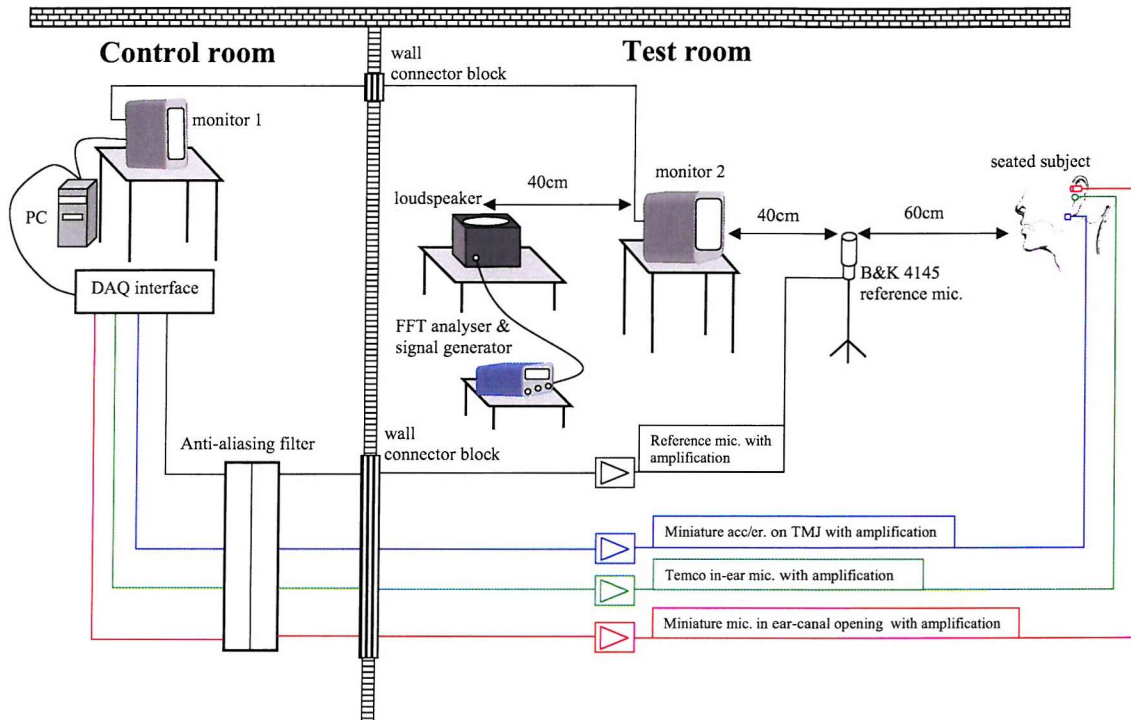


Figure 6.19: Experimental setup for the second experiment.

The subject sat in front of monitor 2 (see Figure 6.19) where she/he could watch the words or the sentences and then repeat them. The monitor was placed 1m from the subject's mouth, 40cm behind the stand with the reference microphone. The reference microphone was located in front of the subject's mouth, with the diaphragm at 0° incidence to the subject's mouth, at a distance of 60cm as shown in Figure 6.19. The phonetic material that was used consisted of the DRT and DAM tests, described in Chapter 4.

6.4 Conclusions

This chapter contributed towards a better understanding of the design of our experimental setup. The main conclusion from the initial experiment, is the advantage of using optimal filtering for the enhancement of speech vibrations, in comparison to high-pass filtering. Band-pass and high-pass filtering with high cut-off frequencies were rejected as inappropriate enhancement strategies. The other important conclusion is the use of the TMJ as a sufficiently good speech vibration pick-up point. This stemmed from an optimisation of factors such as distance from the ear, comfort of the speaker, and applicability of this location to a commercial product. Another point is the experimental verification of the theory from Chapter 3, as it was seen that spectrally “strong” phonemes such as vowels, nasals and voiced fricatives presented better body-conduction properties than their unvoiced counterparts.

Additionally, the setup used in the revised experiment proved to be improved as far as electronic noise suppression and ergonomics are concerned, and formed the basis of the main experiment discussed in the next chapter. This involved the use of a new in-ear commercial microphone, and some modifications in the rig of the initial experiment (recording location, new reference microphone, new amplification settings).

Chapter 7

Subjective Assessment of Optimal Filtering Performance

7.1 Introduction

The aim of this chapter is to present the formal experiment that followed the initial experiments of speech vibration recordings and their enhancement, presented in Chapter 6. The chapter starts with a description of the recording session. The next section includes the signal processing stage, where the principles behind the Matlab programs are explained. This leads to the final part, where the subjective listening tests are described, accompanied by the statistical analysis of the results and overall conclusions about the effectiveness of the optimal and high-pass filters.

(Part of the work included in this chapter was presented as a conference paper at the 1st meeting of the Hellenic Institute of Acoustics, Patras, Greece 2002).

7.2 Experiment Aims

After showing in the preliminary experiment of Chapter 6 that optimal filtering has the potential to improve the overall quality of speech vibration signals in comparison to band-pass, and high cut-off frequency high-pass filtering, it became necessary to verify this with a larger scale experiment. This would include standardised procedures for speech intelligibility and quality assessment, more than one speaker, and more listeners. Therefore, as will be seen in this chapter, it was decided to compare the subjective responses of 20 listeners regarding the intelligibility and quality of acoustic and vibration signals acquired from 2 speakers, by using the standard DRT and DAM tests (see Chapter 4).

The signals that were assessed included the reference microphone, the unfiltered accelerometer, the output from the optimal filter and the output from one high-pass filter, in both quiet and noisy conditions. The general objective of this experiment is the demonstration of the optimal filter's advantage over high-pass filtering in enhancing speech vibration signals for both conditions. What we expect to find are significant statistical inferences concerning the improvement of the accelerometer signal's intelligibility and quality after using optimal filtering, compared to the use of high-pass filtering, for both quiet and noisy cases.

7.3 Speech Recordings

The aim of this section is to depict the formal recordings of the DRT and DAM corpora, including descriptions of the computer programs that were used and representations of the whole process, from the moment the speakers entered the clinic to the end of the sessions.

7.3.1 Acquisition and Randomisation Programs

The first step was to write some routines for the automated, simultaneous data acquisition, using Matlab with its dedicated Data Acquisition Toolbox. The data acquisition program employed four channels (see section 6.3.3), each acquiring data at 24 kHz. For the DRT test, a duration of 3s was used for each word, while for the DAM test, the duration was increased to 4s. Those times were considered large enough to capture all the phonemic information of words or phrases respectively.

The reader will remember that there are 16 pairs of rhyming words in each category of the DRT words list. If this is transferred to a real recording situation, we can imagine the speakers getting tired, but also biased, since they would anticipate the style of the next word to be acquired. Consequently, at the end of the recording, the expression of the recordings would sound fatigued, something that is not desired, since all speech samples must be uttered in a clear, “flowing” way. For this reason a randomisation routine was written, which read all the words for every feature from a *struct* array. It randomised them and then picked up one, subtracted it from the array, and displayed it after the pressing of a key. Hence all words could be displayed once without the same word appearing twice. The above acquisition routine was incorporated in the randomisation program with an additional feature: that for every word or sentence (since the same program was used for the DAM test as well), data were saved in a folder with the same name as the word or the sentence. So, at the end of the recording session, there were two root folders, one for the male and one for the female speaker, each containing subfolders with the appropriate names of the features (voicing, nasality, etc.) or the sentence groups (group 1, group 2, etc.), each one including smaller subfolders with the saved data, named after the words or the sentences.

The acquisition program was controlled by the experimenter from the PC in the control room. Monitor 1, as appears in figure 6.19, communicated with monitor 2 so the speech material could be seen on both monitors simultaneously. The session would start with a welcoming message on a blank screen, followed by an instruction to the subject to repeat each word after she/he saw it. The experimenter pressed a key on the keyboard of the PC, and the session started. The word (or sentence) came into view on monitor 2, and the only task the subject had to carry out was to read the word or phrase from the screen, while data were being recorded from the transducers.

7.3.2 Recording Session

A typical recording session would start by advising the subject to speak clearly and without any fluctuations. The subject would sign the safety and ethics form and then enter the test room, sit on the chair and the miniature transducers would be attached to her/his face. The miniature microphone was attached to the opening of the ear canal, such that it was not blocking it. The miniature accelerometer was attached to the skin surface over the temporo-mandibular joint (TMJ) (see Figure 6.10), and the in-ear Temco microphone was secured in the speaker's ear-canal. Medical tape was used for the positioning of the miniature microphone and accelerometer. The Temco transducer was cleaned with medical alcoholic swabs when a speaker had finished a session, so that it could be used safely by the next speaker.

All equipment was already linked to the connector blocks on the wall leading to the data-acquisition PC and the anti-aliasing filters in the control room. Preliminary recordings of 5-10 words were performed so that the subject would become accustomed to the setup. After that, the experimenter closed the partitioning door, went to the control room and with a wave of a hand to the subject, gave the signal to start the recording. The recording of both DRT and

the DAM tests was done twice in order to avoid any problems such as extraneous coughs, clicks, laughs, etc. At the end, all the recordings were listened to and the tokens without any problems were chosen. No major problems were noted in the first session except some coughs and deep breaths. After the recording of the DRT words, the subject had a rest for 2-3min and the DAM sentence list was recorded.

Since we would test the effect of different filters on quiet and noisy speech, it was necessary to record a short segment of noise that would be added with Matlab to the speech signals recorded from each transducer. For this purpose, the signal generator of the FFT analyser was used, connected to a loudspeaker resting on the table behind monitor 2, as seen in the schematic diagram in Figure 6.19. The subject's exposure to this noise complied with the Safety and Ethics committee regulations. The generated noise was broad-band, with a bandwidth 20 Hz-20 kHz, and amplitude 7V. Five seconds were recorded, with all the transducers attached on each subject. The sound pressure level as recorded by a sound level meter placed 5cm from the subject's ear was approximately 80 dBA. The recording of noise for each subject marked the end of the first part of this experiment.

Figure 7.1 shows experimentally the principle of immunity of accelerometric speech to acoustic noise up to about 1.5 kHz. While the air-conducted signals have an overall low signal-to-acoustic noise ratio (-9 dB for the reference microphone), the body-conducted signals present a higher S/N ratio (21 dB), since the speech vibrations are not significantly degraded by the presence of noise. The above values were calculated as a ratio between the power of the speech over the power of the acoustic noise signal. The peak at 4 kHz in Figure 7.1(b) could be partly explained as an interaction between the rising frequency response of the transducer and the falling spectrum of the noise signal, to which the accelerometer is responding in some way. The outer ear-canal resonance (see section 2.4.1) could also play a role, as it could be picked-up

by the accelerometer placed close to the subject's ear-canal.

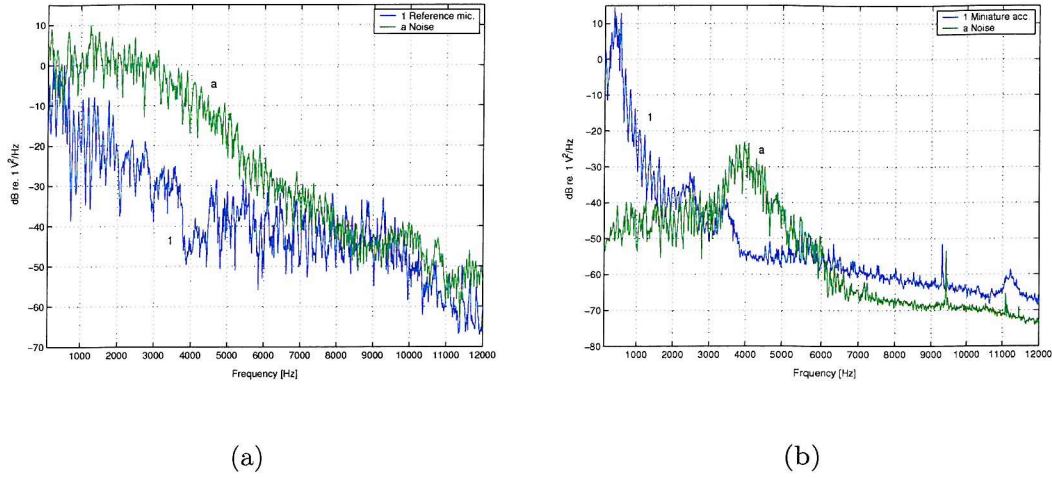


Figure 7.1: Spectra of the sentence “The boat sailed off the edge” (1), and noise (2), as recorded from the reference microphone (left) and the accelerometer (right).

7.4 Processing of the Signals

All the recorded signals were initially saved as Matlab data files, so from this moment onwards it was a matter of processing them and generating the sound files for the listening tests. The program initially added the recorded noise to the reference, the miniature accelerometer, and the Temco voiceducer signals (see Figure 7.2). There are two weaknesses of this additive technique used to simulate a noisy environment. These are: (1) that the subject would not actually talk in the same way in a noisy environment, and (2) that in spectral regions where the signal recorded from the subject is comparable to the electronic noise floor, adding the signals would result in raising the effective noise floor, and this could result in a slight underestimate of the utility of the enhancement filters. The program then calculated the coefficients of the Wiener filter, which was applied to the signals from the miniature transducers. The program calculated the auto- and cross-correlation matrices between the inputs from the miniature transducers and the clean reference signal and re-

trieved the Toeplitz (Hessian) matrix, with the help of which an optimal filter was calculated. Six hundred coefficients were used, and for each input (word or phrase) a different optimal filter was generated.

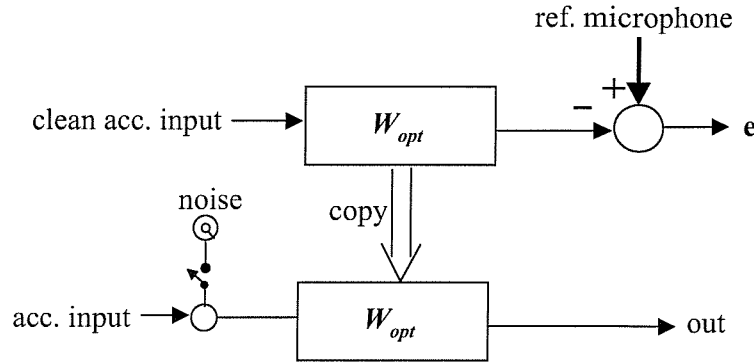


Figure 7.2: Schematic diagram of the Matlab program for optimal filtering. (Top part) Optimal filter calculation. (Bottom part) Generation of filtered output with/without noise

In addition to the optimal filter, a simple second-order high-pass filter (Butterworth response) with a cut-off frequency of 800 Hz was created and employed in the same program (see Figure 7.3). The choice of this filter with this specific cut-off frequency was based on the results of the preliminary experiments, where it was shown that the low-pass cut-off point of the accelerometric speech for most of the cases was around 800 Hz.

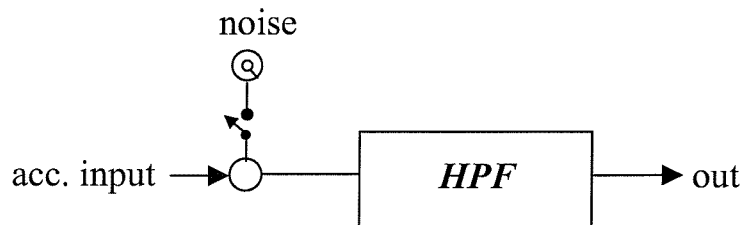


Figure 7.3: Schematic diagram of the Matlab program for high-pass filtering with/without noise.

In order to save time, the program employed a loop function so that during each run it would create all the filter outputs at once, by retrieving the data from the folders created during the acquisitions. Simultaneously, sound files (.wav format) of all the original and processed signals were created. All sound files of the filter outputs were normalised to the power of the reference signal

so that they could be played at the same level.

7.5 Subjective Listening Tests

After the signal processing, it was time to perform the subjective listening tests as described in Chapter 4. At this time, for each word or sentence, there were sixteen different versions: clean and noisy reference signal, clean and noisy accelerometer signal, clean and noisy Temco signal, optimal filter outputs for clean and noisy accelerometer and Temco signals, and finally high-pass filter outputs for clean and noisy accelerometer and Temco signals. For the subjective tests, it was decided to use the following eight signals, since they would provide a good representation of the signal processing outputs. After the description of each signal-pair, the abbreviations by which we will refer to them are given in parentheses.

1. Clean and noisy reference microphone signal (Ref. mic./Noisy Ref.).
2. Clean and noisy miniature accelerometer signal (Acc./Noisy acc.).
3. Optimal output for clean and noisy accelerometer signal (Opt./Noisy Opt.).
4. High-pass filter output for clean and noisy accelerometer signal (HP/Noisy HP).

The reason Temco signals were not used was that in some preliminary listening of the signals, they were found to have a higher noise level caused by clicks of the moving jaw, something that was not present in the accelerometer data. Also, we were restricted to the number of overall options, as we wanted a listening test with a reasonable length. The main difficulty was to organise

and present the signals for the listening test. For this reason some routines were written in Matlab in order to facilitate the tests.

We have to remember that in the DRT test, there are two rhyming words that appear visually to the subject while only one is heard, and the subject must decide which one it is. Hence, the aim was to present mixed male or female signals, randomised in such a way so that for each word pair, four versions of the first and four versions of the second word would be presented in a random manner. The restrictions imposed by the experimenter were that for a single listener, the same version of the same word should not be heard twice; the same word should not be heard twice in a row; half of the presentations of words would be male and half female; and finally, those presentations would be randomised with regard to which word in the pair and which feature were presented to a given subject.

For purposes of illustration, let us choose the first word pair of voicing (see Table 4.1) as an example and call it 1. Then, assign the letters a-h for the eight output versions stated above. Also, name the first word (veal) as w1, and the second (feel) as w2. So, what the program does is to present the subject with 1,w1,a ... 1,w2,f ... 1,w2,h ... 1,w1,b ... etc. The dots represent versions of other word pairs that may have been presented in between. Twenty subjects (10 males, 10 females) were used, all of British origin. In total, every subject listened to 32 words in each category \times 6 categories \times 4 versions for each word = 768 words. The overall DRT test lasted about 45 minutes, during which subjects were allowed to take a break if they felt tired.

In the training phase, prior to the main listening test, subjects heard a set of reference speech signals that exemplified the high, low, and middle judgment categories. This process is known as “anchoring” and is meant to give all listeners the same subjective range and origin in their quality ratings. Hence, certain samples of the clean reference, the noisy reference, the clean accelerom-

eter signal and clean optimal output were presented.

The subject watched the word pair on the screen of the PC, while one word was heard. When the presentation finished she/he pressed the keys 1 or 9, which represented the first (left) word and the second (right) word respectively, according to which one she/he thought has been played back through the headphones connected to the sound-card of the PC. At the end of the test, all results were gathered and a small database was created with the history of presentations, including the true word that had been presented, the subject's answer, the version of the output, the speaker (male or female), and the feature category. Then the program calculated an intelligibility score for each feature based on equation 4.1.

For the DAM test, the constraints were not so strict, as for each version of filter output a whole group had to be presented to the subject. The only randomisation was between the sentences of each group. After each group presentation had finished, the subject was given the quality evaluation form to fill in, as shown in Figure 4.1. For each of the quality scales (signal, background, total effect), a sub-score was calculated by simple averaging. The DAM test lasted about 20 minutes, as each sentence group lasted approximately 1 minute and the subject had to listen to 8 groups (one for every signal version) and assess each one by completing the DAM response scale form. The overall listening test did not take more than 1.5h, including some pauses when the subjects took a break.

7.6 Results from Listening Tests

All the results from the listening tests were gathered in a database in the statistical software SPSS. The statistical test that was used for all subsequent

analysis was the *repeated-measures* ANOVA with the “*differences*” type of contrast. By a combination of ANOVA’s and t-tests, we were able to derive some conclusions about the effect that some factors (type of filter, condition, transducer, speaker) had on the intelligibility and quality results.

7.6.1 Results from DRT Test

The results of the DRT test, averaged across listeners, are plotted in Figure 7.4 and also appear in the following Table 7.1. The legend of the graph describes sensor, quiet/noise condition, and filter combinations. Each point represents an average of 20 listeners’ responses to 16 DRT word pairs, giving a total of 320 values.

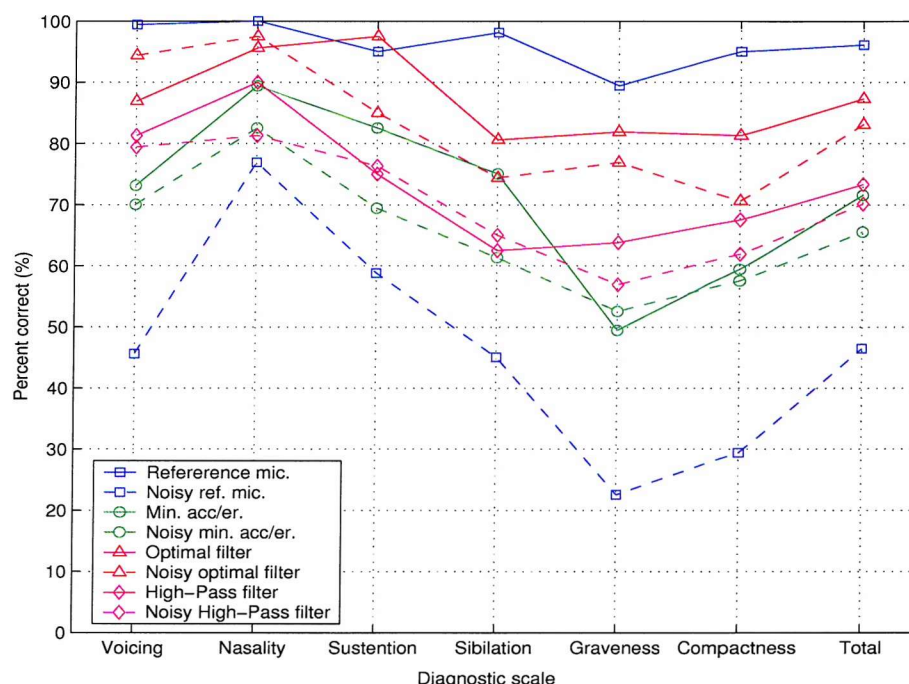


Figure 7.4: Results from the DRT intelligibility test. Scores have been corrected for chance performance.

As can be seen, the quiet reference microphone signals provide the highest intelligibility scores, on average, for all features except for sustention, where the optimal filter results are slightly higher. The results from the optimal

	Ref. mic.	Noisy Ref.	Acc.	Noisy acc.	Opt.	Noisy opt.	HP	Noisy HP
Voicing	99.4	45.6	73.1	70.0	86.9	94.4	81.3	79.4
Nasality	100.0	76.9	89.4	82.5	95.6	97.5	90.0	81.3
Sustention	95.0	58.8	82.5	69.4	97.5	85.0	75.0	76.3
Sibilant	98.1	45.0	75.0	61.3	80.6	74.4	62.5	65.0
Graveness	89.4	22.5	49.4	52.5	81.9	76.9	63.8	56.9
Compactness	95.0	29.4	59.4	57.5	81.3	70.6	67.5	61.9
Avg. per system	96.1	46.4	71.5	65.5	87.3	83.1	73.3	70.1
Avg. across cond.	71.2		68.5		85.2		71.7	

Table 7.1: Intelligibility scores for each feature and every transducer configuration.

filter for both quiet and noisy input signals follow the trend of the reference microphone results, suggesting that the optimal filter improves intelligibility both in quiet and noise compared to unfiltered signals. An interesting point is that the results from the unfiltered accelerometer are in a similar range to the results of Giua (1998) (see section 3.2).

It is also interesting to note that for voicing and nasality the noisy optimal signals have a higher intelligibility score than the quiet ones. This could be due to the fact that the presence of noise may add some ‘liveliness’ to the signals, which results in improved intelligibility. The results from the quiet and noisy accelerometer signal as well as from the quiet and noisy high-pass outputs all follow the same pattern with the highest scores for nasality and voicing. This is due to the fact that the other features are mostly based on consonants, which are more prone to noise or filtering changes than voiced or nasal sounds. The noisy reference microphone signal is, on average, the least intelligible. Nevertheless, there is a need for statistical tests in order to see which of these means is significantly different.

7.6.2 The Variables in the Statistical Analysis

The purpose of the following statistical analysis is to draw some general conclusions about the effect of noise, transducer and filter type on the intelligibility

results. This could potentially show some significance among the DRT results, which is not clearly seen from a first glance of their graphical representation and the mean values.

The independent variables under investigation are the speech features (6-level factor), the transducers (reference microphone/ accelerometer: 2-level factor), the filters (optimal/ high-pass: 2-level factor), and the condition (quiet/ noise: 2-level factor). Our dependent variable is the intelligibility which is tested for different combinations of the independent variables.

Initially we grouped the results and investigated the effect of condition vs. transducer, and the effect of filter on the accelerometer signal vs. the condition.

7.6.3 Effect of Transducer and Condition on Specific Features

This test was conducted in order to look on the interaction of transducer and condition on specific features. It was therefore a $6 \times 2 \times 2$ level ANOVA and the factors were the features (6-level factor), the condition (2-level factor) and the transducers (2-level factor). Data were averaged across subjects. Averages across condition and transducer, for every feature can be seen in Table 7.2.

The null hypothesis H_0 therefore, is that the intelligibility scores for all features and all transducers were the same for all conditions. The alternate hypothesis H_1 is that the scores among the three dimensions are different. The results from the SPSS analysis showed that the main effects of all three factors are significant: Feature ($F = 46.3$, $p < 0.01$), Transducer ($F = 6.67$, $p = 0.018$) and Condition ($F = 546.3$, $p < 0.01$). Interactions vary: Feature \times Transducer is not significant ($F = 0.38$, $p = 0.8$), but Feature \times Condition is significant ($F = 5.84$, $p = 0.002$) as is Transducer \times Condition ($F = 263.5$, $p < 0.01$).

Features		Reference mic.	Min. acc/er	Avg. by condition
Voicing	Quiet	99.4	73.1	86.25
	Noise	45.6	70	57.8
	Avg. by transducer	72.5	71.55	
Nasality	Quiet	100	89.4	94.7
	Noise	76.9	82.5	79.7
	Avg. by transducer	88.45	85.95	
Sustention	Quiet	95	82.5	88.75
	Noise	58.8	69.4	64.1
	Avg. by transducer	76.9	75.95	
Sibilation	Quiet	98.1	75	86.55
	Noise	45	61.3	53.15
	Avg. by transducer	71.55	68.15	
Graveness	Quiet	89.4	49.4	69.4
	Noise	22.5	52.5	37.5
	Avg. by transducer	55.95	50.95	
Compactness	Quiet	95	59.4	77.2
	Noise	29.4	57.5	43.45
	Avg. by transducer	62.2	58.45	

Table 7.2: $6 \times 2 \times 2$ level ANOVA matrix for effect of transducer vs. condition for every feature.

Three-way interaction of Feature \times Transducer \times Condition is significant ($F = 28, p < 0.01$).

We thus conclude that the intelligibility of the reference microphone and unfiltered accelerometer signals varies significantly with each of the three factors. Considering the factor transducer, the results averaged across quiet and noise seem to indicate that the intelligibility for the overall reference microphone is significantly better than that of the overall accelerometer; however, we must consider the fact that the Transducer \times Condition interaction was also significant. We also cannot necessarily conclude that the quiet reference microphone is significantly better than the quiet accelerometer, and that the noisy accelerometer is better than the noisy reference even if the means of Table

7.1 seem to imply this. In order to support this, a t-test with Bonferroni correction was performed comparing noisy microphone to noisy accelerometer results. The result was significant ($p < 0.01$), suggesting that the noisy accelerometer performs better than the noisy reference microphone.

The significance of the interaction of feature and condition can be understood to mean that the main effect of noise varies significantly according to the feature being tested. The significance of the interaction of transducer and condition appears mainly to be due to the fact that noise affects the intelligibility of the reference microphone far more than that of the accelerometer.

A 2×2 ANOVA was also performed for every feature in order to test the main effect of transducer (reference microphone/accelerometer) and condition (quiet/noise) specifically for every feature. The data were averaged across subjects. Results appear in Table 7.3.

Feature	Transducer	Condition	Transducer \times Condition
Voicing	$p = 0.74$	S	S
Nasality	$p = 0.25$	S	S
Sustention	$p = 0.26$	S	S
Sibilant	$p = 0.77$	S	S
Graveness	$p=0.076$	S	S
Compactness	$p=0.069$	S	S

Table 7.3: Effect of transducer (reference/accelerometer), condition (quiet/noise) and interaction between them for specific features. The letter **S** indicates significance at $p < 0.01$ level; the entry in bold indicates significance at $p < 0.1$ level.

Interpretation of the Results for the Effect of Transducer

The reasoning behind the above results is dependent on the physical properties of every feature and on the extent that these are influenced by a change of either transducer or condition. The following points, therefore, form an attempt to understand the trends that were revealed with the ANOVA test.

For *Voicing*, the subject is asked to judge between a word which starts with a voiced phoneme and one which starts with its unvoiced counterpart, while the last part of each word stays the same. The low-pass filtering characteristic of the accelerometer would be expected to leave most of the words which started with a voiced phoneme unaffected, as most of the energy of voiced sounds is primarily concentrated below 1 kHz, where there is little attenuation due to body-conduction. In addition, noise would be expected not to interfere significantly with the accelerometer because of its noise-immunity characteristics.

For *Nasality* the subject is asked to discriminate a word which starts with a nasal phoneme. In this case, the pairs compare words which start with phonemes with the same place categorisation, while they differ in manner (nasal vs. voiced stop) and are presented in an alternating pattern; /m/ vs. /b/ and /n/ vs. /d/. For the same reasons presented in the Voicing category above, we would not expect the accelerometric effect and noise to interfere in this case.

For *Sustention*, one can notice in the DRT list (see section 4.2) that the initial phonemes of the words in the left column are all fricatives. They are contrasted with stops and affricates. In each pair the voicing is kept the same, and place is approximately the same. In a low-pass filtered condition like accelerometric speech, fricatives and stops get confused because the noises whether continuous or transient, are all relatively high frequency. We should therefore expect this feature to be affected by the difference in transducer as well as the difference in condition (broadband noise will obscure the noise differences).

For *Sibilant*, the initial phonemes in the left column are sibilant fricatives and affricates while in the right column they have the same voicing and place but are either non-sibilant fricatives or stops. The sibilant fricatives are much louder than their counterparts; the affricates on the other hand, have a transient aspect and sibilance in the following noise, whereas the stops have only

the transience. Again we would expect both the low-pass filtering of the accelerometric speech and the noisy condition to interfere with this feature substantially.

Graveness contrasts labial sounds (left column) with dentals. Labials will generally have a concentration of energy around 1 kHz; dentals are diffuse (first formant is low, F2 and F3 are high and burst frequencies are spread across the spectrum). Some of these pairs (/f-θ/) are confusable in normal speech with good conditions, which should mix up the intelligibility scores.

Compactness contrasts palatal-velar sounds (left column) with labials and dentals. The palatal and velar sounds should have a concentration of energy around 2 kHz, which we expect would be somewhat affected by the low-pass effect of the accelerometer. Graveness and compactness have to do with spectral shape and therefore place identification, and thus might be more sensitive to the effect of transducer (accelerometer vs. microphone).

So overall the predictions for the effect of the the low-pass filtering characteristic due to the accelerometer, are that it:

- does not interfere significantly with voicing or nasality; this was demonstrated, as seen in Table 7.3.
- interferes moderately with graveness and compactness, which was again partly shown by the insignificance as seen in Table 7.3.
- interferes significantly with sustention and sibilation. This point is difficult to relate to the result of Table 7.3, because the accelerometer effect will favour the microphone signal, while the noise effect will favour the accelerometer signal.

The effect of noise as opposed to quiet condition was expected to,

- interfere insignificantly with voicing and nasality,
- interfere moderately with graveness and compactness,
- interfere significantly with sustention and sibilation.

As it appears here, the main effect of noise was significant for all features, which seem to contradict the above predictions. However, the significance of all interaction terms indicates that we should look further. From the means in Table 7.1, it is clear that the intelligibility of the reference signal is strongly reduced in noise (most in Graveness and Compactness). The accelerometer signal intelligibility also decreases in noise, and appears to do so least for Voicing and Nasality and most for Sibilation and Sustention, as predicted.

7.6.4 Effect of Filter and Condition on Specific Features

This test was conducted in order to look on the main effect of filter and condition on specific features. It was therefore a $6 \times 2 \times 2$ level ANOVA and the factors were the features (6-level factor), the condition (2-level factor) and the filters (2-level factor). Data were averaged across subjects. Averages per filter and condition for every feature can be seen in Table 7.4.

The null hypothesis H_0 is that the intelligibility scores for all features and all filters were the same for all conditions. The alternate hypothesis H_1 is that the scores among the three dimensions are different. The results showed that the main effect of Feature is significant ($F = 42.4$, $p < 0.01$), as are those of Filter ($F = 69.7$, $p < 0.01$) and Condition ($F = 16$, $p < 0.01$). This was also verified with two separate t-tests, comparing the two filters in quiet and in noise for data averaged across features and subjects. The analysis showed that optimal filtering is significantly better than high-pass filtering for both conditions ($p < 0.01$).

Features		Optimal filter	High-Pass filter	Avg. by condition
Voicing	Quiet	86.9	81.3	84.1
	Noise	94.4	79.4	87.15
	Avg. by filter	90.65	80.35	
Nasality	Quiet	95.6	90	98.2
	Noise	97.5	81.3	89.4
	Avg. by filter	96.55	85.65	
Sustention	Quiet	97.5	75	86.25
	Noise	85	76.3	80.65
	Avg. by filter	91.25	75.65	
Sibilant	Quiet	80.6	62.5	71.55
	Noise	74.4	65	69.7
	Avg. by filter	77.5	63.75	
Graveness	Quiet	81.9	63.8	72.85
	Noise	76.9	56.9	66.9
	Avg. by filter	79.4	60.35	
Compactness	Quiet	87.3	73.3	80.3
	Noise	83.1	70.1	76.6
	Avg. by filter	85.2	71.7	

Table 7.4: $6 \times 2 \times 2$ level ANOVA matrix with average intelligibility scores for effect of filter vs. condition.

None of the two-way interactions were significant ($p > 0.05$), but the three-way interaction was ($F = 5.9$, $p < 0.01$). Post-hoc t-tests were conducted to check on the main effect of filter in each quiet/noise condition separately. The quiet optimal and quiet high-pass intelligibility scores differed significantly ($p < 0.01$), and noisy optimal and noisy high-pass also differed significantly ($p < 0.01$), which is consistent with the lack of filter-condition interaction. More important, this means that the optimal filter outperforms the high-pass filter in both quiet and noisy conditions.

The above results show that the main effects of transducer, filter or condition as individual factors, were strong enough to create significant differences in intelligibility scores. However, the insignificant interaction between filter and



condition, suggests that the effect of noise on both the optimal and the high-pass filter was similar.

A 2×2 ANOVA was therefore performed for every feature in order to test the interaction between filter (optimal/high-pass) and condition (quiet/noise) at a more detailed level. The data were averaged across subjects. Results appear in Table 7.5.

Feature	Filter	Condition	Filter×Condition
Voicing	S	$p = 0.12$ (*)	S
Nasality	S	S (*)	S
Sustention	S	$p = 0.16$	S
Sibilant	S	S	S
Graveness	S	$p = 0.15$	$p = 0.65$
Compactness	S	S	$p = 0.34$

Table 7.5: Effect of filter (optimal/high-pass), condition (quiet/noise) and interaction between them for specific features. (*) shows that noisy condition is better than quiet. The letter **S** indicates significance at $p < 0.01$ level.

7.6.5 Effect of Speaker

In order to test whether the particular speaker had an effect on intelligibility, four ANOVA tests were performed.

A 2×2 level ANOVA was performed testing the effects of speaker (female, male) vs. transducer (reference microphone, and accelerometer) for quiet signals. Data were averaged across features and can be seen in Table 7.6.

	Male speaker	Female speaker	Avg. by transducer
Reference microphone	96	96.3	96.15
Accelerometer	71	71.9	71.45
Avg. by speaker	83.5	84.1	

Table 7.6: 2×2 level ANOVA matrix for speaker vs. transducer for quiet condition.

The main effect of speaker was not significant ($F = 0.014$, $p = 0.9$); the main

effect of transducer was ($F = 105.8$, $p < 0.01$), and the interaction was not significant ($F = 0.7$, $p = 0.41$).

A similar analysis was made for the signals in noise. The data are shown in Table 7.7.

	Male speaker	Female speaker	Avg. by transducer
Reference microphone	46.3	46.5	46.4
Accelerometer	65.4	65.6	65.5
Avg. by speaker	55.85	56.05	

Table 7.7: 2×2 level ANOVA matrix for speaker vs. transducer for noisy condition.

In this case, all main effects were significant: that of speaker ($F = 6.1$, $p = 0.023$), transducer ($F = 41.4$, $p < 0.01$), and their interaction ($F = 8$, $p = 0.011$). The intelligibility of the female speaker in noise is on average only slightly higher than that of the male speaker, but that difference is significant.

Similar analyses were also made investigating the effect of speaker vs. filter (optimal, high-pass). For both quiet and noisy signals, main effect of speaker was not significant (quiet: $F = 2.57$, $p = 0.125$; noise: $F = 0.91$, $p = 0.35$), main effect of filter was significant (quiet: $F = 65.2$, $p < 0.01$; noise: $F = 18.3$, $p < 0.01$), and the interaction was not significant (quiet: $F = 3.86$, $p = 0.064$; noise: $F = 0.9$, $p = 0.35$).

We thus conclude that overall both speakers are equally intelligible in nearly all of the test conditions, with just the small but significant difference for the unfiltered signals in noise. We do not then need to break down results by feature in terms of speaker.

7.6.6 Results from DAM Test

The results from the DAM test can be seen in Figure 7.5. The legend describes sensor, quiet/noise condition, and filter combinations. They are averaged across listeners and sub-categories (9 for Signal qualities, 7 for Background qualities and 3 for Total Effect). Each bar represents an average of 20 listeners' responses to 8 DAM sentence groups, giving a total of 160 values.

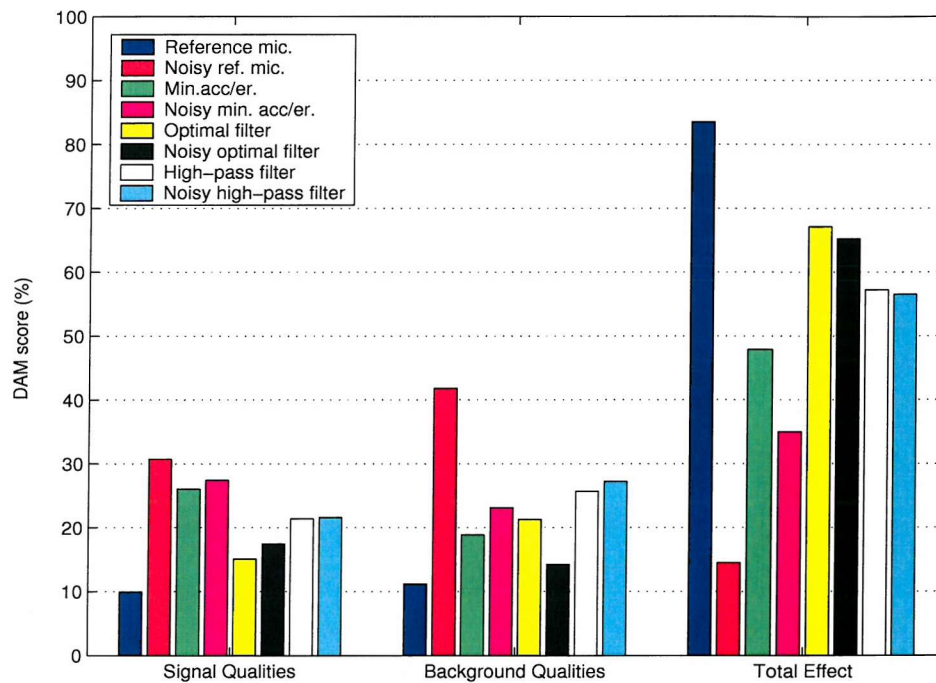


Figure 7.5: Results from the DAM quality test.

This plot is quite revealing since it shows directly the preferences of the listeners during the listening tests. It should be noted that for the first two categories (signal and background qualities), a small score shows a good quality signal (i.e. an absence of a “negative” quality).

Therefore, the reference microphone signal is rated as having the best Signal Quality, followed immediately by the optimal filter output for both quiet and noisy cases. High-pass filter output comes next in preference judgements, and last come the signals from quiet and noisy accelerometer and noisy reference microphone cases.

For the Background Qualities, again the quiet reference microphone is assessed as the best concerning background conditions, while quiet miniature accelerometer and quiet optimal filter signals come next. Those are followed by the noisy optimal and the quiet and noisy high-pass filtered signals, while the worst in the subjects' preferences comes the noisy reference microphone.

For the third category concerning the total effect, the reference microphone is considered to be the best regarding intelligibility, pleasantness and acceptability. The optimal filter for both quiet and noisy conditions comes next in overall judgement, followed by the quiet and noisy high-pass filtered signals. Then we have the miniature accelerometer signals and finally the noisy reference signal which is judged as the signal with the worst total quality.

The following statistical analysis will uncover the different trends in more detail, as we examine the different sub-categories of the three main quality teams.

7.6.7 Signal Qualities

The results of all the sub-categories in the Signal Qualities main category can be seen in Figure 7.6 and also appear in Table 7.8. Each point represents an average of 20 listeners' responses to 8 DAM sentence groups, giving a total of 160 values.

The aim is to investigate the importance of those sub-categories, and their effect in relation to transducer, condition and filter. Therefore, the analysis that will be presented here will be the same as in the DRT test, i.e. we will examine the main effect of transducer vs. condition first, and then the main effect of filter vs. condition. One interesting point in Figure 7.6 is the high score of 'muffling' assigned to the accelerometer signals, both for quiet and

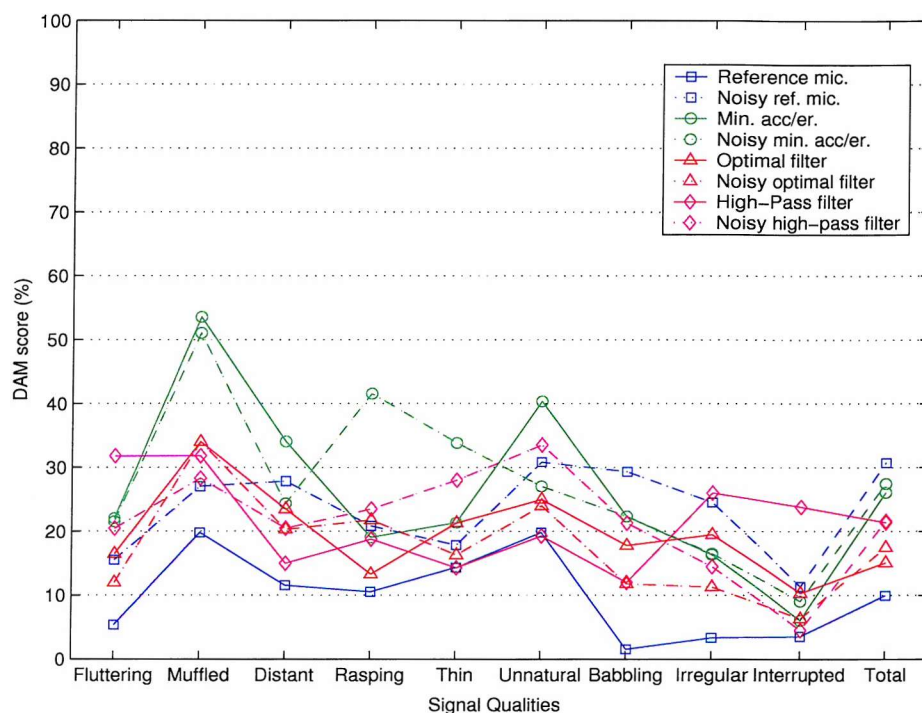


Figure 7.6: Results from the DAM test for Signal Qualities. DAM score is extent to which test sentences exhibit the particular signal quality.

	Ref. mic.	Noisy Ref.	Acc.	Noisy acc.	Opt.	Noisy opt.	HP	Noisy HP
Fluttering	5.3	15.5	22	21.5	16.5	12	31.8	20.5
Muffled	19.8	27	53.5	51	34	34	31.8	28.3
Distant	11.5	27.8	34	24.3	23.5	20.3	15	20.5
Rasping	10.5	20.8	19	41.5	13.3	21.8	18.8	23.5
Thin	14.3	17.8	21.3	33.8	21.3	16.3	14.3	28
Unnatural	19.8	30.8	40.3	27	25	24	19.3	33.5
Babbling	1.5	29.3	22.3	22.3	17.8	11.8	12	21.3
Irregular	3.3	24.5	16.3	16.5	19.5	11.3	26	14.5
Interrupted	3.5	11.3	6	9	10.3	6.3	23.8	4.5
Individual means	9.92	30.72	26.06	27.42	15.11	17.5	21.39	21.61
Avg. per system	20.32		26.74		16.3		21.5	

Table 7.8: DAM scores for each Signal sub-category and every system configuration

noise, as it reveals the accelerometric signal's low-pass filtered nature.

Signal Qualities: Effect of transducer and condition

A $9 \times 2 \times 2$ ANOVA test was performed, investigating the effect of sub-categories (9-level factor), transducers (2-level factor) and condition (2-level factor). Data were averaged across subjects. The main effect of sub-category was significant ($F = 21.5$, $p < 0.01$), as were those of transducer ($F = 18.9$, $p < 0.01$) and condition ($F = 13.9$, $p < 0.01$). The interactions between Sub-category \times Transducer and Sub-category \times Condition were significant ($p < 0.01$) unlike the interaction between Transducer \times Condition which was not significant ($F = 6.47$, $p = 0.02$). The three-way interaction was found significant ($F = 5.37$, $p < 0.01$).

Significance by condition suggests that the quiet case produces significantly better quality results than the noisy case (results are averaged for both transducers and all sub-categories). The only obvious conclusion therefore, is that the quiet signals overall are preferred to the noisy ones. Significance by transducer (where results are averaged for both conditions) indicates that the perceived signal quality for the overall reference microphone is significantly better than that of the overall accelerometer. We can also conclude that the quiet reference microphone is significantly better than the quiet accelerometer. However, we cannot conclude that the noisy accelerometer is significantly better than the noisy microphone even if this can be seen from the total result of Figure 7.6. In order to investigate this point, a t-test was performed comparing Noisy Reference to Noisy accelerometer results. Data were averaged across sub-categories and subjects. The result was not significant ($p = 0.13$), suggesting that the quality of the noisy accelerometer is not significantly different to that of the noisy reference microphone. Insignificant interaction between transducer and condition suggests that the effect of noise on the reference

microphone is similar as on the accelerometer for all sub-categories.

Signal Qualities: Effect of filter and condition

A $9 \times 2 \times 2$ ANOVA test was also performed, investigating the effect of sub-categories (9-level factor), filters (2-level factor) and condition (2-level factor). Data were averaged across subjects. The main effect of sub-category was found significant ($F = 8.66$, $p < 0.01$). However, the effects of filter ($F = 1.2$, $p = 0.28$), and condition ($F = 1.13$, $p = 0.3$) were found not significant.

Significance by sub-category suggests that the signal quality results averaged across all filters and conditions varied significantly across the sub-categories. Insignificance by filter (results are averaged for both conditions) indicates that the signal quality rating for the overall optimal filter is rated close to that of the overall high-pass filter, a point that is challenged with the Total Quality rating, below. Insignificance by condition (results are averaged for both filters) suggests that the subjects did judge signal and not background qualities.

The interaction between Sub-category \times Filter was significant ($F = 2.7$, $p < 0.05$), suggesting that for all conditions, the main effect of filter, was significant for all sub-categories. The interaction between Sub-category \times Condition was found significant ($F = 8.17$, $p < 0.01$), suggesting that for the two filters, the main effect of noise created significant differences when rating different sub-categories. The interaction between Filter \times Condition was not significant ($F = 1.38$, $p = 0.25$), indicating that for all sub-categories, the main effect of noise was similar for both filters. The three-way interaction Sub-category \times Filter \times Condition was also significant ($F = 6.9$, $p < 0.01$).

7.6.8 Background Qualities

The results of all the sub-categories in the Background Qualities main category can be seen in Figure 7.7 and also in Table 7.9. Each point represents an average of 20 listeners' responses to 8 DAM sentence groups, giving a total of 160 values. An interesting point for the same figure, is the exceptionally high score for 'hissing', 'roaring' and 'rumbling' assigned to the noisy microphone signal. The quiet reference signal also has a relatively higher score of 'roaring' compared to the quiet and noisy high-pass filter outputs. Those points can only be attributed to the dependence of the test on subjective responses.

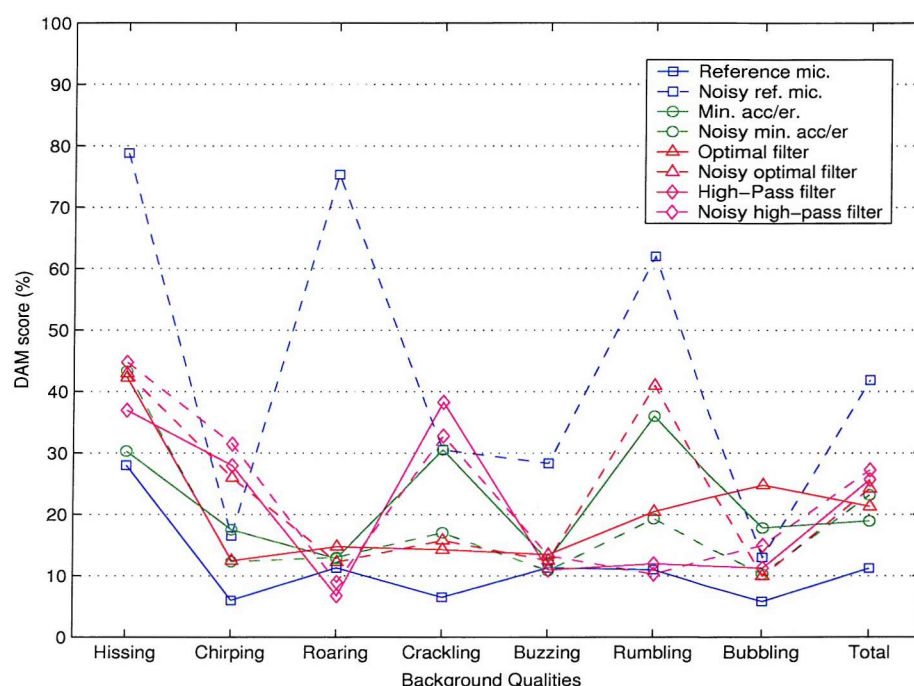


Figure 7.7: Results from the DAM test for Background Qualities. DAM score is extent to which test sentences exhibit the particular background quality.

The aim in this case is to investigate the importance of the sub-categories, and their effect in relation to transducer, condition and filter. The analysis is the same as the one that was performed for the Signal Qualities.

	Ref. mic.	Noisy Ref.	Acc.	Noisy acc.	Opt.	Noisy opt.	HP	Noisy HP
Hissing	28	78.8	30.3	43.3	42.3	43	37	44.8
Chirping	6	16.5	17.5	12.3	12.5	26	28	31.5
Roaring	11.3	75.3	12.8	13	14.8	12.3	6.8	9
Crackling	6.5	30.5	30.5	17	14.3	15.8	38.3	32.8
Buzzing	11.3	28.3	12.5	10.8	13.5	12.5	11	13.3
Rumbling	11	62	36	19.3	20.5	41	12	10.3
Bubbling	5.8	13	17.8	10.3	24.8	10	11.3	15
Individual means	11.21	41.87	18.91	23.15	21.33	24.25	25.7	27.26
Avg. per system	26.54		21.03		22.79		26.48	

Table 7.9: DAM scores for each Background sub-category and every system configuration

Background Qualities: Effect of transducer and condition

A $7 \times 2 \times 2$ ANOVA test was performed, investigating the effect of sub-categories (7-level factor), transducers (2-level factor) and condition (2-level factor). Data were averaged across subjects. All three individual effects were found significant: sub-category ($F = 35.8$, $p < 0.01$); transducer ($F = 18.5$, $p < 0.01$), and condition ($F = 40$, $p < 0.01$).

Significance by transducer (results are averaged across conditions) indicates that the perceived background quality for the overall reference accelerometer is significantly better than that of the overall microphone. The above results were verified with two separate t-tests between microphone and accelerometer for quiet and noisy cases, where the results were found significant ($p < 0.01$). Here, significance of the main effect of condition suggests that the quiet case produces significantly better results than the noisy case. This indicates that the background quality of the quiet microphone is significantly better than that of the quiet accelerometer, and that the noisy accelerometer is significantly better than the noisy microphone as it can be seen from the total result of Figure 7.7.

All two-way interactions were significant ($p < 0.01$) as was the three-way inter-

action ($p < 0.01$). It appears that noise affected the background quality of the reference microphone more strongly than it affected that of the accelerometer, but the amount of this effect varies greatly by sub-category, which is consistent with the significant interactions.

Background Qualities: Effect of filter and condition

A $7 \times 2 \times 2$ ANOVA was also performed, in order to check the effect of sub-categories (7-level factor), filters (2-level factor) and condition (2-level factor). Data were averaged across subjects. The main effect of sub-category was significant ($F = 30.3$, $p < 0.01$). However, the main effects of filter ($F = 0.006$, $p = 0.93$) and condition ($F = 1.34$, $p = 0.26$) were not significant. From the insignificance of the filter factor, we can therefore conclude that the quality rating for the overall optimal filter is not significantly different to that of the overall high-pass filter, a point that is also challenged with the Total Quality rating. Insignificance by condition suggests that the optimal filter achieves some suppression of noise, therefore making the background quality of the noisy signals to be perceived close to that of the quiet ones.

The interactions between Sub-category \times Filter ($F = 19.9$, $p < 0.01$), and Sub-category \times Condition ($F = 4.55$, $p < 0.01$) were significant. The interaction between Filter \times Condition was not ($F = 0.06$, $p = 0.8$). We can thus conclude that for all conditions, optimal filtering had significantly better background quality ratings than high-pass filtering. The three-way interaction was also significant ($F = 7.2$, $p < 0.01$), suggesting that the main effect of changing filters and conditions created significant subjective differences when assessing different sub-categories.

7.6.9 Total Qualities

This assessment in effect is the most direct evaluation of our processed speech, as the subjects had to rate the best signals with the highest score (perceptually easier for a listener). The results of all the sub-categories in the Total Qualities main category can be seen in Figure 7.8 and in Table 7.10, where the optimal filter comes out to be the best rated signal compared to all others. Each point represents an average of 20 listeners' responses to 8 DAM sentence groups, giving a total of 160 values.

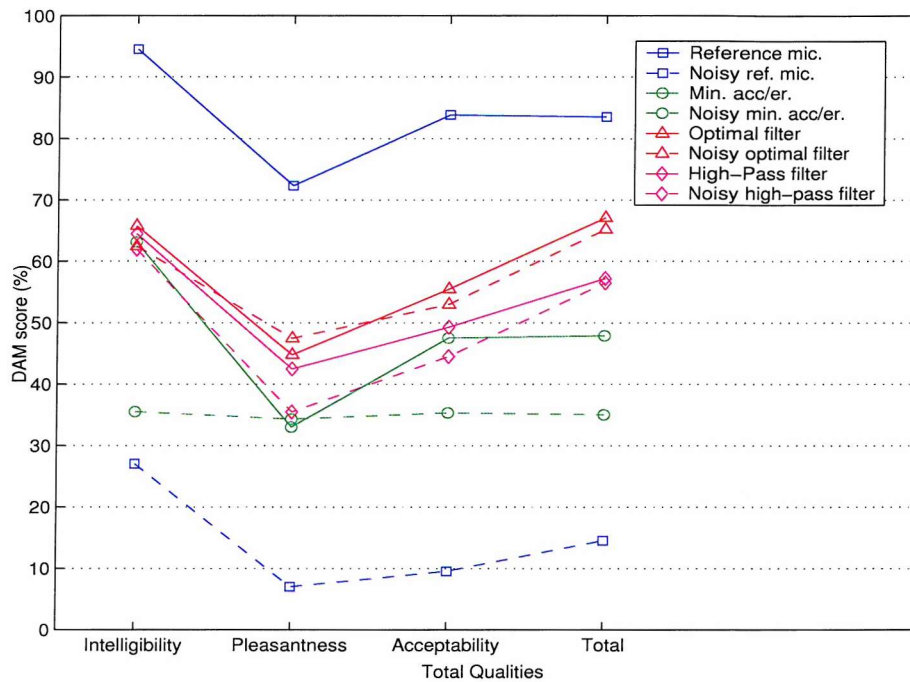


Figure 7.8: Results from the DAM test for Total Qualities. DAM score is extent to which test sentences exhibit the particular total quality.

	Ref. mic.	Noisy Ref.	Acc.	Noisy acc.	Opt.	Noisy opt.	HP	Noisy HP
Intelligibility	94.5	27	63.1	35.5	65.8	62.5	64.5	62
Pleasantness	72.3	7	33	34.3	44.8	47.5	42.5	35.5
Acceptability	83	9.5	47.5	35.3	55.5	53	49.3	44.5
Individual means	83.5	14.5	47.87	35	67.11	65.22	57.22	56.5
Avg. per system	49		41.43		66.16		56.8	

Table 7.10: DAM scores for each Total sub-category and every system configuration

What we will investigate here, is the effect of changing transducer, filter and

condition, on the direct evaluation of the signals. The whole procedure is the same as the Signal and Background Qualities.

Total Qualities: Effect of transducer and condition

A $3 \times 2 \times 2$ ANOVA test was performed, investigating the effect of sub-categories (3-level factor), transducers (2-level factor) and condition (2-level factor). Data were averaged across subjects. The individual main effects of sub-category ($F = 35.8$, $p < 0.01$), transducer ($F = 51.4$, $p < 0.01$), and condition ($F = 308.8$, $p < 0.01$) were all significant. Here, significance by transducer indicates that the perceived total quality for the overall reference microphone is significantly better than that of the overall accelerometer. We can also conclude that the quiet microphone is significantly better than the quiet accelerometer, and that the noisy accelerometer is significantly better than the noisy microphone, as it can be seen from the total result of Figure 7.8. The above results were verified with two separate t-tests between microphone and accelerometer for quiet and noisy cases, where the results were found significant ($p < 0.01$).

All two-way interactions were found significant ($p < 0.05$), suggesting that the main effect of noise was strong enough to create significant differences among the sub-categories, and between the microphone and accelerometer signals.

Total Qualities: Effect of filter and condition

A $3 \times 2 \times 2$ ANOVA was also performed, in order to check the effect of sub-categories (3-level factor), filters (2-level factor) and condition (2-level factor). Data were averaged across subjects. The main effects of sub-category ($F = 82$, $p < 0.01$) and filter ($F = 4.9$, $p < 0.05$) were significant, unlike that of

condition which was not significant ($F = 3.6, p = 0.073$). Interaction between Sub-category \times Filter ($F = 4, p < 0.05$) was significant. All other interactions were not significant.

The insignificance of the main effect of condition suggests that the quiet case is not significantly different from the noisy case for the filtered signals. However, the filtering does make a significant difference, even though this varies somewhat by sub-category as indicated by the significant Sub-category \times Filter interaction. Thus, we conclude that the total quality rating for the optimal filter over both conditions and all sub-categories is significantly better than that of the overall high-pass filter.

Further 2×2 ANOVA tests, comparing the effects of filter vs. condition, were conducted only for the three individual sub-categories of the Total Effect. Data were averaged across subjects only. Results appear in Table 7.11. As seen, the optimal filter improves the overall Pleasantness and Acceptability of the signal compared to the high-pass filter, while the Intelligibility is shown not to have significantly changed.

	Filter	Condition	Filter \times Condition
Intelligibility	$p = 0.17$	$p = 0.17$	$p = 0.85$
Pleasantness	S $p < 0.01$	$p = 0.3$	$p = 0.05$
Acceptability	S $p < 0.01$	$p = 0.08$	$p = 0.58$

Table 7.11: 2×2 level ANOVA results for filter vs. condition (Individual sub-categories of Total Effect) (**S**: significant; p = value: not significant).

7.7 Conclusions

The results from both the DRT intelligibility, and the DAM quality tests were encouraging, since after a statistical analysis using ANOVA's and t-tests, it has been shown that the optimal filter performs overall better than either the high-pass or no-filter condition. More specifically, one of the main points for

the DRT test is that the noisy accelerometer was significantly more intelligible than the noisy reference microphone emphasising the noise-immunity of the accelerometer. It was also shown that the optimal filter was significantly more intelligible than the high-pass filtered and unfiltered accelerometer in both quiet and noisy conditions. Additionally, it was shown that the main effect of speaker (female/male) was not significant when tested with both transducers and filters. For the DAM test, the total effect of the optimal filter's output was rated higher than the high-pass filter output for both conditions. Its signal and background qualities were assessed high enough to be rated second in the subjects' preferences, but were not significantly different from those of the high-pass filter. At the same time the optimal filter was shown to improve the overall acceptability and pleasantness of processed speech.

Chapter 8

Quantitative Assessment of Optimal Filtering Performance

8.1 Introduction

In chapter 7 we showed that, overall, optimal filtering gave better intelligibility and quality results than high-pass filtering. The aim of this chapter, therefore, is a more elaborate analysis of the optimal filters, in order to understand the factors that influence their performance. This is approached initially by presenting and discussing their magnitude and phase responses, in parallel with the spectra of the microphone and accelerometer signals, as well as the output error spectra. The next part is dedicated to the analysis of the block diagram of section 5.3, which is now expanded in order to look into the influence of electronic noise of the system. Finally, the last section investigates the effect of non-causality on the optimal filter performance, by examining the delays imposed by the acoustic/anatomic paths and the equipment.

8.2 Variability of Speech Spectra and Optimal Filter Responses

The aim of this section is to study the variability of the acquired speech spectra and the responses of the optimal filters from the main experiment of Chapter 7. The motivation is to investigate whether a fixed filter across all speech material or across speakers could be a reasonable alternative to a filter generated each time for a different input, as has been done so far.

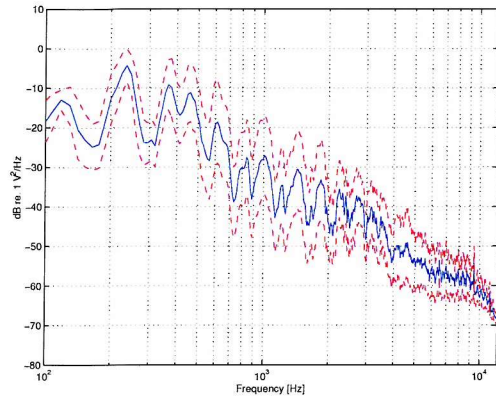
8.2.1 Reference Microphone Signals

Figures 8.1(a) and 8.1(b) show the average spectra of the DRT words for the male and female speakers, respectively. Figures 8.1(c) and 8.1(d) show the average spectra from the DAM sentences for the two speakers. For purposes of convenience, any average spectrum $\overline{S_{xx}}$ that appears in this chapter is the logarithmic mean (see equation 8.1) of the spectra from 192 DRT words or 48 DAM sentences, ± 1 standard deviation, acquired from the transducer under discussion. The mean is represented by a solid, blue line, while the standard deviation is represented by a dashed, red line. Only logarithmic frequency scale plots are presented in this chapter. The corresponding linear frequency scale plots of each figure can be seen in Appendix A.

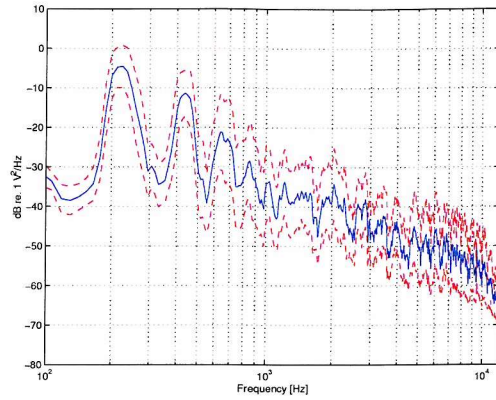
$$\overline{S_{xx}} = 10 \log_{10} \left(\frac{1}{N} \sum_{i=1}^N S_{xx_i} \right) \quad (8.1)$$

where S_{xx} is the power spectral density of the signal x and N is the total number of words or sentences.

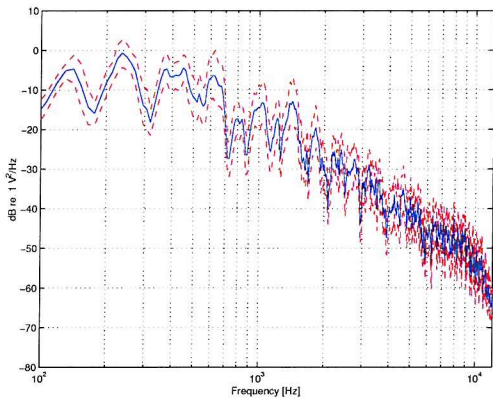
Considering the spectra of the DRT words in Figure 8.1(a,b), peaks corresponding to F0 are clear for both male and female speakers, with typical values at 120 Hz and 220 Hz, respectively (see section 2.2.1). The second and



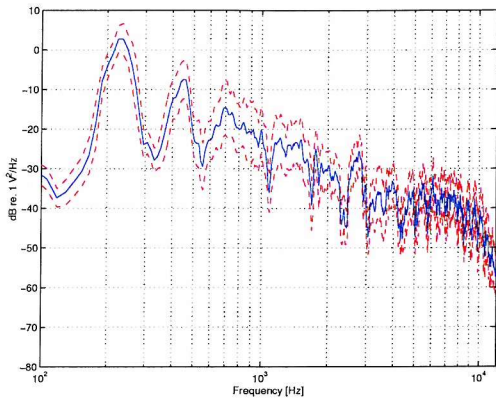
(a)



(b)



(c)



(d)

Figure 8.1: Reference microphone signal spectra for the DRT (top) and DAM (bottom) tests, for male (left) and female (right) speakers. (a) DRT male; (b) DRT female; (c) DAM male; (d) DAM female.

third harmonics are clear for both speakers as well; 2F0 values are at 240 Hz and 440 Hz for male and female, while 3F0 values are around 360 Hz and 650 Hz. The values above were also checked with the Speech Filing System (courtesy of UCL, London, UK), a software package which performs F0 tracking, and confirmed.

The spectra for the DAM sentences follow the same harmonic pattern as the DRT words list, with one main difference. That is, for both speakers, the spectral amplitude for the DAM sentences is higher by 5-8 dB compared to the

DRT words. This can be explained by the higher ratio of speech-to-quiet time in the DAM sentences than in single DRT words. This forces the averaging process to detect more silent segments in the DRT material, therefore lowering the amplitude of the average DRT spectrum. The deviation from the mean is also smaller in the DAM spectra for both subjects for most of the frequency range, since the phonemic content of the sentences is mixed. This results in masking spectral differences of individual sentences, and in essence, of different sentence articulations, making the standard deviation smaller.

8.2.2 Accelerometer Signals

The difference in amplitude that was observed between the spectra of DRT and DAM microphone signals is the same for the accelerometer case. Therefore, only the DAM spectra will be presented here.

Quiet Signals

Figures 8.2(a) and (b) show the average spectra of the quiet accelerometer signals for the male and female speaker, respectively.

As can be seen, the low-pass characteristic of the body-conduction path (part of $G_{an.}$, defined in section 5.3) starts having an effect at about 900 Hz-1 kHz, where the spectrum is attenuated by about 30 dB in comparison to its magnitude at F0, for both male and female (for the reference microphone signals in Figure 8.1, this difference is of the order of 15 dB). The harmonics are still present, although they do not have as high amplitude as in the microphone signal, due to this low-pass filtering effect.

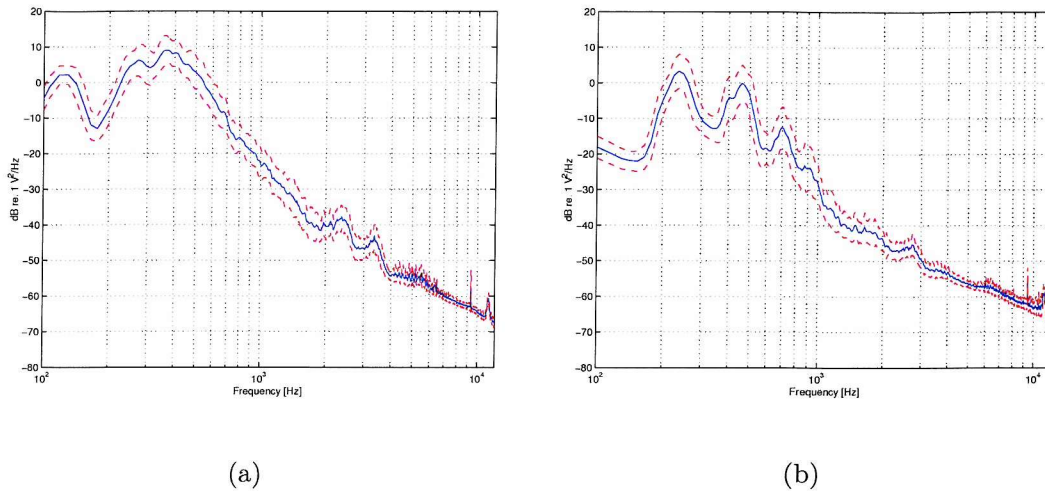


Figure 8.2: Average quiet accelerometer DAM spectra for male (left), and female (right) speakers.

Noisy Signals

The spectra from noisy accelerometer signals, as they appear in Figure 8.3, are also interesting to analyse, since they provide information about the performance of the optimal filters in noise. The amplitude responses are surprisingly similar to the quiet ones up to 1.5 kHz, above which the noise effect becomes apparent. We have to remember that the noise was recorded with the miniature accelerometer placed on the TMJ of the subjects, and then added electronically using Matlab, to the quiet accelerometer signals.

A detail that is present for both subjects, is the distinct peak at 4 kHz. A possible explanation could be the resonance of the subjects' ear canal, when the noise was played back from the loudspeaker. As described in Chapter 2, the ear canal has a resonance at about 3.5 kHz, which could be picked up by the accelerometer that is attached very close to the subject's ear canal. The amplitude of this peak is about 20 dB relative to surrounding frequencies, a point that is supported by the information given in Chapter 2, describing the Transfer Function of the Open Ear.

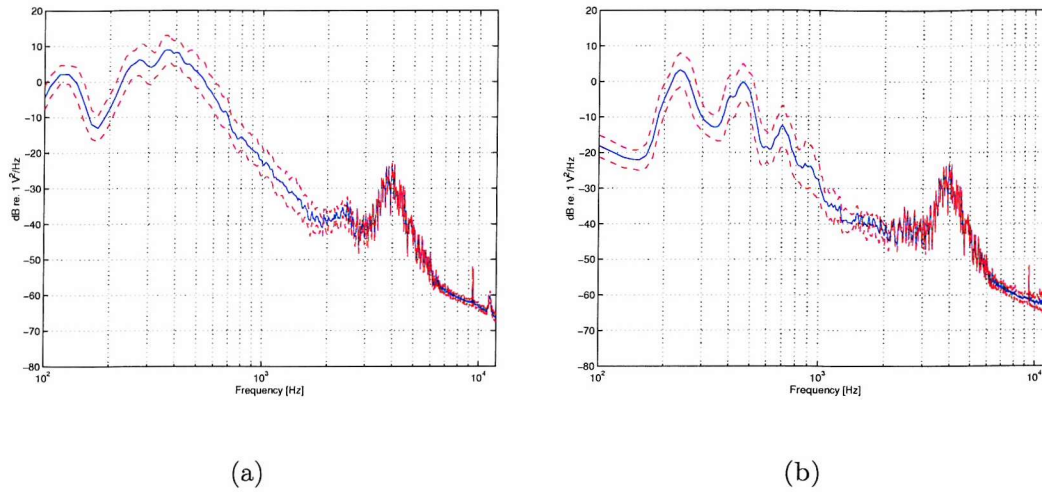


Figure 8.3: Average noisy accelerometer DAM spectra for male (left), and female (right) speakers.

8.2.3 Optimal Filter Responses

The aim of this section is to present the magnitude and phase responses of the optimal filters in parallel with the output error spectra, generated for the quiet and the noisy signals. The motivation is to investigate the performance of the filters in different frequency regions, to interpret a reduced performance, and to find some common characteristics that could potentially set a goal for a better filter design.

Quiet Case

Figure 8.4 shows the average magnitude and phase responses of the optimal filters for the DAM quiet accelerometer signals for both speakers.

It is not difficult to see that the optimal filter magnitude responses have, approximately, the structure of a high-pass filter up to about 1 kHz, below which they try to reconstruct the harmonics in the accelerometer signal. W_{opt} 's are derived for every sentence individually. Each one, therefore, takes into account the impulse responses of the two transfer paths, $G_{acoustic}$ and $G_{body-conducted}$

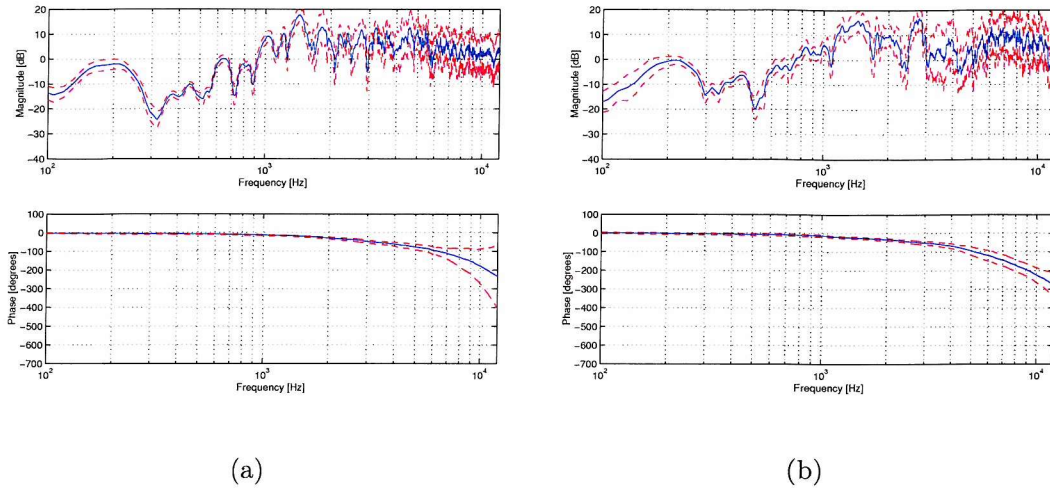
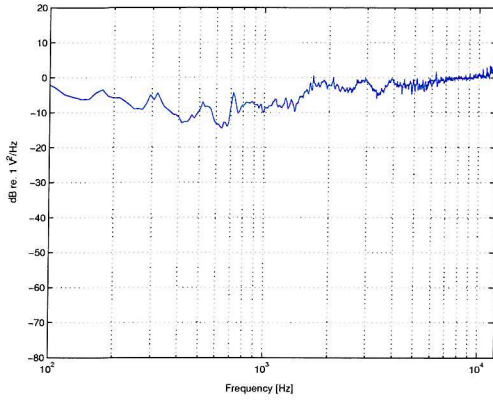


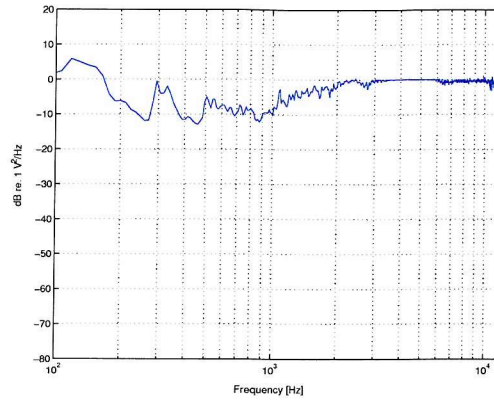
Figure 8.4: Average quiet W_{opt} magnitude and phase responses for the DAM test for male (a), and female (b) speaker.

as in Figure 5.2. It eventually tries to match the accelerometer signal to the reference microphone signal.

Overall, for both speakers, the variability of the filters' magnitude and phase responses increases above 900 Hz-1 kHz, indicating that W_{opt} 's are very different at higher frequencies. A reason is that optimal filtering is based on a mean-square error minimisation criterion. Higher frequency accelerometer signals above 1-1.5 kHz have very little energy. Therefore, W_{opt} 's perform better at lower frequencies, where the mean square error is concentrated. This is shown in Figure 8.5, which shows the average error spectra from the two subjects. As seen, the filters have actually enhanced the accelerometer signal by matching it reasonably well to the microphone signal (indicated by a negative error), from about 100 Hz-2 kHz, for both speakers. Above approximately 2 kHz, the match is not good, therefore the performance is expected not to be good. Small peaks in the error spectra, indicating an imperfect match, may be due to the presence of an anti-resonance in the accelerometer speech spectra.



(a)

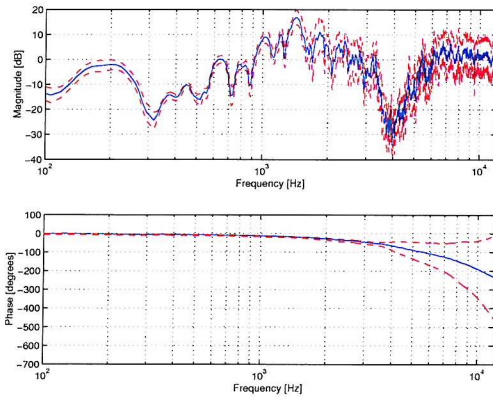


(b)

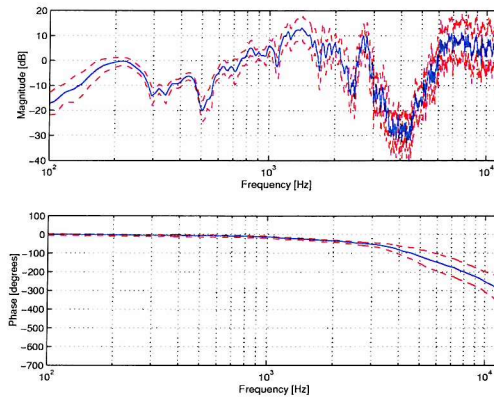
Figure 8.5: Average spectrum of quiet \bar{e}^2 for DAM sentences for male (a), and female (b) speaker.

Noisy Case

Figure 8.6 shows the average magnitude and phase responses of the filters generated for the DAM noisy accelerometer signals for the two subjects.



(a)



(b)

Figure 8.6: Average noisy W_{opt} magnitude and phase responses for the DAM test for male (a), and female (b) speaker.

What was inferred for the quiet W_{opt} 's is valid for the noisy case as well, with the only difference being a high-bandwidth trough in the filter's magnitude response at around 3.5-4 kHz. This can be ascribed to the presence of the peak

that was observed in Figure 8.3. However, it does not affect the performance of the optimal filter, which, from the error spectra in Figure 8.7, appears to enhance the accelerometer signal in the frequency range 100 Hz-2 kHz, similar to the quiet case.

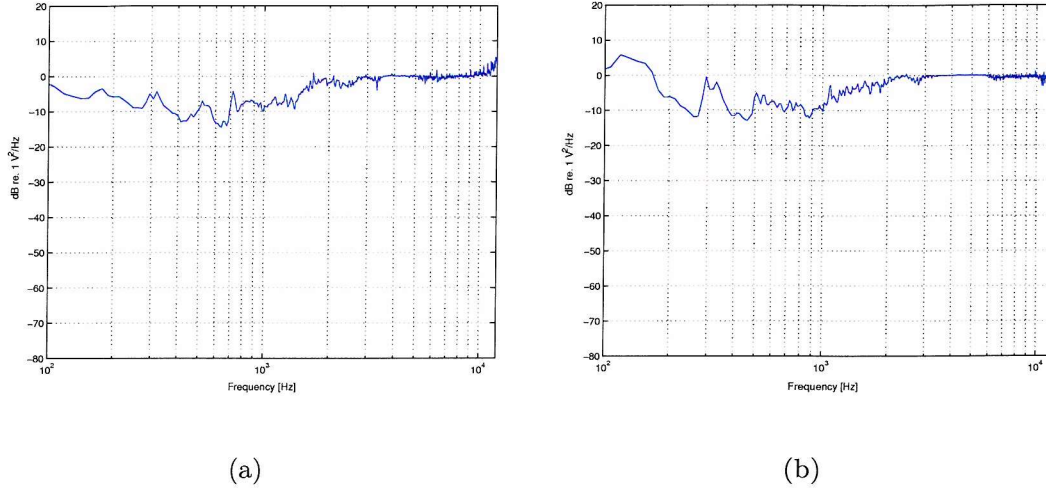


Figure 8.7: Average spectrum of noisy \bar{e}^2 for DAM sentences for male (a), and female (b) speaker.

For both quiet and noisy cases, W_{opt} 's appear to be effective in the range 100 Hz to 1-2 kHz, where their variability is small. The limit in this range (100 Hz-1 kHz) is due to the low amplitude of body-conducted speech at higher frequencies (above 1-2 kHz), and from the effect of electrical noise of the system, a point that is investigated in the next section. Their main function in this range is high-pass filtering of the accelerometer signals, and reconstruction of the harmonics. However, an important question that arises is whether a fixed W_{opt} in this range has the potential to do equally well, since the filters vary much more by corpus elements than by speaker.

8.3 Factors Influencing the Performance of Optimal Filtering

A question that has arisen from section 8.2.3 concerns the possible factors that influence and limit the performance of W_{opt} 's above 1 kHz. The factors to be investigated are the following:

1. Effect of electronic noise on the coherence of the system.
2. Effect of non-causality of the system.

The aim of this section is to provide answers to these points. By verifying (1) and excluding (2), we can say that (1) is the main factor limiting performance.

8.3.1 Effect of Electronic Noise

The model of Figure 5.3 did not include the electronic noise of the accelerometer. A revised model including the noise is shown in Figure 8.8. The motivation here is to investigate whether the accelerometer's electronic noise is the significant factor affecting the system's coherence, and in extent, the performance of W_{opt} above 1 kHz. For our analysis, the noise $r(n)$ is assumed uncorrelated to the signals $d(n)$ and $x(n)$.

According to Bendat and Piersol (1993), the following frequency-domain rela-

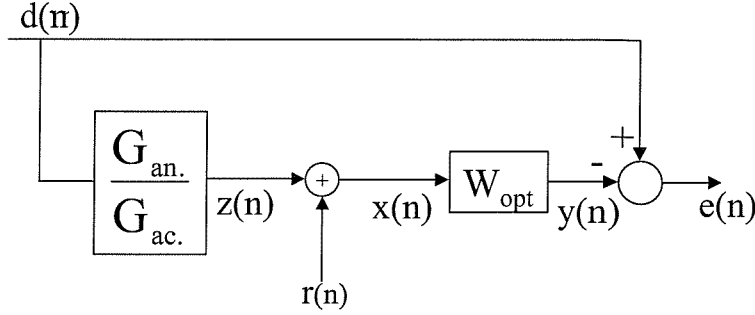


Figure 8.8: Physical model of the optimal filtering problem including the effect of electronic noise.

tionships can be applied:

$$\begin{aligned}
 S_{dr} &= S_{zr} = 0 \\
 S_{xx} &= S_{zz} + S_{rr} \\
 S_{zz} &= \left| \frac{G_{an.}}{G_{ac.}} \right|^2 S_{dd} \\
 S_{dx} &= S_{dz} = \frac{G_{an.}}{G_{ac.}} S_{dd} \\
 S_{xd} &= \left(\frac{G_{an.}}{G_{ac.}} \right)^* S_{dd}
 \end{aligned} \tag{8.2}$$

For this model, the coherence function between the two output records $x(n)$ and $d(n)$ is given by:

$$\gamma_{xd}^2 = \frac{|S_{xd}|^2}{S_{xx}S_{dd}} \tag{8.3}$$

Based on equation 8.2 and after some mathematical manipulation, we end in a generalised coherence function expression for our model:

$$\gamma_{xd}^2 = \frac{S_{zz}}{S_{xx}} = 1 - \frac{S_{rr}}{S_{xx}} \tag{8.4}$$

The optimal filter expression from equation 5.22 will become (Rafaely and Furst, 1996):

$$W_{opt} = \frac{S_{xd}}{S_{xx}} = \frac{G_{ac.}}{G_{an.}} \cdot \frac{S_{zz}}{S_{zz} + S_{rr}} \tag{8.5}$$

The measured coherence between $x(n)$ and $d(n)$ was compared to the model's coherence based on equation 8.4. The accelerometer's electronic noise, as

measured in Chapter 7, was used as the signal $r(n)$. Figure 8.9 shows the average measured coherences for the two subjects and tests, compared to the model's coherences.

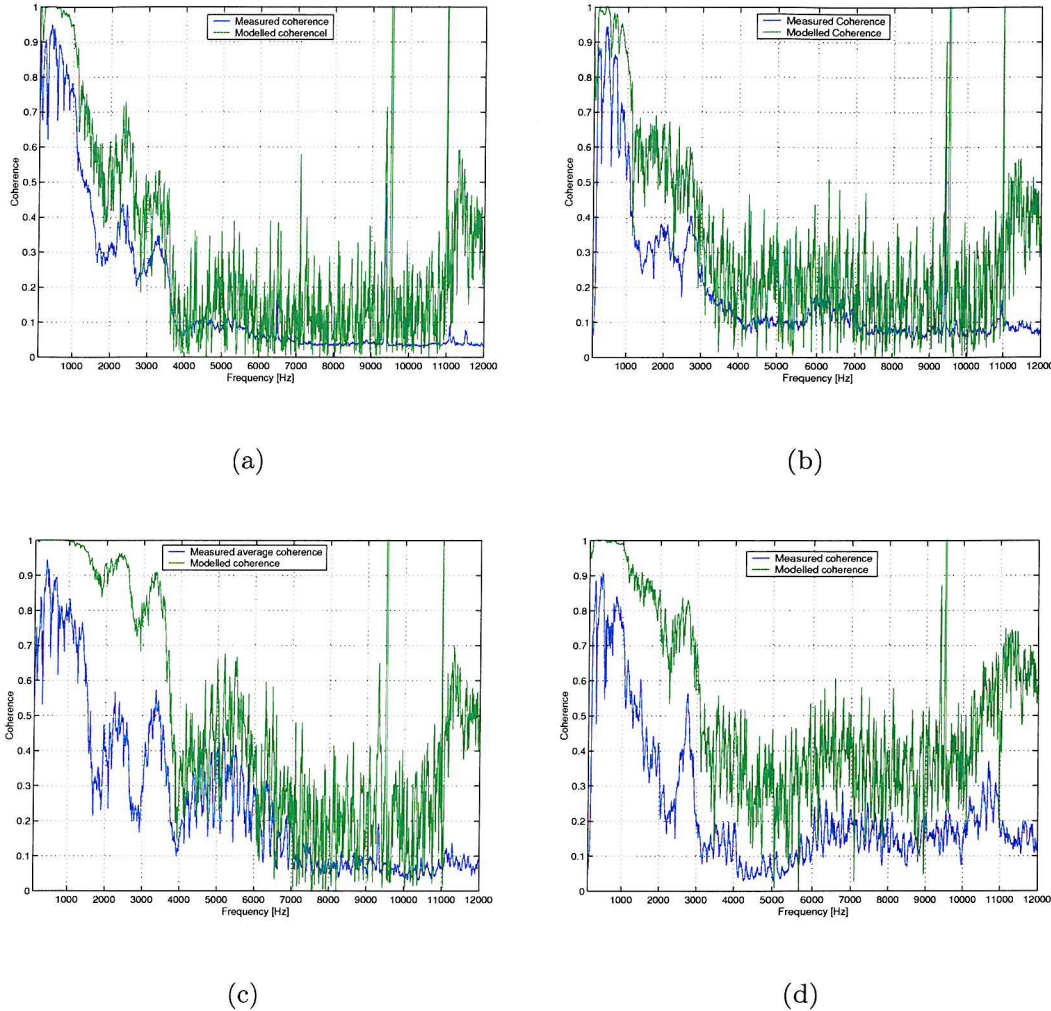


Figure 8.9: Average measured vs. modelled coherence functions for DRT words: (a) Male speaker, (b) Female speaker. For DAM sentences: (c) Male speaker, (d) Female speaker. Blue line: measured coherence; green line: modelled coherence.

The coherence plots can be split into three main regions. The first region is from the lower frequency end up to about 1 kHz, where both coherences are satisfactory, with values approaching 1. This is also reflected in the good performance of W_{opt} 's in this band. From 1 kHz up to about 3 kHz, the coherence drops due to the increasing effect of electronic noise. At higher frequencies, both the measured and modelled coherences are low, corresponding

to the continuous fluctuation of the filters' magnitude responses. The above observations support our hypothesis, that the main limitation of performance at high frequencies is the added electronic noise and not the non-linearity of the system. The conclusion, therefore, is that even though the model predicts our case quite well, we cannot really attain a better match, as above 1 kHz the electronic noise is dominant. A solution, therefore, is to obtain a better accelerometer and amplifier with a lower noise floor.

8.3.2 Effect of Non-Causality

A non-causal optimal filter, resulting from an acoustic transfer function group delay shorter than the anatomical transfer function's group delay, is more difficult to implement (as W_{opt} in practice is restricted to be causal), since the problem in this case is not filtering but *prediction*. The objective of the optimal filter is then to minimise the mean-square error by making the output of the filter as close as possible to the current desired signal, using only past values of the signal, i.e. to act as a linear predictor for the desired signal. The aim of this section is the investigation of the effect of the system's path-delays. The motivation is to see whether non-causality was a factor for the reduction in the W_{opt} performance.

As was seen in Chapter 5, W_{opt} is based on physical signals and was initially constrained to be causal. The first step was to investigate the effect of an added delay on the reference signal, effectively modifying that constraint. The first delay value that was chosen was the acoustic delay from the speaker's lips to the microphone (see Appendix B for details concerning the calculations). The resulting impulse responses of an unmodified and a W_{opt} artificially delayed by that value (1.74ms), appear in Figure 8.10.

The added delay can be clearly seen on the right plot, verifying the system's

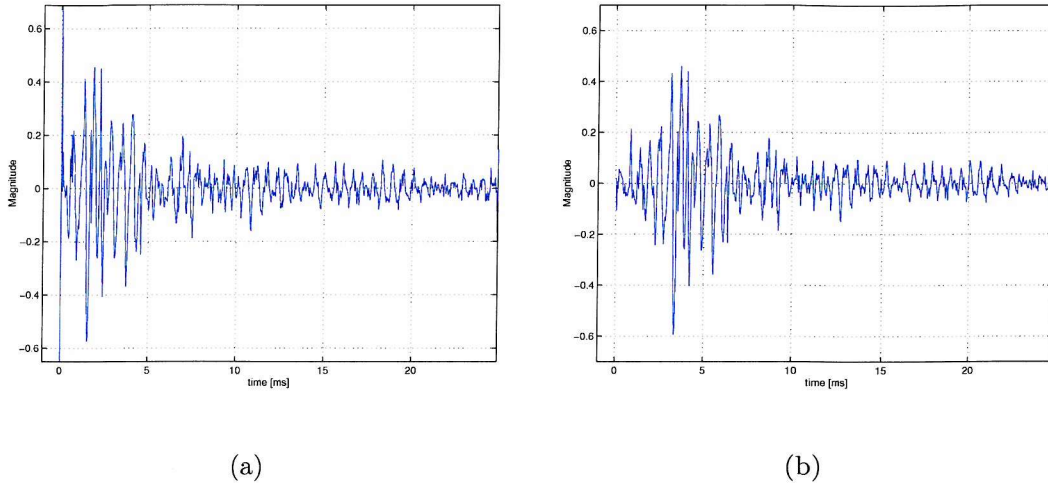


Figure 8.10: Impulse responses of (a) an original, and (b) an optimal filter artificially delayed by 1.74ms, generated for the sentence ‘Dirt was blown in my face’.

causality. It has to be noted though, that the two impulse responses are not exactly the same. This is due to the fact that the cross-correlation S_{xd} of the delayed version is different to that of the unmodified version, since the reference signal is delayed, while the accelerometer signal is not. Therefore, some components which are seen on the left plot, such as the spike at the very beginning of the response, are removed in the delayed version.

The second step, therefore, towards an understanding of the delay characteristics of the filters was to investigate the average phase response of W_{opt} in Figure 8.4. Approximate values of group delays ranging from 1.7ms-2.2ms were calculated with Matlab within the useful range 100 Hz-1 kHz. This shows that W_{opt} delayed the signal, suggesting that it was not constrained by non-causality.

A further step was to investigate the cross-correlation between the microphone and the accelerometer signals. Figure 8.11 shows two characteristic plots of the cross-correlation function R_{xd} , derived from the same sentence (‘Dirt was blown in my face’), for both male and female. Cross-correlations were calculated after the two signals were low-pass filtered with an FIR Butterworth filter, $f_c = 600\text{Hz}$), whose characteristics approached $\frac{G_{an}}{G_{ac}}$ as it appeared in

Figure 5.4. Both plots have a fairly broad envelope, with characteristic fluctuations, indicating delayed harmonics in the air-conducted signal. Not all cross-correlation plots are presented here; however, they have been examined and in all cases the air-conducted signal is delayed relative to the accelerometer by an average of 2.2ms for both male and female for the DRT words, while for the DAM sentences the average delay was 1.9ms for the male and 1.8ms for the female speaker.

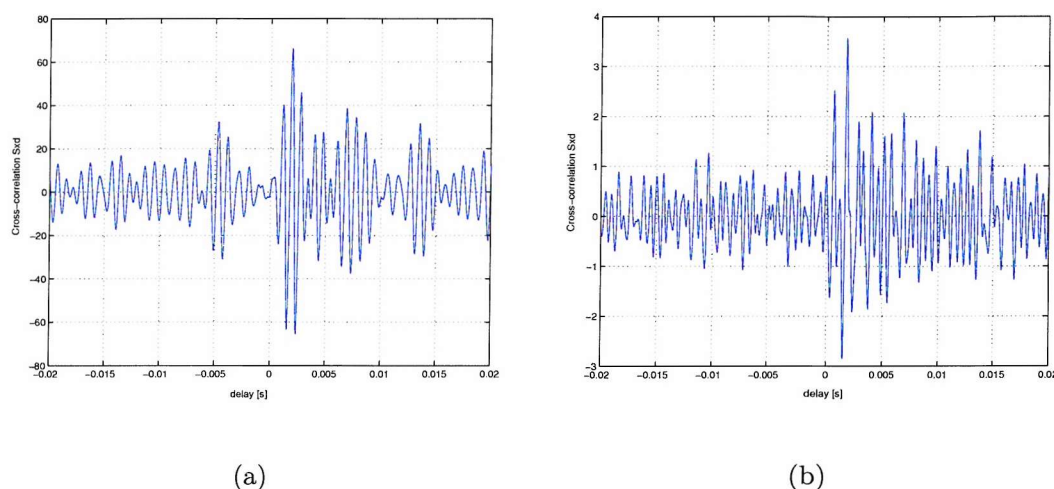


Figure 8.11: R_{xd} 's from the sentence 'Dirt was blown in my face' for male (a), and female (b) speakers.

Another approach was the theoretical computation of the delays for the individual paths of $G_{ac.}$ and $G_{an.}$ (see description in section 5.3).

The total theoretical delay of $G_{ac.}$ is the sum of the individual delays due to:

- The distance between the speaker's larynx and lips.
- The distance between the speaker's lips and the microphone.
- The group delay $-\frac{\partial \varphi}{\partial \omega}$ of the microphone's amplifier, and
- The group delay of the anti-aliasing filter.

Additionally, the total delay of $G_{an.}$ is the sum of the individual delays due to:

- The distance between the speaker's larynx (Adam's apple) and TMJ.
- The group delay $-\frac{\partial\varphi}{\partial\omega}$ of the accelerometer's amplifier, and
- The group delay of the anti-aliasing filter.

The delays due to distance were calculated approximately, with the kinematics law:

$$time = \frac{distance}{speed} \quad (8.6)$$

Typical values for the speed of sound in different materials, considered in our calculations, appear in Table 8.1. Values for the average length of the vocal tract that we used were 17cm for men, and 15.5cm for women.

air (20°C)	vocal tract (37°C)	human tissue (37°C)
343	359	1540

Table 8.1: Speed of sound (m/s) in different materials.

The results from the above approaches appear in Table 8.2. The first column shows the results obtained from the study of the cross-correlations, for the two subjects and tests. The second column gives the results obtained from the phase of W_{opt} . The last two columns give the results from the theoretical calculation of the delays in the two main transfer paths (details of the calculations are shown in Appendix B). The smaller delay values in the DAM test may possibly be due to a slight change of the subjects' seating position after the completion of the DRT test.

		R_{xd}	$-\frac{\partial\varphi}{\partial\omega}$ of W_{opt}	$G_{ac.}$	$G_{an.}$
Male	DRT	2.2ms	2.1ms	2.2ms	0.18ms
	DAM	1.9ms	1.8ms		
Female	DRT	2.2ms	1.7ms	2.2ms	0.18ms
	DAM	1.8ms	1.8ms		

Table 8.2: Total delays from three measurement approaches.

8.4 Summary and Conclusions

This chapter has contributed towards a deeper understanding of the physical mechanisms that govern the performance of the optimal filters. More specifically, an initial investigation of the input signals showed that the low-pass filtering effect of body-conduction attenuates the signals strongly above 1 kHz. Results from a comparison between measured and modelled coherences showed that electronic noise was the most dominant limiting factor above 1 kHz, therefore reducing the performance of W_{opt} . The model also predicted the acoustic path delays, with delay values approaching those calculated from cross-correlations. In fact, the calculated delay from G_{ac} was found to be higher than the delay of $G_{an.}$, as expected. W_{opt} 's appeared to have a small variability at lower frequencies for the same speaker, supporting the fact that the main factor that determines their performance is the ratio $\frac{S_{xd}}{S_{xx}}$. Table 8.3 encompasses all the above points.

	$f < 1 - 2kHz$	$f > 2kHz$
Coherence	Good	Poor
High-Pass filtering	Yes	No
Harmonic Reconstruction	Yes	No
Variability of W_{opt} 's	Small	Large
Variability between speakers	Large	Large

Table 8.3: Main conclusions from the study of W_{opt} 's.

It was also shown that non-causality did not constrain the function of W_{opt} . This point was approached with path-delay measurements, which showed that the air-conducted signals were delayed with reference to the accelerometer signals. The overall conclusion, therefore, is that a filter that captures the characteristics of a speaker appears to have a better performance than a filter which is based on the spectral information of a speech corpus, or particular speech sounds.

Chapter 9

Design and Assessment of a Fixed Optimal Filter

9.1 Introduction

The analysis of the optimal filtering model in Chapter 5 posed the question of whether an optimal filter that is designed for specific phonetic material performs better than a fixed optimal filter that is designed for one speaker. This question was also raised in the objective analysis of the optimal filters' responses in Chapter 8, where the small variability of W_{opt} for one speaker suggested that a fixed filter across a set of speech material or a number of speakers could be useful. The aim of this chapter, therefore, is to present the design and subjective assessment of different types of fixed optimal filters. The first section discusses the motivations behind the need for each type of filter. The next section presents the speech recordings and the design of the filters. The chapter then proceeds with the presentation of the subjective assessment of the filtered and unfiltered speech signals, while the last section is devoted to the statistical analysis of the results.

9.2 Motivation for the Design of a Fixed Filter

The main motivation for the development of a fixed optimal filter is to find a good compromise between the complexity of the filter's implementation and its performance. The filter tested in Chapter 7 which is optimised for every different word, might not be easy to implement in practice. Assuming that in practice a clean reference microphone is not available, then a word-specific filter will require a complex system to identify the word and adapt or select an appropriate filter. As an alternative, a fixed filter which might only require initial tuning to the speaker, is significantly simpler to implement. We propose the following hypotheses:

1. A different W_{opt} for every speaker and different speech material will perform well. This is the type of optimal filter we have been testing up to this point, i.e. different W_{opt} 's are created for different words or sentences recorded by a specific speaker. Implementation complexity is high due to the rapid adaptation/selection required as discussed above.
2. A W_{opt} designed specifically for a speaker will perform almost as well. This type of fixed filter is tested in this chapter and consists of an average across all W_{opt} 's for the speech material recorded by a specific speaker. Implementation of a fixed filter is simple, and requires only initial off-line tuning.
3. A fixed W_{opt} for a number of speakers and speech material will not perform as well as the previous filters. This type of fixed filter is also tested, and consists of an average across all W_{opt} 's generated for words or sentences recorded by different speakers. Implementation here is simpler as the fixed filter does not even require a speaker-specific tuning.

The above points are encompassed in Table 9.1. The objective, therefore, is to show that option (2) and possibly option (3) are good compromises of performance vs. complexity.

	Across speakers	Within a speaker/ speech material	Expected performance	Complexity of implementation
(1)	Variable	Variable	Good	High
(2)	Variable	Fixed	Average	Medium
(3)	Fixed	Fixed	Poor	Low

Table 9.1: Main hypotheses for the design of a fixed W_{opt} .

9.3 Speech Recordings

In order to validate the above hypotheses, it was decided to follow the same experimental framework as in Chapter 7, but this time involving more speakers. The experimental setup was the same as that described in section 6.3.3. Eight native British-speaking subjects were used for the recording of the speech material (4 males, and 4 females). This was considered a sufficient number in order to test the performance of a fixed filter across one or more subjects (points 2 and 3 in Table 9.1), and to keep the statistical analysis to a reasonable length. The phonetic material that was used consisted of the DRT and DAM tests, as in Chapter 7. The recording procedure of the material was the same, i.e. the same acquisition and randomisation programs were used (see section 7.3.1). Only quiet signals were recorded at this stage as it was shown in Chapter 7 that the effect of noise was not a significant factor for the performance of optimal filtering.

9.4 Design of the Optimal Filters

The aim of this section is to present the main design procedures of the five different types of optimal filters. The rationale is to understand the mechanism with which these filters were generated, since this will assist in the interpretation of the results from the listening tests. For the processing of the acquired speech signals, we used Matlab programs which generated the filters, performed the filtering operations and produced the output sound files which were used in the listening tests. The general filtering procedure is shown in Figure 9.1, where $W_{opt\ i}$ indicates the different types of optimal filters described below.

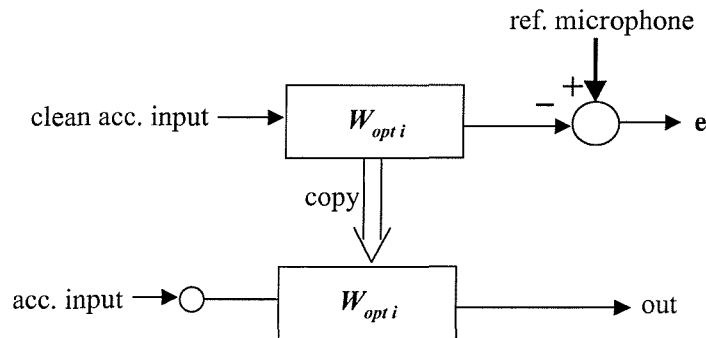


Figure 9.1: Schematic diagram of the Matlab program for optimal filtering. Top part: Optimal filter calculation. Bottom part: Generation of filtered output.

W_{opt} : Optimal Filter for Single Words or Sentences

This optimal filter is the type of filter we have been using up to this point. The mechanism is the same as the one used in Chapters 6 and 7, i.e. the filter tries to match the TMJ accelerometer signal to the reference microphone signal, in a mean square error sense. The processing program calculates the auto- and cross-correlation matrices between the inputs from the miniature accelerometer and the reference microphone signals, and retrieves the Toeplitz matrix, with which an optimal filter is calculated. Six hundred coefficients are

used, and for each input (DRT word, or DAM sentence) a different optimal filter is generated. An example of the magnitude and phase response of W_{opt} derived from a male and a female subject, for one DAM sentence (*'Dirt was blown in my face'*), is shown in Figure 9.2.

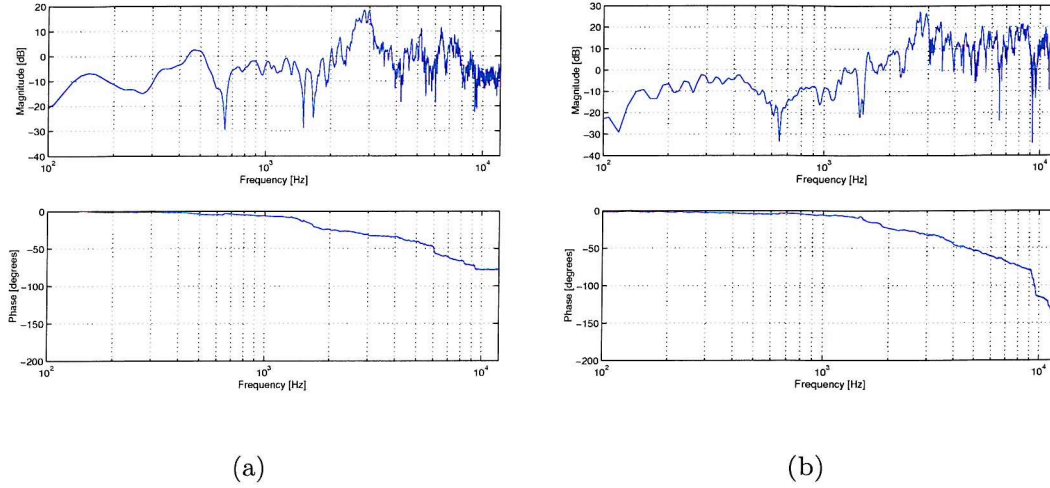


Figure 9.2: Magnitude and phase responses of a W_{opt} for (a) a male subject and (b) a female subject, for the same DAM sentence, (*'Dirt was blown in my face'*).

The main advantage of this filter is good performance, since a different filter is generated for a specific input. However, in the context of a real-life communication system it is fairly complex, since it requires a highly adaptive implementation. In order to switch different filters in and out, a recognition system would need to be used, which would increase the computational complexity substantially.

$W_{opt1sub}$: Optimal Filter for One Speaker

For the derivation of a fixed filter for a single speaker, we calculated the average S_{xx} of the accelerometer signal from all 192 DRT words, and 48 DAM sentences spoken by one speaker, by averaging the individual S_{xx} 's. The same procedure was performed in order to find an average S_{xd} . By calculating their Inverse Fourier Transforms, the average auto- and cross-correlation matrices

were found. The last stage involved the calculation of the Toeplitz matrix using those average R_{xx} and R_{xd} matrices, and the calculation of the filter coefficients. The same filter length was used, i.e. six hundred coefficients. At the end, we had 8 different $W_{opt1sub}$'s, one for every speaker. An example of $W_{opt1sub}$ for a male and a female subject is shown in Figure 9.3.

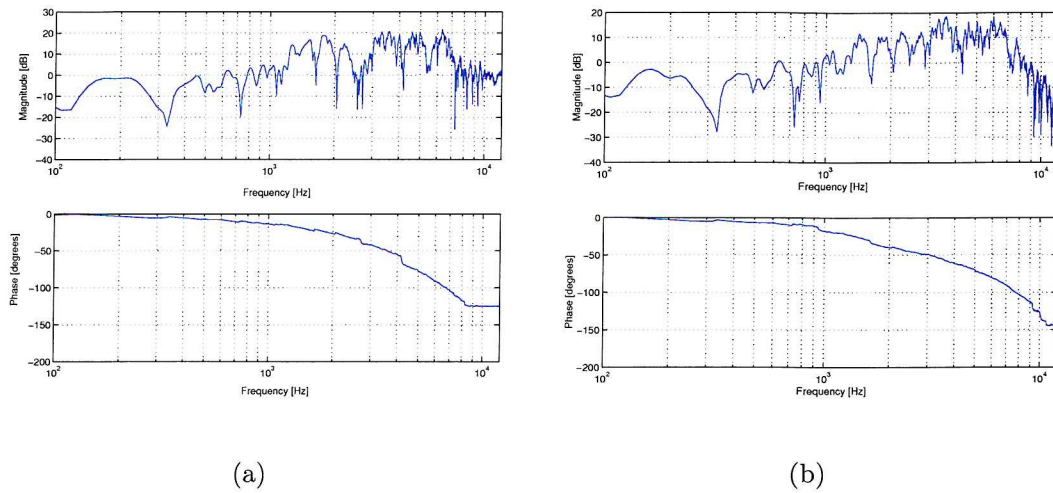


Figure 9.3: Magnitude and phase response of a $W_{opt1sub}$ for (a) a male and (b) a female subject.

The motivation for the design of this filter is to see whether it can perform equally well to W_{opt} . The advantage is that it could be easily implemented in a real-life communication system. The system could include a microphone in front of the speaker's mouth serving as the desired signal, and an accelerometer as the main speech vibration pick-up element. Subsequently, the speaker could recite a pre-defined speech corpus, a process that could generate the filter coefficients specifically formulated for his/her articulatory and anatomical characteristics. We anticipate that its performance will not be as good as W_{opt} , something that is tested with the listening tests.

$W_{opt8sub}$: Optimal Filter for Eight Speakers

This type of fixed filter was generated in the same way as $W_{opt1sub}$. The only difference is that for the averaging of the auto- and cross-spectra we used the material from all 8 speakers, i.e. 1536 DRT words (8 speakers \times 192 words each), and 384 DAM sentences (8 speakers \times 48 sentences each). The filter's response is shown in Figure 9.4.

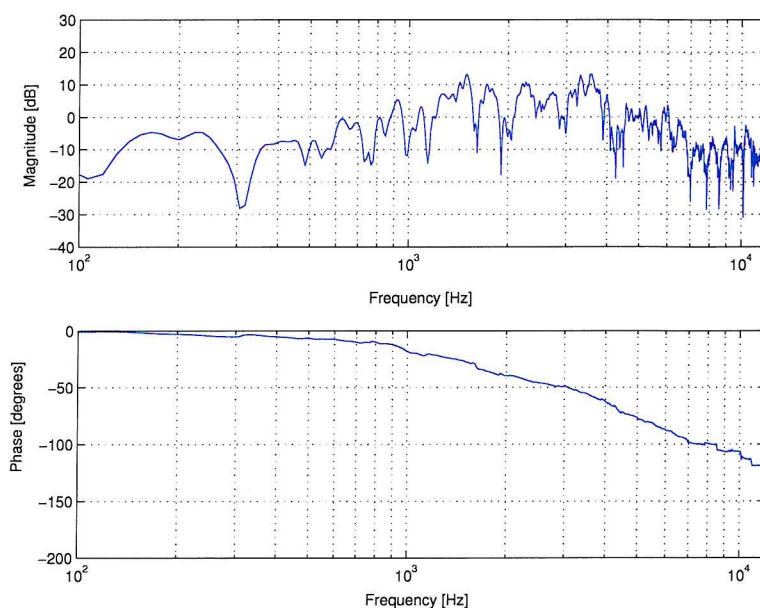


Figure 9.4: Magnitude and phase response of $W_{opt8sub}$ for eight subjects.

The motivation for the design of this filter is the great simplicity in a real-life implementation as it does not require speaker-specific tuning. The main advantage is that it can be directly used as an enhancement algorithm in a communication system which involves an accelerometer without a reference microphone. The reason is that it is designed to encompass the articulatory and anatomical characteristics of a number of speakers. This, however, is expected to perform poorly, since each speaker has identifiable characteristics which are unique for him/her. Therefore, a filter like this is not expected to present the performance of W_{opt} or even $W_{opt1sub}$, but nevertheless, is a viable option which deserves some investigation.

W_{optmal}/W_{optfem} : Optimal Filters for Male and Female Speakers

W_{optmal} and W_{optfem} were generated in the same way as $W_{opt1sub}$, but the averaging was done for the male and female speakers, respectively. The averaging process, therefore, used 768 DRT words (4 male/female speakers \times 192 words each), and 192 DAM sentences (4 male/female speakers \times 48 sentences each). The responses of the two filters are shown in Figure 9.5. The motivation for the

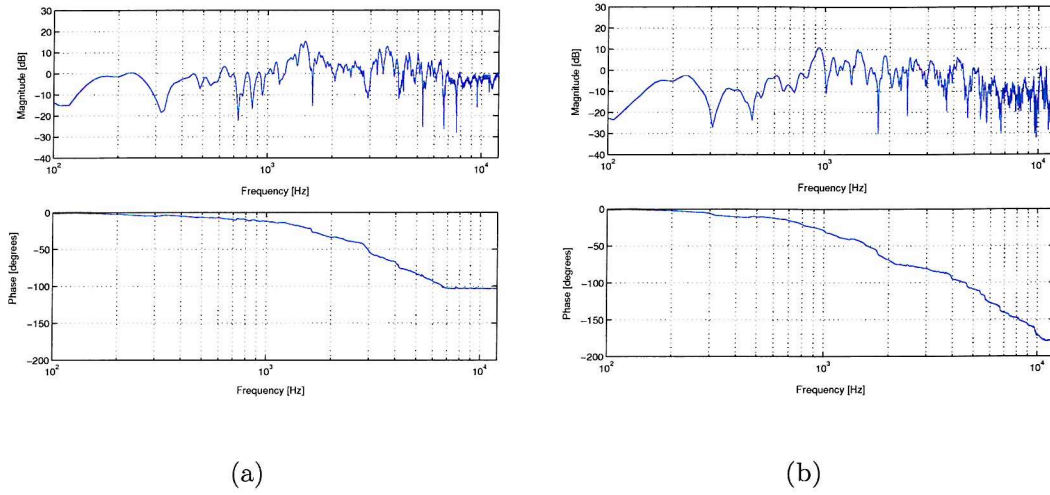


Figure 9.5: Magnitude and phase responses of W_{optmal} (a), and W_{optfem} (b).

design of this type of filter is to investigate the optimal filter's dependence on the gender of the speaker. Since we saw in Chapter 5 that the filter is defined by the anatomical characteristics of the speaker, it would be interesting to investigate the performance of a filter designed specifically for gender-dependent G_{ac} and G_{an} .

9.5 Subjective Assessment of the Optimal Filters

After the signals were processed according to Figure 9.1, we performed the same subjective listening tests as in Chapter 7, using twenty British-speaking listeners. By this time, for every word/sentence there were six different versions:

1. Unfiltered reference microphone signal.
2. Unfiltered miniature accelerometer signal.
3. Output from W_{opt} .
4. Output from $W_{opt1sub}$.
5. Output from $W_{opt8sub}$.
6. Output from W_{optgen} . W_{optgen} , standing for gender-specific filter means that W_{optmal} was used to filter speech from the four male speakers, and similarly W_{optfem} from the four female speakers.

The listening test procedure was designed to present mixed male and female signals, randomised in such a way so that for each word pair the following conditions applied: the same version of the same word should not be heard twice; the same word should not be heard twice in a row; half of the presentations of words would be male and half female; subjects would listen to the material from all 8 speakers; for the gender-specific filters, half would be the outputs from W_{optmal} and half from W_{optfem} ; and finally, those presentations would be randomised both as far as the word in the pair as well as the feature.

The overall listening procedure was the same as the one followed in Chapter 7. In total, every subject listened to 32 words in each category \times 6 categories

$\times 8$ speakers $\times 7$ versions for each word = 10752 words. The overall DRT test lasted about 4.5h during which, subjects took a number of breaks if they felt tired. The same training phase (i.e. “*anchoring*”) was performed prior to the test so that the subjects had an overall impression of the sound samples. A reward of £5 was given to them upon completion of the test.

For the DAM test, the constraints were not so strict, as for each type of filter output a whole group had to be presented to each subject. The only randomisation was between the sentences of each group and the speaker from which they were cited. After each group presentation had finished, the subject was given the quality evaluation form to fill in, as shown in Figure 4.1. The DAM test lasted about 30 minutes, as each sentence group lasted approximately 1 minute and the subject had to listen to 7 groups (one for every signal version) including the time to assess each one by completing the DAM response scale form. The overall listening test lasted for about 6h (about 15 sessions for each subject), including the pauses when the subjects took a break.

9.6 Results from Listening Tests

As in Chapter 7, the results from the listening tests were gathered in a database in SPSS. The use of ANOVA’s and t-tests enabled us to derive conclusions about the effect that the type of filter and transducer had on the intelligibility and quality results.

9.6.1 Results from DRT Test

The results of the DRT test, averaged across listeners, are shown in Figure 9.6 and Table 9.2. The legend of the graph describes sensor and filter combina-

tions. Each point represents an average of 20 listeners' responses to 16 DRT word pairs from 8 speakers, giving a total of 2560 values.

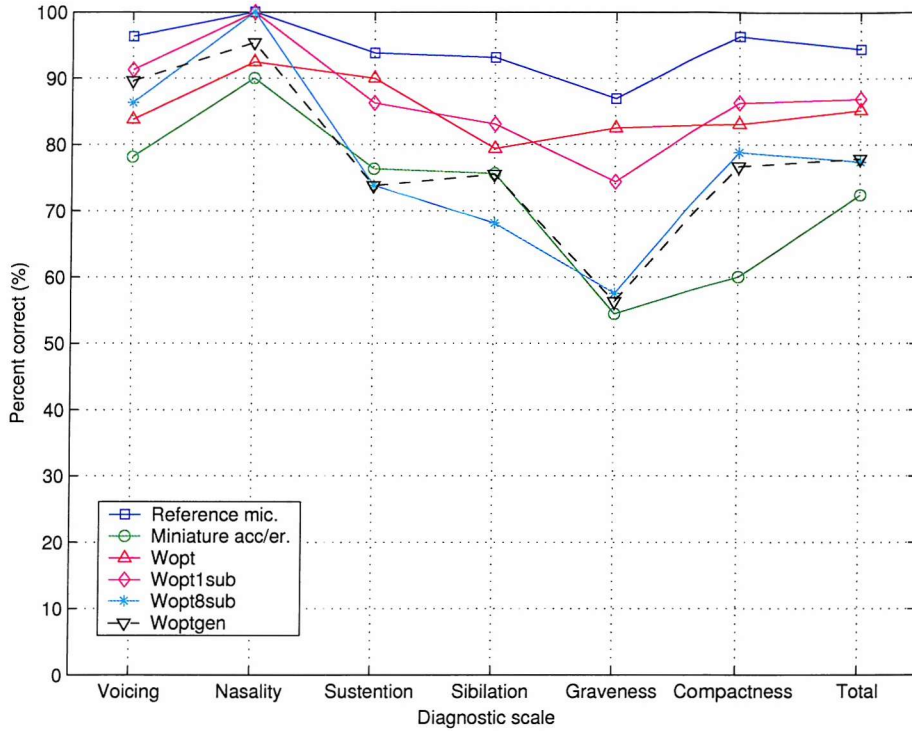


Figure 9.6: Results from the DRT intelligibility test. W_{optgen} is the average of W_{optmal} applied to male speech, and W_{optfem} applied to female speech.

	Ref. mic.	Acc.	W_{opt}	$W_{opt1sub}$	$W_{opt8sub}$	W_{optgen}
Voicing	96.3	78.1	83.8	91.3	86.3	89.6
Nasality	100	90	92.5	100	100	95.4
Sustention	93.8	76.3	90	86.3	73.8	73.8
Sibilant	93.1	75.6	79.4	83.1	68.1	75.5
Graveness	86.9	54.4	82.5	74.4	57.5	56.3
Compactness	96.3	60	83.1	86.3	78.8	76.7
Avg. per system	94.4	72.4	85.2	86.9	77.4	77.9

Table 9.2: Intelligibility scores for each feature and every transducer/filter implementation.

As can be seen from the mean values in Figure 9.6, the reference microphone presents the highest intelligibility scores. W_{opt} and $W_{opt1sub}$ follow with fairly high results in the same range, something that supports our hypotheses concerning $W_{opt1sub}$'s performance. The score from $W_{opt8sub}$ follows next, suggesting that the performance of an average filter across subjects may not be as good as the one of a word- or speaker-specific filter. The similarity in scores

leads to the same suggestion concerning the performance of a filter designed and applied to speakers of same gender (W_{optgen}). The lowest score is presented by the unfiltered accelerometer. It is interesting to note that the scores from the first three columns of Table 9.2 are in the same range as the results of the same configurations from the previous DRT test in Chapter 7.

The independent variables for this test are the speech features (6-level factor), the transducers (reference microphone/ accelerometer: 2-level factor), and the filters (W_{opt} , $W_{opt1sub}$, $W_{opt8sub}$, W_{optgen} : 4-level factor). Our dependent variable is the intelligibility which is tested for different combinations of the independent variables.

9.6.2 Effect of Transducer on Specific Features

The first test was conducted in order to investigate the interaction of transducer and specific features. It was therefore a 6×2 level ANOVA and the factors were the features (6-level factor) and the transducers (2-level factor). Data were averaged across subjects (see Table 9.3).

Features	Ref. mic.	Acc.
Voicing	96.3	78.1
Nasality	100	90
Sustention	93.8	76.3
Sibilant	93.1	75.6
Graveness	86.9	54.4
Compactness	96.3	60
Avg. per system	94.4	72.4

Table 9.3: Intelligibility scores for each feature and every transducer configuration.

The null hypothesis H_0 is that the intelligibility scores for all features and all transducers were the same. The alternate hypothesis H_1 is that the scores among the two dimensions are different. The results showed that both factors are significant: Feature ($F = 20.2$, $p < 0.01$) and Transducer ($F = 283.9$,

$p < 0.01$). Interaction was also significant: Feature \times Transducer ($F = 6.8$, $p < 0.01$).

We can therefore conclude that the intelligibility of the reference microphone and unfiltered accelerometer signals varies significantly with each of the two factors. Considering the transducer factor, the significance indicates that the intelligibility of the reference microphone is significantly better than that of the accelerometer, a point that was also shown in Chapter 7.

9.6.3 Effect of Filter on Specific Features

The next test was conducted in order to investigate the effect of different types of W_{opt} on specific features. The initial test was a 6 \times 4 level ANOVA and the factors were the features (6-level factor) and the filters (4-level factor). Data were averaged across subjects. The results can be seen in Table 9.4.

Features	W_{opt}	$W_{opt1sub}$	$W_{opt8sub}$	W_{optgen}
Voicing	83.8	91.3	86.3	89.6
Nasality	92.5	100	100	95.4
Sustention	90	86.3	73.8	73.8
Sibilation	79.4	83.1	68.1	75.5
Graveness	82.5	74.4	57.5	56.3
Compactness	83.1	86.3	78.8	76.7
Avg. per system	85.2	86.9	77.4	77.9

Table 9.4: Intelligibility scores for each feature and every filter implementation.

The results showed that both factors are significant: Feature ($F = 41.4$, $p < 0.01$), Filter ($F = 11.1$, $p < 0.01$), and their interaction ($F = 5$, $p < 0.01$). The above results show that the effects of feature and filter as individual factors, were strong enough to create significant differences in intelligibility scores. However, we still do not know the significance among the filters for specific features.

We therefore grouped the filters in two categories: The first one included the

filters W_{opt} and $W_{opt1sub}$. The second category included W_{optgen} and $W_{opt8sub}$. The reason for this grouping is that during the DRT test, subjects listened to cross-filtered signals (i.e. outputs from W_{optmal} and W_{optfem} applied to female and male speech, respectively), as well as to the same-filtered signals (i.e. outputs from W_{optmal} and W_{optfem} applied to male and female speech, respectively). Hence, the statistical analysis for those two filters needs a specific handling.

As far as the first category is concerned (W_{opt} and $W_{opt1sub}$), an ANOVA was performed testing the significance among the two levels of the factor Filter. Data were averaged across subjects and features. The calculated overall F-value ($F = 1.4, p > 0.05$) showed non-significance. We can therefore infer that the speaker-specific filter $W_{opt1sub}$ has similar performance to W_{opt} , verifying our initial hypothesis. Individually, W_{opt} and $W_{opt1sub}$ were found significantly better than the unfiltered accelerometer ($p < 0.01$).

Comparisons using ANOVA's were also performed individually for every feature in order to investigate the significance between the two levels of the factor filter in more detail. The results appear in Table 9.5:

Features	W_{opt} vs. $W_{opt1sub}$
Voicing	NS
Nasality	NS
Sustention	NS
Sibilation	NS
Graveness	NS
Compactness	NS

Table 9.5: Effect of filter for specific features. (NS indicates non-significance with $p > 0.05$).

As far as the second category is concerned (W_{optgen} and $W_{opt8sub}$), ANOVA results showed non-significance ($p > 0.05$) between those two levels of the factor Filter. We can thus conclude that an average filter across a number of speakers of mixed gender ($W_{opt8sub}$), performs equally well to a filter that is de-

signed and applied to speakers of same gender (W_{optgen}). Individually, W_{optgen} and $W_{opt8sub}$ were found significantly better than the unfiltered accelerometer ($p < 0.01$).

However, the listening test also included cross-filtered signals, which can be seen individually in Figure 9.7.

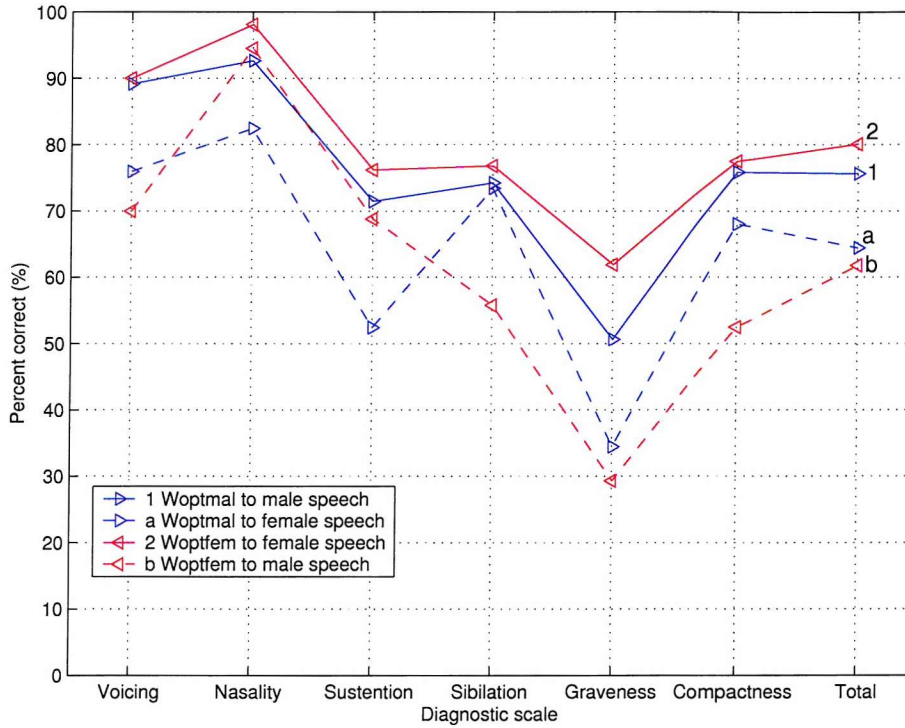


Figure 9.7: DRT results from presentation of the cross-filtered signals.

The significance among the results of Figure 9.7 was investigated with separate t-tests comparing the outputs of the two filters applied on subjects of the same gender, vs. the outputs of the two filters applied on subjects of different gender. Data were averaged across features. The results are the following:

- W_{optmal} to male vs. W_{optfem} to female: Significant ($p < 0.01$).
- W_{optfem} to male vs. W_{mal} to female: Significant ($p < 0.05$).
- Overall, same-filtered signals vs. cross-filtered signals: Significant ($p < 0.01$).

We thus conclude that the application of a gender-specific filter to subjects of the same gender produces significantly more intelligible results than when applying the same filter to subjects of the opposite gender. Another point is that it appears to be more important for female voices to be processed by W_{optfem} , than the male voices by W_{optmal}

9.6.4 Results from DAM Test

The results from the DAM test are shown in Figure 9.8. The legend describes sensor and filter combinations. They are averaged across listeners and sub-categories (9 for Signal qualities, 7 for Background qualities and 3 for Total Effect). Each bar represents an average of 20 listeners' responses to 6 DAM sentence groups, giving a total of 120 values. It has to be noted that during the DAM test, no cross-filtered signals were presented to the subjects. The statistical analysis, therefore, will only include W_{optgen} among the other filters, i.e. the average across W_{optmal} applied to male speech, and W_{optfem} applied to female speech.

This plot is quite interesting as it shows the direct preferences of the subjects. For the Signal Qualities we can see that the reference microphone signal is the best rated signal (a low score indicates a good quality signal). The next best is the output from $W_{opt1sub}$ which is followed by the outputs of W_{opt} and $W_{opt8sub}$ which have similar scores. The gender-specific filter W_{optgen} follows in the overall signal rating, while the signal with the worst score is the unfiltered accelerometer.

For the Background Qualities, again the reference microphone is rated the best, followed by $W_{opt1sub}$ and $W_{opt8sub}$. The rest of the signals all have approximately the same score. The scale, however, that shows the direct assessment of the signals is the Total Effect. Here, we can see that the reference

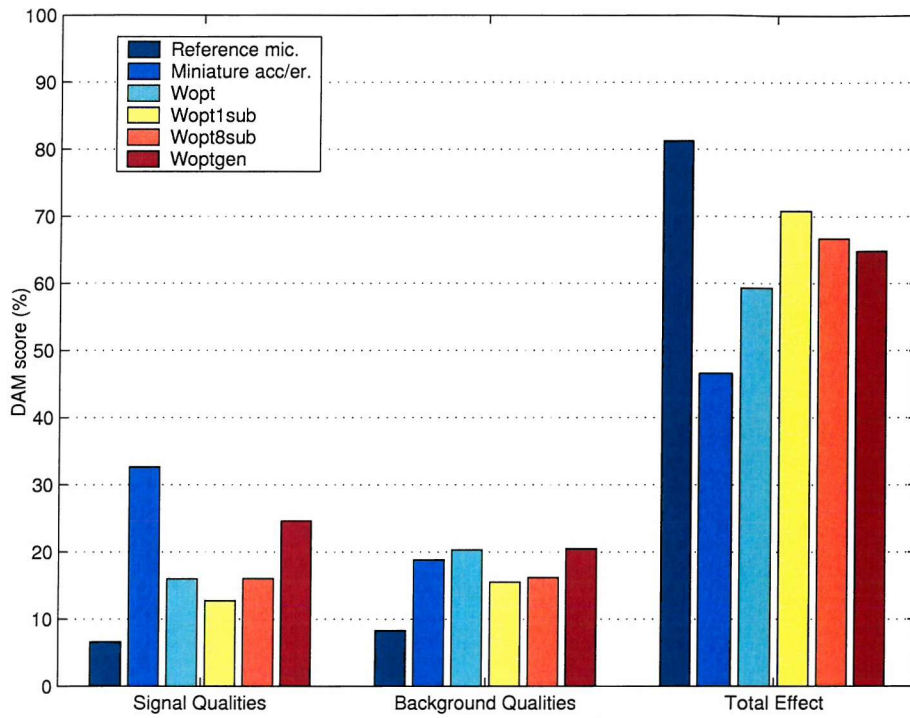


Figure 9.8: Results from the last DAM quality test.

microphone is considered to be the best regarding intelligibility, pleasantness and acceptability. It is followed by the outputs of the four optimal filters $W_{opt1sub}$, $W_{opt8sub}$, W_{optgen} and W_{opt} of which the highest score is presented by the subject-specific filter $W_{opt1sub}$, while the output from the unfiltered accelerometer has the lowest score.

9.6.5 Signal Qualities

The results from all the sub-categories in the Signal Qualities can be seen in Figure 9.9 and in Table 9.6. Each point represents an average of 20 listeners' responses to 6 DAM sentence groups, giving a total of 120 values.

The aim is to investigate the importance of those sub-categories, and their effect in relation to the filter type. Therefore, the analysis that will be presented here will be the same as in the DRT test, i.e. we will examine the effect of the filter on the intelligibility using ANOVA and t-tests. The effect of transducer

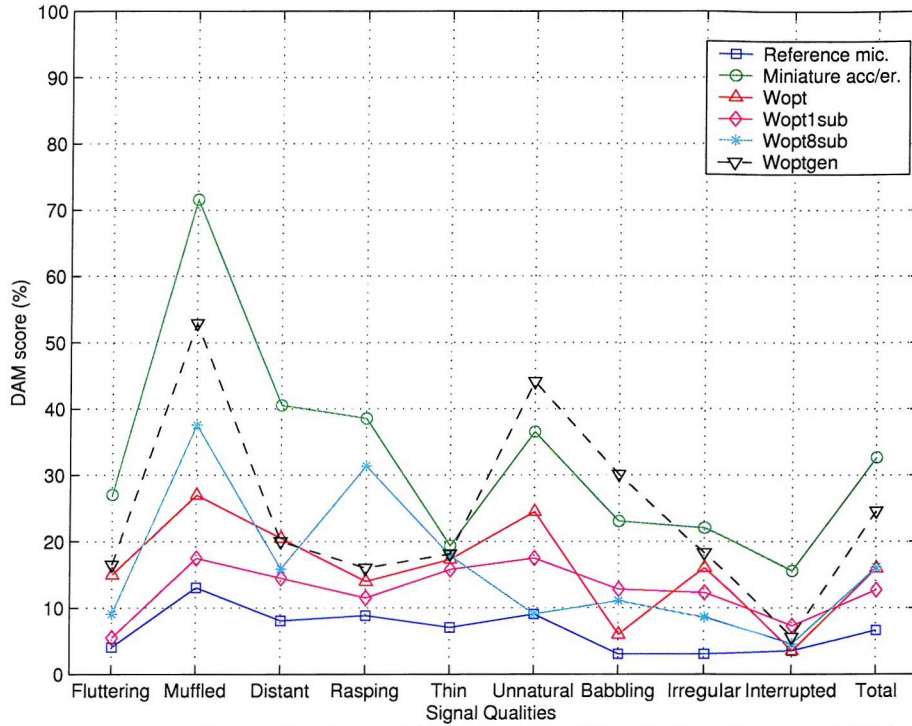


Figure 9.9: Results from the last DAM test for Signal Qualities. DAM score is extent to which test sentences exhibit the particular signal quality.

is not investigated here, as it has already been shown at different instances that, overall, the quality rating of the reference microphone is significantly better than that of the unfiltered accelerometer.

Signal Qualities: Effect of filter

A 4×9 ANOVA was performed in order to investigate the effect of filter (4-level factor) vs. sub-category (9-level factor). The results showed that the individual effects of Filter ($F = 11$, $p < 0.01$) and Sub-category ($F = 26.3$, $p < 0.01$) are significant, as is their interaction ($F = 6.5$, $p < 0.01$). The above result was verified by comparing the four levels of the factor Filter, revealing a significant difference ($F = 12.2$, $p < 0.01$). Post-hoc t-tests also revealed significant differences between the results of the unfiltered accelerometer and each filter. The individual differences among W_{opt} , $W_{opt1sub}$ and $W_{opt8sub}$ were found not significant ($p > 0.05$). Individually, W_{opt} , $W_{opt1sub}$ and $W_{opt8sub}$ were

	Ref. mic.	Acc.	W_{opt}	$W_{opt1sub}$	$W_{opt8sub}$	W_{optgen}
Fluttering	4	27	15	5.5	9	16.5
Muffled	13	71.5	27	17.5	37.5	52.9
Distant	8	40.5	20.5	14.5	15.8	20
Rasping	8.8	38.5	14	11.5	31.3	16
Thin	7	19.3	17.3	15.8	17.8	18.1
Unnatural	9	36.5	24.5	17.5	9	44.1
Babbling	3	23	6	12.8	11	30.1
Irregular	3	22	16	12.3	8.5	18.3
Interrupted	3.5	15.5	3.5	7.3	4.5	5.6
Individual means	6.6	32.6	15.9	12.7	16	24.6

Table 9.6: DAM scores for each Signal sub-category and every system configuration

found significantly better than W_{optgen} ($p < 0.01$).

9.6.6 Background Qualities

The results from all the sub-categories in the Background Qualities can be seen in Figure 9.10 and in Table 9.7. Each point represents an average of 20 listeners' responses to 6 DAM sentence groups, giving a total of 120 values.

	Ref. mic.	Acc.	W_{opt}	$W_{opt1sub}$	$W_{opt8sub}$	W_{optgen}
Hissing	6.3	25	33	11.3	44	41.3
Chirping	14	13	10	20.3	7	4
Roaring	7	6.8	19	11.8	1	13.9
Crackling	8.8	31	22.5	19.3	9.5	14
Buzzing	0	15.5	16.5	12.5	18	25.8
Rumbling	13.5	11	24	22	21.5	22.6
Bubbling	10	15.6	21.5	14.3	12.5	17.9
Individual means	8.2	18.8	20.3	15.5	16.2	20.5

Table 9.7: DAM scores for each Background sub-category and every system configuration

The aim in this case is to investigate the importance of the sub-categories, and their effect in relation to the filter type. The analysis is the same as the one that was performed for the Signal Qualities.

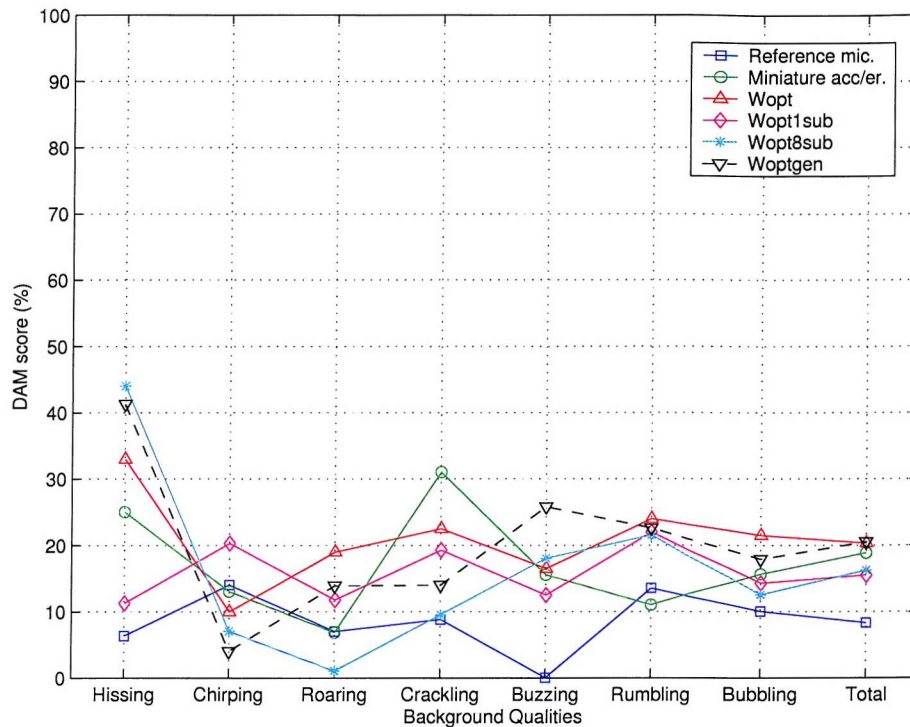


Figure 9.10: Results from the last DAM test for Background Qualities. DAM score is extent to which test sentences exhibit the particular background quality.

Background Qualities: Effect of filter

A 4×7 ANOVA was performed in order to investigate the effect of filter (4-level factor) vs. sub-category (7-level factor). The results showed that the individual effect of Sub-category ($F = 23$, $p < 0.01$) was significant. However, the effect of Filter ($F = 2.5$, $p > 0.05$) is not significant, unlike the interaction ($F = 5.8$, $p < 0.01$) which was found significant. The above result was verified with the investigation of the factor Filter, which showed non significance ($p > 0.05$) among its four levels. Results from post-hoc t-tests also showed that the effects of all filters were not significant ($p > 0.05$) when compared among them, and also when compared to the unfiltered accelerometer.

9.6.7 Total Qualities

This direct evaluation of the processed speech, is effectively the most straightforward scale to see the overall preference of the subjects, as the way it is scored is perceptually easier (the higher the score, the better the signal). The results of all the sub-categories in the Total Qualities main category can be seen in Figure 9.11 and in Table 9.8, where $W_{opt1sub}$ presents the highest score overall after the reference microphone. Each point represents an average of 20 listeners' responses to 6 DAM sentence groups, giving a total of 120 values.

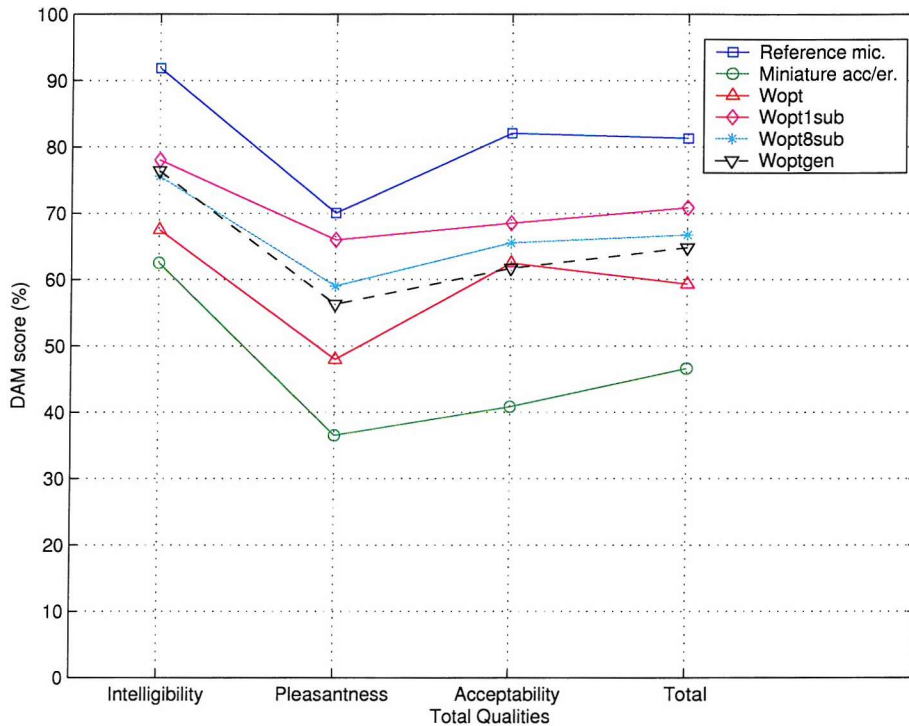


Figure 9.11: Results from the last DAM test for Total Qualities. DAM score is extent to which test sentences exhibit the particular total quality.

	Ref. mic.	Acc.	W_{opt}	$W_{opt1sub}$	$W_{opt8sub}$	W_{optgen}
Intelligibility	91.8	62.5	67.5	78	75.5	76.4
Pleasantness	70	36.5	48	66	59	56.3
Acceptability	82	40.8	62.5	68.5	65.5	61.7
Individual means	81.2	46.6	59.3	70.8	66.6	64.8

Table 9.8: DAM scores for each Total sub-category and every system configuration

What is investigated here, is the effect of filter type on the overall quality

assessment of the signals. The whole procedure is the same as in the Signal and Background Qualities.

Total Qualities: Effect of filter

A 4×3 ANOVA was performed in order to investigate the effect of filter (4-level factor) vs. sub-category (3-level factor). The results showed that the individual effects of Sub-category ($F = 41.9$, $p < 0.01$), and Filter ($F = 18.4$, $p < 0.01$) were significant. The interaction, however, was not significant ($F = 2.2$, $p > 0.05$), suggesting that the effect of different types of filters created similar overall impressions for all sub-categories. Results from post-hoc t-tests showed that W_{opt} , $W_{opt1sub}$ and $W_{opt8sub}$ gave, individually, significantly better quality results compared to the unfiltered accelerometer. However, the difference between $W_{opt8sub}$ and W_{optgen} was found not significant ($p > 0.05$), verifying that the average quality across the two gender-specific filters (W_{optmal}/W_{optfem}) is rated close to that from a mixed-gender filter ($W_{opt8sub}$). The individual comparisons among W_{opt} , $W_{opt1sub}$ and $W_{opt8sub}$ all gave significant results ($p < 0.05$).

9.7 Conclusions

The DRT intelligibility test results verified our initial hypothesis concerning the usefulness of an optimal filter based on the phonetic material of one speaker ($W_{opt1sub}$). This presented similar intelligibility scores when compared to the optimal filter that we have been using up to now, which was designed based on specific words or sentences (W_{opt}). In addition, the optimal filter designed for a number of subjects ($W_{opt8sub}$) showed some potential in enhancing accelerometer signals, although its performance was not as good as that of the

two previous filters. As was also seen, an optimal filter based on the phonetic material of a number of subjects of specific gender seemed to perform adequately well only when it was applied on speech recorded from subjects of same gender, otherwise its performance dropped significantly.

The DAM quality test was the last check on the subjective preference of the processed speech. Here, as in the DRT, $W_{opt1sub}$ and $W_{opt8sub}$ were surprisingly rated higher in the overall quality assessment than W_{opt} , something that is encouraging as far as the filters' implementation on a communication system is concerned. In addition, the gender-specific filter W_{optgen} was rated, overall, close to $W_{opt8sub}$. The above points can be attributed to the smoothing of the filters' response during the averaging across the phonetic material of one or more subjects of mixed or specific gender, resulting in the removal of artifacts from the processing of single words or sentences.

Chapter 10

Conclusions and Suggested Further Work

10.1 Summary

The study of an inter-disciplinary topic such as speech vibrations is fairly intricate to approach as it combines elements of speech physiology, human anatomy, and vibrations which, on their own, are complicated areas. Moreover, since we are investigating the detection, enhancement and assessment of such signals, the task becomes quite challenging, since areas such as digital signal processing, psychoacoustics and statistics are brought together. The main aims of this closing chapter, therefore, are to present the general conclusions inferred from this study, and to propose some ideas for a future expansion of the project.

10.2 General Conclusions

The general conclusions that are presented in this section can be split into two broad categories: The first category includes the conclusions related to the generation and detection of speech vibrations, while the second includes conclusions about optimal filtering and its enhancement properties.

As far as the first category is concerned, body-conducted speech has been shown to be detected efficiently within the circumaural area of a speaker. More specifically, in the context of a mobile communication system, the usefulness of the TMJ as a good pick-up point for speech vibrations has been verified. It was also shown that an important factor which influences the quality of the detected speech vibration signals is the movement of the lower jaw. The modelling of our experimental setup has shown that speech, detected by a microphone at a certain distance from a speaker's mouth, and an accelerometer on the TMJ, is influenced by the speaker's acoustic and anatomical transfer functions, respectively. These include the various paths from any acoustic or vibratory speech-generation point to the equivalent transducer, including the effects from the transducers' responses. The filtering effect due to the two paths, which also includes the low-pass filtering due to body-conduction, was shown to have a main cut-off at about 500-600 Hz for both male and female speakers, while the roll-off rate was approximately 18-20 dB/oct.

As far as the second category is concerned, optimal filtering has been shown to be superior to high-pass and band-pass filtering, regarding the improvement of intelligibility and quality of accelerometric speech. The performance of the optimal filter is good in the range 100 Hz-1 kHz, where it high-pass filters the accelerometer signals and reconstructs the harmonics attenuated due to body-conduction. However, the electronic noise of the accelerometer was found to be the most dominant factor above 1 kHz, contributing to the reduction of the

optimal filter's performance. The causality of the optimal filtering setup has been verified with a model which showed that the acoustic path delays were higher than the anatomical path delays. The model also assisted in demonstrating that the main factor determining the optimal filter's performance is the ratio of a speaker's acoustic to anatomical transfer functions, as defined above. From that, it was suggested that it is more important to design a filter based on the anatomical characteristics of a speaker, than on the spectral information contained in a speech corpus. This was verified experimentally with the design of a fixed optimal filter across a number of speakers and/or speech materials. The speaker-specific filter presented, overall, similar intelligibility but better quality than the corpus-specific filter, as well as greater simplicity in its practical implementation.

10.3 Suggestions for Further Work

One fruitful way to extend this work would be to use both miniature transducer signals (microphone and accelerometer) as inputs to an optimal filter. This would use the theory of multiple input optimal filters as explained in Elliott (2001). Another idea could involve the use of the miniature microphone at the ear as the desired signal (since this has been shown to be not so different in amplitude and spectrum from the reference microphone), and thereafter the miniature accelerometer signal would be optimised with reference to that. This would result in a more compact system where the reference microphone, which has been always present up to now, would not be needed.

The two types of filters proposed above are easily implementable by using the speech material from the recordings of Chapters 7 and 9. The filters' performance can also be tested with noisy signals. For this case, real noises from e.g. a car or a fighter plane cockpit environments could be added in the

same way as appeared in Figure 7.2. The filters' performance could then be assessed subjectively with the DRT and DAM tests as in Chapters 7 and 9. The above filters are also easily implementable in a real-life communication system which could consist of a compact, in-ear device incorporating the two miniature transducers.

It would be useful to consider different ways of attaching the miniature transducers around the aural area more efficiently, without the dependence of any kind of medium like the medical tape which was used in our experiments. Future work could investigate the design of a commercially applicable product that would incorporate one (miniature accelerometer) or more transducers in a single device that could be placed at the aural area in order to detect body-conducted and/or air-conducted speech. The device could eventually have a single chip programmed with the enhancement algorithm.

The last idea concerns a new set of experiments that should be devised in order to understand and quantify the speech vibration transmission paths, from the generation point to the pick-up location. This could exploit the use of a number of accelerometers attached to the face of a speaker, in order to draw a 'map' of speech vibration intensities.

References

- Bekesy, G. (1960). *Experiments in Hearing*. McGraw-Hill Book Company, United States of America.
- Bendat, J. and Piersol, A. (1993). *Engineering Applications of Correlation and Spectral Analysis*. John Wiley and Sons, New York, 2nd edition.
- Black, R. (1957). Ear-insert microphone. *Journal of the Acoustical Society of America*, 29:260–264.
- Crystal, D. (1991). *A Dictionary of Linguistics and Phonetics*. Blackwell Reference, 3rd edition.
- Dunn, H. and Farnsworth, D. (1939). Exploration of pressure field around the human head during speech. *Journal of the Acoustical Society of America*, 10:184–199.
- Elliott, S. J. (2001). *Signal Processing for Active Control*. Academic Press, London.
- Everest, F. (1994). *The Master Handbook of Acoustics*. TAB books, New York, 3rd edition.
- Flanagan, J. (1960). Analog measurements of sound radiation from the mouth. *Journal of the Acoustical Society of America*, 32:1613–1620.
- Franke, E., Gierke, H. V., Grossman, F., and Wittern, W. V. (1952). The

- jaw motions relative to the skull and their influence on hearing by bone conduction. *Journal of the Acoustical Society of America*, 24:142–146.
- Giua, P. (1998). Voice transmission through vibration pickups. *Journal of the Acoustical Society of America*, 103:2773.
- Howell, P., Williams, M., and Dix, H. (1988). Assessment of sound in the ear canal caused by movement of the jaw relative to the skull. *Scandinavian Audiology*, 17:93–98.
- Huopaniemi, J., Kettunen, K., and Rahkonen, J. (1999). Measurement and modeling techniques for directional sound radiation from the mouth. *IEEE Workshop in Applications of Signal Processing to Audio and Acoustics, New York*, pages 183–186.
- Ishimitsu, S. and Kitazake, H. (2001). Study for constructing a recognition system using the body-conduction speech. *Autumn Meeting of the Acoustical Society of Japan, Oita, Japan*.
- Khanna, S., Tonndorf, J., and Queller, J. (1976). Mechanical parameters of hearing by body-conduction. *Journal of the Acoustical Society of America*, 60:139–154.
- Kinsler, L., Frey, A., Coppens, A., and Sanders, J. (1982). *Fundamentals of Acoustics*. John Wiley & Sons, New York, 3rd edition.
- Kuttruff, H. (1979). *Room Acoustics*. Elsevier Applied Sciences Publishers Ltd., New York, 3rd edition.
- Maurer, D. and Landis, T. (1990). Role of body-conduction in the self-perception of speech. *Folia Phoniatrica*, 42:226–229.
- McKendree, F. (1986). Directivity indices of human talkers in english speech. *Proc. Inter-Noise, MIT, Cambridge, MA*, pages 911–916.

- Meakawa, Z. and Lord, P. (1994). *Environmental and architectural acoustics*. E&FN Spon, London.
- Moser, H. and Oyer, H. (1958). Relative intensities of sounds at various anatomical locations of the head and neck during phonation of the vowels. *Journal of the Acoustical Society of America*, 30:275–277.
- Mullendore, J. (1949). Relative amplitudes of vowels on the body. *Speech Monographs*, 16:163–177.
- Olson, H. (1967). *Music, Physics and Engineering*. Dover Inc., New York, 2nd edition.
- Omologo, M., Svaizer, P., and Matassoni, M. (1998). Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication*, 25:75–95.
- Ono, H. (1977). Improvement and evaluation of the vibration pick-up-type ear microphone and two-way communication system. *Journal of the Acoustical Society of America*, 62:760–768.
- O'Shaughnessy, D. (2000). *Speech Communications: Human and Machine*. IEEE Press, Canada.
- Owens, F. (1993). *Signal Processing of Speech*. MacMillan Press Ltd, London.
- Porchmann, C. (2000). Influences of bone conduction and air conduction on the sound of one's own voice. *Acustica-Acta Acustica*, 86:1038–1045.
- Prinz, J. (1998). Resonant characteristics of the human head in relation to temporomandibular joint sounds. *Journal of Oral Rehabilitation*, 25:954–960.
- Quackenbush, S., Barnwell, T., and Clements, M. (1988). *Objective measures of speech quality*. Prentice Hall, New Jersey.

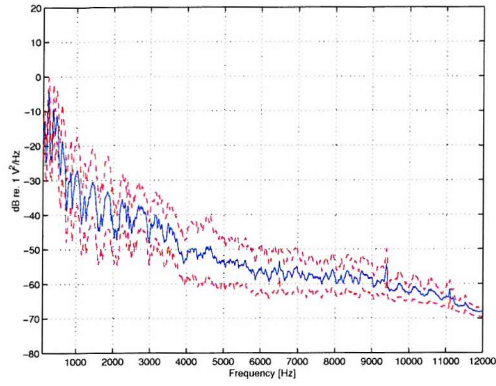
- Rafaely, B. and Furst, M. (1996). Audiometric ear canal probe with active ambient noise control. *IEEE Transactions on Speech and Audio Processing*, 4(3):224–230.
- Roederer, J. (1995). *The physics and psychophysics of music*. Springer-Verlag, New York, 3rd edition.
- Sadaoki, F. (1989). *Digital Speech Processing, Synthesis and Recognition*. Marcel Dekker Inc., NTT Human Interface Laboratories, Tokyo, Japan.
- Schroeter, J. and Poesselt, C. (1986). The use of acoustical test fixtures for the measurement of hearing protector attenuation. part ii: Modelling the external ear, simulating bone conduction, and comparing test fixture and real-ear data. *Journal of the Acoustical Society of America*, 80:505–527.
- Sebesta, G., Mellen, A., and Carlisle, R. (1970). An inertial head-contacting audio communications headset system. *Journal of the Audio Engineering Society*, 18:418–423.
- Shiga, H. and Kobayashi, Y. (1995). Quantitative evaluation of tmj sounds by frequency analysis. *Transactions on fundamentals of electronics, communications and computer sciences*, E-78-A(12):1683–1687.
- Shigeaki, A., Kazumasa, M., and Nishino, Y. (1999). Super-compact microphone/receiver unit for noisy environments. *Journal of the Acoustical Society of Japan*, 20:381–383.
- Stenfelt, S. (1999). Hearing by body-conduction. physical and physiological aspects. Technical Report 358, Doktorsavhandlingar vid Chalmers Tekniska Hogskola.
- Tomatis, A. (1987). *L'Oreille et la Voix*. Robert Laffont, S.A, Paris.
- Tonndorf, J. (1972). *Foundations of modern auditory theory*, volume 2, chapter 5, pages 760–768. Academic Press Inc., New York. edited by Jerry.V Tobias.

- Viswanathan, V., Henry, C., and Derr, A. (1985). Noise-immune speech transduction using multiple sensors. *Proc. IEEE ICASSP, Tampa, Florida, USA*, pages 712–715.
- Voiers, W. (1977a). Diagnostic acceptability measure for speech communication systems. *Proc. IEEE ICASSP, Hartford, Connecticut, USA*, pages 204–207.
- Voiers, W. (1977b). *Diagnostic Evaluation of Speech Intelligibility*, chapter IX, pages 374–387. Benchmark Papers in Acoustics. Dowden, Huthenson and Ross. edited by M.E. Hawley.
- Watkinson, J. (1995). *The art of digital audio*. Focal Press, London, 2nd edition.
- Wee, A. and Ashley, R. (1990). Transmission of acoustic or vibratory signals from a contracting muscle to relatively distant tissues. *Electromyography and Clinical Neurophysiology*, 30:303–306.
- Westermann, S. (1987). The occlusion effect. *Hearing instruments*, 38(6):43.
- Wiener, F. and Ross, D. (1946). The pressure distribution in the auditory canal in a progressive sound field. *Journal of the Acoustical Society of America*, 18:401–408.
- Wise, C. (1932). Rank-ordering the conductivity of different tissue types. *Quart. J. Speech*, 18:446–452.

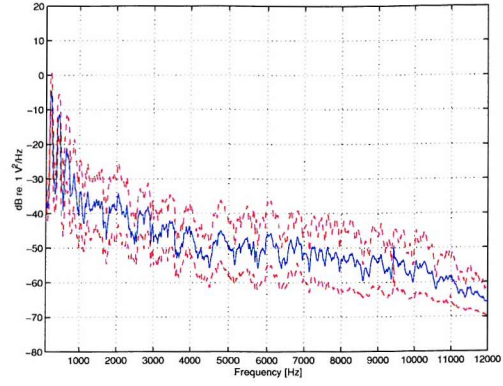
Appendix A

Linear Frequency Scale Plots of Chapter 8

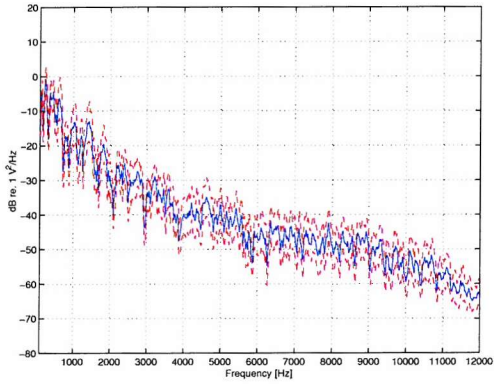
A.0.1 Reference Microphone Signals



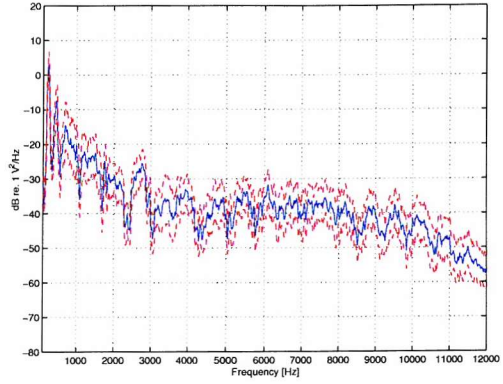
(a)



(b)



(c)



(d)

Figure A.1: Reference microphone signal spectra for the DRT (top) and DAM (bottom) tests, for male (left) and female (right) speakers. (a) DRT male; (b) DRT female; (c) DAM male; (d) DAM female. Same data as log-frequency graphs in Figure 8.1.

A.0.2 Accelerometer Signals

Quiet Signals

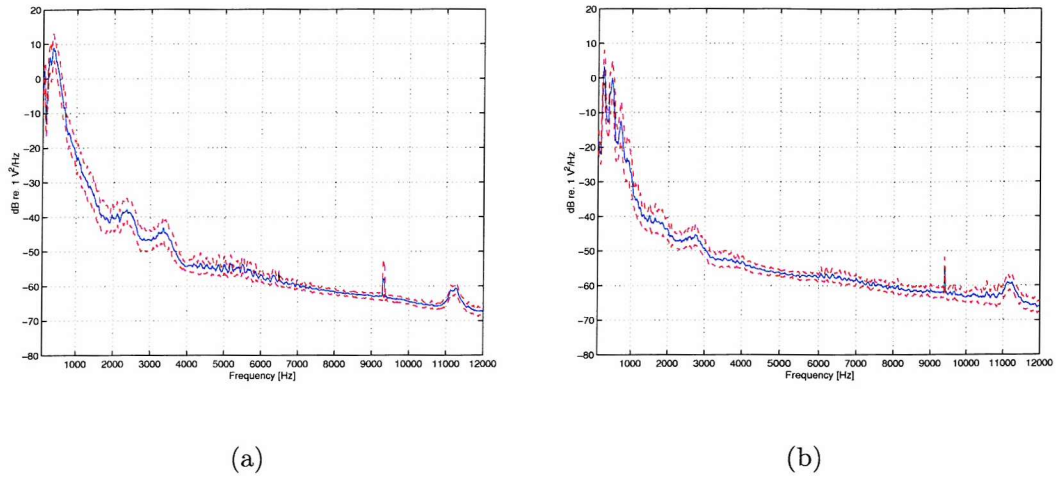


Figure A.2: Average quiet accelerometer DAM spectra for male (left), and female (right) speakers. Same data as log-frequency graphs in Figure 8.2.

Noisy Signals

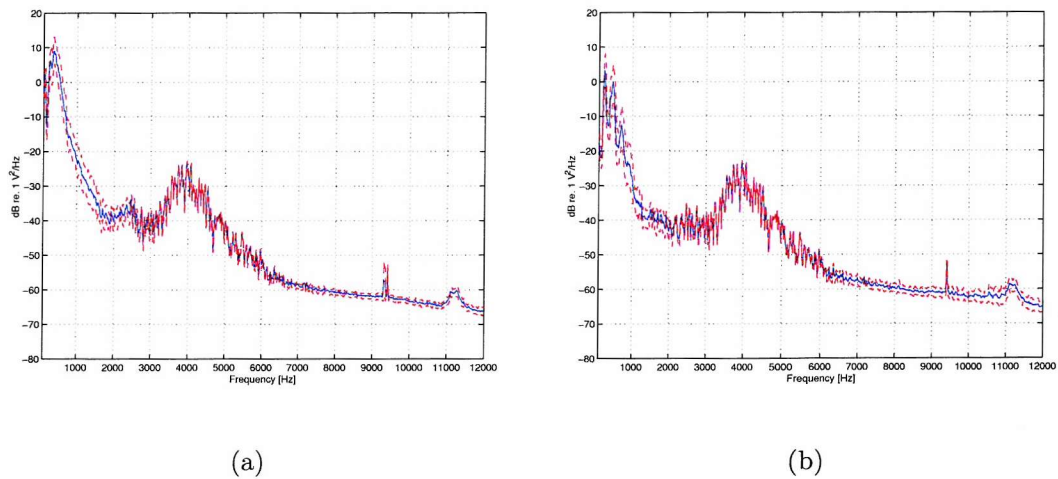


Figure A.3: Average noisy accelerometer DAM spectra for male (left), and female (right) speakers. Same data as log-frequency graphs in Figure 8.3.

A.0.3 Optimal Filter Responses

Quiet Case

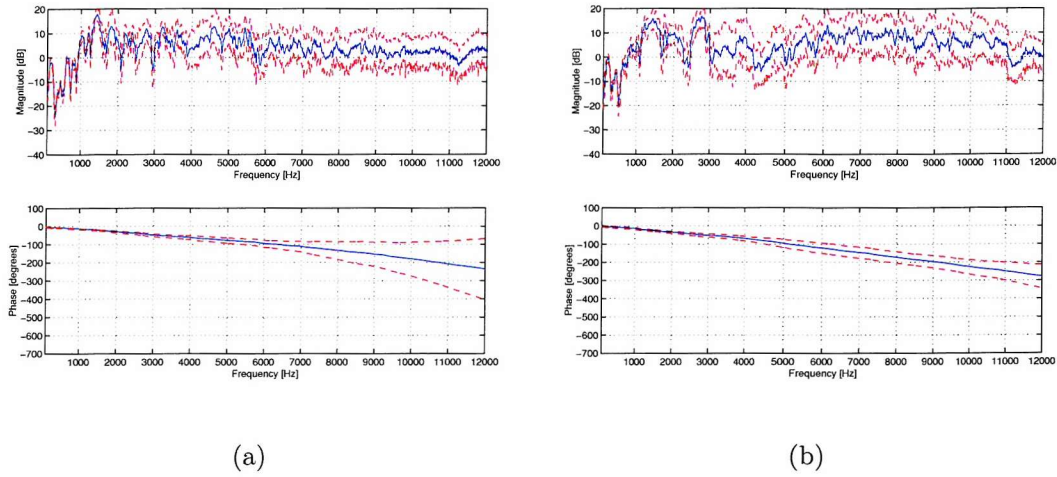


Figure A.4: Average quiet W_{opt} magnitude and phase responses for the DAM test for male (a), and female (b) speaker. Same data as log-frequency graphs in Figure 8.4.

Noisy Case

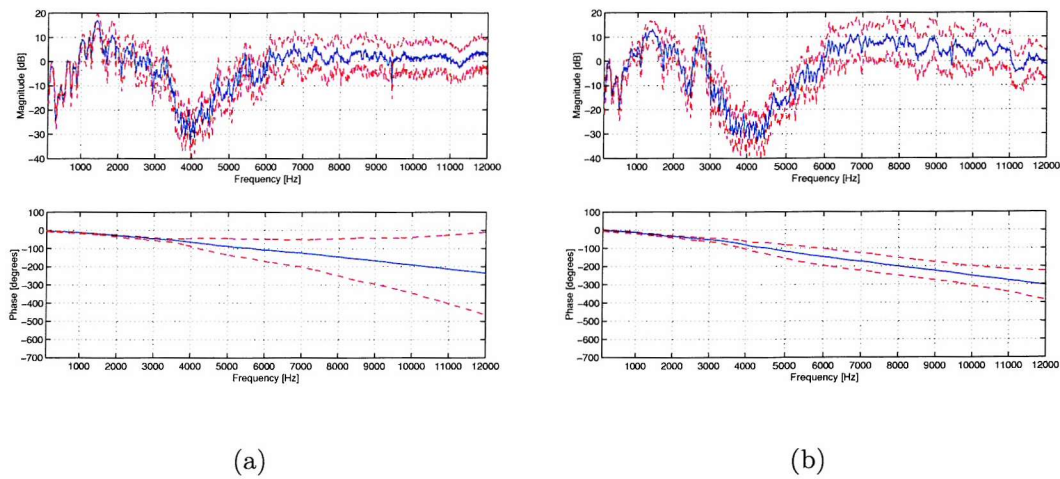


Figure A.5: Average noisy W_{opt} magnitude and phase responses for the DAM test for male (a), and female (b) speaker. Same data as log-frequency graphs in Figure 8.6.

Appendix B

Acoustic and Anatomical Path Delay Calculations

The delays in $G_{ac.}$ and $G_{an.}$ were calculated approximately, with the kinematics law:

$$time = \frac{distance}{speed} \quad (B.1)$$

Typical values for the speed of sound in different materials, considered in our calculations, appear in the Table below. Values for the average length of the vocal tract that we used were 17cm for men, and 15.5cm for women.

air (20°C)	vocal tract (37°C)	human tissue (37°C)
343	359	1540

Table B.1: Speed of sound (m/s) in different materials.

The delay in $G_{ac.}$ due to the distance between the male speaker's larynx and lips is equal to:

$$time = \frac{0.17m}{359m/s} = 0.47ms \quad (B.2)$$

The delay in $G_{ac.}$ due to the distance between the male speaker's lips and the

microphone is equal to:

$$time = \frac{0.6m}{343m/s} = 1.74ms \quad (B.3)$$

where 0.6m is the distance from the speaker's lips to the microphone (see Figure 6.19). The group delay of the reference microphone's amplifier was calculated from the data-sheets to be approximately 1.4μ s. The group delay of the anti-aliasing filter was calculated from the data-sheets as approximately 69μ s. All the above added together give a total delay of about 2.2ms for both the male and the female speaker.

The delay in $G_{an.}$ due to the distance between the speaker's larynx (Adam's apple) and TMJ is equal to:

$$time = \frac{0.1m}{1540m/s} = 65\mu s \quad (B.4)$$

where 0.1m is the approximate distance between the larynx and the TMJ. The delay of the accelerometer's amplifier was measured and found equal to 50μ s. The delay of the anti-aliasing filter was equal to 69μ s. Those add up to approximately 0.18ms for both the male and female speaker.