UNIVERSITY OF SOUTHAMPTON

Faculty of Physical Sciences and Engineering

School of Electronics and Computer Science

Modern Standard Arabic Phonetics for Speech Synthesis

Nawar Halabi

Supervisor: **Prof Mike Wald**

Internal Examiner: **Dr Gary B Wills**

External Examiner: **Assoc Prof Nizar Habash**

Thesis for the degree of Doctor of Philosophy

July 2016

ABSTRACT

Arabic phonetics and phonology have not been adequately studied for the purposes of speech synthesis and speech synthesis corpus design. The only sources of knowledge available are either archaic or targeted towards other disciplines such as education. This research conducted a three-stage study. First, Arabic phonology research was reviewed in general, and the results of this review were triangulated with expert opinions – gathered throughout the project – to create a novel formalisation of Arabic phonology for speech synthesis.

Secondly, this formalisation was used to create a speech corpus in Modern Standard Arabic and this corpus was used to produce a speech synthesiser. This corpus was the first to be constructed and published for this dialect of Arabic using scientifically-supported phonological formalisms. The corpus was semi-automatically annotated with phoneme boundaries and stress marks; it is word-aligned with the orthographical transcript. The accuracy of these alignments was compared with previous published work, which showed that even slightly less accurate alignments are sufficient for producing high quality synthesis.

Finally, objective and subjective evaluations were conducted to assess the quality of this corpus. The objective evaluation showed that the corpus based on the proposed phonological formalism had sufficient phonetic coverage compared with previous work. The subjective evaluation showed that this corpus can be used to produce high quality parametric and unit selection speech synthesisers. In addition, it showed that the use of orthographically extracted stress marks can improve the quality of the generated speech for general purpose synthesis. These stress marks are the first to be tested for Modern Standard Arabic, which thus opens this subject for future research.

Table of Contents

List of Tables

List of Figures

List of Accompanying Materials

1. www.arabicspeechcorpus.com is a single-speaker, Modern Standard Arabic speech corpus made for high quality speech synthesis. The License is available on the website.
2. https://github.com/nawarhalabi/Arabic-Phonetiser is the Modern Standard Arabic phonetiser developed and used in this work based on the phonetisation rules devised in this work and stress markings of Halpern, 2009.
3. https://bitbucket.org/nhalabi/numbers-to-words is an Arabic number-to-word converter. An equivalent in English, for example, would be a program that converts the number "123" to the text "one hundred and twenty three". It can deal with symbols such as $ and £ and converts them to their textual form. Along with the phonetiser above and the MADAMIRA diacritiser, these can be used to build a high quality Modern Standard Arabic TTS front-end.

DECLARATION OF AUTHORSHIP

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

Modern Standard Arabic Phonetics for Speech Synthesis

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. [Delete as appropriate] None of this work has been published before submission [or] Parts of this work have been published as: [please list references below]:

Signed: Nawar Halabi .....

Date:

Acknowledgements

I would like to thank the experts who helped throughout this work both with their expertise and hard work. I also would like to thank each member of my family for their continuous belief in me. Finally, I would like to thank my supervisors and my sponsor, without whom none of this would have been possible.

It is overdue. My mom, brother, father and the rest of my family are certainly why I made it here. Living abroad away from them and not being able to contact them regularly made it difficult to thank them for every step I made. And being me does not help. So, since I'm not too good at doing this in person, thank you from the bottom of my heart.

I learned to abhor prejudice from you. I learned to doubt myself from you. I learned how to be happy from you without needing to wreck myself. I learned how to accept others from you. And you kept all those childish dreams in me as long as you could. So thank you again, for everything.

My friends in Syria, Southampton and the ones scattered around the world, Thank you for all what you have done for me. You know who you are.

It is all good ☺

شكراً جزيلاً عائلتي

Definitions and Abbreviations

When the following terms first appear in the main text, they are written in **bold** to indicate to the reader that this section may be referred to for their definition.

Back-end: Part of a complete **TTS** system which converts a sequence of phonemes with linguistic features to a speech signal.

Bootstrapping (HMM models): Training **HMM** models by using a manually segmented and aligned speech corpus for potential use to segment another speech corpus by forced alignment.

Buckwalter Transliteration (Buckwalter, 2002): Is a one-to-one mapping between Arabic characters and Latin letters and symbols. Mainly used in this work because HTK (Young et al., 1997) cannot handle Arabic script as input.

C means geminated consonant (unless otherwise stated).

Corpus Design/Speech Corpus Design: The process of gathering prompts for recording by the speech talent. This also involves optimising the phonetic coverage of the speech corpus.

Deep Neural Network (DNN): In simple terms, Neural Networks which have a more complicated and layered structure that requires different methods of training.

Diacritics and Diacritisation: Diacritics are symbols added to letters. In Arabic, they correspond to short-vowel phonemes, gemination or absence of short-vowel phonemes (sukoon). Diacritisation is the process of adding those diacritics to Arabic script.

Emphasis (Arabic Language): The velarisation or pharyngealisation of consonants in Arabic (Laufer and Baer, 1988). They are secondary articulations that correspond to changes in the pharynx or epiglottis from the primary articulation. For convenience these movements are called 'emphasis' in this work.

Front-end: Part of a complete TTS system which converts raw text to a phoneme sequence with linguistic features, which is used as the input to a speech synthesiser (Back-end).

Gemination: In Arabic, it is the doubling of a consonant. Usually the effect is dependent on the consonant's articulation category. This work demonstrates that gemination in Arabic is more accurately described as the lengthening of part of the consonant. Linguistically, a geminated consonant is treated as two consecutive consonants when syllabifying a word (Halpern, 2009).

Hidden Markov Model (HMM): A sequential probabilistic speech model for speech recognition and synthesis that is used to predict acoustic and linguistic features.

International Phonetic Alphabet (IPA) (Cambridge University Press, 2014): A set of symbols used to describe phones or phonemes. Published by the International Phonetic Association (Cambridge University Press, 2014).

Mel Frequency Cepstral Coefficients (MFCC): A parametric representation of the speech signal's power spectrum in a short interval (Jurafsky and Martin, 2009).

Modern Standard Arabic (MSA): A standardised variety of Arabic which is used across the Arabic-speaking world in official documents, news, etc.

Normalisation and Normalised Script: In Speech Synthesis, this refers to the input text after all irregular content in it has been converted into a form that can be phonetised by a machine. For example, abbreviations such as HMM could be converted to "Hidden Markov Models" or "Aitch Em Em" by the normalisation process, making it easier to generate the phoneme sequence to be synthesised. The normalisation process also includes numbers, punctuation (such as brackets) and – in some cases – spell-checking (Taylor, 2009).

Orthographic Transcript: In this work**,** the raw text extracted from the web and divided to utterances (sentences) based on punctuation and then manually corrected after size reduction.

Phoneme: Not to be confused with phone, this is the smallest unit of phonology in a language which – when changed – could change the meaning. Phonemes can be seen as classes of phones meaning that a phone is a realisation of a phoneme in a certain context (Taylor, 2009).

Phonetic Transcript: In this work**,** the phoneme sequences corresponding to each utterance in the orthographic transcript, which are generated by running the phonetiser on each utterance separately.

Phonetic Unit: Phone, Diphone, Triphone, Syllable… is a phonetic or phonological segment which, in corpus design, is used to define the phonemic content required to be covered by the transcript.

Phonetisation or Vocalisation: The conversion of a normalised script to a phoneme sequence.

Phonotactics: The rules that govern the types of phonemes, syllables, consonant clusters, etc. that are allowed to occur in speech (Habash, 2010; Biadsy and Hirschberg, 2009).

Pronunciation Dictionary: A list of pronunciations (phoneme sequences) used mainly in speech recognition and phonetisation. Every entry in a pronunciation dictionary contains the orthographic transcript of a word with the corresponding phoneme sequence describing how the word should be pronounced. Orthographic transcripts of words can repeat in different entries showing different possible pronunciations for the same word.

Statistical Parametric Speech Synthesis: A statistical model, trained on a speech corpus, is used to predict the acoustic features needed to generate the desired speech signal.

Stress (Word Stress, Lexical Stress or Syllable Stress): The emphasis on a certain syllable in a word for the purpose of emphasising the word itself to indicate that it has more semantic importance over the rest of the sentence. Emphasis here does not necessarily correspond to a certain articulation process as stress could be realised in different ways (increased loudness, pitch, vowel length…) (de Jong and Zawaydeh, 1999).

Talent or Speech Talent: The person whose voice is recorded for the speech corpus.

Text-to-Speech (TTS): Text To Speech, a complete system for converting raw text to spoken utterances. This involves normalisation, phonetisation and synthesis.

Unit Selection Speech Synthesis: A type of concatenative speech synthesis. In this method of speech synthesis, the utterances are generated by concatenating units, possibly of different lengths, to produce the desired utterance. These units are usually taken from a large, boundary-annotated speech database, and are chosen based on identity and contextual linguistic features.

Utterance: A short script containing a small number of sentences (2 to 6) or a short recording corresponding to that script. The latter is sometimes referred to as “recorded utterance”.

V means long vowel; v means any vowel.

Viterbi Algorithm: A dynamic programming algorithm for finding the most probable sequence of states of the HMM which generated the observation sequence.

Chapter 1 Introduction

1.1 Text to Speech Synthesis

Natural sounding speech synthesis has improved significantly in recent years. The quality of speech generated by **Unit Selection** synthesis is high. Unit Selection uses large segmented and annotated speech corpora and combines segments from the corpora that produce the best possible natural sound. Subjective tests have shown that Unit Selection speech is pleasant, comprehensive and natural to listen to (Black, 2002; Muthukumar and Black, 2014; Indumathi and Chandra, 2012). These tests have been conducted for many languages, including English.

Still, many under-resourced languages (such as Arabic) suffer from lack of good quality text-to-speech synthesisers. Partly, the most natural method of synthesis is Unit Selection which requires the availability of large speech corpora (from 2 hours to over 16 hours in the systems reviewed) (Black et al., 2008; Kumar et al., 2007; Prahallad et al., 2007; Taylor, 2009). These are segmented into speech units of different sizes (syllables, for example) and aligned with a transcript. Ideally, the transcript should contain the phonetic representation of the speech signal rather than the actual script that was recorded, because the alignment should be with a transcript that actually describes the content of the recording (Prahallad et al., 2007).

Other issues such as text **normalisation** also prevents good quality speech synthesis to be available in under-resourced languages (Jurafsky and Martin, 2009; Taylor, 2009). Usually text normalisation is the first stage of processing conducted on the input text. Its job is to resolve issues such as numbers, abbreviations and currency symbols. Research conducted on creating or using speech corpora for segmentation usually assume that the transcripts available are normalised before conducting segmentation, which makes acquiring speech corpora more difficult, as texts accompanying audio books and newscasts have not been normalised. Although this step might seem straightforward, it is a difficult process. Language is changing and it is difficult to keep track of the rules or generate statistical models which require up-to-date data (Jurafsky and Martin, 2009; Taylor, 2009).

The output of text normalisation usually requires more processing to determine how words should be pronounced. This step, called **Phonetisation**, generates the phonetic representation of the text and is highly dependent on the language. For example, in English there are heteronyms like the verb “desert” (meaning abscond) and the noun “desert” (meaning arid place) written similarly and

pronounced differently and so a decision is needed as to which of the possible pronunciations of each heteronym is suitable for the context. Also, co-articulation of letters on word boundaries must be resolved because human speech does not contain a silence between every word. In fact there are fewer silences than co-articulations in normal human speech (Rodero, 2012). Foreign Named Entities written in Arabic letters are commonly pronounced irregularly and this is usually dealt with, in some languages, by finding those entities after normalisation and carrying out dictionary lookups to find the correct pronunciation (Jurafsky and Martin, 2009).

Other issues related to Phonetisation: Names (Ewan, Keith...) may be pronounced in a way that is related to the language of origin of those names. It is also possible to have different spellings (Eamonn and Eamon) which makes it difficult to map from letters to phones as there are no formal rules. This issue and others make the task of phonetisation not a simple mapping from one or many letters to a **phoneme**. Phonemes are highly context sensitive. Another challenge related to phonetisation, specific to Arabic, (although a highly phonetic language, phonemes or phones strongly correspond to letters), is the long vowel /aa/ which has a phonetic realisation that is directly affected by the consonant preceding it. A modified version of the **Buckwalter transliteration** will be used to refer to phonemes (see Table 4-7). A mapping to the **International Phonetic Alphabet** (IPA) (Cambridge University Press, 2014) along with Arabic letters is provided in Table 4-7. The reason for not directly using IPA is the fact that the corpus was to be used with software which might have problems dealing with IPA characters in Unicode.

In English, words like “enough” and “bought” are good examples of the difficulty of converting text to phonetic representation. Even phonemes in sequence might be merged to yield different phones. This is called co-articulation, which happens inside and between words in different forms, for example in the phrase “Good dog”, the two “d” letters at the word boundaries are merged together into one /d/ phoneme. Usually, every version of the phone in a context is called an allophone. These subtle changes in the way a phone is pronounced in different contexts increases the difficulty of speech synthesis as a whole and mainly increases the speech corpora required for the different methods of synthesis. A more detailed explanation of the challenges of speech normalisation is found in a study by Jurafsky and Martin (2009).

Another challenge is prosody. Prosody (changes in rhythm, temp, **stress** and intonation to express different emotions, meaning or moods in speech) has a great effect on the naturalness of synthesised speech (as will be shown in this work) which is reflected in its wide discussion in the literature (Black, 2002; Strom et al., 2006; Black, 2006; Szaszák et al., 2015) and it has even been used to help segment speech corpora (Essa, 1998). Speech corpora have to include a recording with high prosodic variability to cater for the occurrences of the same phones in different parts of

the sentence. Pausing (an element of prosody) can change the meaning of the uttered speech fundamentally. As demonstrated by Kim Silverman (principal research scientist at Apple) (Silverman, 2012), the sentence "the nights that we met you were awful" could be understood in two different ways depending on prosody and specifically pausing. If one pauses after "met" then the speaker would be asserting that the person being met is awful, while if the pause is after "you", the person would be asserting that the nights were awful. This also affects changes in pitch as pausing and pitch are correlated. Methods to generate prosodic speech have been used and they have an effect on the size of the corpora required, as different versions of the same phones in phrases with different emotions are required. This is yet another reason to increase the corpus size, as shown earlier with context dependence of phones (Jurafsky and Martin, 2009).

Figure 1-1 shows an overview of the steps involved in speech synthesis, for which issues have been introduced in this section.

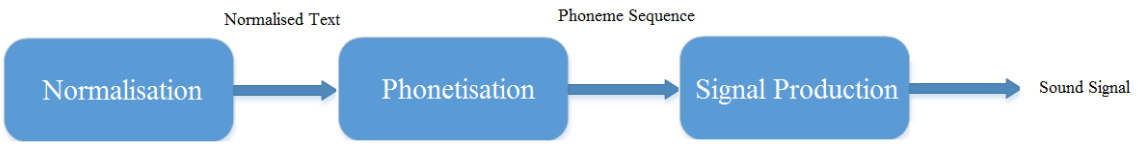

Figure 1-1. Overview of process of speech synthesis

The main issue this work focuses on is the difficulty of creating new speech corpora for under-resourced and under-researched languages, which lies in the rightmost rectangle in Figure 1-1. All the issues presented above influence what should be considered when creating such a speech corpus, despite the issues with normalisation and phonetisation. This is explained in more detail in Sections 1.2, 1.3 and 1.4.

In general, the following should be taken into consideration when creating a speech corpus.

- Corpus size: The length of the corpus can vary. The longer it is the better quality is produced. Usually a few hours of speech are required. This is to cover as much as possible of the language phonetics and the different stress and intonation features. Arabic is pitch-accented, stressed-timed language (Bertrán, 1999). Languages with these two features were focused on when conducting this research.
- Corpus content: The corpus should be diverse and should not be biased towards some phones and ignore others. This problem can be formulated as finding the smallest subset of the text corpus in which each letter, phoneme, diphone or triphone appears a certain minimum number of times (François and Boëffard, 2002). This could be done automatically or semi-automatically using different methods to ensure that the corpus contains as diverse a **phonetic**

transcription as possible. In this work, different methods for text selection will be reviewed, one of which will then be applied.

- Corpus Quality: Quality (not just sound quality) is very important. Uniform mood, loudness and speed of speech are often required for a general purpose speech corpus.

These factors are just a brief summary and the criteria for a good speech corpus (for each part of that corpus) will be explained in detail from Chapter 4 to Chapter 7.

In order to efficiently create such a speech corpus, a set of tools must be available to speed up the process and avoid manual labour which might take months or even years to complete. Building these tools requires knowledge about the target language's phonetics and phonology which this work claims to be lacking in Arabic in general (all dialects). Because of this, and all the difficulties outlined above, it was decided to carry out a process of building an Arabic single-speaker speech corpus in the **Modern Standard Arabic** (MSA) dialect for the purpose of speech synthesis, and solve the problems and knowledge gaps faced during the process. The choice of Arabic was mainly because it is less researched than other more popularly spoken languages such as English, French and German, specifically on the web. The study was conducted so that the knowledge gaps and challenges of building such a corpus in MSA were identified while conducting the different stage of **corpus design** and construction. This meant that the corpus design and construction had to start early to determine the problems, research questions, methodology and contributions involved in this work. In Section 1.4, the process of creating a speech corpus is explored in more detail. An overview of the knowledge gaps and problems are presented next.

1.2 Lack of Speech Corpora in Arabic

The challenges of speech synthesis are numerous and most of these challenges differ from one language to another. These challenges increase in difficulty the scarcer the resources available for the target language.

The aim of this work is to focus on the "signal production" stage of speech synthesis which is the stage that transforms the sequence of phonetic labels to sounds. It could be argued that "signal production" is a solved problem for English in particular dialects but this research aims to tackle the "signal production" task when the resources are scarce or not as rich in structure and to enable the construction of new speech corpora without the need for many human experts and man/hours.

The main problem tackled in this work is the lack of speech corpora. Only a few Arabic speech corpora are freely available: The first corpus was created by Almeman et al. (Almeman et al.,

2013), which has been acquired from its authors. Their speech corpus was built for speech recognition as it contains transcribed phrases with no granular segmentation at phone level. This corpus could be a valuable resource to this research and the possible uses of it will be investigated. The second corpus is the "KACST Arabic Phonetics Database" (KACST) (Alghamdi, 2003). This corpus was recorded by a group of 8 subjects as part of an experiment that was aimed at airflow, air pressure, linguapalatal contact, nasality, perception, side and front facial images and stroboscopic images of the glottis rather than speech synthesis and recognition. This corpus was made for helping research in speech recognition and synthesis among other things, and it is useful for this work as the recorded phones can be used in **bootstrapping** and training initial language models. It contains **utterances** of the form cvCvc, cvCCvc, Cvc and cvC where "c" is always the Arabic consonant /z/, "v" changes based on the utterance (/a/ for the first two utterance forms and /a/,/i/,/u/ for all of them), and "C" represents each of the 28 Arabic consonants (see Table 4-7 excluding foreign and additional realisations). Note that "CC" means that the consonant is geminated (doubled), but this notation will NOT be used in the rest of the work. The usage of "v", "c" and "C" to represent vowels and consonants in syllables will be explained when used.

There are works for single speaker Arabic speech corpora in the literature (Almosallam et al., 2013). They created a corpus using methods very similar to the ones used in this PhD work. This PhD work expands their phoneme set – which is missing some phonemes such as the emphatic consonants – and conducts a more comprehensive evaluation of the corpus quality. Their corpus is longer than the one produced in this PhD work (7 hours) and includes the electroglottograph signal (EGG). They did not specify how they collected the text in detail before the recordings. This PhD work benefits substantially from the work of Almosallam et al., 2013, and expands it to not only produce a single speaker speech corpus, but also introduce the steps, rules, guidelines and tools to create a single speaker speech corpus, specifically for MSA Arabic. Almosallam's corpus was not compared with the corpus produced here using listening tests. The fact that the corpus in this PhD work has certain improved features does not mean that Almosallam's corpus is not adequate for speech synthesis.

There are other Arabic voices either multi-speaker with no clearly declared purpose (Alsulaiman et al., 2011, 2013) or only available commercially but those were dependant on larger, accurately annotated corpora that are not available for free nor are the methods for creating them. An example of which would be the speech ocean Arabic **Text-to-Speech** (TTS) databases (Ocean, 2016).

There are other Arabic voices available commercially but they are dependent on larger, accurately annotated corpora that are not available for free nor are the methods for creating them published.

For example, even with the availability of recorded MSA from sources like audio books or newscasts, there are significant hurdles in using these for synthesis (Prahallad, 2010). First, the recordings might have background music, noise or other artefacts. Secondly, the recordings are not uniform in loudness, intensity and pitch change, even if recorded by the same person depending on their mood and state of health. Thirdly, the textual transcripts are not normally available and if they are, there is no guarantee that the anchors have adhered to the textual transcript and whether their pronunciation and co-articulation matches the assumptions made for this research. Also, going through the corpus to check for errors and peculiarities in pronunciation is an onerous, labour-intensive and complicated task.

Phonetisation, which is required for corpus design and construction, not just speech synthesis, also includes generating annotations like stress, pause and prosody in speech to make the speech more natural. This increases the size of the corpus required and poses another challenge if the method used is not Unit Selection as the prosody might be generated rather than recorded (Black, 2002).

When a professionally recorded corpus is available, the task of segmentation becomes easier as the transcript is available and is read accurately and uniformly. However, most segmentation methods (Amith, 2012; van Niekerk and Barnard, 2009) rely on previously trained models that are either used for speech recognition or speech synthesis. For low resource languages these are not available and if they were, there would not be much need for new ones.

To summarise, there is a lack of fully phonetically annotated, MSA speech corpora based on sound, published and comprehensive phonetic studies of MSA. This work addresses this problem.

1.3 Knowledge Gaps (Arabic Phonetics and Phonology)

In order to create the speech corpus proposed in this work, a set of tools is required. More on why these tools are required is explained in Chapter 4. These tools are needed to perform the following:

- Extract text from the digital medium to produce the orthographic transcript of a number of utterances.
- Reduce the orthographic transcript's utterance count until it is possible to process with the resources available (studio time and man-hours) with minimal effect on the phonetic coverage.
- Phonetise the reduced orthographic transcript's utterances. This generates the phoneme sequences (phonetic transcript).

- Segment and align the phonetic transcript with the recorded utterances.

In order to build these tools, a comprehensive study of Arabic phonology and phonetics (focusing on MSA) must be conducted in order to answer the following questions.

- What is the phonemic inventory of MSA? Put in a different way: Which phonemes are found in MSA connected speech?
- What are the rules which govern how the orthographic transcript is to be converted to the phonetic transcript to reflect how the utterances are going to be pronounced during the recordings?

The above led to the decision to conduct a study of MSA phonetics and phonology, which is discussed in Chapter 3.

Next, an overview of the process of creating a speech corpus is presented. This process is based on the preliminary study conducted during the first 9 months of this project. This process overview summarises this work, and the structure of this thesis.

1.4 Creating a Speech Corpus

Based on the preliminary study (Section 2.3 gives detailed support to this section), the process of creating the speech corpora involves four stages: Gathering the orthographic transcript, recording the corpora, generating the phonemic representation, and aligning the three together.

- Gathering the orthographic transcript involves collecting the text to be read and recorded by the **talent**. This also involves reducing the size of the collected text before recording in order not to exceed the allocated resources for creating the corpora and minimising the effect of this on phonetic coverage. This step also involves correcting orthographic and syntactic errors.
- Recording these corpora. This is not a trivial effort. Several hours of speech is usually required for Unit Selection and all the parts of the recording should be reasonably uniform in terms of speed (words per minute), loudness (the average amplitude of the speech signal) and mood (happy, sad, angry, singing…). In addition, the recording has to be of a good quality and thus recorded in a studio. Black (2002) explains in general the factors that must be understood when creating a speech corpus. This step also involves a second round of orthographic transcript corrections to ensure that the recorded speech matches the content of the transcript.
- Attempts have been made to create Unit Selection voices from recordings that were originally produced for different purposes, such as newscasts and audio books, because of the

availability of a transcript (Prahallad, 2010; King, 2013). This introduces issues of consistency, noise, background music, and sounds, which are not easy to remove. The transcripts of these recordings also do not necessarily correspond to the actual recording, for example, an error in a newscast that might include the word "apologies" with a correction. This is less likely the case when the transcript has been created prior to the recording and chunked into short sentences, which enables the talent (person whose voice is recorded), to record without silences and repeat on errors. Also, usually the whole recording has to be done by one voice talent and this puts a great strain on the talent's vocal tract and requires a considerable amount of time and resources as the talent needs to take breaks.

- Generating the phonetic representation of the transcript. This could be done automatically depending on the language of the recording. In the case of English, this requires a dictionary of phonetically transcribed words. In Arabic, this is an easier task with less ambiguity as words and multi-words are usually pronounced as written given that the transcript has the diacritics, which is not always the case.
- Aligning the recording with the transcriptions (lower part of Figure 1-2). This consumes the most time and resources out of the three stages and is the focus of this work. In this stage, each phoneme, syllable or other type of **phonetic unit** is assigned beginning and end time stamps in the recording. This is done in many ways and is heavily covered in the literature (Hosom, 2009; Van Bael et al., 2007; Stolcke et al., 2014; Yuan et al., 2013; van Vuuren et al., 2013).
- Transcription can either be done automatically and then optionally revised by a group of human experts, or done by a group of experts in the first place. Even the use of experts will not be 100% accurate because there are always disagreements between experts (Hosom, 2009; Van Bael et al., 2007; Zue and Seneff, 1996). These disagreements mostly arise on boundaries between a consonant and vowel, a consonant and a glide, or a glide and a vowel. The goal of automatic alignment systems is to achieve a close accuracy to the experts.
- Alignment can also be carried out manually and using this alignment to train a model which is used to align the rest of the corpus (**bootstrapping**). This step also could include boundary refinement (Peddinti and Prahallad, 2011; Jakovljević et al., 2012; Hoffmann and Pfister, 2010) after forced alignment (with or without bootstrapping) which has been shown to increase the accuracy of forced alignments.
- Evaluating the resulting speech corpus. This involves metrics to evaluate the quality of the corpus. Objective metrics assess how good the corpus is in terms of phonetic coverage and how accurate the alignments are (Barros and Möbius, 2011; Bonafonte et al., 2008). Subjective measures are used to assess the quality of speech generated using this corpus,

which give a 'black-box' conclusion of the overall quality of the corpus (Boros et al., 2014; Chalamandaris et al., 2013; Almosallam et al., 2013; Wester et al., 2015; Dall et al., 2014).

It is important to note here that the double ended arrow in the lower part of

Figure 1-2 is meant to represent the fact the forced alignment could accept previously trained models as input and will always produce trained models as output.

Figure 1-2 shows the overall process adopted here for producing the speech corpus.

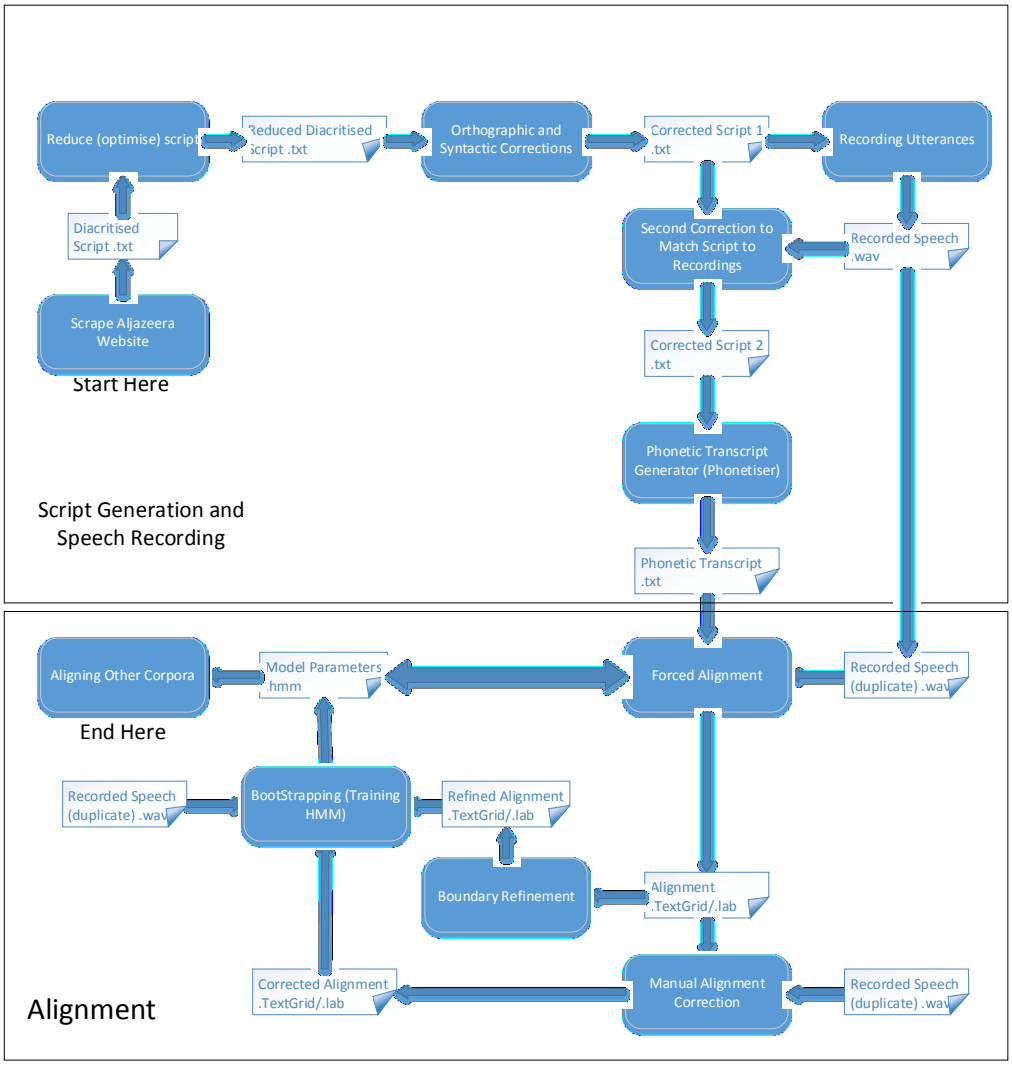

Figure 1-2. Speech Corpus Construction Workflow

Subjective and objective evaluations are conducted after each of the stages above.

1.5 Scope

The main purpose of this work is to tackle the issues of Arabic speech corpus design. To be more specific, the dialect chosen was MSA for two reasons:

- The availability of sufficient MSA digital content to be used for creating the orthographic transcript.
- The nature of usage of other dialects, when it comes to digital content, is beyond the author's expertise. To conduct a comprehensive study in orthography, phonetics and phonology of the different Arabic dialects, specifically when written dialects on the web are in question is a complex work in itself (Watson, 2007).

Further, the corpus to be designed in this work is to be targeted at the main methods of speech synthesis: Unit Selection Speech Synthesis and **Statistical Parametric Speech Synthesis**. The choices of the methods lead to the introduction of some criteria for the desired speech corpus, such as length (in minutes or words), number of speakers (single or multiple), content of the orthographic transcript, and types of metadata to be included with the corpus. All of this is discussed in more detail when necessary.

Finally, the domain had to be specified for which a speech synthesiser based on this speech corpus would be used. By domain it is meant the purpose of the application as used by the user. It was decided to use the term "Open Domain" or "General Purpose" speech synthesis as described in the literature (van Niekerk and Barnard, 2009), as there were no specific applications in mind before this project started.

1.5.1 Target Synthesis Methods

In order to evaluate the corpus after it is built, Unit Selection and Statistical Parametric Speech synthesis were set as a combined target application. The Unit Selection method is one of the types of more general "Concatenative Methods" in speech synthesis. Other methods include diphone concatenation, which produces less natural sounds but requires less (less than an hour) recorded speech and segmentation (Lenzo and Black, 2000).

In concatenative speech synthesis a sequence of speech units are chosen from a unit database that is populated from the segmented and aligned speech corpus. The units are chosen by phone identity and other criteria such as prosody, position in phrase, position in word, etc. Then, after performing acoustic modifications on the individual segments, they are concatenated to produce the desired utterance (Black, 2002).

In addition, there are statistical parametric methods. The naming here is not always consistent. What is usually meant by statistical parametric methods is the assumption that the data exhibits a probability distribution and the goal is to find the parameters for this probability distribution that optimises some criteria. The data in this case is the recorded speech with the aligned phonetic

representation and the features extracted from the phonetic representation. An example is **HMM**-based speech synthesis. Here, the input text is converted into a sequence of phones and features representing context which are extracted (part of speech, adjacent phones, pitch, prosody, …). Based on these phones and features, a sequence of context-dependent HMMs are chosen from the trained HMM database and these in turn generate the speech parameters (for example, mel-cepstral coefficients and the excitation). Then, the speech is synthesised from this low dimensional set of parameters using a vocoder such as STRAIGHT (Zen et al., 2007).

Other types of statistical parametric speech synthesis methods can be used but are not covered in this work. The literature review shows that the HMM-based speech synthesis is the most popular method used (Zen et al., 2007; Kim et al., 2006; Qian et al., 2008; Lu et al., 2011; Maia et al., 2007). However, recently, **Deep Neural Networks** (DNNs) have been used to synthesise speech successfully with good results (Zen et al., 2013), and these generally use the same type of corpora as HMM based systems.

1.5.2 Project Description

In this work, an MSA, single-speaker speech corpus for the purpose of speech synthesis was created alongside the tools required to create such a corpus. An MSA phonetics and phonology study was conducted to build those tools, and finally the corpus and theses tools were evaluated (objectively and subjectively).

1.5.3 Research Questions

It is important to note that not all the research questions were taken from the problems and gaps discussed in this work. Some were added during the process of producing the speech corpus. The questions follow:

1. What is the phonemic inventory of MSA? Alternatively: Which phonemes occur in connected MSA speech?
2. What rules govern how the orthographic transcript is to be converted to the phonetic transcript to reflect how the utterances are going to be pronounced during the recordings?
3. How can an accurate segmentation and alignment system be achieved, that has been trained on the proposed corpus? How can this be answered comparatively and take into account all the possible parameters that can be changed in that system?
4. How much coverage does the reduced orthographic transcript achieve for MSA phonetics and phonology?

5. How would a speech synthesis system based on unit selection and built with the proposed speech corpus perform when subjectively evaluated? The corpus is segmented and aligned by the system devised to solve research question 3.
6. Would adding stress markings to the speech corpus – based on the orthographically generated word stress markings from the work of Halpern (2009) – improve the results of the subjective tests conducted to answer question 5?

1.6 Research Contributions

The contributions of this work mainly lie in the field of phonetics, phonology and corpus design.

- MSA phoneme set and phonetisation rules: This contribution was an outcome of a triangulation of the review undertaken in Chapter 3 and expert opinions.
- An MSA, single-speaker, fully-evaluated speech corpus: This was the main outcome of this work. The corpus is published and freely available for research purposes and is the first of its kind to be built based on scientific research.
- An assessment of the quality of the speech corpus and the impact of stress features on it: This contribution was an outcome of the triangulation of data analysis of the results of laboratory experiments (objective and subjective). See Chapter 5 and Chapter 7.
- Ways of conducting and limitations of listening tests: This contribution was a triangulation of the data analysis of the results of a review and laboratory experiments (subjective only) and the literature. See Section 7.6.5.
- An ordered list of steps of how to create a speech corpus inferred by the results of this work. See Section 2.2.
- A software tool to phonetise MSA script (to convert normalised MSA script to a phoneme sequence).
- A software tool for segmenting and aligning MSA sound recordings with the phonetic and orthographic transcripts.

1.7 Summary

Now that the problem, scope and questions are set, Chapter 2 will discuss the different methods used for the various parts of this work. Chapter 3 discusses Arabic phonetics and phonology before starting the design and construction of the speech corpus. Chapter 4 to Chapter 7 will discuss the steps of creating and evaluating the speech corpus chronologically. Finally, Chapter 8 concludes this work by summarising the contributions and suggesting future work.

Chapter 2 Methodology

Chapter 1 introduced the gaps in the literature and the research questions which – when answered – may fill these gaps. In this chapter, the "How" and "Why" questions are answered, illustrating the methods used and why they were chosen.

Section 2.1 describes the literature search and selection process, while the methodology is presented as both a chronological summary of this work (Section 2.2), and then a list of the research methods used to answer the proposed research questions (Section 2.3), or to deliver a contribution, and an explanation for the choice of each method is given.

Section 2.4 addresses the criteria used for choosing experts, while Section 2.5 describes the ethics processes and guidelines that were adhered to during the listening tests.

Overall, this work applies a mix of methods and triangulates the results from each to confirm the conclusions. Previous work showed that the most prominent ways for collecting data for evaluating speech corpora were objective statistics taken from the corpora like phone or diphone counts, average utterance length and coverage (Barros and Möbius, 2011; Kominek and Black, 2014); and subjective listening tests like MOS (Mean Opinion Score) and preference tests (Boros et al., 2014; Sainz et al., 2012). The term "Triangulation" will be used to describe the use of several methods together to answer a single research question or produce a contribution. All parts of this work contain a "Traditional Review" (Jesson et al., 2011) providing a broad reach of both qualitative and quantitative studies that allow the gaps in the area of research to be explored.

2.1 Literature Selection Process

2.1.1 Keywords used in Search Engines

To select the literature required, the following list of search keywords was used as input to the search engines. The terms are related to the fields of Speech Synthesis, Speech Corpus Design and Phonetics, focusing on Arabic and MSA. These keywords were found during the preliminary study in the literature used in this work. This was after looking through the journals and proceedings chosen to be reviewed.

Speech Corpus Statistical Parametric Speech Synthesis

Corpus Segmentation	Arabic Speech Corpus
Corpus Alignment	MSA Speech Corpus
Corpus Selection	Listening Tests
Corpus Design	Research Methods
Corpus Evaluation	Statistical Testing
Corpus Construction	Speech Corpus Evaluation
Unit Selection	Arabic Prosody, Stress, Phonetics and Phonology
HMM Synthesis	MSA Prosody, Stress, Phonetics and Phonology

2.1.2 Literature Selection Process

The three most popular academic sources related to corpus design and automatic speech segmentation and alignment were selected, along with other less popular ones, which are related to the scope of this work.

In addition, some papers were included that were referenced by papers read previously and thought to be important, as they contained results that directly affected decisions made in this work. Google Scholar was also used when a very specific query was used (such as 'speech segmentation with ten month old infants').

The reason for the process presented in this section is the topic-sparsity and number of results returned by search engines like Google Scholar, ACM and IEEE. The sources selected contained the most relevant material, and the bibliographies of these works were also used to enrich the sources discussed.

Google Scholar was used for queries where it was known precisely what it was that was required because it includes a wide selection of conferences and journals. Books on machine learning and signal and speech processing were included. The year 2008 was chosen as the limit to how far back to look when reading through the literature, but older material was included if explicitly needed. Publications in 2015 were the most recent literature and conferences queried.

If a search query returned more than 100 results, the results were ordered by relevance and only the top 100 were considered in this study.

2.1.3 Literature Sources

The following is a list of the academic sources which were included as part of the literature review. They were chosen based on the preliminary study conducted and the recommendation of

peers in the field. The main focus was to find credible sources related to speech technologies and corpus design. The ranking of each of these conferences and the impact factor (3 years citations per doc) of each of the journals is stated below, but these rankings vary from one source to another. The sources used for these rankings are Scimago Lab (2016) for impact factors of journals, and Education (2016) for conference rankings.

- Computer, Speech and Language (Journal). Impact Factor 2.55
- Conference of the International Speech Communication Association (INTERSPEECH) (Conference, Big). Rank A
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Conference, Medium). Rank B
- Speech Communication (Journal). Impact Factor 2.38
- Spoken Language Technology workshop IEEE (Conference, Small). Ranking not available.
- International Conference on Language Resources and Evaluation (LREC) (Conference, Big). Rank C
- Journal of Language Resources and Evaluation (JLRE). Impact Factor 1.44

Literature was also gathered using several authors' names, including those whose work was considered critical for this research. For Arabic and MSA phonetics and phonology, specifically in the application of computational linguistics, Nizar Habash's work was used as a core source as Habash has been working in the general field of Arabic computational linguistics for over 15 years. Habash's book "Introduction to Arabic Natural Language Processing" contains a chapter about Arabic Phonology and orthography aimed at computational purposes (Habash, 2010). This was – alongside the work of Janet Watson (Watson, 2007) – the main source for Arabic phonetics and phonology. Another important author whose work has been thoroughly investigated for Arabic phonetics and phonology was Fadi Biadsy, who has been working in Arabic computational linguistics for 10 years.

Alan Black's and Heiga Zen's works on speech synthesis were used as the main source of knowledge about the advances of computerised speech output based on written text. Their work was important since the speech corpus built here is designed for speech synthesis. Alan Black has been working in this field for over 20 years, and Heiga Zen is currently employed by Google and has been working in speech synthesis (recently using Deep Neural Networks for synthesis) for over 15 years. They are both leading researchers in their field.

Corpus design is not usually considered a subject which researchers dwell on for a long period. The only author who has had career-long dedication to corpus design has been Jindrich Matousek.

Matousek has been working in this field for over 15 years and his work is considered seminal, as it is mainly in MSA speech corpus design.

All the authors mentioned above have had a significant impact and have been cited hundreds if not thousands of times in the literature.

2.2 Sequential Process Methodology

This work started with a "Traditional Review" (also called a preliminary study) to explore the problems and research questions to be answered. Throughout this work, 'Expert Opinion' was taken in the form of semi-structured interviews (Radermacher, 2006; Bryman, 2006). The reason for choosing this method is that some issues arose during this work that were not determined from the preliminary study and so it was not possible to prepare structured interviews with predefined questions. Therefore, the experts were informed of this semi-structured nature of the interviews and they were contacted when issues arose. The experts were either directly asked questions about MSA phonetics and phonology or they were instructed to perform a task (corpus segmentation and alignment). They were also asked to review two of the main research contributions of this work, the MSA phonetisation rules and the phoneme set (first contribution), and to provide feedback in further interviews until the rules and phoneme set had been agreed upon. Finally, a set of subjective lab experiments were conducted to evaluate the resulting speech corpus.

The specific nature of these formal interviews is explained in detail depending on the part of this work in which they were conducted. This work is made up of seven parts, each corresponding to a contribution defined in Section 1.6.

Following the preliminary study, the corpus design process began with the orthographic transcript collection and reduction, moved on to the recording, and then the segmentation of the speech corpus. This included further literature reviews to choose the most suitable methods.

The resulting speech corpus was then evaluated objectively and subjectively. This was done in the form of laboratory experiments. The objective tests showed how accurate the segmentations and alignments were, and the percentage of phonetic coverage of the corpus. The subjective tests were used to show the overall quality of the speech corpus as a form of black box testing. They were also used to show the impact of using orthographically extracted stress features on the generated speech.

The following list shows the different activities (after the preliminary study) in more detail and ordered sequentially. This process methodology is considered to be one of this work's contributions so that future works can refer to it when designing their research.

1. **Research Arabic and MSA phonetics and phonology:** The best name to describe the method here is a "Traditional Review" (Jesson et al., 2011). This was conducted to identify the main issues and research gaps from the literature.
2. **Creating a set of phonemes and phonetisation rules for MSA:** A follow-up from part 1. This involved expert opinion triangulated with the conclusions taken from part 1's review to create a set of phonemes and phonetisation rules for MSA.
3. **Research speech corpus design:** This is a "Traditional Review" (Jesson et al., 2011). This was conducted to find a set of guidelines for building a speech corpus for speech synthesis for low resource languages.
4. **Designing and building a speech corpus:** With the outcomes of part 3 and the opinions and practical help of experts, a speech corpus was designed and built. More details about parts 3 and 4 are given in Chapter 4 to Chapter 6.
5. **Research ways of evaluating speech corpora:** This was conducted to find the different methods for evaluating a speech corpus for speech synthesis and choosing the ones suitable for the scope of this work. The "Traditional Review" highlighted a lack of consensus about speech corpus evaluation.
6. **Evaluating the speech corpus objectively**: The evaluation procedure was informed by the review in part 5. By "Objective" is meant a process that does not involve people. The objective measures included statistics concerning phonetic coverage, length, alignment accuracy and number of errors expected in the corpus.
7. **Evaluating the speech corpus subjectively:** The evaluation procedure was informed by triangulating outcomes of expert opinion and the outcomes from part 5. Parts 6 and 7 produced quantitative data that was used to support the final conclusions.

The order of these parts corresponds to the order in which this work was undertaken, mirrored by the sections of this thesis.

2.3 Research Methods

This section introduces the methods used and decisions made for each part of this work. The organisation used here for research methods is taken from the Computing Research Methods taxonomy (CRM as referred to by Holz et al. 2006) developed by Glass et al. (2004). CRM is heavily used in the computing literature and the publication which introduced this taxonomy was cited by 55 others in Google Scholar.

After deciding on the problems, research gaps and scope of this work, the main decision-making points were as follows. **First**, formalising a high-level corpus design process to be adopted

throughout the work. **Secondly**, deciding which methods to use for acquiring and optimising the orthographic transcript. **Thirdly**, deciding which methods to use for segmenting the speech recordings to individual phones and aligning these to the phonemes of the phonetic transcript. **Fourthly**, finding the suitable methods for evaluating the resulting speech corpus (all of its elements which are relevant to the scope) objectively and subjectively.

These are now discussed in detail.

2.3.1 Methods for the Corpus Design Process

The high-level corpus design process used in this work was formalised as a result of the preliminary study. There is little disagreement in the literature on the general structure of the process of corpus design, but each publication fails to describe the process completely from beginning to end. Therefore, the literature was reviewed and a complete process formulated in this work which served as a guideline for the steps to be followed. One disagreement found in the literature was the required corpus length. As will be shown in Section 4.1, a number of different sizes have been suggested for the corpora reviewed in the literature. Because of this, and the limit of resources available in this work, it was decided to impose an upper limit on the desired corpus size and increase it if evaluation results were later unsatisfactory.

Another decision made as a result of the preliminary study was to conduct a phonetical and phonological review of Arabic (MSA in particular). This research was targeted specifically towards speech synthesis. This review was necessary for creating a phoneme set and building a phonetiser for MSA, which were essential for many of the following parts of this work. The phonetiser was used to generate the phonetic transcript of the scraped utterances to be able to calculate the utterance scores for reducing (optimising) the orthographic transcript. These phonetic transcripts were also necessary to align the recorded speech with the phonemes, as carried out in previous work (Bonafonte et al., 2008).

2.3.2 Methods for Acquiring and Optimising Orthographic Transcript

This work used a ‘greedy method’ for reducing (optimising) the initial orthographic transcript obtained from scraping a website containing fully diacritised text (Al Jazeera, 2015). The decision to use greedy methods here, was made because all reviewed publications used a form of greedy methods for corpus design (François and Boëffard, 2002; Bonafonte et al., 2008; Kawai et al., 2000; Kawanami et al., 2002; Tao et al., 2008). It was not the intention in this work to achieve 100% phonetic coverage (diphone coverage), which is sometimes the case in previous literature (Kominek and Black, 2014; François and Boëffard, 2002). Rather, it was intended to show that

the achieved phonetic coverage can be used to produce high quality speech. This is justified by the fact that the increase in quality of speech synthesis methods and systems resulted in loosening the requirement for complete phonetic coverage (Chalamandaris et al., 2013) and that the problem of coverage could be solved by recording more data later in the process (Black, 2002). What is meant by "loosening" here, is enough to produce good quality speech from the recorded and segmented data, which has been shown in this work to be true by using the greedy method, and then conducting the evaluations explained in Section 2.3.4.

One decision that remained to be made was the choice of the equation which calculates the score of every utterance in the orthographic transcript (see equation 1 section 4.2). There is no agreement on the form of the equation which calculates this score. But all the works agreed that the sentence (utterance) score should be proportional to the number of new diphones it covers (François & Boëffard 2002; Bonafonte et al. 2008; Kawai et al. 2000; Tao et al. 2008). In this work, another element was added to the score of the sentence, which is the number of times the diphones in the utterance appear in the corpus selected so far. The reason for selecting these criteria is to prevent very long utterances from being chosen unless they have a number of new diphones proportional to their length.

The process of utterance reduction is not included recent literature, and no comparisons have been conducted between the approaches suggested in previous literature. This could be due to the advances in speech synthesis technology which have made the need for complete phonetic coverage less critical, further confirmed by this work.

Chapter 4 contains more detail on the decisions made for corpus reduction and transcript preparation in general.

2.3.3 Methods for Segmenting and Aligning Recordings and their Evaluation

The HMM method and its parameters used for segmenting and aligning the recorded transcript are explained in detail in Section 5.3 and Section 6.3 respectively.

In this work, HMM forced alignment was used to segment and align the speech with the phonetic transcript. After the first attempt, the accuracy was not high enough, which led to manually aligning about 10% of the corpus and then using the manually aligned portion to train a model to align the rest of the corpus (bootstrapping).

The choice of this method was mainly its popularity in the literature (Stolcke et al., 2014; Karnjanadecha and Zahorian, 2012; Hosom, 2009) and ease of implementation. Other methods employ boundary refinement (Peddinti and Prahallad, 2011) or neural networks (Hosom, 2009) to

improve performance, which have proven to be effective. These were suggested for future development and were the fall-back case where the bootstrapped HMM model's alignment accuracy was not enough to produce high quality voice. Later it was shown that the bootstrapped HMM model's alignments were sufficient to produce high quality speech.

A widely used standard for precision was used to evaluate the accuracy of the alignments, which is the percentage of boundaries in the same position as the manually corrected ones with a certain tolerance (Stolcke et al., 2014; Yuan et al., 2013; Jakovljević et al., 2012). Results were presented for several tolerance levels but 20 ms has been considered as standard in the literature.

The precision was presented both before and after bootstrapping, and a measure of expert agreement showed the reliability of the gold standard alignments (manual alignments). All these results were compared with the work of Hosom (2009) to highlight that even a slightly less accurately segmented corpus, with fewer human resources and less expert agreement, can achieve high quality speech synthesis.

2.3.4 Methods for Objective and Subjective Corpus Evaluations

This is where most important decisions were made. Evaluating the speech corpus resulted in two of the main contributions of this work. The task was to conduct a full, black box test of the corpus when used for its designated purpose. This was to confirm that the corpus construction methodology in this work was at least sufficient for constructing a high quality speech corpus, and also to confirm that the orthographically extracted stress features improved the quality of speech generated using the corpus.

2.3.4.1 Objective Evaluation Methods

No objective measures were made on the acoustic signal generated for the listening tests. This was mainly due to the vulnerability of these measures to noise (Chevelu et al., 2015; Wester et al., 2015; Buchholz and Latorre, 2011; Latorre et al., 2014) (see Section 7.2).

On the other hand, objective measures were used to calculate the phonetic coverage, reduced corpus size and manual correction statistics of the phonetic transcript. Section 4.4 shows these results in detail. This works' speech corpus transcript identified over 89% of diphones appearing at least once, and over 86% appearing at least three times. This was within the two hour limit allowed for the recording of utterances, which was imposed by the researcher. This is a *de facto* measure usually employed in the literature to measure the phonetic coverage of a transcript (Kominek and Black, 2014; Kawai et al., 2000).

What is “sufficient” phonetics coverage? The answer is full coverage, but some corpora in the literature start from an utterance set that has 100% coverage and reduce it until the coverage starts decreasing making sure they are sacrificing no coverage in return for reduced recording costs (Kominek and Black, 2014; François and Boëffard, 2002). This was not possible here, due to the small size of the initial utterance set available, and the resource limits. In addition, it is more interesting – from a research point of view – to show the quality of this corpus, which is made from scarce resources for a low resource language, rather than going for the approach of recording a large amount of data, and using more resources. Full coverage does not help in overcoming one of the main challenges presented in this work, which is lack of resources.

2.3.4.2 Subjective Evaluation Methods

In Section 7.1, a full review of previous work on subjective speech corpus evolution is conducted and the different methods for subjective testing methods are analysed and compared. This resulted in choosing three different methods for subjective tests: preference tests, MOS, and DMOS (Degradation Mean Opinion Score) tests. MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor), SUS (Semantically Unpredictable Sentences) and other types of test were excluded for several reasons. One is the fact that preference and MOS-like tests are most commonly used in the literature. Another reason, specifically related to MUSHRA tests, is that listeners have to choose a score between 1 and 100 which has been shown to be more difficult, and the increase in accuracy in results diminishes after a certain score precision (Nunnally and Bernstein, 1994). SUS tests were rarely used in the literature and are used for evaluating different features of speech synthesisers to the ones required in this work (Benoît et al., 1996).

Section 7.5 shows in detail the setup of this work’s listening tests and why they were conducted the way they were (listening conditions, required number of subjects, chosen prompts and the presence of an instructor). It also highlights which parts of the tests were randomised and which were not and the reliability of each test’s results. A critique of MOS tests is also presented showing the weakness of these types of test. It is important to note here that the weaknesses of MOS tests – in spite the fact that they are the most popular in the literature – were the main reason why three types of test were conducted. Triangulating the results from these three tests consolidated the conclusions made in Chapter 8.

Decisions had to be made on which statistical significance tests to conduct, based on the analysis of the resulting data from the listening tests. Since most literature did not include these tests, it was necessary to consult experts in statistics to help make these decisions. The works in the literature which have carried out statistical tests differ on which tests to use (Clark et al., 2007a; Hirose and Tao, 2015; Mohammadi et al., 2014; van Niekerk, 2014). Because of this and after

consulting statistical experts (Gilbert, 2015; Green, 2016), it was decided to conduct both a Wilcoxon signed-rank test and two-way ANOVA for analysing the MOS and DMOS tests as there was a disagreement between the expert opinion and the literature. This emphasised the necessity of triangulating the results from the different tests to confirm the conclusions made.

A Pearson's chi-squared test was used for the preference tests as there was no disagreement between the experts and the literature. For more detail on the methodology and decisions made for the subjective laboratory experiments, see Section 7.6.

2.4 Criteria for Choosing Experts

The two experts chosen for this study had to comply with certain criteria, but are labelled "Expert 1" and "Expert 2" for anonymity. Table 2-1 shows the experts and their experience. The experts adhered to the following criteria:

1. A native Arabic speaker preferably with working-level English.
2. Over 18 years of age.
3. Obtained at least one degree in higher education.
4. Knowledgeable about Arabic language phonetics (Arabic teacher, translator, writer) and fluent in Arabic Grammar.
5. Has over 3 years work experience using the Arabic language.
6. Comfortable with using computers.

Table 2-1. Details of the two experts who participated

	Expert 1	**Expert 2**
Education	Arabic language	Arabic language
Age	51	31
Highest degree	PhD in Linguistics and Grammar	BA in translation of English and Arabic
Work experience	• 12 years teaching • 8 years translation • 5 years Arabic language consultant • 8 years working with Pearson's as an expert examiner	• 5 years in one-to-one Arabic tutoring • 1 year as a bilingual assistant working with children in the classroom • 10 years interpreting and translating experience between Arabic and English
First Language	Arabic (Moroccan born)	Arabic (Iraqi born)

Both were part of all the stages of this work and contributions, except for the evaluation.

2.5 Ethics

The subjective listening tests conducted as part of this work involved human participants. This work was conducted in the UK and hence has to comply with the rules of health and safety related to research in the UK (University of London, 2016). An application was submitted on 11 November 2015 for ethics approval with the specifications of the listening tests, the way the participants would be approached and the dates of the beginning and end of the study. The application was approved on 24 November 2015. The data collection was finished before the deadline of 29 February 2016. The speech stimuli used in the listening tests did not contain any explicit or abusive language, and was collected from public news data. The tests were conducted in normal rooms which were relatively quiet, as opposed to conducting them in studios or soundproof rooms. This helped to keep the atmosphere non-threatening for the participants.

Appendix C shows the participant test instructions, information sheet and the consent form, which give further details about the rights of the participants.

2.6 Summary

In this chapter, a summary of the methodology used and research design has been discussed, alongside the details of the ethics process conducted for the subjective listening tests. The relevant chapters and sections have been referred to where more in-depth discussions of the methods are available.

Following the ordering outlined in Section 2.2, the next chapter discusses the first stage of this work (after the preliminary study), which is a study of MSA phonetics and phonology aimed at applications for speech synthesis.

Chapter 3 MSA Phonetics and Phonology

A study of Arabic phonetics is required, mainly for choosing the criteria on which the optimisation of phonetic coverage in Chapter 4 will be based. This includes creating a list of all possible units to be covered by the corpus, and what metrics should be used to determine how good a text corpus is in covering the phones or combinations of phones (diphones, triphones,…).

3.1 Stress

A study of MSA stress is essential as phones (especially vowels) are articulated differently when the syllable they belong to is stressed (Barros and Möbius, 2011; Halpern, 2009; Kenworthy, 1987). Substituting a stressed syllable for a non-stressed syllable (and *vice versa*) in a speech signal will generate an unnatural utterance even if the concatenation points are optimal (Yi, 2003). Stress was covered in many of the publications reviewed. It was used as a feature of segments in speech corpora for both optimising the phonetic coverage before the recording and to help with choosing the best unit for concatenation in speech synthesis (Barros and Möbius, 2011; Kominek and Black, 2014). In the former, stress is usually a feature of vowel phones as stress affects vowels more than consonants (Biadsy et al., 2009; de Jong and Zawaydeh, 1999) (pitch is altered and vowel length is changed). Thus, a stressed vowel is considered a different phone to that of the same vowel non-stressed when optimising phonetic coverage of a text corpus for recording. This is sometimes referred to as vowel reduction, which affects vowels in unstressed syllables in Arabic (Kenworthy, 1987).

The difference in articulation of a stressed syllable varies from person to person and from dialect to dialect (Kenworthy, 1987; de Jong and Zawaydeh, 1999). Prosody has even been used in some publications to classify dialects successfully from an input recording (Biadsy and Hirschberg, 2009). This shows that the differences in prosody between dialects could affect the stressed syllables' realisation characteristics. Despite studies that show stress correlates with longer vowels, higher pitch and higher intensity, this is relative to the adjacent phonetic units and may not always be true between syllables from different utterances with different sentiments (Halpern, 2009; de Jong and Zawaydeh, 1999). Automatically changing the pitch of a syllable will – after a certain threshold – cause the natural recording to become robotic and unnatural (Kawai et al., 2000). Therefore it was decided to make sure all syllables, both stressed and non-stressed, were included, or to make sure that stressed and non-stressed syllables could be generated from smaller

units in the corpus which belong to stressed or non-stressed syllables (of different identity) accordingly.

The algorithm for determining stressed syllables in a text transcript is based on a set of rules presented in Halpern (2009) which is the latest stress study with MSA as the target language. Halpern (2009) showed how previous work in MSA stress does not take into account the different dialects and how stress varies in both its realisation and location in the words between dialects. He presents a set of rules based on research on how stress in MSA is actually realised and interviews with experts. This assumes 3 different types of syllable in MSA shown in Table 3-1, light, heavy and super-heavy, from which branch 6 subtypes.

Based in the rules in Halpern (2009), an algorithm was developed to extract potential stress from an input word assuming there is only one stress in a word. The steps are a series of conditional statements as follows:

1. If last syllable is super-heavy, then the stress falls on it, otherwise
    2. If the word has only one syllable then this syllable is stressed, otherwise
        3. If the word has two syllables then stress falls on the penultimate syllable, otherwise
            4. If the word has more than two syllables and the penultimate syllable is heavy, then the stress falls on the penultimate syllable, otherwise
                5. The stress falls on the antepenultimate syllable.

It is important that MSA **phonotactics** do not allow a super-heavy syllable of sub-type 6 (see Table 3-1) to appear anywhere except the end of a word. If this super-heavy syllable were to occur in the middle of the word, it means that there should be three or more consecutive consonants which do not exist in MSA phonology, but it is not known if this rule applies to other dialects.

Table 3-1. MSA syllables

Sub-type	**Syllable**	**Example**		**Type**
1	cv	لَ	Consonant + short vowel	Light
2	cvv	لا	Consonant + long vowel	Heavy
		لَو	Consonant + diphthong	
3	cvc	لُبْ	Consonant + short vowel + consonant	Heavy

Sub-type	Syllable	Example		Type
4	cvvc	عامْ	Consonant + long vowel + consonant	Super-heavy
		لَوم	Consonant + diphthong + consonant	
5	cvcc	شَدٌّ	Consonant + short vowel + geminated consonant	Super-heavy
		كَبْتْ	Consonant + short vowel + consonant + consonant	
6	cvvcc	شادّ	Consonant + long vowel + geminated consonant	Super-heavy

To avoid any differences between our formalisation of stress in MSA and what is actually recorded, a brief had to be prepared informing the talent of our concept of stress and give them feedback throughout the recording about their realisation of stress.

Other works have discussed stress in Arabic. Gadoua (2000) presented similar rules to the ones in (Halpern, 2009) and added that each full word in Arabic (they did not use MSA and their scope was Quranic Arabic) has one or more stresses, but did not detail how a word could have more than one stress. They also claim that stress could have two levels in Arabic, primary and secondary. They also did not elaborate that.

3.2 Prosody

Prosody is loosely defined as the changes of rhythm, intonation, stress (including loudness, also called intensity) that express the speaker's emotion or state and the type of utterance being spoken (declarative, interrogative…).

For prosodic coverage, it is important to define the scope of the corpus to be produced. The terms "domain specific" and "open domain" (sometimes called "general purpose") voices have been used occasionally to distinguish between corpora recorded for a specific domain (automated phone answering services is an example) or corpora recorded with no specific usage in mind (Black, 2003; Clark, Richmond, et al., 2007; Krul et al., 2007), mainly used in screen readers to narrate text on a personal computer. Speech synthesisers could incorporate prosody generation (generating pitch contour and optimal pause locations) before generating the speech (Lindstrom et al., 1996; Malfrère et al., 1998; Xijun Ma et al., 2004). This is to make the voice more natural and expressive.

Prosodic features include stress, intonation (pitch), rhythm, pausing and sometimes word accent. In this work:

- Stress has been covered in Section 3.1. This work's corpus includes all light syllables with and without stress and all endings of heavy syllables (nucleus and mora) with and without stress.
- Rhythm: Arabic is a stressed timed language (Bertrán, 1999). This means that in Arabic, stressed syllables occur with roughly equal intervals between them no matter how many non-stressed syllables occur in between. The longer the unstressed syllable sequences between the stressed syllables are, the shorter these unstressed syllables are usually pronounced. This phenomenon is related to vowel reduction as the vowels of these unstressed syllables are shortened in this case. In Arabic though, reduction is done on a smaller scale in contrast to English where light syllables are sometimes pronounced very quickly and almost disappear (Kenworthy, 1987). Function words and some suffixes (Dual forms) in Arabic (pronouns, preposition, conjunctions with some exceptions,…) receive stress unlike English (Halpern, 2009), which hints that the effect is weaker on unstressed syllables than English. This is based on a study of Arabic that focuses on the Iraqi and Egyptian dialects and not MSA. In this work, it is uncertain what the effect on rhythm will be before recording the talent's dialect (Levantine speaker in MSA). The talent is instructed to speak with a consistent speed (words per minute), but this does not mean that they avoid stress-timing, rather speak in what is to them MSA rhythm.
- For pausing, every phoneme in our phonemic vocabulary was included before a word boundary. To make sure that the effect of co-articulation does not reduce the coverage of consonants followed by word boundaries, the talent was instructed to utter some of the short word-final consonants followed by a "sokoon" with a short pause after. All vowels are included before phrase boundaries so the same procedure was not instructed for vowels (see Appendix F).
- Sentence stress, sometimes referred to as contrastive stress, is giving a word a certain emphasis to make it stand out as a more important part of the utterance. The realisation of this type of stress is usually a rapid change in pitch and/or intensity and/or adding a pause after the word (Kenworthy, 1987). This phenomenon was considered too strong emotionally and context sensitive in this work, so the talent was instructed not to emphasise any word in the utterances in the transcript (see Appendix F).

In domain-specific corpora, neither phonetic nor prosodic coverage is important as specific utterances are recorded and then the same utterances are used in production with no modifications. In open-domain corpora, even though the prosodic scope is not specified when

building corpora, minimising the prosodic effects on the recording without affecting the comprehensibility and naturalness of the speech is usually recommended. This means that the talent is instructed not to speak with excess emotion of any kind (Black, 2003). Generating emotional speech is not part of this work and the talent was instructed to use declarative intonation for all utterances (see Appendix F).

There is agreement that the more automatic the changes to the pitch and speed (duration) of natural human speech recording, the more unnatural it becomes (Bozkurt et al., 2002; Clark, Richmond, et al., 2007; Maia et al., 2007). Since it is not known in advance in which method of speech synthesis this corpus will be used (parametric or concatenative), it is intended to make prosodic coverage as high as possible without making the recording difficult for the talent. The number of unreliable recordings is thus reduced and without making the size of the transcript for recording unfeasibly long.

In this work, the voice talent is instructed in the brief not to express strong emotion and to speak in consistent pitch patterns that include roughly three main pitches of their choice (their comfortable pitch, one higher and one lower), so the higher pitch and the lower pitch being the upper and lower bounds accordingly and the comfortable (neutral) pitch is roughly in between. It is impossible for the talent to speak in the exact range of pitch given in the instructions, but they are continuously given feedback from the sound engineer (who can see the pitch changes on a screen) about their pitch changes after each recording. At the beginning of each recording session the talent listens to their recordings from the day before so that they keep pitch and speed consistent between sessions.

To estimate the pitch range that the talent should stay within, this work reviews publications in speech synthesis in which the authors attempted to automatically modify the f0 (fundamental frequency) of human speech segments to make them more suitable for context. The reason for reviewing these works is to see if it is possible to find a threshold of the ratio of change in f0, above which any change in f0 would cause the segment to become unnatural or incomprehensible. Kawai et al. (2000) carried out a perceptual test where users had to give a score out of five of how natural ten words sounded when they modified their duration and fundamental frequency in different ratios. When considering the score 4 as the minimum acceptable, any ratio of change between –0.2 and +0.2 is considered acceptable. Kawanami et al. (2002) based their work on Kawai et al. (2000) and decided to record the corpus 9 times with f0 and phone duration altered by the talent. f0 had three variations (natural, 0.4 octave higher and 0.4 octave lower), and phone duration had also three variations (natural, 0.5 octave higher and 0.5 octave lower). The talents had to be instructed not to change their pitch and speed for every recording.

In this work, exact numbers for ratio of pitch or phone duration change are not given to the talent. This is because the talent said it is difficult to accurately maintain a pitch they are not comfortable with. For phone duration, vowels appearing in different contexts and in different syllables (stressed or non-stressed) will have different length (usually longer for stressed syllables) (de Jong and Zawaydeh, 1999). This work does not intend to produce a corpus that is suitable for multiple speed synthesisers, but it is possible to change speed of generated speech using signal processing with high accuracy and little effect on naturalness (Karrer et al., 2006).

In this work, the whole orthographic transcript was intended to be recorded twice, once with a declarative intonation and another with an interrogative intonation (question). We claim that this is easier to explain to the talent and will cover a wide pitch range because it is usually assumed that declarative intonations are falling, while interrogative intonations are rising (Malfrère et al., 1998). Even if this does not apply to all dialects, as different dialects in Arabic have different phonotactic (Watson, 2007) and prosodic properties, the talent was instructed to use rising intonation for questions and natural speech for the declarations. However, the corpus was not recorded twice but only once, with declarative intonation unless there was a question mark in the transcript. This was because of limited resources but does not mean that recording with a rising intonation is not necessary in corpus design.

The talent was always required to start speaking from their comfortable pitch and speak naturally without rapid changes in frequency that exceeded 0.4 octave change from the starting pitch.

3.3 Gemination

Gemination or “shadda” (“tashdeed”) in MSA and in Arabic is generally described as the doubling of a consonant so that the resulting segment is double the length of its non-geminated counterpart (Selouani and Caelen, 1998). Gemination as term is used in different ways in the literature, but here a geminate consonant is defined phonetically as an elongated consonant that is phonemically different from the same non-geminate consonant (Newman, 1986). In Arabic orthography, gemination is represented by adding the “shadda” diacritic (ّ) above the consonant with an optional short vowel diacritic appended above or below the “shadda”.

Linguistically and pedagogically, “shadda” (ّ) is considered to be a letter rather than a diacritic as it causes the addition of another consonant to the word which is used for syllabification (Halpern, 2009). It is common practice for Arabic language teachers to describe gemination as the doubling of a consonant, performed by repeating the consonant twice with the first instance *not* being followed by any short or long vowels (followed by silence or “sokoon”) and immediately

followed by the second instance of the consonant which is to be followed by a short vowel (determined by the optional short vowel diacritic appended to the "shadda"), long vowel or a stop (usually at end of phrases).

Phonetically, gemination is not simply the consonant doubling process described above. But this description is useful for its simplicity and also determining the syllabic structure of a word, because – unless the geminate consonant is not followed by either a short or a long vowel (followed by a stop) – geminate consonants occur on syllable boundaries, the first (hypothetical) consonant belonging to the leading syllable and the second (hypothetical) consonant belonging to the following syllable. For example, the word كَتَّبَ – which roughly means "he dictated" – is made up of the following sequence of phones types (left to right): "cvCvcv" where we use the capital letter to represent the geminated consonant. To analyse its syllable structure, we convert the geminated consonant "C" to "cc" which is equivalent to a consonant cluster, which makes the sequence "cvccvcv" yielding three syllables "cvc", "cv" and "cv". And the geminated consonant is split between the first and second syllables.

In practice, gemination is not realised simply as a doubling of a consonant, but by increasing the duration of the articulation of the consonant, and this realisation differs depending on the type of consonant. For plosives (stops), the length (duration) of the low energy region before the explosion is increased in gemination. This region is sometimes called the "plosive closure". For all other consonants (fricatives, nasals, approximates,…), the length of articulation of the spectrally stable section of the phone is increased (Selouani and Caelen, 1998; Essa, 1998). The geminated consonant is not merely a repetition but rather it is a new phoneme to be added to the vocabulary to be considered for optimising the phonetic coverage of the corpus. Even though some consonants are rarely geminated (and sometimes only in very restricted contexts), they have been included in this study and missing (non-allowed) contexts were added as nonsense sentences (see Section 4.4).

This adds to the 28 consonant phonemes established earlier another 28 geminated consonant phonemes to our phonemic vocabulary.

3.4 Nasalisation

Nasalisation occurs in classical Arabic in specific contexts. This excludes the nasal consonants "n" and "m" that are always voiced in Arabic. Nasal consonants and also vowels adjacent to them in Arabic can be nasalised. It is not imperative that a nasal consonant be nasalised. Nasalisation, among other phenomena, is part of "Tajwid", which are the rules governing pronunciation of verses from Quran. It is not imperative that text is read with these rules in MSA to be correct or

comprehensible. For the purpose of corpus design, all nasalisation should be avoided as it has been shown that nasalised segments in concatenative speech synthesisers can cause a mismatch at phonetic boundaries when trying to use the nasalised segments in contexts where it is not appropriate (Yi, 2003). This would require adding a feature to each segment determining its nasality, which increases the corpus size and is not needed for MSA as nasalisation is not considered to affect semantics nor is it part of prosody. The talent was instructed to avoid nasalisation (see Appendix F) and nasalised vowels and consonants were not added to the phonemic vocabulary.

After revising the recorded speech corpus (see Section 4.5), no nasalisation was noticed by the experts except for three instances of the same foreign Arabised word يونيو (roughly phonetically "j u: n j u") which means "June", where the nasalisation affects the second "j" consonant following the nasal consonant "n". Since this is a very specific context, it does not affect the overall phonetic coverage of our corpus.

3.5 Emphasis (emphaticness)

Consonants in Arabic (not just MSA) can be phonologically divided into three categories. Always emphatic, always non-emphatic, and two-state (could be either emphatic or non-emphatic depending on context). Note that by emphatic here is not meant the *phonetic* class of consonants but rather a *phonological* feature of some Arabic consonants (Laufer and Baer, 1988).

Some vowels in Arabic can also become emphatic if followed or preceded by an emphatic consonant. The changes in articulation of the **emphasised** vowel are enough to make it a different phoneme all together. 4 vowels in Arabic may be emphasised (shown in Table 3-3); these include the two diphthongs as they start with a "fatha" or (ـَ) or /a/ which is the emphasised part of the diphthongs. Not all the consonants emphasise the preceding or following vowels in dialects of Arabic (including MSA) (Watson, 2007). Even in classical Arabic, Sibawayh did not include (ق) or /q/ in the emphatic set of consonants as it is not velar nor emphatic (sometimes called pharyngeal) when it comes to the place of articulation (Laufer and Baer, 1988). This follows the rule that emphaticness is either achieved by velarisation or pharyngealisation and /q/ falls under neither of those categories. In practice in MSA, some speakers, depending on dialect, would pronounce vowels proceeding or following /q/ emphatically. The talent was instructed to pronounce vowels around /q/ emphatically all the time in order not to get incompatible concatenation points between emphatic consonants and non-emphatic vowels in case the corpus is to be used for concatenative synthesis. From the above, the 4 emphatic vowels shown in

Table 3-3 were added to the phonemic vocabulary for optimisation.

The two consonant phones (ل) and (ر) (/l/ and /r/ respectively) on the other hand could appear emphasised and non-emphasised depending on the following vowel whose emphaticness in turn is determined by the identity of the following consonant (whether it is emphatic or not). This was deduced by the experts from listening to this work's recordings. This means that emphaticness would propagate through these two phones to the adjacent vowels. This was not anticipated before the recording, the effect of which will be discussed in Section 4.3.

Emphasis spread in MSA is not the subject of study. Speakers tend to apply rules of emphasis spread from their local dialects when speaking MSA. Most studies of emphasis and emphasis spread have presented different and often overlapping sets of rules for different dialects (Laufer and Baer, 1988; Watson, 2007), sometimes without specifying the dialects and with little evidence that only comes from classical Arabic (Laufer and Baer, 1988). For example, in Cairene Arabic the rules of emphasis spread are many and could span the whole word. In this work, to simplify the optimisation process, only adjacent vowels are affected by the propagation and the talent was informed not to spread emphasis any further unless it is difficult to articulate (see Appendix F).

This does not mean that the synthesisers developed by this dataset will not be able to produce speech in which emphasis is spread. It only enables the optimisation algorithm to deterministically predict whether a vowel is emphatic or not in order to choose an optimal subset of utterances. Concatenative synthesisers, for example, can use different emphasised or non-emphasised vowel segments from different sections of the corpus to produce the required utterance with the required emphasis spread.

In this work, the talent is presented with a set of rules about emphasis with examples before the recording. When an agreement about emphasis rules has been reached, more sentences are added to the final recording session as required.

Table 3-2. Emphatic consonants in MSA

Phonetic Class	**Always emphatic**		Two-state	Always non-emphatic
Emphatic-Dental		tˤ ط		All other consonants
Emphatic-Alveolar	sˤ ص	dˤ ض ðˤ~zˤ ظ	ɬ~l ل rˤ~r ر	
Uvular	q ق	ɣ~ʁ غ		
Velar		x~χ خ		

Table 3-3. Vowels and diphthongs affected by emphatic consonants in MSA

Vowel	**/a/**	**/aa/**	**/aw/**	**/aj/**
Arabic Representation	◌َ Possibly Pharyngealized	ا ى Possibly Pharyngealized	وَ	يَ

3.6 Diphthongs

Arabic (not just MSA) theoretically only allows two diphthongs (وَ) and (يَ) with corresponding IPA representations /aw/ and /aj/ respectively. When preceded by an emphatic consonant, both diphthongs are realised from a different phonemes. These phonemes are the same as the ones above but with the /a/ part becoming emphasised. The IPA symbol for this diphthong or the emphasised /a/ is not known in this work. All four phonemes were included in our phonemic vocabulary.

Two extra diphthongs were added to the final text transcript because they were found in the automatically generated script after optimisation, but not in all contexts. They were not added to our optimisation as they are not phonotactically valid but still pronounceable. The diphthongs are (وِ) and (يُ).

3.7 Summary

This chapter reviewed MSA phonetics and phonology. Stress, prosody, gemination, nasalisation, emphasis and diphthongs were studied in order to formalise the phoneme set and phonetisation rules which will be carried out in Chapter 4 and Chapter 5. These features of MSA were chosen as they could potentially affect the realisation (acoustic features) of certain phones.

Nasalisation was not considered to be essential to this work and the talent was instructed to avoid nasalising vowels around nasal consonants. This way the identity of a vowel phoneme would not be affected by the identity of the nasal consonant.

Stress was considered important as it has a strong effect on the realisation of phones. Replacing a stressed vowel phone with a corresponding non-stressed vowel phone (or *vice versa*) in synthesis leads to distinguishably unnatural speech (Biadsy et al., 2009; de Jong and Zawaydeh, 1999) and vowel phonemes corresponding to stressed phone realisations were annotated as stressed in the corpus.

In a similar fashion to stressed vowels, geminated consonants were treated as separate phonemes. This is because of the significant change in articulation and change in meaning when changing a non-geminated consonant to a geminated one (Newman, 1986).

Emphasis (pharyngealisation) was considered as well. Emphasised vowels were treated as separate phonemes as emphasis has a strong effect on realisation of phones (Laufer and Baer, 1988; Watson, 2007).

It can be argued that being too specific when annotating the identity of phonemes in the corpus – in terms of emphasis, stress and gemination – might lead to issues in phonetic coverage as there would be more phonemes with different emphasis, stress and gemination states to cover. But this can be dealt with by merging the identities of these phonemes later if coverage is low.

All this, and the effect of it on transcript collection and corpus segmentation, will be discussed in more detail in Chapter 4 and Chapter 5.

Chapter 4 Transcript Collection, Reduction and Recording

The transcript was collected from Aljazeera Learn (Al Jazeera, 2015), a language learning website which was chosen because it contained fully **diacritised** text which makes it easier to phonetise. The transcript was split into utterances based on punctuation, to make it easier for the talent during the recording sessions.

After splitting the transcript into short utterances, the transcript was reduced (see Section 4.1) while maintaining acceptable phonetic coverage, then inspected to normalise the text and correct errors. This inspection was completed after reducing the transcript in order to decrease the manual labour required to clean the text, but this meant that the numbers and abbreviations were not included in the phonetic optimisation as they were not previously normalised.

After the reduction and before the recording, the text transcripts extracted from Aljazeera Learn were inspected and normalised. Abbreviations and numbers written in digit form were converted to word form. This is because the talent said that it was difficult to produce the correct inflection for numbers phrases while reading, if they were not written as words. In this phase, unwanted characters were removed and replaced with their word representation; for example '$' was converted to دولار which means "Dollar". After the inspection, only Arabic words were left in the transcript.

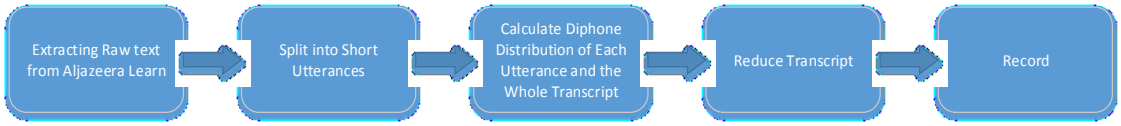

Figure 4-1. Collection and Reduction of Transcript

4.1 Corpora and Transcript Size

In order to make a decision about the size of the speech corpus needed (orthographically transcribed, phonetically transcribed, segmented, or a combination of these), previous work was looked at, most importantly, the size of the manually segmented corpus since it requires the most resources to create.

Many of the systems reviewed did not explicitly provide the duration of the training speech corpora used. However, some did and some provided the number of utterances, words or sentences. Also, the sizes of the corpora between training and testing are different and it is important to look at these numbers so that this work is easier to compare with the literature.

4.1.1 TIMIT

The TIMIT corpus (Zue and Seneff, 1996) was used in a number of systems reviewed. It contains 10 sentences spoken by 630 American-English speakers. The size of the TIMIT corpus is 6300 utterances (sentences). Karnjanadecha et al. (2012) used the TIMIT corpus to build their HMM models and for evaluation they used a subset if the OSMLA database that contains 636 utterances after they manually segmented it to utterances. Obin et al. (2013) used the TIMIT corpus for evaluation of their system, which did not need a training corpus as it does not use machine learning methods. However, they used 4620 utterances for training other supervised methods they compared their own against, where all the tests were done using the remaining 1680 utterances for all systems. Mporas et al. (2010) used the TIMIT corpus to evaluate their system that used several regression methods to choose the optimal boundary locations from the output of 112 different HMM models that vary in their setup (number of states, Gaussians, context dependency and features used). Hosom et al. (2009) used 3.145 hours of speech from the TIMIT corpus for training and 1344 files (sentences) corresponding to 49261 phonemes for testing.

4.1.2 Other Corpora

Many other works reviewed used the TIMIT corpus (Brugnara et al., 1993; Hoffmann et al., 2010; Yuan et al., 2013; Kalinli, 2012; Amith, 2012; van Vuuren et al., 2013), but the details are not mentioned because they are redundant.

van Niekerk & Barnard (2009) bootstrapped HMM models for three different African languages using the TIMIT database after mapping all phones used in the TIMIT corpus to broader categories. Their evaluation sets were relatively small. They used 21 minutes (12341 phones) of speech for Afrikaans, 20 minutes (8559 phones) of speech for isiZulu, and 46 minutes (26010 phones) of speech for seTswana.

Peddinti et al. (2011) used a relatively long evaluation, single speaker set (5 hours), manually segmented for the Telugu language. They did not provide details of the number of sentences in these 5 hours of speech.

Jarifi et al. (2008) used a long corpus for evaluation of 7300 sentences in French and 8900 for English. For training they used a 100, 300 and 700 sentence corpus for each language and tested for each separately.

Stan et al. (2012) used an audio book that contained 155,261 words divided into 7498 utterances for the first stage of their segmentation system.

Mporas et al. (2009) used a speech corpus of 5500 words divided into 500 utterances manually segmented for evaluation. They did not use a training corpus as their method only used flat start training.

Jakovljević et al. (2012) used a corpus that contained 900 sentences to train their HMMs but only segmented 50 of them manually for testing a segmentation system for Hebrew. They used flat start training so did not need a training corpus.

Other methods reviewed did training and testing with corpora with similar characteristics to the ones above or gave insufficient explanation of their training and testing materials and have not been included here because their work has no impact on the decisions to be made.

4.2 Optimisation (Orthographic Transcript Reduction)

All the works reviewed for corpus optimisation for speech synthesis use greedy methods (François and Boëffard, 2002; Bonafonte et al., 2008; Kawai et al., 2000; Kawanami et al., 2002; Tao et al., 2008). Greedy methods, as explained in the "National Institute of Standards and Technology" (Black, 2005), are methods that apply a heuristic that finds a local optimal solution that is close to an initial solution. The initial solution and the heuristics were different between works in the literature. Also the unit of choice for optimisation (triphone, diphone, phone…) varies. Greedy methods do not guarantee the production of a globally optimal solution as the corpus selection problem is Non-deterministic Polynomial-time hard (NP-hard) (François and Boëffard, 2002), which needs a brute force search to find the optimal solution. This requires astronomical processing power as the number of possible solutions is 2^n, where n is the number of sentences. In this case the number of solutions is 2^{2092} which is more than 10^{600}.

François & Boëffard (2002) classified greedy algorithms into three categories:

- Greedy: The initial solution is the empty set and then utterances that increase coverage the most (relative to solution at iteration) are added to the solution, until a certain target coverage is achieved or a limit is reached.

- Spitting: The initial solution is the whole sentence set and then sentences that contribute least to coverage are removed iteratively until a utterance removal would damage coverage in some way.
- Exchange: Starting from a specific solution (could be the output of one of the two methods above) exchange one of the solution's utterances with one of the utterances excluded from the solution if this exchange increases coverage, until no increase in coverage is possible. This maintains a static set size.

François & Boëffard (2002) used diphone as their unit and did not mention prosody or stress in units. The criteria for the three different approaches above are simple. They used unit counts from each sentences to give a score. "Useful units" in a sentence are units that would contribute to the corpus coverage (taking into account the need to have multiple units with the same identity. 3 in their case) while "useless units" are those that are redundant as the set already has a number of units with the same identity that equals or is higher than the limit (3 is the limit chosen in this work. See Table 4-1). They used unit counts with the sentence cost (length) in different ways which they compared. They have shown that using "Spitting" after "Greedy" methods improves coverage cost (number of chosen sentences and their average length) but does not necessarily increase phonetic coverage. The way they combined the two methods is by running "Greedy" and then running "Spitting", restricting its choice of sentences to the output of "Greedy".

Table 4-1. Statistics of this work's Al Jazeera transcript after reduction

Minimum number of occurrences for each diphone	1	2	3 *(The chosen limit)*	4
Number of utterances	468	700	884	1025
Number of Words	5624	8982	11560	13479
Recording length (hours)	~ 1.1	~ 1.6	2.1 (3.7 hours with nonsense sentences (see Section 4.4))	~ 2.5

Since the primary concern in this work is coverage and not the length of the corpus, but the length of the generated speech (2 hours maximum for proper utterances), the "Spitting" method was chosen to reduce the transcript to a size that would potentially generate between 1.5 and 2.5 hours of speech. In future work, a combination of the above methods could be used.

In short, an equation was used (see below in the current section) to give all the utterances a score. The utterance with the lowest score was excluded and the equation used again to re-score the

utterances. This is repeated until a stopping criterion is reached. In this PhD work, the stopping criteria was that "no utterance can be removed without jeopardising the coverage of one or more diphones". In other words, a diphone occurring a number of times less than a controlled threshold. See Table 4-5.

To select criteria for iteratively choosing utterances, a simple count was adopted where each utterance is scored by the following formula:

$$\begin{aligned} US(U,C) &= \sum_{k=0}^{n} \frac{UUF_k(U)}{CUF_k(C)} \quad if\ CUF_k(C) > UUF_k(U)\ for\ all\ k \\ US(U,C) &= -1 \qquad\qquad\qquad otherwise \end{aligned} \tag{1}$$

where $US(U,C)$ is the "Utterance Score" of the utterance U relative to corpus C; $UUF_k(U)$ is the "Utterance Unit Frequency", which is the number of times a specific unit indexed by k appears in the sentence U; $CUF_k(C)$ is the "Corpus Unit Frequency" which is the number of times a specific unit indexed by k appears in the corpus C at a certain stage of the optimisation.

It is important to note that no claim is made that the equation used for optimisation is optimal. In the literature, several methods have been used and no comparison has been made between them. The choice of this equation did not seem to be considered critical in previous work (Kelly et al., 2006; Kominek and Black, 2003; Barros and Möbius, 2011; Bonafonte et al., 2008; Matoušek and Romportl, 2007a).

The optimisation process started from the initial solutions being the whole set of 2092 utterances and iteratively removing those utterances which had the lowest $US(U,C)$, excluding utterances which have a score of –1. The processes stopped when removing any utterance would cause at least one phonetic unit to occur in the transcript below the allowed limit. The allowed limit was a controlled parameter.

In this work, diphones were used as basic phonetic units for optimisation. The reason for using diphones is that was the most choice in the literature reviewed (Kelly et al., 2006; Kominek and Black, 2003; Barros and Möbius, 2011; Bonafonte et al., 2008; Matoušek and Romportl, 2007a), and the numbers of possible units for phones, diphones and triphones (see Table 4-2) favoured choosing diphones. Optimising using phones as units is trivial as there are only 67 chosen for the optimisation (see Section 4.4) and phone optimisation is not ideal as it is well known that some co-articulation effects between phones spoken in sequence are not reproducible when using phone segments from different contexts, which is the case when phone optimisation is ignored. Triphone

optimisation has been reported by the literature (Matoušek and Psutka, 2001), but no coverage measure was given to enable a comparison against diphone optimisation.

In this work, 3 occurrences of each unit as a target are assumed and diphones are chosen as the target unit. Triphone optimisation was excluded as it meant there would have to be at least 3 * 300763 = 902298 triphone instances occurring in the corpus, and this is too good to be true as the unit distribution always follows biased distributions in human generated transcripts. If this perfect scenario is assumed, in the target 12000 word corpus every word would have to contain more than 50 unique and novel triphones. This is a very unrealistic constraint as shown more clearly in Table 4-2 containing all the possible frequencies of each phone type and the corresponding value in the corpus before optimisation.

Table 4-2. Theoretical Unit frequencies for different types of units

Phonemes	Diphones	Triphones
67	$67^2 = 4489$	$67^3 = 300763$

4.3 Optimisation Vocabulary

Table 4-7 lists full information about MSA phonemes used in this work.

Not all diphones were included in the optimisation. The optimisation only included "short syllable diphones" and "half syllable diphones" (see Table 4-3). A short syllable is a syllable starting with a consonant (could be geminated) and ending with a vowel (could be long), while a half syllable is the second part of a syllable ending with a consonant (a vowel followed by a strictly non-geminated consonant). Both of these terms are used in this work for convenience and are not defined elsewhere.

Table 4-3. Diphones included and excluded from optimisation

Short syllable diphones	**Half syllable diphones**	**Excluded Diphones**
cv	Vc	Cc
cV	Vc	
Cv		
CV		

V means long vowel and C means geminated consonant.

4.3.1 Short Syllable Diphones

Some short syllable diphones were excluded for the following reason. Emphatic consonants cannot be followed or preceded by a non-emphatic diphthong or a non-emphatic /a/ or /aa/ which are (ا) and (ـا) in Arabic script respectively. This excludes 14 * 2 + 14 * 2 = 56 diphones of this form.

The validity of these exclusions was only theoretical and based on rules of Arabic phonology before the recording (Watson, 2007), but were found to be true in the talent's speech, as the experts found during the correction phase, after the recording. The talent never emphasised a diphthong after a non-emphatic letter or *vice versa*.

According to the above, theoretically, there are 56 * 10 = 560 possible short syllable diphones. 56 represents the number of consonants doubled to include geminated consonants. The number 10 represents the number of vowels. This exclusion leaves 560 – 56 = 504 short syllable diphones included in the optimisation.

4.3.2 Half Syllable Diphones

The above short syllable diphone set, explained earlier, covers syllables of the form "cV", "CV", "cv" and "Cv". Where syllables of the form "cvc", "Cvc", "cVc" or "CVc" are to be synthesised by a concatenative speech synthesiser, which have a consonant coda (syllable ending) that is not followed by a vowel (otherwise the coda would have belonged to the following syllable), it would be useful to have segments of the form "vc" or "Vc", where the consonant ("c" part) is followed by a pause or another consonant rather than a vowel. This is because consonants which are followed by a vowel are highly co-articulated with the following vowel (Yi, 2003) making them unfeasible to use for concatenatively creating syllables which end with a consonant as these are not followed by a vowel and hence should not include this co-articulation effect. Half syllable diphones of the form "vc" were added to the phonemic vocabulary. Table 4-4 shows how a concatenative speech synthesiser would hypothetically create each of the heavy and super-heavy syllables ending with a consonant coda. It is important to note that in the "vc" diphones, the vowel could either be long or short as its identity is merged just for optimisation. This assumes that when concatenating "cv" and "vc" diphones to create a heavy or super-heavy syllable, the length of the vowel in the syllable is determined by the vowel in the first syllable.

Ignoring long vowels in half syllable diphones (as explained above) leaves 6 vowels (one of which is emphatic) and excludes 4; and leaves 56 consonants (geminated and non-geminated). A further exclusion would be of diphones which are made up of a non-emphatic vowel /a/ followed by an emphatic consonant. This leaves the exclusion of 4 * 56 – 1 * 14 = 210 half syllable

diphones. (The minus term because the diphones of the non-emphatic vowel /a/ followed by an emphatic consonant have been excluded earlier and should not be excluded twice.)

Table 4-4. Generation of heavy syllables from short and half syllable diphones

Short syllable	Half syllable (the vowel corresponds to the vowel in the short syllable)	Heavy and super-heavy syllable
Cv	vc or Vc	Cvc
Cv	vc or Vc	Cvc
cV	vc or Vc	cVc
CV	vc or Vc	CVc

We also included consonants at phrase endings (before a pause) as part of the phonemic vocabulary. Silence (represented as /sil/ in this work) is considered a phone in its own right. This is to avoid any effect of co-articulation on the consonant being followed by another phone (consonant or vowel). This consonant can be used at the end of phrases by concatenative speech synthesisers and the concatenation point would be the region of low amplitude before the consonant (Yuan et al., 2013). This adds 66 * 2 = 132 diphones in the optimisation. 66 is the number of phonemes (excluding /sil/), and 2 is the pause (/sil/) phoneme both succeeding and proceeding the consonant or vowel. The diphone /sil sil/ was excluded.

4.3.3 Consonant Clusters and Vowel Clusters

Two consecutive consonants "cc" that occur in MSA Ali & Ali (2011) were excluded from the optimisation for three reasons:

1. "cc" diphones constitute a big part of Arabic diphones. Theoretically, there are $28 * 28 = 784$ possible "cc" diphones in Arabic out of 4489 total diphones. So being able to exclude some of them from the optimisation process, increases the possibility of reducing the dataset size and simplifies the problem. But the question is: How much would this damage the phonetic and prosodic coverage in the corpus?
   It is important to note that "Cc", "cC" and "CC" diphones are not possible in MSA (but could be in other dialects). This is because a consonant cluster of more than 2 items is forbidden. This further excludes 3 * 28 * 28 = 2352 diphones from the total 6724.
2. The 784 theoretically possible "cc" diphones do not all occur in Arabic, ignoring foreign imported words. 246 "cc" diphones are either do not occur or very rare in Arabic

(Alderete and Frisch, 2009). The study from which these numbers were taken does not state which "cc" diphones these are, but does say to which consonant class (articulation type) each of the consonants in the diphone belongs. It is safe to assume that many of these clusters will not be found in the corpus transcript used for this work before optimisation.

3. Yi, Jon Rong-Wei (2003) shows how certain concatenation points between specific types of phone are better than others and would generate natural sounding speech when used in concatenative synthesisers. One of these is the very brief period of silence and gathering of pressure before the release of a stop letter, and other consonants which involve the same phenomenon on a different scale (Tench, 2015; Yi, 2003). This would make it possible to construct those consonant clusters from smaller units by concatenating at the low amplitude region before the consonant. Following the recording, it was noticed that the region of low amplitude is clear before stop consonants and less significant before other consonants. To try to alleviate this issue, a consonant from each of the articulation categories was chosen and for each an utterance from the recordings selected. The low amplitude before these consonants was further attenuated and no effect on naturalness was noticed by the experts. Subjective testing will be conducted later to further justify this finding. The attenuation of the low amplitude period shows that these points can be used as concatenation even when the consonant is not a stop.

"vv" vowel clusters were excluded as they do not occur in this work's model MSA, and also following the syllable structure introduced by de Jong & Zawaydeh (1999) and Halpern (2009). This excludes 10 * 10 = 100 diphones.

Both types of cluster exclude 784 + 100 + 2352 = 3236 phonemes.

In summary, the diphones left (the diphone "sil sil" was excluded as well, hence the extra 1 in the formula below):

$$\text{Diphones remaining} = 4489 - 56 - 3236 - 210 - 1 = 986$$

4.4 Results of Reduction

Table 4-6 lists results based on all **986** diphones included in the optimisation. More detailed results are given in Halabi (2015). After running the optimisation script, 884 utterances were left in the data set out of the complete 2092. The optimisation process was run through several times with changing the threshold for the minimum number of diphone occurrences allowed. The threshold 3 was chosen because of resource limitations (15 hours recording studio time and talent

time); more utterances were planned for recording should extra studio time remain (see Table 4-1).

Even with the threshold set at 3 it does not guarantee that all diphones have occurred at least 3 times in the optimised corpus. Diphones that occur below the chosen threshold – before the optimisation started – were not included in the optimisation process and any utterance that includes them at all is never excluded.

Table 4-5. Optimisation results

Threshold	Number of words	Number of utterances
Before optimisation	23531	2092
1	5284	463
2	8407	700
3	10958	884
4	12785	1025
5	14397	1150
6	15554	1245
7	16653	1334
8	17575	1414

The row in blue was chosen based on resources available.

To cover the gap of these underrepresented diphones, 896 nonsense utterances were recorded. Nonsense utterances have been used before in the literature to study language phonetics (including Arabic) (Alderete and Frisch, 2009; Kain et al., 2007; Laufer and Baer, 1988). The benefit of using them is being able to cover many units with less material, but a talent may find them more difficult to pronounce and this could potentially slow the recording time and cause more errors in the final recording output. The absence of syntax makes the prosody of the generated utterances potentially random. The nonsense utterances used here are experimental and after recording them, the talent did state that they were more difficult than news transcripts. The fact that they were generated by a template made the effort easier as the talent recorded more of them since they were similar in length and orthographic structure, and utterances from the same template were grouped together on the prompt shown to the talent. The nonsense utterances were

automatically generated using 4 templates. The parenthesised entries are replaced by a short syllable diphone to generate a nonsense utterance and underlining represents stress. Some stress depends on the diphone, which is not shown. The templates are shown next in Buckwalter transliteration:

1. /(cv)Sbara wata(cv)S~ara watu(cv)SA(c)un taSar~u(cv)/
2. /(cv)sbara wata(cv)s~ara wati(cv)sU(c)in tasar~u(cv)/
3. /ta(Cv)Saw~ara wata(Cv)Sara watu(Cv)Sa taSa(Cv)/
4. /ta(Cv)saw~ara wata(Cv)sara watu(Cv)sa tasi(Cv)/

Templates 1 and 3 guarantee that all short syllable diphones with emphatic vowels are included, and templates 3 and 4 guarantee that all short syllable diphones with geminated consonants "C" are included, and in templates 1 and 2, mild /u1/ and /i1/ short syllable diphone are included. All the templates repeat the same diphone in different locations in the word to include stressed and non- stressed diphones. Note here that the replacement is only done orthographically. The eight vowels in Arabic, the 28 consonants and the 28 geminated consonants were used to replace "v", "c" and "C" respectively. But those vowels (including in diphthongs) are uttered emphatically or non-emphatically depending on the context in the template. This generated a total of 28 * 8 * 4 = 896 nonsense utterances that cover all the short syllable diphones (four times each at least) with different stress. Half syllable diphones were not included as this would have doubled the amount of recoding required.

It is suggested that future work could add emphatic, non-emphatic, stressed and non-stressed vowels (a stressed vowel being a vowel in a stressed syllable) as separate phonemes in the optimisation process. This would require much more data as shown in the results.

Table 4-6. Coverage statistics for different parts of the transcript

Part	Aljazeera before optimisation	Aljazeera after optimisation and normalisation	Nonsense utterances	Aljazeera after optimisation with nonsense utterances
Number of diphones covered at least once	561	544	547	669
Percentage of diphones covered at least once	74.70	72.44	72.84	89.08
Number of diphones covered at least three times	492	476	545	646
Percentage of diphones covered at least three times	65.51	63.38	72.57	86.02

Finally, the short and half syllable diphones left a total of 896 + 884 = 1780 utterances for the recording (see Section 0 for more information on the recording and error correction procedures). The coverage of these utterances is shown in Table 4-6 for each of the nonsense utterances and the news transcript and both combined. Table 4-7 shows the complete set of phonemes used in this work excluding geminated consonants which are represented by doubling the consonant phoneme's symbol. The symbols on the right of the columns will be used to refer to phonemes hereafter.

Table 4-7. Final Phoneme set (82 in total)

أ	<	ر	r	غ	g	ي	y	ـُ	u0	[ـُ]	i1
ب	B	ز	z	ف	f	ڤ	v	ـِ	i0	[و]	uu1
ت	T	س	s	ق	Q	پ	p	(ا)	AA	[ي]	ii1
ث	^	ش	$	ك	k	ج	G	(و)	UU0	([ـُ])	U1
ج (ʒ)	J	ص	S	ل	l	ج (d͡ʒ)	J	(ي)	II0	([ـِ])	I1
ح	H	ض	D	م	m	ا	aa	(ـَ)	A	([و])	UU1
خ	X	ط	T	ن	n	و	uu0	(ـُ)	U0	([ي])	II1
د	D	ظ	Z	ه	h	ي	ii0	(ـِ)	I0	pause	sil
ذ	*	ع	E	و	w	ـَ	A	[ـُ]	u1	distor-tion	Dist

*For simplicity, geminated consonants are not included in the table. The left hand column in each section represents the phoneme in Arabic script while the right hand column is the **Buckwalter** representation.*

Phonemes Revisited for clarification (Left: Arabic. Middle: IPA. Right: Buckwalter) except for last section where there is no IPA available

أ	ʔ	ʿ	ر	r	r	غ	ɣ	G	ي	j	y	ـِ(i)	i	i0
ب	b	B	ز	z	z	ف	f	F	ڤ	v	v	[ـُ]	ǫ	u1
ت	t	T	س	s	s	ق	q	Q	پ	p	p	[ـِ]	ɪ	i1
ث	θ	^	ش	ʃ	$	ك	k	K	ج	d͡ʒ	J	[ا]	ɑ:	AA
ج	ʒ	J	ص	sˤ	S	ل	l	L	ا	æ:	aa	[ـَ]	ɑ	A

ح	ħ	H	ض	dˤ	D	م	m	M	و	u:	uu0	[و]	ọ:	uu1
خ	x	X	ط	tˤ	T	ن	n	N	ي	i:	ii0	[ي]	ɪ:	ii1
د	d	D	ظ	ðˤ	Z	ه	h	H	ـَ	a	a	sil	N/A	sil
ذ	ð	*	ع	ʕ	E	و	w	W	ـُ	æ	u0			

Diphthongs for general knowledge (Left: Arabic. Right: IPA)							
ـَي	/æj/	ـَو	/æw/	و(ـَ)	/ɑw/	ي(ـَ)	/ɑj/

Green means: Only in foreign words used in Arabic like فيديو

Blue means: Vowels

Black means: Consonants

Red means: distortion (not included and only used if experts find useless and noisy segments)

4.5 Recording Utterances

The recording of the corpus was spread over 5 days. Each day involved a 3 to 4 hour session including one or two breaks to avoid straining the talent's voice. This is the same time as reported by Matoušek & Romportl (2007b) and two hours more than Oliveira et al. (2008). The fifth recording day was used to go through the recordings and rerecord unreliable utterances. A sound engineer, the voice talent and at least one expert were always present during the recording. The expert provided feedback to the talent about speed, emotion, loudness and pitch consistency and errors in pronunciation. The sound engineer started each session with a sound check to test if the talent was an acceptable distance from the microphone for human voice recording and to produce recordings with consistent loudness. Loudness and speed were less of an issue as long as the talent spoke within a comfortable range set by the sound engineer. The sound engineer was able to change the speed and intensity (loudness) of recordings based on the experts' opinion and the readings from the software used (Pro Tools 11) without affecting the naturalness of the recordings.

The sound engineer also played recordings from previous sessions to the talent at the start of each session, and when the expert felt that the talent was deviating from the acceptable ranges described above. The talent was a native Arabic speaker and recordings were repeated on request if he felt it wasn't suitable for our purpose.

The recording was done in a studio. The equipment used was Neumann TLM 103 Studio Microphone known to be used for high quality human speech recordings. It had a pop shield to reduce the sound impact of exhaled air on the microphone. The talent sat in a soundproof anechoic recording booth. The booth only contained a prompt screen and the microphone. After the recording was finished, the sound engineer went through the whole recording in order to perform the following edits:

- Adding short silences at the beginnings and ends of utterances. This is needed to give each recorded phone a context (a preceding and trailing phone) as pauses are modelled as phones in HMM forced alignment, which is used later.
- Performing "Dynamic Range compression" for the intensity (loudness) of all the utterances. This is used to make intensity as uniform as possible with a dynamic gain that is multiplied by the signal to keep the signal within a set limit. –12 db was chosen by the sound engineer but it is possible to re-export the output with different limits.
- Reduce the length of speech pauses that are too long. No specific length was agreed but the sound engineer was given feedback about reducing long pauses which keeps acceptable variability in pause length without jeopardising the automatic alignment whose precision might be affected by long pauses.
- Normalise speed (change speed of each utterance separately to a predefined speed).

The sound engineer was also given feedback after the second error correction phase about the errors still in position, in order to fix them and redeliver the recordings. The errors included only clipped phones next to pauses, unreliable edits (recording radically different from transcript) and speed inconsistencies.

The recordings were delivered in separate files for each utterance (1780 files in total with 33 extra utterances because of residual studio time) which correspond to 17040 words overall after transcript corrections. 896 of the utterances correspond to the sentences that were automatically generated. The rest correspond to the utterances chosen from Aljazeera Learn (Al Jazeera, 2015) and optimised automatically (see Section 4.4). Each utterance starts and ends with a short pause of about 100 ms. The speech was not delivered in one large file as it is known that sequence models align shorter utterances more accurately than longer ones (Moreno et al., 1998). Having utterances with different lengths is usually considered a goal as it may enrich the prosodic coverage of the corpus (Umbert et al., 2006a; Vetulani, 2009). This corpus's utterance statistics are shown in Table 4-8. It is not claimed here that these statistics are optimal. The lack of diacritised Arabic text constrained the choice of utterances and the optimising utterance length distribution is not covered in this work.

Table 4-8. Recording Statistics

	Total Utterances	**Nonsense Utterances**	**Proper Utterances**
Count	1780	896	884
Average duration (sec)	7.5	5.9	9.0
Mode duration (sec)	5	5	5
Maximum duration (sec)	36	8	36
Minimum duration (sec)	1	3	1
Total Duration (hours)	3.7	1.5	2.1

After completing the recording sessions, two experts went through the corpus in sequence (for more scrutiny) to correct orthographic errors in the transcript and to change the transcript so that it reflected what was actually pronounced by the talent. All punctuation was removed and a special symbol was used to represent a pause. Most pauses were easy to detect as they were long enough (over 0.3 seconds). Since some pauses or errors were hard to detect at normal speed, the speed of the recordings was slowed down in this correction phase. Even with the speed reduced, it was hard to detect some hesitations in word boundaries and to decide whether to classify them as a pause or not. The decision was made to classify as a pause any word boundary that could be pronounced more naturally if the two phones surrounding the boundary were uttered closer to each other. This is justified by the fact that if these two phones do not naturally follow each other, the pause mark would tell the synthesis system that these two phones do not naturally follow and their concatenation (in case of concatenative speech synthesis) would be give a high cost (less likely to be chosen) (Yi, 2003).

Later, in the phonetic "manual corrections" phase (see Section 5.4), the experts were allowed to remove or add pauses that were incorrectly added or missed in the transcript. In this phase, it is easier to classify a segment as a pause or not, because the signal's spectrum and amplitude is visible to the expert.

4.6 Summary

This chapter presented the process and methods of transcript collection and reduction. This included a review of previous works on corpus design and construction, which resulted in using a greedy algorithm to reduce a transcript scraped from a language learning website (Al Jazeera, 2015).

The discussion regarding the phoneme set, formalised for the reduction process, is considered the most important element of this chapter as it is one of the main contributions of this work. After formalising these phonemes, the exclusions from the complete set of diphones (pairs of phonemes) were presented. These exclusions were part of the reduction process as only non-excluded diphones were considered necessary to be covered by the corpus.

Chapter 5 Corpus Segmentation and Alignment

The terms segmentation and alignment are used interchangeably in the literature to describe the general processes of annotating a speech corpus with phone labels and finding the timestamps of the boundaries that delimit those phones. This could involve annotating pauses and stress (Braunschweiler, 2006).

In this work, the term segmentation will be used to refer to annotation of the speech corpus with a sequence of phone labels taken from the phonetic transcript of this corpus which is in turn automatically generated from the textual transcript as described in Section 5.1. Segmentation also involves finding boundaries that surround these phone labels. The timestamps of these boundaries do not have to be 100% accurate (or anywhere close to that) in segmentation, but the sequence of phone labels must match the audio. Both the creation of the phonetic transcript from the textual transcript, and the segmentation of the corpus, are performed automatically in this work. The former was completed by an algorithm developed in this work and the latter using HMMs built using the **Hidden Markov Model** Toolkit (HTK) framework (Young et al., 1997).

Alignment is the determination of the exact timestamps of the phone boundaries. This can be done either automatically (using boundary refinement techniques or HMM models as described above) or manually by a group of experts whose job it is to correct the boundaries generated from the segmentation (or they could perform segmentation and then alignment manually, which is known to be very time consuming). Note that the segmentation process could produce high precision alignments as shown in previous works (Hosom, 2009). This depends on the quality of the recording, speech, text transcript, phonetic transcript and the algorithm used for segmentation (and alignment in this case).

The experts manually corrected a portion of the corpus in order to assess whether they needed to align the whole corpus manually. This correction was used to assess the quality of the automatic segmentation, how closely the experts agreed, and the quality of any further alignments carried out using the same algorithm with different parameters or using the manually aligned data to bootstrap the automatic segmentation process.

The size of our created corpus exceeds 3 hours of speech. To avoid manual segmentation of the corpus, forced alignment (Murphy, 2012) was used in different modes to create an initial segmentation of the corpus (see Section 5.3). This was carried out after the corpus transcript had

been revised twice by the experts, so at this stage, the corpus transcript had to be converted into a phonetic transcript to be used for segmentation and alignment. The following is a summary of the steps of the segmentation and alignment process.

1. Generating the phonetic transcript: The text transcript is automatically converted to a phonetic transcript which includes phonemes from Table 4-7. In this work, the phonetic transcript was in the form of a **pronunciation dictionary** because the software used for alignment requires a pronunciation dictionary as input with the textual transcript.
2. The dictionary contained several possible pronunciations of each word.
3. Automatic Segmentation: The phonetic transcript and the speech corpus audio are used as input to forced alignment that produces the segmentation with initial boundaries.
4. Manual corrections: Three experts inspect a portion of the segmented corpus to correctly align the boundaries with the speech. This could be repeated, where in each iteration the corpus is automatically realigned after the system is trained on the manually aligned data (leaving some for evaluation). The precision of this alignment is calculated to determine if an acceptable precision has been reached or if more iterations will not increase it further.

5.1 Generating the phonetic transcript

This was done automatically using: classical Arabic orthography rules (Elshafei, 1991; Thelwall and Sa'Adeddin, 2009; Watson, 2007; Ali and Ali, 2011; Gadoua, 2000; de Jong and Zawaydeh, 1999; Halpern, 2009), the nature of the text transcript harvested from the web, and the dialect of the speech talent (Levantine from Damascus). The experts noticed that different segments of the text taken from different articles applied different rules for orthography. This was dealt with by creating a list of all these rules. During the text transcript's error correction stage (see Section 5.4), the experts discussed and assembled what they found and added a list of rules as they corrected the script. The complete list of rules for generating the phonetic transcript is given below.

1. All characters that are not MSA letters or diacritics are omitted. Even letters in classical Arabic that are no longer used need to be omitted. Letters to be excluded are shown in Table 5-1.

Table 5-1. Classical Arabic characters excluded from the transcript

Description	Unicode	Arabic Script

Description	Unicode	Arabic Script
Arabic Tatweel	U+0640	ـ
Subscript Alif	U+0656	◌ٖ
Superscript Alif	U+0670	◌ٰ
Alif Wasla	U+0671	ٱ

2. All punctuation characters are omitted because the experts located the pause locations during the manual correction of the textual transcript. This renders the punctuation characters useless as the locations of pauses are already known. However, the punctuation could be used later for prosodic feature extraction as the prosodic features of utterances correspond strongly with punctuation (Taylor, 2009).
3. Arabic orthography is described as a phonemic orthography (sometimes Arabic script and alphabet are called "phonetic", having the same meaning) and the correspondence between letters and phones has been studied in the literature (Watson, 2007; Elshafei, 1991; Newman, 1986). This allows Arabic letters to be thought of as phonemes. However, as will be shown, this is not always the case. Arabic symbols (letters and diacritics in this case) usually correspond to phonemes in a regular manner, although in rare instances there are alternative pronunciations in Arabic. Arabic (including MSA) includes a set of words (nouns and function words) which have an implicit "Alif" vowel (ا) which is not written and corresponds to the /aa/ vowel phone in Table 5-2. This set of words is small and unchanging. The system uses a table lookup method to resolve those words when they are encountered, where the phonemic transcriptions of each of these words is predetermined by the experts. Note that these words could be affixed or suffixed but their pronunciation stays the same.

Table 5-2. Irregularly pronounced words in Arabic

Arabic word	Pronunciation	Arabic word	Pronunciation
هَذا	/h aa TH aa/	ذَلِكُمْ	/TH aa l i0 k u1 m/
هَذِهِ	/h aa TH i0 h i0/	أولَئِكَ	/AH u l aa AH i0 k a/
هَذانِ	/h aa TH aa n i0/	طَهَ	/T aa h a/

Arabic word	Pronunciation	Arabic word	Pronunciation
هَؤُلاءِ	/h aa AH u0 l aa AH i0/	لَكِنْ	/l aa k i1 n/
ذَلِكَ	/TH aa l i0 k a/	رَحْمَنْ	/r a H m aa n/
كَذلِكَ	/k a TH aa l i0 k a/	لله	/l AA h/

4. Manually annotated silences were represented by the phone /sil/ in the phonetic transcript.
5. All consonant letters except Waw and Ya' (و and ي respectively) are simply converted to their phonetic representation without ambiguity (see Table 4-7). An exception is when the consonant is followed by a Shadda (ّـ), when it is represented by a doubling of the consonant's phonetic representation. For example, /b/ (ب) becomes /bb/ (بّ).
6. Ta' marboota (ة) is converted to "t" if followed by a diacritic, otherwise it is ignored.
7. Madda (آ) is converted to a glottal stop /</ followed by /aa/ or /AA/ long vowels based on the amount of emphasis.
8. Vowels are emphasised if they follow or precede an emphatic consonant with the exception of /x/ (خ) and /g/ (غ) which only affect following vowels and not preceding ones. Emphasis is represented by capitalising the vowel's phonetic transcription's representation (see Table 4-7).
9. Short vowels /i/ and /u/ corresponding to diacritics (ـِ) and (ـُ) have – in addition to the possibility of being emphasised – the possibility of being *leaned* towards /a/ or (ـَ). This means that the pronunciation of the /i/ or /u/ will be closer to a Schwa. The phenomena is not documented anywhere and was noticed by the experts after recording the corpus. The talent leaned towards /i/ and /u/ when these vowels preceded a word-ending consonant which is not followed by a short vowel. In the phonetic transcription this is represented by the numbers 0 and 1. 0 meaning "not leaned" and 1 meaning "leaned". For example, the /i/ in the word مَغْرِبْ , which means "west" or "morocco" (pronounced as /m a g r i1 b/), is phonetically represented as /i1/. (See Table 4-7).
10. Waw and Ya' (و and ي) are transcribed phonetically as either vowels or consonants, determined by their context. If followed by a vowel, they are identified as consonants. If followed by a consonant then the preceding phone determines their identity; if preceded by a vowel, they are consonants, otherwise vowels.

11. Alif (ا) is transcribed as a vowel /aa/ or /AA/ depending on emphasis. An exception is a type of Alif called Hamzat Alwasel which is not pronounced in Arabic (including MSA). Also, Hamzat Alwasel becomes a glottal stop /</ at the beginning of sentences or phrases (after silences). Alif is realised as a Hamzat Alwasel when it is the first letter in the word, or the second (after an affix).

The phonetic transcription produced was in the form of a pronunciation dictionary similar to the ones used in speech recognition systems, for example HTK and Sphinx (Young et al., 1997; Lamere et al., 2003). The dictionary is a long list of orthographic representations of words each followed by their corresponding phonetic transcript. Note that multiple repetitions of the same orthographic representation can occur showing different possible pronunciations. "Hamazt Alwasel" in rule 11, when not in the beginning of the word, is ambiguous and could be pronounced or not. Both pronunciations were added to the dictionary to be resolved in the forced alignment stage, as HTK will choose the most probable sequence of phonemes that generates the speech signal. Other instances of ambiguity are Alif ا after Waw و at the end of a word. Here the Alif is not pronounced if the Waw is a plural Waw, which is difficult to automatically determine with high precision (as in foreign words transliterated into Arabic). For example, the word "Nicaragua" is written نيكاراغوا in Arabic and the final Alif represents a long vowel phoneme /aa/. Both possible pronunciations for each word ending with a Waw followed by an Alif were included. Word-ending long vowels were also optionally shortened in the pronunciation dictionary due to the phenomena of vowel reduction (de Jong and Zawaydeh, 1999; Biadsy and Hirschberg, 2009) which was noticed in this corpus as well.

5.2 Automatic Segmentation

The automatic segmentation was done using flat start forced alignment in a similar way to the method described in The HTK Book (Young et al., 1997). HTK version 3.4.1 was used, which was the most recent version at the time the segmentation was conducted. HTK contains several tools to perform tasks such as: extracting acoustic features like the **Mel Frequency Cepstral Coefficients** (MFCCs) (see Appendix A) from the raw speech signal; constructing (training) HMM models (HCompV, HRest and HERest) from aligned and non-aligned data (the former being flat start training); using previously trained HMM models to align new data with the transcript using **Viterbi** decoding (Murphy, 2012) (HVite); and performing other text processing tools (HHEd, HCopy…). These tools were built mainly for speech signals but HTK has been used for other purposes. More depth on what each of these tools do is contained in The HTK book (Young et al., 1997).

Because of the complexity of the HTK training scheme, because it requires manual manipulation of the text files between stages of training and alignment, and the complexity of HTK's syntax used to write the HMM topology, a python wrapper was used to script the different stages and tasks. Another motivation for using this wrapper was the fact that the training and alignment were conducted several times, due to changes in parameter values and errors found in the results, that required some alteration to the data. The wrapper used was Prosodylab-Aligner (Gorman et al., 2011), developed by the Department of Linguistics at McGill University. The aligner contained two main features before modification in this work: HTK's flat start training scheme (that generates an HMM model and also aligns the training data), and alignment using previously trained models. Flat start training is the term commonly used in the literature when the initial training stage is not done with manually-labelled data and the input utterances are uniformly segmented. For example, an utterance that is 10 seconds long with 100 labels would be split into 100 segments each being 100 milliseconds long. Flat start trained HMMs usually produce less accurate alignment than aligning using HMMs trained with manually aligned data (Brognaux et al., 2012; van Niekerk and Barnard, 2009). It was used in this case as an initial alignment for the experts to use when creating the manual alignments.

A third feature was added to the python wrapper which was to bootstrap (train) the HMM models using previously aligned data (data with timestamps of phone boundaries). All three features in the wrapper were modified to optionally allow different HMM topologies for different phones (it is possible to use a default topology for all phones). The three features were used in the following general stages.

1. HTK alignment: The output phonetic transcription system described in Section 5.1, along with the raw audio, is input to the python wrapper which in turn uses the HTK flat-start training scheme to generate the automatic alignments of the corpus.
2. Manual corrections: The output alignments are given to the linguistic experts for manual inspection and correction. The correction involves adjusting the boundaries of phones and correcting false phone labels, deleting labels for phones that did not exist in speech or adding labels for phones that were missed by the phonetic transcriber. The corrected alignments are used to calculate the precision of the automatic alignment of the different runs of stage 1 (with different parameters).
3. HTK bootstrapping: The output of stage 2 is used for further refining the automatic alignments by bootstrapping the training with manually corrected boundaries. This could be done iteratively until an acceptable precision is reached. The precision was considered acceptable if it was within 2.5% of the expert agreement value for the same tolerance (see and compare the tolerances in tables 19 and 20). More iterations could have been

conducted to make the precision value closer to the expert agreement value, but this was considered necessary only if the final quality of the voice was unsatisfactory which was not the case as explained in Chapter 7.

4. Boundary Refinement: Optionally, before or after stage 3 (or both) a novel approach to boundary refinement was performed. It was inspired by the results of the evaluation of the first stage (see Table 6-4). The results showed a strong tendency by certain predicted boundaries of predominant boundary types to deviate from the correct boundary location in a regular manner (delta) both in magnitude and direction. For example, over 80% of the time boundaries (that were moved by the experts) between vowels and consonants showed a negative delta, with an average delta of –0.01811 seconds. This tendency is detected by high positive or negative delta values. An absolute delta value of more than 0.01 was considered high enough to conduct a corrective shift on the boundaries of that type by 0.005 seconds.

The following is a more detailed description of the two first stages. The text processing performed between those stages is not described here because it is redundant. More detail is contained in the code (Halabi, 2015).

5.3 HTK Alignment

After calculating the MFCC acoustic features of the raw audio signal using HCopy, HCompV is used to calculate the initial means and variances of the Gaussians, whose mix makes up the observation probability distributions (Ghahramani, 2001). These means and variances are the same for all phone models initially and is the global mean and covariance for all the data points (all frames for the audio signals). HCompV also generates Variance Floors (VFloors) which are lower bounds for the variances that can be used later to prevent over-fitting by prohibiting the variance from going below those values at each training iteration. By default, the variance floors are taken to be 0.01 times the global variance which was used in this work. Note that there are other ways of calculating variance floors that are not included in HTK (Young et al., 1997) and are not covered here.

The global means and variances generated by HCompV, alongside a default initial transition matrix, are used to create the initial HMM definition. This is similar to HTK's initial model state as shown in Young et al. (1997).

The generated initial HMM models are then used iteratively as input to HRest. HRest updates the means and variances discussed above and also the transition matrixes of HMM states. HRest uses the Baum-Welch algorithm (Jurafsky and Martin, 2009) which is based on the more general

Expectation Maximisation (EM) method for probabilistic model parameter estimation when the probabilistic model contains hidden variables. In EM, the goal is to maximise the data log-likelihood function, which is given by:

$$L(\theta) = \sum_{i=0}^{N} P(X_i, Z_i|\theta) \quad (2)$$

where $L(\theta)$ is the log-likelihood function, X_i are the observed variables, Z_i are the hidden variables, θ generally represents the model parameters and N is the number of data entries in the dataset.

Since no values for the hidden variables are known during training, the current estimates of the HMM parameters are used to find the expectation of the log-likelihood function, which is given by:

$$Q(\theta, \theta^{t-1}) = E[L(\theta)|\mathcal{D}, \theta^{t-1}] \quad (3)$$

where E is the expectation of the log-likelihood function given the current estimate of the model parameters θ^{t-1} and the dataset $\mathcal{D}$.

Then the function $Q(\theta, \theta^{t-1})$ is itself maximised with respect to θ which yields the new estimate of the model parameters θ^{t}as shown here:

$$\theta^{t} = argmax_{\theta} Q(\theta, \theta^{t-1}) \quad (4)$$

The same is repeated either a predefined number of times or until convergence is reached. It can be shown that each iteration of the EM algorithm will either increase the value of the log-likelihood function or keep it the same for new estimates of the parameters θ. The monotonic increase of the log-likelihood function at each iteration in the EM algorithm has been proven (Murphy, 2012).

In the HTK implementation of Baum-Welch (EM algorithm), the number of epochs (iterations) is adjustable and different numbers were used to test precision increase.

After HRest has finished optimising the parameters of the phone models, HVite is used to generate the final alignment using the Viterbi algorithm (Murphy, 2012), which is a dynamic programming method that finds the most probable sequence of states (hidden variables) that generate the observed variables in an HMM.

After one third of the Baum-Welch iterations have elapsed (a total of 15 was conducted), the aligner adds optional "pause" models between words in the script. HTK chooses the most probable pronunciation – given the HMM, the pronunciation dictionary and the silence (pause) models – which finally includes the boundaries and the detected pauses. This is done using finite

state word networks that model all the possible pronunciations of the utterances (Young et al., 1997). These networks are built from the pronunciation dictionary which contains possible multiple pronunciations of the same word. This work's phonetic transcription system (see Section 5.1) deals with ambiguous pronunciations of words by adding multiple entries in the output pronunciation dictionary. More on using finite state networks for multiple pronunciations in speech recognition can be found in Pereira & Riley (1996) and Young et al. (1997).

5.4 Manual corrections

According to Yuan et al. (2013), segmenting and aligning speech given only the phonetic transcript and raw audio with no initial segmentation could require as much as 400 times the audio time to finish with acceptable precision. This means that every minute of speech would take over 6.5 hours to segment. Segmenting and aligning the whole corpus produced (which is 3.5 hours long) would require around 1400 hours of work. It has already been noted that there is considerable emphasis on the difficulty of segmentation and alignment in the literature (Van Bael et al., 2007; Malfrère et al., 2003; Mporas et al., 2009).

The alignments produced at stage 1 were used to decrease this time required. It is assumed here that correcting automatic segmentation and alignments is quicker than creating them from scratch. This manual correction stage has been undertaken in previous work and is usually recommended for speech synthesis corpora (Peddinti and Prahallad, 2011; Jakovljević et al., 2012; Black, 2002) but it suffers from:

- the huge effort required to segment and align even a medium-sized corpus, of 3-4 hours, for speech synthesis. The suggested solution in this work is an iterative method where experts correct small parts of the corpus, which are then used to realign the corpus automatically after bootstrapping with manually corrected data. This is done iteratively with accuracy calculated at every step to check if improved precision can be obtained without manually correcting the entire corpus.
- requiring a team of qualified linguists with good knowledge of the target language's phonetics and training them on the conventions, phone sets, boundary types, potential errors and the software used for correction. Three 3-hour training sessions for the experts were conducted before the segmentation and alignment tasks were distributed. The experts kept in contact throughout the alignment to report common errors and enquiries.
- agreement between experts. This is due to the subjective nature of some phone boundaries or phone boundary types (Yi, 2003). To solve this issue, each utterance was corrected at least twice by at least two different experts. This helped in two ways. One is calculating inter-

expert agreement, which a measure of how close the experts' corrections were to each other (see Section 6.4.2). The other is to increase the precision of the alignment, with more experts analysing the same set of utterances.

Each expert was given batches of 50 utterances per iteration and which were then exchanged for the second correction. The software used for the correction of the boundaries was Praat (Boersma and Weenink, 2015) which accepts the file format that Prosodylab-Aligner is able to generate. Praat was chosen as it was the only freely available tool for this purpose at the time of the experiment. Figure 5-1 shows the interface of Praat used by the experts. Tier 1 contains the phone labels and boundaries to be corrected by simple keyboard and mouse actions. Tier 2 contains the Buckwalter representation of words in the original transcripts. These were not to be changed by the experts.

5.5 Summary

This chapter presented the process and methods for segmenting and aligning the speech corpus recordings with their phonetic transcripts. This included a review of previous work and tools available in order to choose the methods and the parameters for those methods to be used for the segmentation and alignment. The process of manually correcting the alignments was also explained. Most important was the discussion of the phonetisation rules used to produce the phonetic transcript to be aligned with the recordings. These phonetisation rules are one of the main contributions of this work.

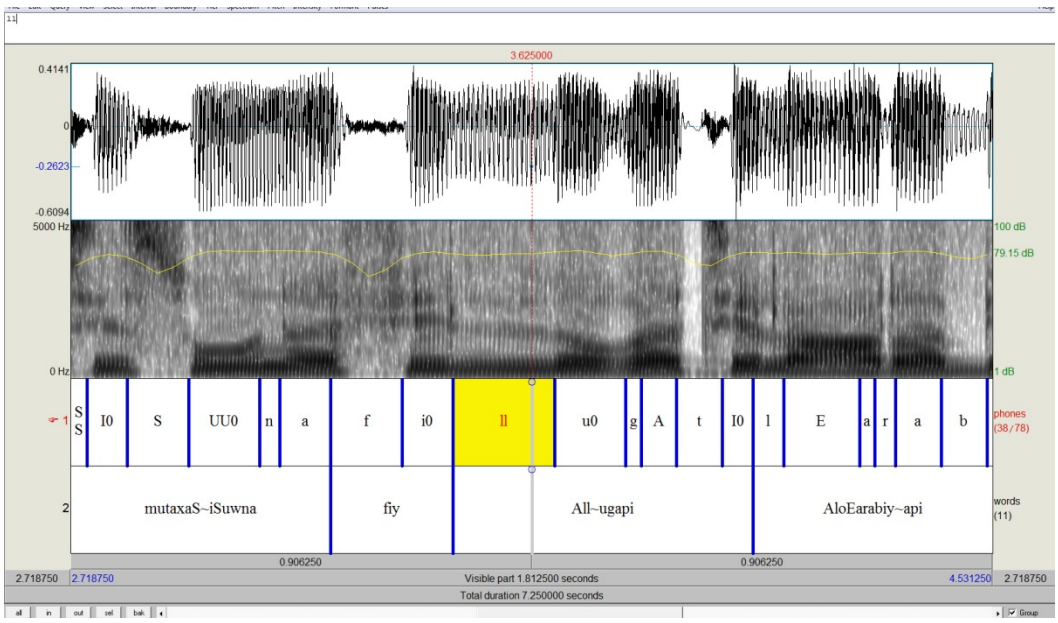

Figure 5-1. Praat interface

Chapter 6 Evaluation of Segmentation and Alignment

6.1 Evaluation metrics

To measure precision, the proportion of boundaries found within a controlled distance from the correct boundary were calculated. This is the metric agreed in the literature (Hosom, 2009; Yuan et al., 2013; Mporas et al., 2009; Jakovljević et al., 2012). In most reports, this value is used on all boundaries combined (Jakovljević et al., 2012; Yuan et al., 2013; Hoffmann and Pfister, 2010), although sometimes boundary types have been excluded (Jarifi et al., 2008; Stolcke et al., 2014). Some reports calculated separately the precision for different boundary types (Hosom, 2009). This inspired the inclusion of boundary types in this work, as this had not previously been undertaken for MSA. The 'distances' are all a multiple of 5 milliseconds. These distances are sometimes referred to as the tolerance T and the percentage of boundaries within a tolerance will be referred to as the precision for that tolerance $P_{T,B}$ where B is the boundary type. In addition, the average absolute value of delta D (absolute value of boundary shift caused by the experts' corrections), the number of positive and negative deltas and the standard deviation of the deltas are calculated for each boundary type. Table 6-1 shows the calculated values and metrics in more detail.

Table 6-1. Metrics used in evaluating segmentation

Value or Metric	**Symbol**	**Formula**
Tolerance	T	-
Number of boundaries of type B	N_B	-
Number of boundaries of type B within Tolerance T	$N_{T,B}$	-
Precision	$\boldsymbol{P_{T,B}}$	$\boldsymbol{\frac{N_{T,B}}{N_B} * 100}$
Predicted time stamp of boundary b	$tp(b)$	-
Expert corrected time stamp of boundary b	$tc(b)$	-
Delta	$D(b)$	$abs(tp(d) - tc(d))$

Value or Metric	Symbol	Formula
Number of positive deltas for boundaries of type *B*	$\boldsymbol{D_B^+}$	-
Number of negative Deltas for boundaries of type *B*	$\boldsymbol{D_B^-}$	-
Average Delta for boundaries of type *B*	$\boldsymbol{D_B^*}$	$\sum_{b \in B} \frac{\mathbf{D(b)}}{\mathbf{N_B}}$
Standard deviation of Delta for boundaries of type *B*	$\boldsymbol{D_B^\sigma}$	$\sqrt{\frac{\sum_{b \in B}(\mathbf{D(b)} - \boldsymbol{D_B^*})^2}{\boldsymbol{D_B^\#}}}$

Metrics shown in bold text are used to assess segmentation quality. They were used to find which types of boundary were most often incorrectly predicted by the system or to identify misunderstandings between the experts. Symbols in column 2 are used in the rest of this work for convenience. As stated above, the $P_{T,B}$ metric is not novel and is the metric used in the literature to evaluate segmentation precision. The four other metrics (D_B^+, D_B^-, D_B^*, D_B^σ) are novel, and as future work, could also be used for boundary refinement (they are called shift metrics in this work).

Table 6-2. Insertion, deletion and update metrics

Value or Metric	Symbol
Number of boundaries added	B^+
Number of boundaries deleted	B^-
Number of phone labels changed	L^c

Table 6-2 shows three metrics that were used for assessing expert performance in alignment. Even though the textual transcript was corrected before generating the phonetic transcript and aligning, the experts were not only required to correct boundary locations but also add missing boundaries, remove unnecessary ones and correct phone labels that did not match the speech.

There could still be errors in the phonetic transcript after manual revision, for several reasons:

- Experts did not detect an error in the first stage of text correction or in second stage when matching text with recorded speech.

- The phonetic transcript generated automatically could contain errors as some parts of the algorithm are non-deterministic and generate multiple possible pronunciations of the same word. HTK's forced alignment would need then to choose the best possible pronunciation for each word and this did not result in matching phonetic sequences in all cases. This was sometimes due to the talent pronouncing letters with some imperfections, which makes the identity of the phone disputed. An example of these imperfections would be emphasising a vowel in a non-emphatic context or *vice versa*.
- Some words were not pronounced according to the rules found in this work for automatic phonetic transcription (see Section 5.1). This includes foreign words that are written in Arabic but pronounced using phones that may not be part of Arabic's (or MSA's) phonetic vocabulary (see Table 4-7). The pronunciations of these words had to be entered manually by the experts.
- The experts were given the option to add a distortion label (see Table 4-7) to segments where pronunciations were not clear.

All the causes listed above were found after correcting the first three batches of utterances (150 utterances with 50 each). Any systematic errors in the transcript were attributed to either a flaw in the algorithm generating the phonetic transcript (see Section 5.1), or to irregular pronunciation caused by a mistake by the talent or the nature of the word (foreign nouns). The former type of error was dealt with by modifying the algorithm and re-running the alignment. Two issues were found in the phonetic transcript after the first corrections phase.

- Geminated consonant letter ي is pronounced inconsistently depending on context. If preceded by the diacritic ِ , the generated phonetic transcript is /ii0 y/. If preceded by a ُ or a ِ it is transcribed as /u0 yy/ and /i0 yy/ respectively. The reason for this issue is that no previous formalisation of the phonetic transcription of geminated ي was reported and it was assumed that the effect of geminating a consonant ي would always result in a separate consonant phone /yy/, but this was seen not to be the case in practice and in the context explained above, geminated ي is pronounced as a combination of a vowel followed by a consonant.
- Likewise, the geminated consonant letter و is pronounced inconsistently based on the preceding short vowel. The only difference is that a preceding ُ would cause the transcription to be a long vowel /uu0/ followed by a non-geminated /w/.

Table 6-3 shows the number of inserted, deleted and altered tags in the three batches used to evaluate the first stage of (flat start) forced alignment.

6.2 Boundary Types

Feedback from experts indicated that correcting certain boundary types was more difficult than others because of strong co-articulation between phones. This led to the idea of categorising boundary types based on the type of articulation of surrounding phones (fricative, stop, trill,…). For example, the boundary between the phones /q/ and /l/ is labelled a "stop/approximate" boundary or "st/ap" boundary or more specifically a "voiceless-stop/approximate" or "vl-st/ap" boundary (the latter being a subset of the former). For stops and fricatives, both the voiced and voiceless subsets were included in the analysis. This means that the boundary types are not disjoint sets and some sets are subsets of others (the above being an example). Vowels were all grouped together under the same articulation category, "vowels" or "vo".

The precision and shift metrics were calculated for each boundary type to show how accurately the forced alignment works for each type and the nature of shift happening in each type. This was inspired by feedback from the experts who found systematic shifts in the boundaries between the predicted and corrected timestamps. Also, it is already established that some boundary types, when realised in speech, correspond to abrupt changes in the acoustic features (intensity and spectrum) and hence could potentially be easier to detect by a machine (Yi, 2003; Hosom, 2009).

Boundary types used in this work are shown in the results available through the web link (Halabi, 2015).

6.3 HTK Parameters

HTK allows several parameters to be changed before running each of its components. The choices for each of these parameters are not optimal since, for some, there have yet to be experiments showing performance for different values. Most of the chosen values for the parameters were based on the HTK segmentation scheme (Young et al., 1997). The parameters are the following.

- Acoustic Features (MFCC, LPC,…): MFCC were chosen with 36 features for each window. HTK allows more specific parameters to be changed when it comes to MFCC feature extraction, such as the number of filterbanks. All the values for these parameters were set according to the HTK segmentation scheme (Young et al., 1997).
- Pre-emphasis Parameter: Determines the extent to which certain frequencies are boosted in the speech signal to decrease the effect of noise (Mporas et al., 2009; Young et al., 1997). The value used was 0.97, which is the one used in the HTK segmentation scheme (Young et al., 1997).

- Hamming Window: A window function which is zero or a very low value outside a certain range used to extract parts of a speech signal for analysis. This was set to true (use a hamming window), based on the literature where Hamming windows were almost always used (Yuan et al., 2013; Young et al., 1997; Prahallad, 2010; Mporas et al., 2009).
- Window Size: The length of the Hamming window used. This determines how long the segments of speech used for MFCC coefficient extraction are. The default value in the HTK segmentation scheme (Young et al., 1997) was used.
- Energy Normalisation: A true or false value indicating whether to normalise the log-energy (log-intensity) of the speech signal before extracting features. It was set to true, which was the default value in the HTK segmentation scheme (Young et al., 1997).
- Topology: The HMM models used had 3 states (in addition to dummy start and end states) which is the most common configuration in the literature. Emission probabilities were modelled as a single Gaussian in 36 dimensions for each of the MFCC coefficients. The transition probabilities between states are multinomial. As future work, it is intended to use more Gaussians for the emission probabilities. There are many possibilities for adjusting the number of mixtures and the number of channels to which each of the MFCC coefficients belong.
- Window Shift Rate: Called TARGETRATE in HTK. It is the shift applied to the window after each calculation of the MFCC coefficients. The default value in the HTK segmentation scheme (Young et al., 1997) was used.

6.4 Initial Evaluation (Flat Start)

6.4.1 Alignment quality

Table 6-4 shows the precision values of all the metrics for the initial 3 batches of alignment. These contained 13166 boundaries (including boundaries with pauses) and 11311 phone boundaries (excluding boundaries with pauses). 1047 boundaries were skipped by the evaluation script out of the complete boundary set because of phone label mismatch between the automatically generated phonetic transcriptions and the experts' corrections. This decision was made because boundaries corresponding to incorrect phonetic transcript affect the precision of alignment and would skew the results since the aligner would try to align script to a non-matching speech signal. The goal of this evaluation is to calculate the precision of the alignment knowing that a certain percentage of phone boundaries and labels were mismatching (7.9% of boundaries in this case). 12119 boundaries were left for analysis as shown in Table 6-4. 68.49% of the predicated boundaries were within 20 milliseconds of the corrected boundaries. This is significantly lower than the precision achieved on the TIMIT corpus in previous work (Hosom,

2009), the difference being that in this work, a different HMM topology was used, and the phonetic transcription was automatically generated by a rule-based algorithm rather than depending on a human-generated pronunciation dictionary. This generated errors that affect the precision of the HMM forced alignment system.

To increase this precision further, analysis of the common errors detected by the experts and retraining of the system based on the manually aligned subset was done (see Section 6.5).

Table 6-3 shows the number of insertions, deletions and updates the experts performed. The "Mismatching boundaries" value does not simply equal the sum of the insertions, deletions and updates, because each one of these could either cause one or two mismatching boundaries. Overall, less than 8% of the boundaries were mismatched between the correction and automatically generated transcript, and mostly due to recording errors or foreign words. This could be improved by adding a foreign-word pronunciation dictionary which currently does not exist for MSA.

Table 6-3. Correction Statistics for three batches

B^+	133 (~1.0%)
B^-	134 (~1.0%)
L^c	534 (~4.0%)
Mismatching boundaries	1047 (~7.9%)

No previous works on transcript corrections have been published with which to compare these numbers, but it is important to note that speech synthesis voices have been built on uncorrected automatically generated and aligned transcript in the past. This work attempts to find whether an uncorrected portion of the corpus, aligned by a system trained on a corrected portion of the corpus, would be suitable for speech synthesis in MSA using a listening test.

The rest of the results are available through Halabi (2015). They show the precision for different boundary types. It is easy to see that some boundary types correspond to significantly higher precision than others.

Table 6-4. Precision of Initial forced alignment for general boundary types

T	<0.005	<0.010	<0.015	<0.020	<0.025	<0.030	>0.050	D_B^*	N_B	D_B^+	D_B^-	D_B^σ
ph/ph	33.42	45.26	57.67	68.49	76.93	83.1	100	–0.00741	11311	2059	6534	0.002695
vo/co	28.48	38.22	50.41	63.01	73.48	80.87	100	–0.01181	4955	580	3325	0.002874
co/vo	37.66	52.06	64.91	74.21	80.69	85.89	100	–0.00231	5075	1277	2480	0.002782
co/co	35.63	45.42	57.01	67.03	75.25	80.58	100	–0.01063	1277	202	727	0.001472
Silence Boundaries												
pa/ph	28.07	29.82	32.89	40.79	57.46	75.44	100	0.002481	228	16	154	0.028729
ph/pa	22.22	37.20	46.38	57.97	67.63	76.33	100	–0.00188	207	41	129	0.015074
pa/co	28.07	29.82	32.89	40.79	57.46	75.44	100	0.002481	228	16	154	0.028729
co/pa	21.05	46.05	60.53	69.74	76.32	80.26	100	–0.00815	76	14	48	0.000440
vo/pa	22.90	32.06	38.17	51.15	62.6	74.05	100	0.001762	131	27	81	0.023527
Reported TIMIT*	48.42	79.30	89.49	93.36	95.38	96.74	100					

* precision (Hosom, 2009) Blue shows this work's system. Red is the TIMIT Result (Hosom, 2009).
"ph" = "phone", "pa" = "pause", "co" = "consonant", "vo" = "vowel"

6.4.2 Expert Agreement

Because the alignment process takes a long time, batches of 50 utterances were distributed to two experts (each expert receiving a different batch of utterances). A third expert later checked their work and corrected any errors remaining. The two experts at the beginning were trained together to make sure that their alignments were as similar as possible but it is useful to know how close their alignments were. This is referred to in the literature as expert agreement or inter-annotator agreement (Hosom, 2009; Romportl, 2010). Each expert was given 5 additional utterances that were part of the other expert's workload giving a total of 10 utterances aligned by both experts to conduct an expert agreement test.

To show how similar the alignments were between the experts, the same metrics were used as in the precision evaluation of the alignment (Hosom, 2009; Romportl, 2010). The only difference is that the number of changes in phone labels was calculated. This number is the sum of the number of labels changed, the number segments added by the experts, and the number of segments removed by the experts. If the resulting phone label sequence does not match, the analysis script skips the boundaries and does not included them in the agreement analysis shown in Table 6-5.

In this test, both experts had to correct the predicted boundaries resulting from forced alignment rather than correcting each other's. This is to estimate the agreement more accurately, because if experts were given each other's alignments, it might be tempting not to change boundaries if the error is too small (smaller than that found in the forced alignment output).

Table 6-5 shows the results of comparing the alignment of 10 utterances between two of the experts. The 10 utterances contained a total of 981 phone boundaries (including ones with a pause) of which 47 had changes in identity (phone label) applied to their adjacent phones by either or both experts which led to non-matching boundaries; these were excluded from the analysis even if accurate. One of the experts inserted 7 new segments that they thought were missing which the other did not, while the other expert inserted 7 segments which the first expert did not include. This resulted in the system ignoring 97 boundaries when calculating precision. 884 phone boundaries remained for agreement analysis. 827 of those boundaries were between two phones (no pause) and the remaining had at least one adjacent pause. Note that two consecutive pauses are possible.

84.28% of all boundaries were within 20 milliseconds of each other. As mentioned earlier, the 20 millisecond tolerance is the *de facto* standard found in the literature for evaluating alignment precision (Hosom, 2009; Stolcke et al., 2014; Yuan et al., 2013) but it is also used for evaluating expert agreement. The highest precision in previous work for expert agreement was on the TIMIT

corpus with 93.49% of boundaries generated by the author within 20 milliseconds of corresponding boundaries in the TIMIT corpus (Hosom, 2009). In the same work, Hosom reviews previous work which shows results in expert agreement. All the reviewed attempts reported agreement of over 90% which poses the question: why is the agreement in this work lower? Hosom excluded two types of boundary from his evaluation because they suggested that they were subjective and should not be included in the precision analysis. No boundaries were excluded in this work, but still this leaves a significant difference in agreement which led to a third expert running through the two experts' alignments (specially the points of disagreement) and normalising the alignment. This is not as laborious a task compared to the initial alignments as it only requires that the expert to review 10% to 20% of the corpus. This mainly occurred at the boundaries that were not included in the analysis, due to experts disagreeing in the segment's label or boundaries of a type corresponding to a lower precision score.

This stage helped identify misunderstandings in the labelling, segmentation and alignment processes by the experts, which they were informed about for more agreement in future manual alignment and corrections.

Table 6-5. Expert Agreement Analysis Results

T	<0.005	<0.010	<0.015	<0.020	<0.025	<0.030	>0.050	D_B^*	N_B	D_B^+	D_B^-	D_B^σ
ph/ph	42.63	59.81	73.93	84.53	90.86	95.01	100	0.010362	821	376	258	0.047647
vo/co	42.35	59.29	74.04	84.15	91.53	95.63	100	0.00566	366	199	84	0.000206
co/vo	43.24	61.54	76.13	86.74	93.1	96.82	100	9.33E-05	377	148	147	0.000176
co/co	41.56	54.55	63.64	76.62	77.92	84.42	100	0.001998	77	28	27	0.000349
Silence Boundaries												
pa/ph	18.75	31.25	37.50	37.50	50	68.75	100	0.019574	16	13	1	0.000204
ph/pa	7.14	14.29	42.86	57.14	64.29	64.29	100	–0.01578	14	2	11	0.000728
pa/co	13.33	26.67	33.33	33.33	46.67	66.67	100	0.020879	15	13	1	0.000191
co/pa	16.67	16.67	50.00	50.00	50.00	50.00	100	–0.01436	6	1	4	0.001432
vo/pa	0	12.50	37.50	62.50	75.00	75.00	100	–0.01685	8	1	7	0.000197
TIMIT Agreement Results	60.38	81.73	89.07	93.49	95.36	96.91	100	-				

6.5 HTK Bootstrapping

At each iteration and after the correction of 150 utterances was completed with a second revision, another automatic segmentation was conducted with the manually corrected data as input to bootstrap the HMM models. HInit is an HTK tool that initialises the phone HMM parameters by using manual segmentations. HInit was used to initialise the parameters of the HMM models used for the different phones. For each phone, all the available segments for that phone in the training data were loaded and used to iteratively update the parameters of the phone's initial HMM using Viterbi training (Jurafsky and Martin, 2009). Viterbi training works in a slightly different way to the Baum-Welch algorithm described in Section 5.3. In it, each of the phone's segments are divided equally between the states of the phone's HMM, then these divisions are used to calculated each of the HMM's states' parameters. The new HMM model with the new parameters was utilised by the Viterbi algorithm to find the most likely sequence of states (under the new model) and the operation is repeated until convergence. These parameters include the means and variances of the Gaussians whose mix makes up the observation probability distributions and the transition matrixes which define the transition probabilities between states.

Then the training process continued in the same way as in stage 1 using the parameters estimated from HInit as a starting point instead of the output of HCompV. Note that HCompV was still run in this stage as it is required to produce the variance floors and HRest is iteratively run (Baum-Welch).

Table 6-6 shows the results after bootstrapping; a significant improvement from 68% to 82% has been achieved. The results in Table 6-6 are based on 50 utterances or about one third of the amount used for the flat start evaluation. This included 3320 phone/phone boundaries. The complete set of results can been viewed through the link (Halabi, 2015).

Table 6-6. Alignment results after bootstrapping

T	<0.005	<0.010	<0.015	<0.020	<0.025	<0.030	>0.050	D_B^*	N_B	D_B^+	D_B^-	D_B^σ
ph/ph	32.77	56.14	71.57	82.50	88.73	92.80	100	–0.00521	3320	961	1921	0.000267
vo/co	30.25	50.76	67.33	80.52	86.95	91.78	100	–0.00862	1448	293	948	0.000266
co/vo	35.78	62.35	77.10	86.32	91.52	94.59	100	–0.00146	1498	562	770	0.000233
co/co	30.38	52.15	65.59	74.73	84.41	89.52	100	–0.00706	372	105	203	0.000303
Silence Boundaries												
pa/ph	27.47	57.14	72.53	84.62	91.21	92.31	100	–0.00456	91	37	49	0.000337
ph/pa	15.29	41.18	51.76	67.06	72.94	77.65	100	–0.00862	85	27	55	0.001062
pa/co	27.47	57.14	72.53	84.62	91.21	92.31	100	–0.00456	91	37	49	0.000337
pa/vo	0	0	0	0	0	0	0	0	0	0	0	0
co/pa	13.33	40.00	46.67	70.00	73.33	80.00	100	–0.01547	30	7	22	0.001185
vo/pa	16.36	41.82	54.55	65.45	72.73	76.36	100	–0.00488	55	20	33	0.000955
TIMIT Agreement Results	48.42	79.30	89.49	93.36	95.38	96.74	100	-				

6.6 Precision Comparison

The next stage involves comparing the precision of this work with the highest precision systems in other published works. Table 6-7 shows that 93.36 P_{20} is the highest precision found in literature (Hosom, 2009). His system used HMM/ANN (Hidden Markov Models paired with Neural Networks for feature extraction) as a baseline to compare it with their modified HMM/ANN system which achieved the higher precision by adding features on top of the MFCC feature set used. These features included energy and burst detection to help give areas of rapid acoustic feature changes more chance of being detected as boundaries.

Hosom (2009) also trained his system on part of the TIMIT corpus and did not perform any forced alignment. He claimed that regular HMM forced alignment (similar to the one in this work) did not perform as well as his. He used two-fifths of the dataset for evaluation and three-fifths for training.

It is easy to see from Table 6-7 that using an HMM/ANN system would improve the precision. It is also possible to infer this from the improvement HMM/ANN systems give to speech recognition relative to pure HMM (Hosom, 2009). It was not possible to obtain or implement a version of it for this work but is suggested for use in future work (see Section Chapter 88.3). Improving either the HMM forced alignment or the HMM/ANN system is not part of this work; it is used to demonstrate the correctness of the phone set and pronunciation rules produced by the automatic phonetic transcript generation system and the overall quality of the corpus.

Table 6-7 shows the system proposed in this work approaches the state of the art HMM forced alignment systems but still lags behind HMM/ANN. The difference in the evaluation setup of each system is detailed in this table.

Table 6-7. Precision comparison

Metric	**Basic HMM forced alignment on MSA**	**Basic HMM forced alignment on MSA (with bootstrapping)**	**Baseline HMM/ANN forced alignment on TIMIT** (Hosom, 2009)	**Hosom (2009) Proposed System**
Feature used	MFCC (basic HTK setting) (Young et al., 1997)	MFCC (basic HTK setting) (Young et al., 1997)	MFCC with mel scale replaced by Bark scale	MFCC with mel scale replaced by Bark scale. Additional energy-based features

Metric	Basic HMM forced alignment on MSA	Basic HMM forced alignment on MSA (with bootstrapping)	Baseline HMM/ANN forced alignment on TIMIT (Hosom, 2009)	Hosom (2009) Proposed System
Model Architecture	1 Gaussian to model emission probabilities. Basic 3-state HMM architecture. (Young et al., 1997)	1 Gaussian to model emission probabilities. Basic 3-state HMM architecture. (Young et al., 1997)	HMM with ANN instead of Mixture of Gaussians	HMM with ANN instead of Mixture of Gaussians. With modifications on the state structure of the HMM.
Dataset used	Recorded as part of this work	Recorded as part of this work	TIMIT	TIMIT
Training data size	Unsupervised	150 utterances; approximately 25 minutes of speech	3696 files (3.145 hours of speech)	3696 files (3.145 hours of speech)
Evaluation data size	150 utterances; approximately 25 minutes of speech	50 utterances; approximately 6 minutes of speech	1344 “si” and “sx” file from TIMIT corpus	1344 “si” and “sx” file from TIMIT corpus
Language	MSA	MSA	English	English
Precision ($\boldsymbol{P_{20}}$)	68.49	82.50	91.48	93.36

6.7 Summary

This chapter evaluated the chosen process of segmenting and aligning the speech corpus recordings with phonetic transcript (HMM) (with and without bootstrapping). The parameters used for the HMM model used for forced alignment were presented and contrasted with the ones of Hosom (2009). An estimate of the precision of these alignments showed that it is lower than the work of Hosom (2009). Expert agreement was also discussed showing less expert agreement compared with the work of Hosom (2009).

Chapter 7 Subjective Evaluation

Several types of listening test are found in the literature and used for different purposes. To choose the correct test for this work's purpose, previous attempts in assessing corpus quality were examined.

Boros et al. (2014) conducted subjective listening tests, which they called the "anonymous preference test", to compare two systems built using their Romanian speech corpus. The difference between the two is that one used the corpus including the ToBI prosodic annotations and the other used the corpus excluding these annotations. 37 randomly selected sentences were synthesised using both systems (19 from news and 18 from novels. All of which were manually labelled using ToBI annotations). Participants were presented with utterances from both systems and asked to give one of five answers: "identical", "first is slightly better", "first is significantly better", "second is slightly better" and "second is significantly better". Participants were encouraged to ignore the naturalness of the synthesised speech and focus on prosodic enhancements, which would help indicate the degree to which the ToBI annotations helped, since the voices were built using parametric speech synthesis methods which are not considered completely natural (van Niekerk, 2014) because of the general averaging feature of HMMs. They did not indicate how many times each participant answered the questions nor did they state the number of participants but there were 587 answers in their study. They did not indicate the listening conditions or the type of participants. They also conducted the same study to show how much an automatic ToBI annotator, trained on their data, improves the prosodic quality of synthesised speech. This study showed that automatic labelling improves the quality, but only slightly less than manual labelling. In this work, therefore, automatic labelling of the stress feature for vowel phonemes was used and evaluated because of the difficulty of manually annotating stress. The stress feature extractor was based on the work of Halpern (2009).

The evaluation in Boros et al. (2014) is referred to as a preference test or an AB test in the literature (Qian et al., 2008; CSTR, 2016), but is usually conducted with a 3 choice answers (sometimes 2) rather than 5 (Buchholz and Latorre, 2011; Mohammadi et al., 2014). The 3 choice answers were used in this work because of their predominance, where the choices were: "First system", "Second system" or "No preference". In these tests, each participant was presented with two stimuli and given the option of choosing which is better, based on a certain factor (naturalness, prosody,…).

Sainz et al. (2012) conducted a Mean Opinion Score test (MOS) (ITU-T, 1996). MOS tests – originally used for evaluating network quality of telephony – are subjective listening tests. In an MOS test, listeners are presented with segment of speech voiced by a male or female over the particular communication medium to be tested. Listeners would score from 1 to 5 the voice's quality, or any other feature based on the experiment in case of speech synthesis evaluation. Sainz et al. (2012) assessed the naturalness of their Ahosyn speech synthesis corpus using MOS tests; a score of 1 being "completely unnatural" and 5 as "completely natural". Each listener was presented with at most 20 signals. 18 participants took part with no hearing impairments and most of them were fluent in the languages in question (Spanish and Basque); half of them did not have previous experience in speech technologies. The experiment was conducted in a quiet environment with high quality headphones (no further specification was given of the listening environment). They were also shown the difference when the MOS tests were bilingual (Spanish and Basque speakers assessing both languages). The voices were built using statistical parametric HMM models using HTS (Zen et al., 2009) and using a vocoder they had developed.

To summarise, MOS is a form of questionnaire recommended by the International Telecommunications Union (ITU-T, 1996). It consists of a set of questions which are answered by giving a rating from 1 to 5 of several factors or elements to be assessed. The questions are answered after the participant listens to speech stimuli through a channel. The MOS is used widely to conduct black box evaluations of the quality of speech synthesisers or vocoders (Boros et al., 2014; Inai et al., 2015; Hu et al., 2014).

Since MOS started being used for evaluating speech synthesis in terms of intelligibility and naturalness, different modifications have been applied to it. Polkosky & James (2003) developed a revised MOS called MOS-X to increase the sensitivity, validity and reliability of MOS. More factors were added and the scale was increased to 7 points.

In this work, a scale of 1 to 5 was used, which has been used most in the literature (Boros et al., 2014; Inai et al., 2015; Hu et al., 2014; Kato et al., 2011).

The questions of the MOS test are subject to the experiment setting and are usually changed between experiments. Dall et al. (2014) showed the importance of carefully writing the questions since the expectation of the listener affects their opinion about the naturalness of the speech, and the questions should be chosen to carefully target the research question.

MOS has also been used as a standard test in the Blizzard Challenge (SynSIG, 2016). The number of participants differs each year but mostly comprises over 100 paid listeners. In the Blizzard

Challenge from 2008 to 2015 (SynSIG, 2016), MOS was used to rate the naturalness of the generated speech form, the different synthesisers, and the similarity to the original speaker.

In this work, the prosodic naturalness and synthetic quality of speech generated using our corpus is to be tested. The speech synthesis engine used for the evaluation is the ILSP/INNOETICS Text-to-Speech System for the Blizzard Challenge 2014 (Chalamandaris et al., 2013).

Because most of the methods mentioned above are geared towards testing the communication channel or the speech synthesis system, in this work they need to be used in a way to highlight the contribution or effect of the speech corpus to the different metrics of the evaluation. This resulted in the need for a 10 minute briefing given to every participant, before conducting each listening test, to explain the research and the purpose of the listening tests. Appendix B shows a summary of these instructions.

A paper by Wester et al. (2015) looked at the studies conducted at the Interspeech 2014 conference a year before and the Blizzard Challenge 2013 (SynSIG, 2016), and collected statistics about the subjective listening tests carried out in those studies. The results shown help in deciding which tests to use and how many participants and data points are needed to carry out reliable listening tests. They created a checklist of factors to be considered when conducting listening tests (which was used to inform the design of this work's test setup), and their data could be used when deciding from which tests to choose.

7.1 Review of other published work

Latorre et al. 2014 conducted two experiments both of which included two listening tests: the first, Preference and ABX, while the second MOS and DMOS.

In both experiments the log-f0 variance was modified by 5 different factors using a mel-cepstral vocoder (one of the factors = 1 which meant no modification). The vocoded speech was then evaluated (MOS and Preference) and compared (DMOS and ABX) with the natural signal.

Both experiments showed that when there is no natural speech reference (in Preference and MOS tests) the listeners do not necessarily prefer the speech which is closer to the natural intonation (factor 1), which tended to change when they were presented with the original signal as a reference. Without reference, they tended to prefer the signal with the highest variance.

This helped inform the decision about which tests to conduct in this work and whether a reference was necessary. Even based on their results, this does not mean that having a reference is always necessary. A reference is necessary when the intonation of the reference is considered to be

optimal, but when there is no reference, and based on the listener's categories (age, nationality and gender) the preference seems to change.

Raitio et al. (2015) conducted listening tests similar to the MUSHRA tests (see 2.3.4.2) to compare three different analysis by synthesis methods, where the method for modelling the phase was altered (the natural signal was also hidden in the tests). 15 Finnish listeners aged 23-37 with no reported hearing problems participated in listening booths, using professional headphones. The experiment took around one hour for each participant. 75 utterances were presented to the participants.

Szaszák et al. (2015) conducted a Comparison Mean Opinion Score (CMOS) test with 5 grades to compare their TBSM and ABSM (automatic stress annotation by either audio or text) which is similar to this work's general methodology. 20 subjects, 8 female and 12 male native speakers, were involved with no known hearing impairments with an average age of 34 (19 - 73). 5 of the subjects were speech experts. 20 pairs of utterances were presented for each subject.

Lu et al. (2015) evaluated their different speech database pruning methods by conducting MOS listening tests. They synthesised 36 utterances from different themes: "General", "News", "Navigation", "Voice Assistant", "Email reading" and "Website reading". 25 native speakers were involved with no further description.

Inai et al. (2015) worked on using natural high band spectra in combination with generated spectra (HMM) to improve the quality of HMM-based TTS. They conducted two types of test:

- MOS tests: 7 participants to evaluate the clarity and smoothness of the three TTS systems being compared. 90 stimuli. All pairs were presented to all participants.
- Preference tests: 8 participants, and the rest is the same as in the MOS tests. The participants had to choose between the two pairs (they were not allowed equality).

Other types of test that reported in the literature were the MUSHRA (ITU-T, 2015) and intelligibility tests based on word error rate (WER) as in the Blizzard Challenge (SynSIG, 2016). Neither of these were included in this work for several reasons. MUSHRA involves choosing a specific score between 1 and 100 which has been shown to increase the effort by participants and the increase in accuracy in results diminishes after a certain score count (Nunnally and Bernstein, 1994). The main benefit of MUSHRA over MOS is that it requires fewer participants for obtaining statistically significant results (Wester et al., 2015; CSTR, 2016), but the prevalence of MOS tests in the literature compared to MUSHRA tests was considered more important. The MUSHRA tests are not inferior, but the scope of this work does not go as far as trying to show which is more suitable. Intelligibility tests were not included simply because this metric is not

considered of interest in this work and it was assumed that the process of corpus design does not have a significant effect on intelligibility used to evaluate vocoders and speech synthesisers rather than speech corpora (SynSIG, 2016; Wolters et al., 2010).

The SUS (Semantically Unpredictable Sentences) test (Benoît et al., 1996) was used rarely in recent work in speech synthesis, so it has not been included in this work. In SUS tests, the stimuli are generated by choosing words with certain parts of speech randomly and plugging them into an utterance template such as Subject/Object/Conjunction/Subject/Verb/Object to generate a syntactically correct but semantically meaningless sentence. This is used in intelligibility to prevent participants from predicting words they did not hear.

Wester et al. (2015) did not include this test in their survey of listening tests at Interspeech 2014. It is considered important that the evaluation conducted here produces results that are comparable with recent results in the literature. SUS tests were used in the Blizzard challenges (SynSIG, 2016) to evaluate intelligibility, which have also been excluded in this work.

The Interspeech 2015 track for speech synthesis was used as a main source of literature about listening tests. Some works were excluded because of redundancy or irrelevance. The testing method in recent Blizzard Challenges was also included in the review (SynSIG, 2016). Other works included were Innoetics (Chalamandaris et al., 2013), whose system was used for evaluating our corpus. The work of Wester et al. (2015), which is a survey of Interspeech 2014 listening tests was also included. Other works closely related to speech corpus evaluation were included.

The final decision regarding the choice of subjective tests in this work was confirmed after the exclusion of MUSHRA, Intelligibility and SUS tests. This resulted in MOS and Preference tests being chosen. This left the decision regarding which form of MOS or Preferences tests to use. The research questions were used as a reference since each one of these tests answers a difference question. It is of interest in this work to answer the following:

- Does the inclusion of the stress features extracted using the rules from Halpern (2009) in the speech corpus improve the naturalness and overall quality of speech generated?
- Which is closer to natural speech (recorded in a studio); speech generated with this work's corpus with or without Halpern (2009) stress features?
- Since the speech corpus built in this work is intended to be "General Purpose", does the "Overall Impression" of the speech quality generated using this work's speech corpus compare to the "Overall Impression" for another speech corpus in another language?

Based on the review conducted in this section and the questions above, it was decided to conduct three experiments.

1. A Preference test, which was carried out once to compare the TTS with and without the stress features added to the corpus. Two factors were included in this test, naturalness and overall impression. This means that the participant answers the preference question twice (once for each factor). The participants are allowed to choose "no preference" as well.
2. A DMOS test, which was carried out once to assess the degradation the TTS causes with and without the stress features given a natural recording reference. Two factors were included in this test, naturalness and overall impression. This means that the participant answers the DMOS twice (once for each factor) on a scale of 1 to 5.
3. An MOS test, to assess the overall impression of the TTS system with and without the stress features. The MOS test only included the overall impression factor and did not include naturalness. This is because it was assumed that the DMOS tests would answer this question.

In addition to the stress features, it was intended to evaluate the effect of including and excluding the nonsense utterances in the speech corpus because the speech synthesiser used segments from the nonsense section of the corpus less than 2% of the time even though this section comprises around 40% of the corpus. The nonsense segments are phonetically extreme and were excluded by Innoetics' system's pruning algorithm, which excludes segments which have outlying acoustic features (Chalamandaris et al., 2013). This does not mean that the nonsense section of the corpus is useless; it just means that for this study, and using Innoetics' system, the effect of the nonsense section of the corpus is negligible. As speech technology is evolving, new, high quality algorithms for modifying acoustic features are being developed and this could be used in the future to normalise the acoustic features of these outlying segments, making them more suitable for unit selection speech synthesis. In addition, they could be used to train statistical parametric speech synthesisers that do not usually require segments that are very similar to one another due to its averaging feature (of statistical parametric speech synthesis) (van Niekerk, 2014). It produces speech that is not very natural and hence there is more room for acoustic modification (Bonafonte et al., 2008).

The choice of "Naturalness" and "Overall Impression" as factors was informed by this work's research questions and the MOS tests that were conducted on the Innoetics speech synthesiser with other languages. This would make the results comparable for analysis. Further factors can be conducted and are suggested as future work.

7.2 Objective tests

No objective metrics were used to evaluate speech generated using our speech corpus. A number of studies have been conducted to show how strongly certain objective tests correlate with subjective listening tests for different purposes (Cerňak and Rusko, 2005; Möller et al., 2010; Chevelu et al., 2015). Objective tests were not used here because they are still sensitive to noisy data and, when the differences between the compared utterances in the test data are not large enough, they are still unable to capture the difference (Chevelu et al., 2015; Wester et al., 2015; Buchholz and Latorre, 2011; Latorre et al., 2014).

In addition, objective tests do not necessarily cover all the factors for assessing generated speech. PESQ (Cerňak and Rusko, 2005) is an example of a widely used family of objective listening tests. Like MOS, PESQ was used in telephony applications but is now used for evaluating speech synthesis systems and can be used to simulate a listening test. In future work will conduct PESQ and compare the results. An analysis of results is not in included in this work due to its sensitivity to noisy data (Chevelu et al., 2015; Wester et al., 2015; Buchholz and Latorre, 2011; Latorre et al., 2014).

It would be useful to see if the results of other objective tests conducted on speech generated by this work's corpus correlate with the result of subjective tests on the same corpus, but it was outside the scope of this work to analyse the effectiveness of objective tests.

7.3 Factors to consider when conducting listening tests

7.3.1 Test Data (utterances to synthesise)

Wester et al. (2015) mention the necessity of having multiple sentences for the tests and these sentences should be different. They do not specify how these sentences need to differ and they report no work carried out before for creating test data (a set of utterances) for evaluating speech corpora or speech synthesis in Arabic.

The Harvard sentences (Rothauser et al., 1969) is a phonetically balanced sentence set, which could be used for listening tests in English. Phonetic coverage is one of the features that is considered desirable for a listening test, and was considered when building our set. The Harvard sentences have a phoneme frequency similar to that found in English. The set contains 72 lists with 10 sentences each, which is much higher than the total number of sentences usually used in listening tests for practicality (Shannon and Byrne, 2009; Inai et al., 2015; Szaszák et al., 2015), so it would need to be reduced. The reduction method used was shown in Section 4.2.

Other works have considered other factors when choosing the test utterance set. Boros et al. (2014) considered the source of the utterances (news, novels,…) and randomly chose the utterances rather than considering phonetic coverage. In this work, phonetic coverage and utterance length were both included as criteria for choosing utterances. Sainz et al. (2008) had varying utterance length between 4 and 14 words, with an average of 8.6 words. The reason why prosodic function (declarative vs. interrogative) and utterance source were not considered is resource limitations. The only source of digital, modern and fully diacritised text is Aljazeera learn (Al Jazeera, 2015), whose content is mainly news-related and with no interrogative content. This remains as possible future work.

Other works either failed to describe how the test data was collected or gave a very brief description of it and only the number of utterances was mentioned (Sainz et al., 2012; Chevelu et al., 2015; Dall et al., 2014).

In summary, in this work, since the prosodic naturalness (intonation specifically) and overall speech quality were to be tested, the dataset for testing has been chosen to emphasise these two factors. The length of the utterances and phonetic coverage were considered. The phonetic coverage of the selected utterances was maximised so that as many boundary types between phonemes were covered in the test. Each utterance is semantically self-contained and the reference version recorded in isolation to eliminate contextual effects when conducting the tests. This means that each utterance is a separate intonational phrase (Du Bois et al., 1992). Eventually, 100 utterances were chosen using the same reduction method as detailed in Section 4.2. For the full creation process and statistics of this utterance set, see Section 4.4.

Some utterances included mid-utterance pauses to enrich the prosodic variation in the test data. This is because intonation patterns differ between utterance (intonational phrase) endings and intonation before pauses.

By prosody is meant the intonation (F0 contour) and pauses in the utterance. Other elements of prosody were not considered when creating the test data set.

7.3.2 Listening conditions

The authors did not specify the listening conditions in most of the works reviewed. In some, the listeners are actually sitting in a booth with professional headphones used to hear the prompts and the whole experiment was fully controlled (Raitio et al., 2015; Wolters et al., 2010; Dall et al., 2014). Others only requested that the users wear headphones and test the loudness and clarity before starting the tests (Buchholz and Latorre, 2011; Kawanami et al., 2002; Sainz et al., 2012).

Controlling the experiment is more difficult when using crowd-sourcing and the only guarantee is to ask the participant to wear headphones and listen in an almost noise-free environment.

The level of control of the experiment depends on the type of test to be conducted. Intelligibility tests that use utterances with artificial noise added to them require highly-controlled listening conditions. Since this work was testing prosodic naturalness and overall quality, as long as the signal was loud enough and the surroundings were reasonably quiet, it was not considered necessary to use soundproof rooms but the experiment was controlled (Wester et al., 2015). Every participant was always accompanied by the researcher while doing the experiment and had the option of asking questions, and the researcher was responsible for making sure the conditions were quiet and distraction-free. Controlling the experiment was recommended by Wester et al. (2015) especially for low number of participants.

The listeners were briefed about the research and then asked to wear headphones (provided by the researcher), to test the loudness before starting the tests, and to make sure they would not be interrupted in a quiet room of their choice.

The pairs were also presented in random order in the Preference test to prevent bias. In all tests, the participants were encouraged not to listen to the audio stimuli more than once, but they had the option to in case they were interrupted, distracted or tired. This was because in MOS tests, repetition might make the listener's judgment more biased towards positive results as they become used to the talent's voice and speaking style (Müller, 2007). This is sometimes referred to as habituation in the literature (Müller, 2007).

7.3.3 Choice and Number of Subjects and Data Points

Listening tests conducted previously have included significantly different numbers of participants: 8 (Inai et al., 2015), 15 (Raitio et al., 2015), 25 (Lu et al., 2015), 33 (Hu et al., 2014). Wester et al. (2015) claimed that the number of participants required in a listening test is dependent on the task. They analysed the Blizzard Challenges listening test results and based on those they recommended using at least 30 paid listeners for MOS naturalness tests in controlled environments. They showed that in general, MOS tests should include more participants than have been reported so far in the literature. They did not show any analysis for Preference tests but they claim that Preference tests are generally more sensitive than MOS and require fewer participants. This is also backed by results of *a priori* tests in this work. The minimum of 20 participants was taken for all experiments here with a minimum of 10 data points per participant for each system or pair of systems, which resulted in at least 200 data points for each experiment, as recommended by Wester et al. (2015).

In this work, 31 participants were invited to conduct the tests, of whom 24 agreed and none were excluded.

Wester et al. (2015) also indicated the importance of the type of participants in the tests. In this work, all participants were native Arabic speakers; all participants had to declare that they did not have any hearing difficulties; all participants were asked whether they considered themselves to be speech experts; and all participants provided their age, with the minimum age being 18 years. The effect of age on listening tests is not discussed as part of this work, but it is not known whether age has an effect on listening test results.

7.3.4 Test Questions

The text presented to listeners before listening to the prompts and providing the scores has a strong effect on the outcome of the experiment and what can be inferred from the results (Dall et al., 2014; Wester et al., 2015). The main issue here is trying to explain to the listener what naturalness and overall quality mean in the context of this study in simple terms, as they are not necessarily speech experts. Because the experiment is going to be performed in a semi-controlled manner, the listeners can be briefed about what is meant by each term. The questions were presented in English and the instructions were also presented in English. They were not translated into Arabic as there were many technical terms that do not exist in Arabic. However, the briefing was carried out in the language preferred by the participant.

The text presented to the participants before and during the survey is detailed in Appendix C.

7.4 Generating the utterances

The listening test utterances were generated by taking the reduced dataset (see Section 4.2) and taking all the utterances from the reduced dataset with limit 5 (minimum phoneme boundary occurrence), which are not in the reduced dataset with limit 4. This resulted in 135 utterances; this needed to be reduced to 100. The number 100 was arbitrarily chosen as it is higher than any size of test set seen in the literature (see Section 7.1). Having more test utterances may not always be better, but to achieve better coverage the test sentences were increased, although this means it was necessary to collect more data points to make sure that most utterances are covered. The reduction from 135 to 100 was done manually by removing utterances to make sure that the resulting utterances had an acceptable length variety (word count). The reason that the reduced datasets from Section 04.2 were chosen, was to ensure sufficient phonetic coverage.

The resulting 100 utterances were 12.41 words long on average, the shortest being 3 words and the longest being 30 words. This is more variable and greater in quantity than examples seen in the literature.

The natural recordings of these utterances were on average 10.38 seconds long (including pauses in speech and short pauses before and after the utterance). The shortest duration of a recorded utterance was 3.15 seconds, while the longest was 28.75 seconds.

After these utterances were synthesised using systems 1, 2 and 3, the following was found:

- For system 1, the average length of the synthesised utterance was 6.99 seconds. The longest being 23.77 seconds and the shortest being 3.06 seconds.
- For system 2, the average length of the synthesised utterance was 9.07 seconds. The longest being 23.94 seconds and the shortest being 3.23 seconds.
- For system 3, the average length of the synthesised utterance was 8.95 seconds. The longest being 23.69 seconds and the shortest being 2.88 seconds.

It is clear that the average length of utterances synthesised by systems 2 and 3 (systems which include stress features) were much closer in duration to the natural utterances. This is an indicator that the stress features help produce more natural speech as vowel length can significantly change based on stress in a syllable in Arabic (de Jong and Zawaydeh, 1999). This assumes that there are no modifications being made to the duration of the segments in the unit selection synthesiser before generating the segments. This was further explored after conducting the listening tests (see results in Section 7.6). Table 7-1 shows the complete duration statistics of the utterances used in the listening tests.

Table 7-1. Duration statistics of the utterances used in the listening tests

	Min Duration	**Max Duration**	**Average Duration**
Natural Utterances	3.15	28.75	10.38
System 1 Utterances	3.06	23.77	6.99
System 2 Utterances	3.23	23.94	9.07
System 3 Utterances	2.88	23.69	8.95

System 2 (which included the nonsense utterances) only used 147 segments from the nonsense part of the corpus out of the total 9059 segments used to synthesise the 100 test utterances. This is

less than 2% of the time where the nonsense utterances constitute about 50% of the corpus. Because of this, system 3 was excluded from the listening tests.

This does not mean that the nonsense utterances are considered useless; it only means that the system used for the evaluation did not use them often and the difference between the utterances synthesised by systems 2 and 3 is not big enough to warrant a separate test. This could be due to the system used being Innoetics unit selection speech synthesiser (Chalamandaris et al., 2013), which clusters segments based on acoustic features like loudness and pitch, and then excludes outliers from the system to avoid generating speech with abrupt changes in loudness and pitch. The phoneme segments extracted from the nonsense utterances could be mostly outliers. More work is required to prove or contradict this hypothesis.

Both the natural and synthesised utterances were converted to mp3 format with a bit rate of 64Kbps before they were presented to the participants, to avoid audio quality having an effect on the results.

7.5 Test setup

Two systems were included in the listening tests. Both of them built using the Innoetics unit selection speech synthesiser (Chalamandaris et al., 2013), which is a high quality speech synthesiser that won the Blizzard challenge in 2013. The systems differed by the type of corpus used to generate them and are as follows:

- Used the whole corpus without stress annotations.
- Used the whole corpus with stress annotations included.
- Used the whole corpus except for the nonsense utterances (see Section 4.4) with stress annotations included (this system was eventually excluded as explained in Section 7.1 and Section 7.4).

The tests were conducted in relatively quiet rooms using soundproof headphones (Sennheiser HD201). The participants were briefly introduced to the experiment where they were briefed about the concept of listening tests, what type of questions to expect, how many times to listen to the audio stimuli and the meaning of some of the terms that they encountered throughout the survey. Each participant had to complete a form stating their age, gender, whether they have a hearing impairment, whether they were speech experts or not, their Arabic dialect and their education level. For the Preference tests, the participants were encouraged to listen to the prompts once, but were allowed to listen as many times as they wanted. For the DMOS tests, the listeners were allowed to listen as many times to both the synthesised prompts and natural ones for

reference, but were gently encouraged not to repeat them if possible. This setup is based on experiments carried out by experts in previous works (Dall et al., 2014; Müller, 2007) where in MOS type tests, participants were allowed to replay the stimuli, unlike in Preference tests. The participants were allowed to take a break at any time if they wanted to, or withdraw from the test.

100 naturally recorded utterances were used as the test set. They were split into two sets of 50, called partial sets (p-sets). Each set was allocated to half the 24 participants, where the participants listened to all the natural and synthesised stimuli in their allocated set with no randomisation.

- The first 10 utterances from each p-set were synthesised by the two systems (1 and 2) to be used in the Preference test. In each question, the order in which each pair was presented was random. The randomisation was done once and the same order was used for all participants but the participants did not communicate with each other as the test was anonymous. The reason not to randomise for every participant was due to the iSurvey platform used (University of Southampton, 2016), which is restricted in terms of its randomisation capability. A bespoke system for listening tests could be implemented, but this could lead to errors in the data. iSurvey has been used at the University of Southampton for a long time (University of Southampton, 2016). Overall, there were 20 pairs for each Preference test for each system (10 from each p-set).
- The following 20 utterances from each p-set were synthesised by one of the two systems to be used in the Preference and DMOS tests. This generated 20 pairs for each DMOS test for each system (10 from each p-set). The synthesis of the utterances by the systems to generate the test pairs were designed as follows:
    1. The first 10 utterances were synthesised by system 1 only and each generated utterance was paired with the natural studio recorded speech. Then each pair was presented as one question in the DMOS listening test.
    2. The second 10 utterances were synthesised by system 2 only and each generated utterance was paired with the natural studio recorded speech. Then each pair was presented as one question in the DMOS listening test.
- The remaining 20 utterances from each p-set were used for the final MOS test. The MOS test was conducted on both systems. This guaranteed that each MOS test (for each system) involved 20 distinct utterances used (10 from each p-set). The synthesis of the utterances by the systems to generate the test utterances was done as follows:
    1. The first 10 utterances were synthesised by system 1 and each generated utterance was presented as one question in the MOS listening test.

2. The second 10 utterances were synthesised by system 2 and each generated utterance was presented as one question in the MOS listening test.

Figure 7-1 shows how the utterances were used in the tests showing only one p-set.

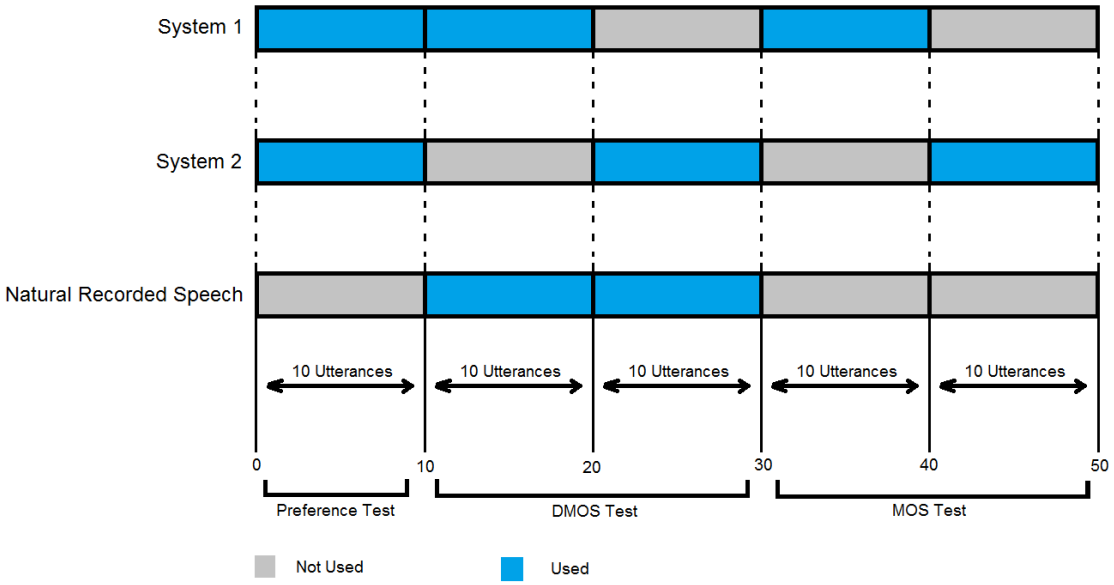

Figure 7-1. Utterance usage in listening tests for each p-set

For the DMOS test, the natural studio recorded utterance was presented before the synthesised utterance. The participants listened to the natural utterance then the synthesised utterance, and then scored how much degradation there was. In total, each participant was presented with 20 DMOS pairs, 10 for system 1 and 10 for system 2.

For the Preference tests, utterances as a pair were presented together (with no natural speech) in random order. Each participant was presented with 10 Preference pairs, and chose the one they preferred or "no preference".

Both in the Preference test and the DMOS test, the participants answered twice for each question, once for each factor: "Naturalness" and "Overall Impression". The choice of these factors was justified in Section 7.3 and was mainly to focus on the effect of the speech corpus rather than the system as a whole and assuring comparability with other works.

For the MOS test, each participant was presented with 10 utterances separately for each system and they were asked to score 1 to 5 each utterance just their "overall impression". Since there was no reference natural recording in this test, they were asked to give the score based on their own expectations. This test was conducted to allow this system to be compared with other systems that had conducted similar MOS tests (SynSIG, 2016; Chalamandaris et al., 2013).

The tests were "semi-controlled"; the participants were accompanied by a supervisor making sure that the environment was quiet, the headphones and computer were properly setup, and the participants had no unanswered questions.

7.6 Results of Tests

7.6.1 Participant performance time and errors

On average, each participant took 34.5 minutes to finish the tests, excluding the time they spent listening to the briefing and reading the information sheet, instructions and consent information. This means that on average, each participant spent 41.4 seconds on each question. Section 0 compares this with the average duration of the stimuli. No participant withdrew from the test after they had started and no participants took any breaks. 31 candidates were invited of which 24 participated. 12 participants conducted the test with p-set 1 and the remaining 12 participants conducted the test with p-set 2. This means that every stimuli pair (or utterance in case of MOS tests) was listened to 12 times. None of the participants' information was excluded from the final results but 4 answers (out of the total 1920) were excluded because they contained empty values, 2 in the Preference test and 2 in the MOS test (one of the missing values from system 1 and the other from system 2). Table 7-2 shows the number of answers and exclusions for each test, system and factor.

Table 7-2. Total and Excluded answers (data points) of each test, system and factor

System/Factor	Preference		DMOS		MOS	
	Included	Excluded	Included	Excluded	Included	Excluded
1/Naturalness	238	2	240	0	N/A	N/A
2/Naturalness			240	0	N/A	N/A
1/Overall	238	2	240	0	238	2
2/Overall			240	0	238	2

7.6.2 Demographics

Figure 7-2 to Figure 7-8 show the demographics gathered from the 24 participants in these tests. These indicate that there is sufficient coverage in some types but a lack of coverage in others. There was good coverage of gender, dialect and education. The most represented dialect was

Levantine. This is considered to be a positive feature as the target dialect in this work is Levantine MSA, but all other dialects were represented except for Sudanese. As the corpus was in MSA, the participants were preferred to be educated to a University level, because it is then guaranteed that participants had received formal MSA education (Habash, 2010). No participant reported having any hearing difficulties and most participants were aged between 25 and 35. Age range coverage was not considered critical as long as all participants had no hearing difficulties. It was merely used as a second confirmation of the participant's compliance with the minimum age restriction.

The participants were asked how they would rate their understanding of the fields of "Arabic Phonetics" and "speech technologies". The participants' understanding of the term "Arabic Phonetics" was slightly biased towards "full understanding" (Not the case for "speech technologies"). This is considered desirable for the purpose of this experiment as the participants had to be briefed about the concepts of "Naturalness" and "Overall Impression" in the context of MSA phonetics, and their understanding of this concept is critical to the success of the experiment (Dall et al., 2014; Wester et al., 2015).

It is important that one of the participants did not provide their demographic information. To make sure that all participants were over 18, they had to approve the content of the participant's form which means that they comply with the age restriction introduced in this study.

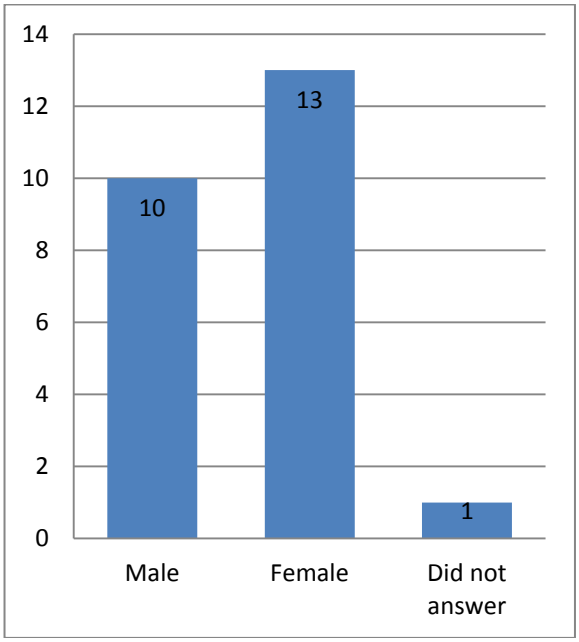

Figure 7-2. Gender of participants

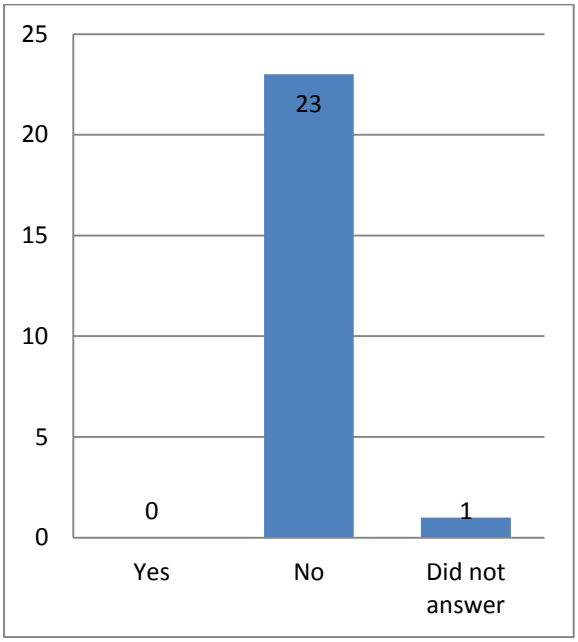

Figure 7-3. Presence of a hearing difficulty

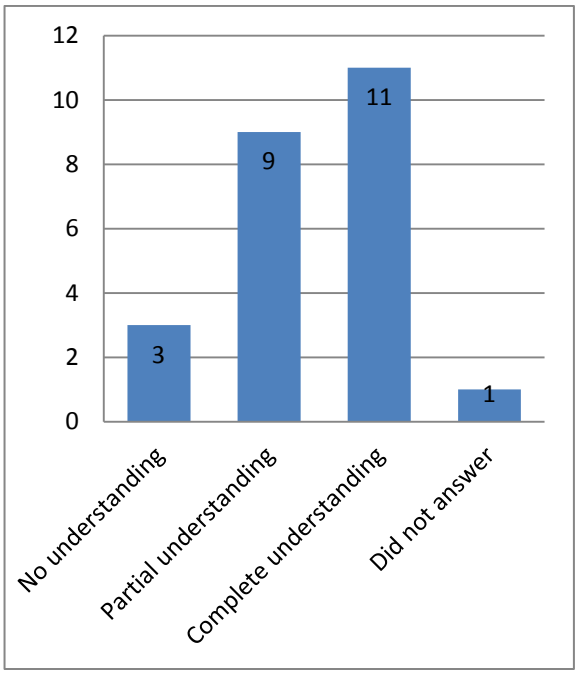

Figure 7-4. Arabic phonetics expertise

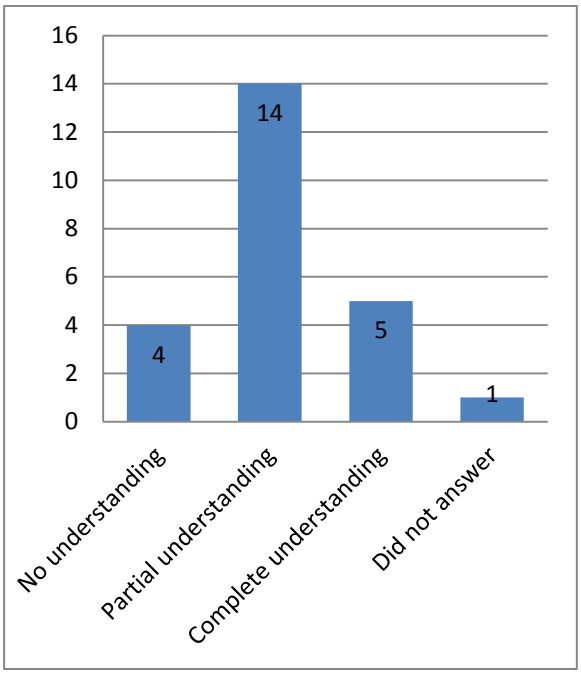

Figure 7-5. Speech technologies expertise

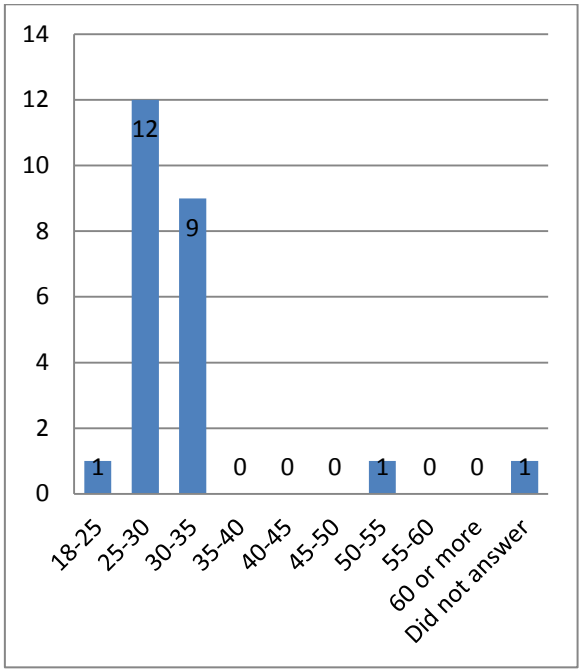

Figure 7-6. Age range

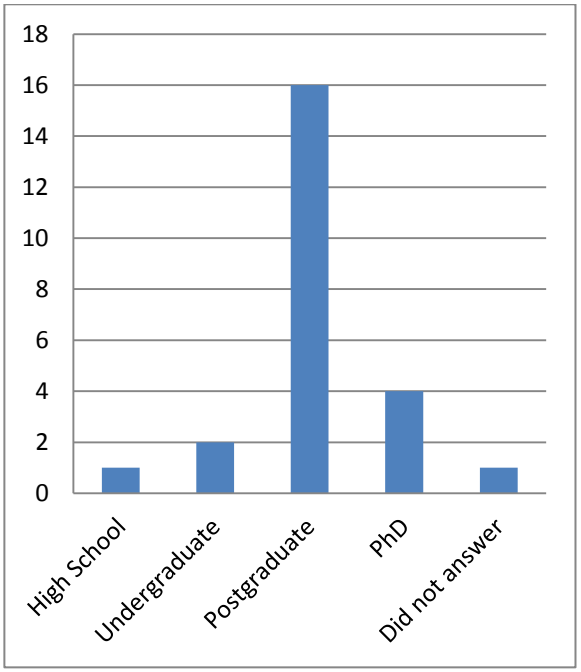

Figure 7-7. Education

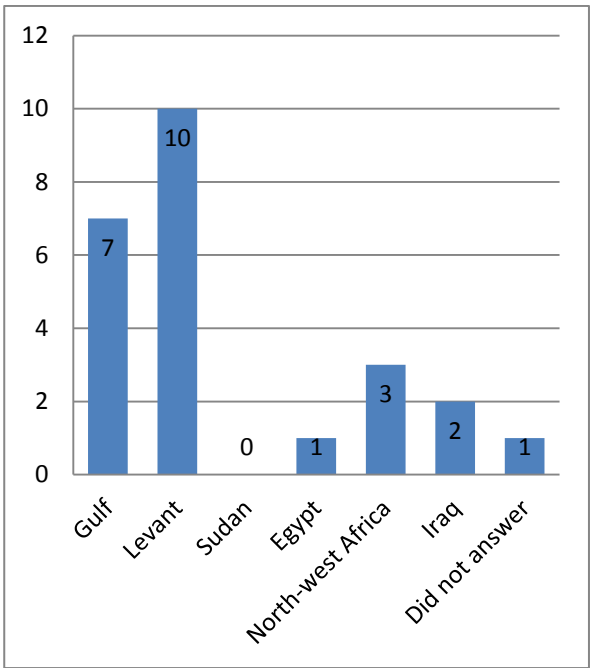

Figure 7-8. Dialect category

7.6.3 Results of Preference, DMOS and MOS tests

7.6.3.1 Analysis of Statistical Significance

So far, this work has established that the data extracted from the test conducted is reliable, and that the test conditions, setup, and participants are suitable for generating data that answers the research question. Next, the statistical significance of the results is assessed.

The three listening tests conducted in this work, MOS, DMOS and Preference tests, can be split in two categories:

- Categorical (Preference test): The participants had to choose from a finite, discrete set of options. Each option cannot be described as being smaller, less-than, greater or higher than the other (First, Second, No preference).
- Ordinal (MOS and DMOS tests): The participants had to choose from a finite, discrete set of options. Each option can be described as being greater or smaller than another (1, 2, 3, 4 and 5; or Very bad, Bad, Medium, Good, Excellent).

To assess the statistical significance of our results, different types of statistical test were chosen for each type of listening test.

- For the Categorical listening test, the Pearson's chi-squared statistical test (Field, 2009; Michael, 2001) (also known as Chi-squared goodness of fit) was used to find how unlikely

the binomial distribution was resulting from the preference test, assuming that a uniform binomial distribution is expected (H0). This test's assumptions are:

a. Data is obtained from a random sample. This assumption could only be satisfied if the target population of this study is reduced. As the demographic data has shown, the target population is quite specific. This means that – except for gender which is fairly equally represented – age range, hearing difficulties, educational level, understanding of speech technologies and Arabic phonetics, and dialect, are all ignored in this test. Their possible effects on the results are suggested as future work.
b. Sufficient sample size. This was satisfied after conducting an *a priori* Chi-squared test to estimate the required sample size.
c. Categories are mutually exclusive, so a participant is not allowed to have "no preference" and also prefer one of the systems simultaneously.
d. Each observation is independent. Any of the participants' choices are independent of the other choices they or others made. This is further confirmed by randomising the order of the preference pairs as they are presented to the participant. This restriction is sometimes stated more strongly by saying that each participant is only allowed to vote once to further guarantee the independence of each vote. In this work, due to the difficulty of gathering large numbers of participants, this was not the case. Each participant was part of 10 data points, so it is assumed here that there was a negligible effect of each participant's vote on their other votes. This assumption is almost always mentioned in published works which include listening tests (Mohammadi et al., 2014; van Niekerk, 2014). This is also related to the lack of diversity in the demographics shown in Section 7.6.2. As these tests are difficult to conduct (except when crowd-sourcing them (Buchholz and Latorre, 2011)) one has to limit the number of participants (with each participant answering the same question multiple times) which restricts the demographic coverage.
e. The expected frequency of each category is at least 5. The data collected satisfies this.

- For the Ordinal listening tests, the Wilcoxon signed-rank test (Field, 2009; Lowry, 2007) was used to find how unlikely it would be that the two samples (system1 and system2) differ or are the same (H0). It assumes the following:
    a. Data comes from the same population. This is sometimes assumed (Lowry, 2007) and is satisfied in this work.
    b. The dependent variable is intrinsically continuous (Lowry, 2007). This is satisfied as the MOS and DMOS value are considered to be continuous, but the participants are only allowed to choose from the discrete Likert scale question for simplicity. This is reflected in the average of the Likert scale answers, which is continuous.

c. The pairs of samples are independent of each other (participants' votes are not affected by other participants' votes but might be by other votes by the same participant).
d. The data is ordinal. The intervals between values are not required to be equal but could be. For example, the difference between "Good" and "Medium" is not assumed to be necessarily the same as the difference between "Excellent" and "Good" in a labelled, Likert scale setting.

For all types of test result (proportions for Preference test, and means for DMOS and MOS tests), the confidence intervals at 95% confidence were calculated for visual justification when examining the graphs, so that the level of overlap between compared values was visible.

The assumptions listed here for the two statistical tests used differ from the publications reviewed. Some used a paired t-tests instead of Pearson's Chi Square for hypothesis testing of Preference tests (Buchholz and Latorre, 2011; Zen et al., 2012), without much explanation of the statistical testing process. Others also used t-tests to hypothesis test MOS type tests (Hirose and Tao, 2015). Others used Chi Square tests on Preference tests as undertaken here (Mohammadi et al., 2014; van Niekerk, 2014).

Most other works did not specify what types of statistical test they used and did not discuss statistical significance of results in their analyses. This informed the decision to rely on the assumptions and expert opinions when choosing the tests. The main reason t-tests were excluded was the ordinal scale used being small (1 to 5) and the assumption of normality was in danger of being broken. It was safer to choose the Wilcoxon signed-rank test instead, which does not make this assumption.

In this work, since the mean of the Likert scale results were taken, the intervals between the each of Likert scales' ranks were assumed equal (Holz et al., 2006).

For each test, an *a priori* and a *post hoc* analysis was conducted. The *a priori* tests were done to determine whether the sample size was enough to achieve the required power value (power analysis), and the *post hoc* test was done to assess the statistical significance of the results (should one reject the null hypothesis). Here are some definitions and assumptions for these tests:

- *Required Power* = 90%: Power is the probability of rejecting the null hypothesis when the observation hypothesis is true. The value 90% was chosen as it was higher than the *de facto* standard in the literature (Field, 2009) but no previous work which included listening tests uses power analysis to estimate sample size.

- Alpha α = 0.05: Alpha is the statistical significance level, which is the maximum probability value allowed for the P-value to conclude that the results are statistically significant.
- P-value: The probability of obtaining results more or equally extreme to the observed results (Field, 2009). A P-value lower than α indicates statistical significance.

7.6.3.2 Hypothesis testing of Preference tests

For Preference tests only, the statistical significant test was formed of three stages (for Naturalness and Overall Impression separately) as the results were slightly more complicated than MOS and DMOS tests. The Preference tests were categorical tests of three categories. To assess the statistical significance of the results, three Pearson's Chi-square tests were conducted in sequence on the following distributions:

- The full, 3-category distribution from the Preference test, including "System 1", "System 2" and "No preference". The expected probabilities were all assumed equal (33.333%) in this test.
- The 2-catgory distribution resulting from grouping the "System 1" and "System 2" categories together. This basically leaves the categories "No preference" and "Had preference". The expected probabilities were 33.333% for "No Preference" and 66.666% for "Had preference" in this test. This was as a result of grouping the expected discrete uniform distribution in the previous test.
- The 2-category distribution resulting from omitting the "No preference" category. This was the conditional probability distribution of the participants' preference knowing that they had a preference. The expected probabilities were all assumed equal (50%) in this test.

Each one of these tests answers the following questions:

1. Is there any statistical significance in the results?
2. If yes, does the statistical significance lie in the fact that more or fewer participants than expected chose "No preference"?
3. If no, or if the number of participants have a preference higher than expected, is there a statistically significant difference between the number of people preferring "System 1" and "System 2"? And which one is higher?

A total of 238 data points were collected for each of these two tests. Table 7-3 shows the results of the Pearson's chi-squared for each of the two tests including the required power, alpha and X^2 values.

Table 7-3 shows that both results of the preference tests indicate statistical significance in the distribution. The P-values are lower than the Alpha and the sample size is enough to achieve the required power.

Table 7-3. Pearson's Chi-square test results on Preference test

Test	Minimum data points for 0.95 power	Data points	Critical X^2	Achieved X^2	Required Power	Actual power	Alpha	P-value
Naturalness preference	200	238	5.991	15.059	0.9	0.945	0.95	< 0.001 ***
Overall impression preference	157	238	5.991	19.193	0.9	0.982	0.95	< 0.001 ***

Assumes the discrete uniform distribution as the expected distribution.

This answers question 1, there is statistical significance somewhere in both results.

Table 7-4 shows that in test 2, there is no unexpected difference between the number of people who had "No preference" and the expected number in both tests. Further, the results indicate that more data points are required to allow for the possibility of rejecting the null hypothesis. Hence the null hypothesis is not rejected and the assumption that 33.3333% of participants in samples from our population would have "No preference".

Table 7-4. Pearson's Chi-square test results for "No preference" and "Had preference" categories

Test	Minimum data points for 0.95 power	Data points	Critical X^2	Achieved X^2	Required power	Actual power	Alpha	P-value
Naturalness preference	4650	238	3.841	0.538	0.9	0.114	0.95	0.463 ns
Overall impression preference	1163	238	3.841	2.151	0.9	0.311	0.95	0.142 ns

Assumes the grouped distribution from the discrete uniform distribution in test 1.

This answers question 2, there is no statistical significant difference between the expected and obtained ratio of people who had "No preference" in our sample.

Table 7-5 shows that in test 3, there is a statistically significant difference between the expected and resulting distributions of choices of preference in "System 1" and "System 2". Both P-values

are lower than Alpha and the number data points acquired for both tests is enough to achieve the power required.

Table 7-5. Pearson's Chi-square test results excluding the "No preference" category

Test	Minimum data points for 0.95 power	Data points	Critical X^2	Achieved X^2	Required power	Actual power	Alpha	P-value
Naturalness preference	123	164	3.841	14.049	0.9	0.994	0.95	< 0.001 ***
Overall impression preference	86	148	3.841	18.270	0.9	0.9997	0.95	< 0.001 ***

Assuming the discrete uniform distribution as the expected distribution.

This answers question 3 and one can conclude that in both tests, System 2 is preferred by the participants, see Figure 7-9 and Figure 7-10. 58 participants preferred "System 1" and 106 preferred "System 2" in the naturalness preference tests. 48 participants preferred "System 1" and 100 preferred "System 2" in the overall preference tests. Both values indicate more than 64% of participants preferred "System 2".

The two tests were considered independent, and the number of users who preferred "System 2" for both tests, one of the tests or none of the tests was not assessed. This was because this is not done in previous work and does not help answer any of our research questions.

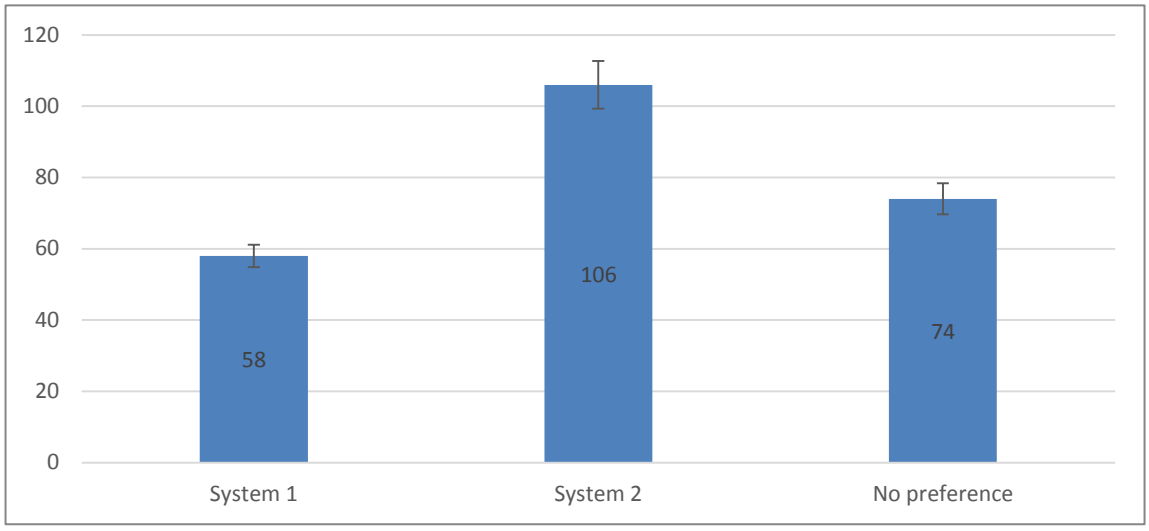

Figure 7-9. Preference test results for naturalness

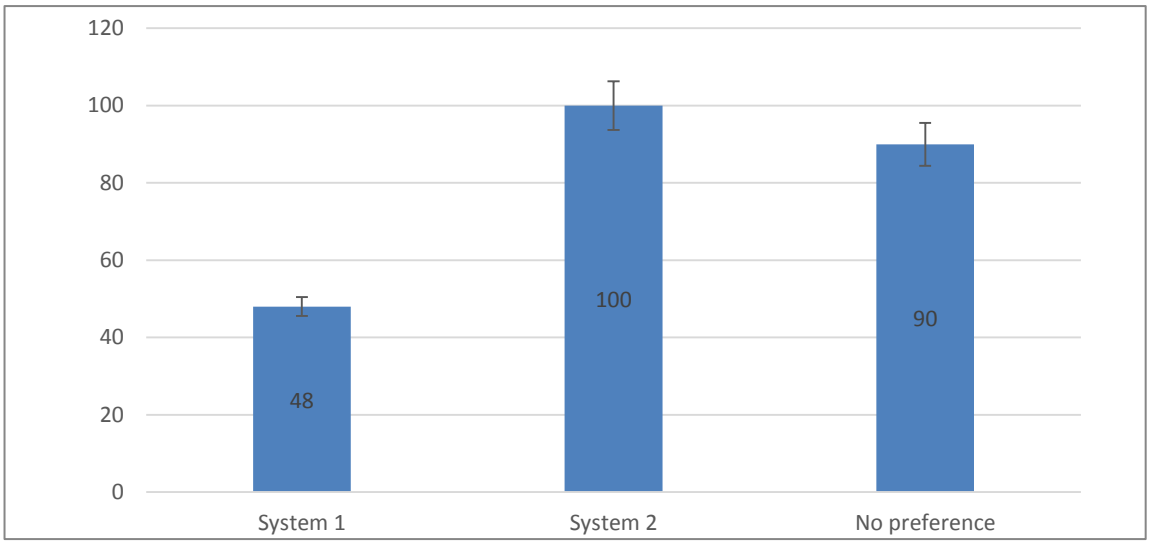

Figure 7-10. Preference test results for overall impression

7.6.3.3 Hypothesis Testing of MOS and DMOS Tests

The Wilcoxon signed-rank test was conducted to compare the means (averages) three times on each of the MOS tests carried out:

- DMOS naturalness test with 240 data points for "System 1" and "System 2".
- DMOS overall impression test with 240 data points for "System 1" and "System 2".
- MOS overall impression test with 238 data points for "System 1" and "System 2".

Each of these 3 tests helps answer one of the following questions in sequence:

1. Which system is further from human recorded speech in terms of naturalness in case there was a statistically significant difference in the DMOS results?
2. Which system is further from human recorded speech in terms of overall quality in case there was a statistically significant difference in the DMOS results?
3. Which system is perceived to be "better quality" by the participants in case there was a statistically significant difference in the MOS results?

Table 7-6 shows the results of the 3 tests. For both the DMOS tests, there is a statistically significant difference between the two means. Both P-values are lower than Alpha and the number of data points is enough to achieve the minimum required power. In the MOS test, no statistically significant results can be inferred as the number of data points acquired does not achieve the require power and the P-value is higher than Alpha.

Table 7-6. Results of the Wilcoxon signed-rank test for both MOS and DMOS tests

Test	Minimum data points for 0.9 power	Data points	Required power	Actual power	Alpha	P-value
DMOS naturalness	181	240	0.9	0.963	0.95	< 0.01 **
DMOS overall impression	217	240	0.9	0.927	0.95	< 0.01 **
MOS overall impression	585	238	0.9	0.545	0.95	0.104 ns

Figure 7-11, Figure 7-12 and Figure 7-13 show the means of each of the DMOS and MOS tests with 95% confidence intervals. The results in the two DMOS tests agree with the preference tests results in Section 7.6.3.2. The participants found "System 2" to have less degradation relative to natural recorded speech in terms of both naturalness and overall impression. In both graphs, the differences are clear with very slight overlapping of confidence intervals. These results not only strengthen the claim of preference for "System 2" (for both naturalness and overall impression), but they enable comparison with past and future works.

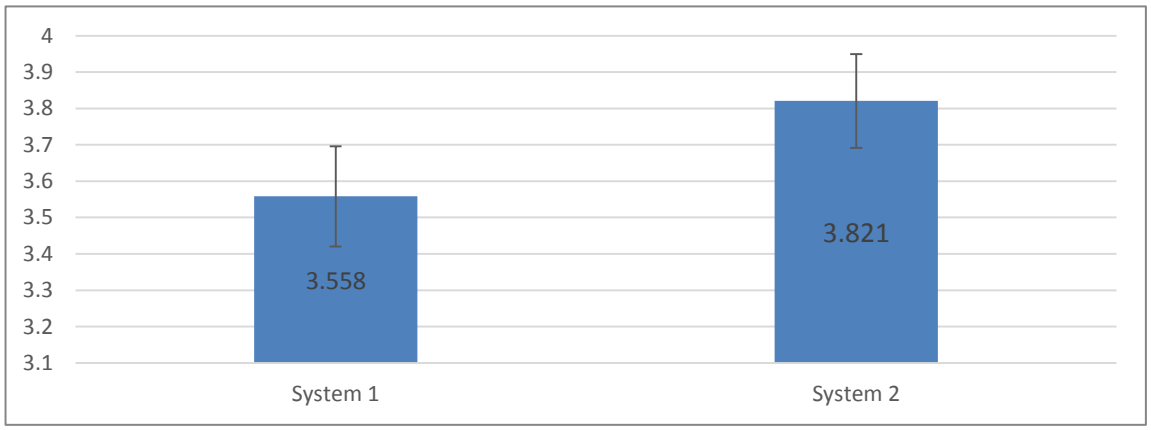

Figure 7-11. DMOS for naturalness test results with 95% confidence intervals

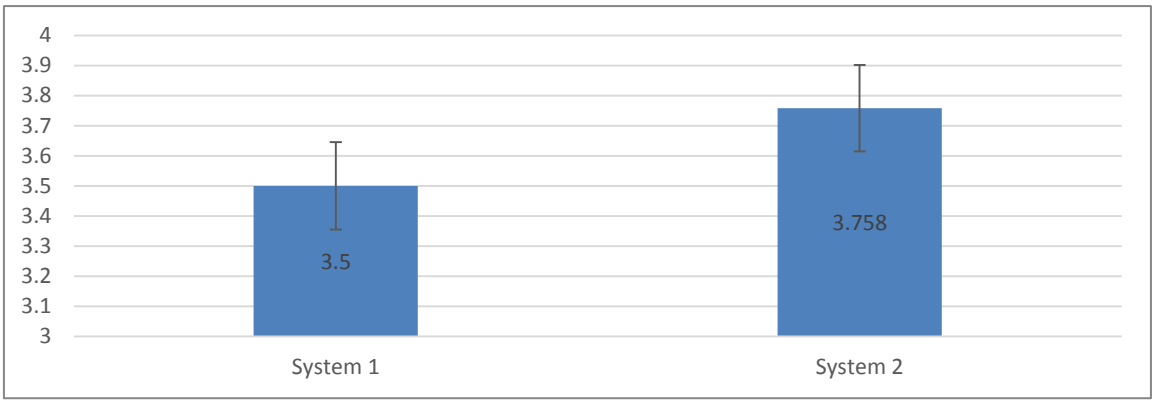

Figure 7-12. DMOS for overall impression test results with 95% confidence intervals

Figure 7-13 further emphasises the statistical insignificance of the MOS test results and the strong overlap between "System 1" and "System 2". This means that when there are no clear criteria to rate the stimuli, the participants tend to have less preference between the systems. In spite of the statistically insignificant MOS tests result, the means can still be used to compare this work's corpus or the Innoetics system to other work.

0 contains the full numerical results of all the DMOS and MOS tests with standard deviation and confidence intervals.

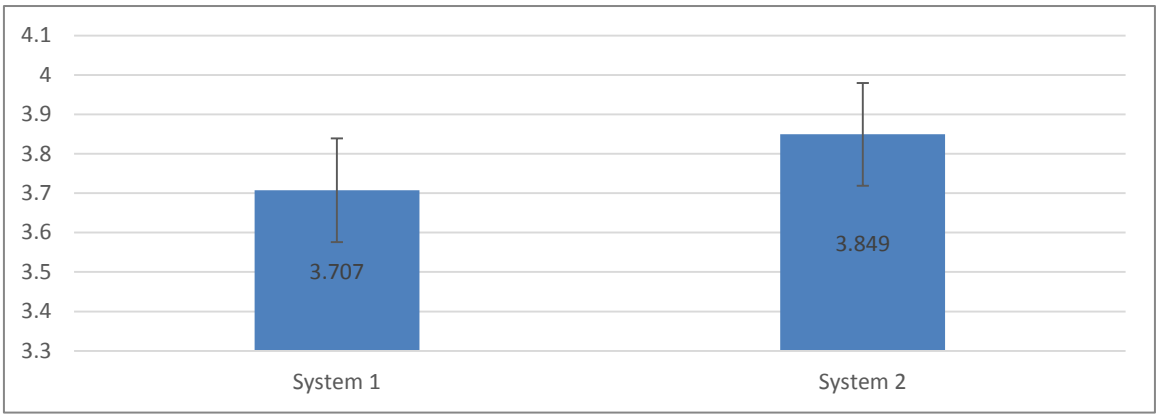

Figure 7-13. MOS for overall impression test results with 95% confidence intervals

7.6.4 Descriptive analysis

From the demographics data in Section 7.6.2, it can be easily seen that the sample used in the evaluation of this work's speech corpus does not cover the demographics information collected in the listening tests. It is argued here that this is not an issue, as the subject dialect being used in the listening tests is MSA and all the participants (except for one) have at least an undergraduate university degree and it is assumed that they have received formal Arabic education (Habash, 2010).

With the numbers of participants usually used in listening tests (Wester et al., 2015), it is impossible to cover these demographics. Further work needs to be done to find better criteria for choosing participants and the number of required data points.

From the outcomes shown in Section 7.6.3, one can draw the following conclusions.

- The number of people who had a preference between the two systems is statistically significant.
- For both naturalness and overall impression, participants preferred "System 2" when presented with the stimuli of both systems together.
- For both naturalness and overall impression, participants gave "System 2" a statistically significant, higher score when there was a baseline stimuli presented to them (DMOS), and they did not have a clear preference when the baseline was missing (MOS).
- The MOS result means that both systems are comparable with other published works (Chalamandaris et al., 2013; Sainz et al., 2012) including winners of the Blizzard challenge (SynSIG, 2016). The Innoetics system (the same system used in this work) achieved 3.1 average on MOS test for "Overall impression" when used with other speech corpora, (Chalamandaris et al., 2013), compared with 3.849 using this work's corpus.
- That 3.1 average achieved previously was using data for audio books in an Indian language, which was not designed specifically for speech synthesis. Further research is suggested to assess this work's corpus relative to other speech corpora, but it can be safely concluded that this speech corpus is suitable for high quality speech synthesis.

This leads to the conclusion that the stress features, extracted using rules introduced by Halpern 2009, can increase the prosodic naturalness and overall quality of Standard Arabic synthesised speech. The naturalness and overall quality here have not been assessed by specific types of speech synthesis users and it is not part of this work as the corpus is aimed at "general purpose" speech synthesis. People with certain types of visual or hearing impairment might prefer voices which are robotic (Moore et al., 1997). This answers the third research question put forth in this work.

A complete comparative analysis is not within the scope of this work and was not conducted due to the lack of Arabic, single-speaker speech corpora and the unfairness of comparing our corpus to other corpora in different languages.

7.6.5 DMOS and MOS Test Reliability

The subjective nature of MOS tests has been critiqued in previous work for being unreliable and subjective. Huckvale et al. (2012) compared their performance-based method of evaluation, which involves participants finding errors in text after listening to the correct synthesised version. The fact that MOS and DMOS tests require more data to yield reliable comparative results (CSTR, 2016; Wester et al., 2015) (which was true also in this work) further emphasises the superiority of preference tests for comparing different speech synthesisers. T-tests are commonly used to assess the statistical significance of MOS tests' results when comparing means of two different samples from two different synthesisers. These tests assume normality, which was a strong assumption in this work, as the Likert scale response is too narrow and discrete. It was also recommended by statistics expert consulted (Green, 2016).

In spite of the above, DMOS and MOS tests were conducted here to allow future works to compare their results to these using MOS tests if preferred, or to conduct preference tests of their own, and to allow the comparison of this work to theirs. DMOS yielded more statistically significant results than the pure MOS test in this work, and these results agreed with the preference test's results. This further emphasises the unreliability and subjective nature of MOS tests.

In the Blizzard Challenge 2007 (Clark, Podsiadło, et al., 2007), a detailed description is presented of the statistical testing process to analyse "MOS tests" results. They also used the Wilcoxon signed-rank test for assessing statistical significance, but they claimed that comparing or calculating means of MOS tests is incorrect for ordinal data where the intervals are not guaranteed to be equal. So far in this work, these intervals were considered to be equal and means were calculated. To further confirm the results, Figure 7-14 and Figure 7-15 show the box plot which contains the medians and the quartiles of the four DMOS samples and the two MOS samples. For all samples the median is 4, and it is clear that "System 2" has its upper quartile at 5 for all samples and it being 4 for all of "System 1" samples. This further assures the statistical significance of the results and the preference of "System 2".

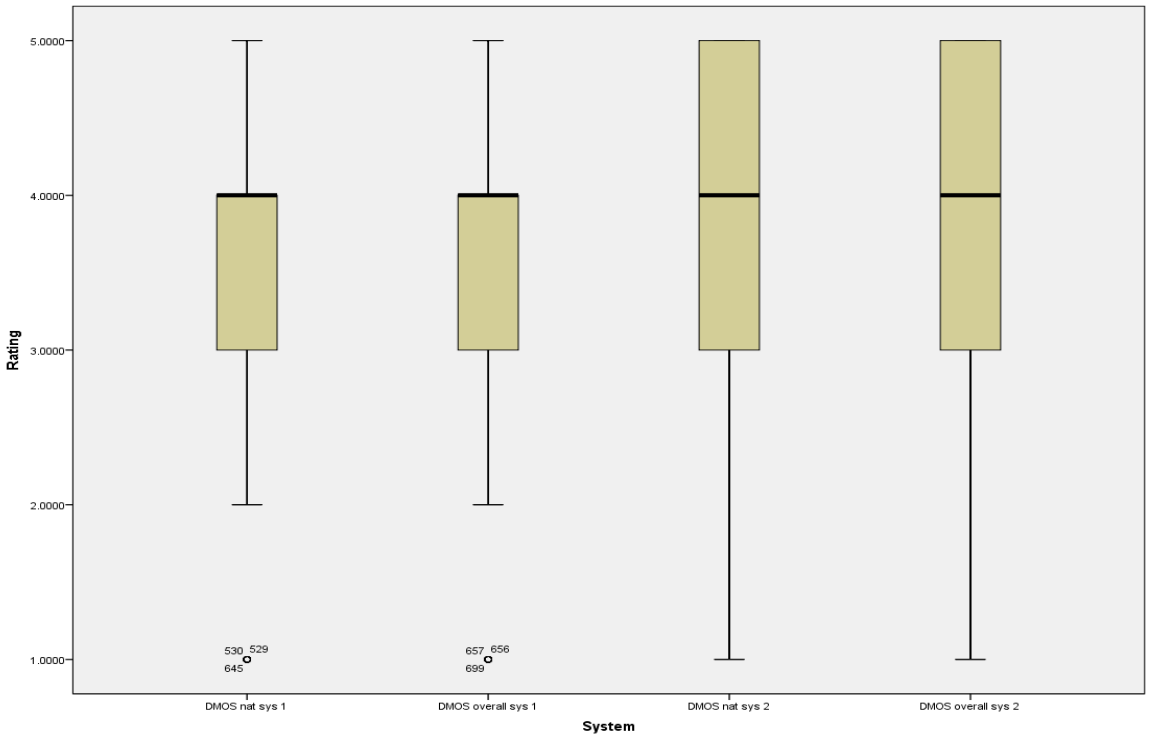

Figure 7-14. Box plot for DMOS tests for overall impression and naturalness

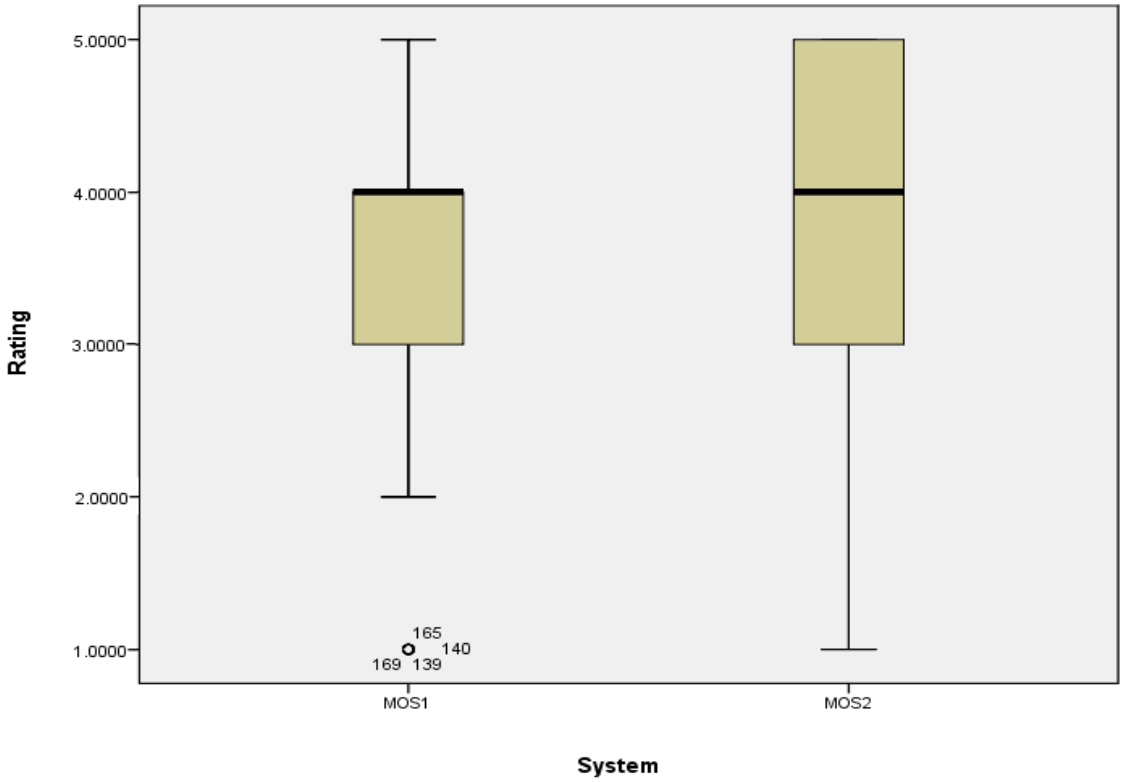

Figure 7-15. Box plot for MOS test for overall impression only

Further questions regarding the reliability of MOS type tests remain. As recommended by Gilbert (2015), the score of each participant should be averaged, which yields much smaller data sets, but means there is a need to conduct the listening tests on a much larger scale (hundreds of

participants). In this work, the assumptions and methods of published work were adhered to and there was no attempt to fix the testing process; rather, just conduct it in the most comprehensive way possible to remove doubt.

According to an expert opinion (Gilbert, 2015), since the DMOS tests for naturalness and overall impression were done together, it is a strong assumption to consider each a separate experiment. The expert suggested using a one-way ANOVA (Analysis of Variance) (Field, 2009) to compare the means of the four samples:

- DMOS test of naturalness for system 1
- DMOS test of naturalness for system 2
- DMOS test of overall impression for system 1
- DMOS test of overall impression for system 2

It was also suggested averaging the 10 scores given by each of the 24 participants in every sample. This left 24 data points in every sample instead of 240. Table 7-7 shows the ANOVA results clearly demonstrating that the 24 data points are not statistically significant (P-value = 0.293 > 0.05) even though the means are different. The same test (one-way ANOVA) was completed as well without averaging for every participant (240 data points for each sample), which was usually reported in the literature. This second test clearly shows that the samples were NOT taken from populations with the same means (P-value = 0.003 < 0.05). This further emphasises the lack of consensus on the reliability of MOS type tests, especially with low numbers of participants.

Table 7-7. General statistics before the ANOVA test on the average for each participant

	N	**Mean**	**Std. Deviation**	**Std. Error**	**95% Confidence Interval for Mean**		**Minimum**	**Maximum**
					Lower Bound	**Upper Bound**		
1	24	3.558	0.651	0.133	3.284	3.833	1.800	4.900
2	24	3.500	0.661	0.135	3.221	3.779	1.800	4.500
3	24	3.821	0.640	0.131	3.551	4.091	2.400	4.800
4	24	3.758	0.740	0.151	3.446	4.071	2.100	4.800
Total	96	3.659	0.677	0.069	3.522	3.797	1.800	4.900

Table 7-8. ANOVA test statistical significance results for the averaged participant scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.715	3	0.572	1.259	0.293 ns
Within Groups	41.796	92	0.454		
Total	43.512	95			

Table 7-9. General statistics before the ANOVA test without averaging

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	240	3.558	1.092	0.071	3.419	3.697	1.000	5.000
2	240	3.500	1.150	0.0742	3.354	3.646	1.000	5.000
3	240	3.821	1.021	0.066	3.691	3.951	1.000	5.000
4	240	3.758	1.135	0.073	3.614	3.903	1.000	5.000
Total	960	3.659	1.107	0.036	3.589	3.730	1.000	5.000

Assuming each participant's answer as a data point

Table 7-10. ANOVA test statistical significance results for the non-averaged participant scores

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17.153	3	5.718	4.718	< 0.01 **
Within Groups	1158.463	956	1.212		
Total	1175.616	959			

The order of presentation of the MOS and DMOS stimuli was not randomised in the tests in this work, nor were stimuli necessarily always randomised in the literature (Dall et al., 2014; Shannon and Byrne, 2009). Another option is to conduct the tests for each system on different dates. In this work, the order of MOS and DMOS tests is not randomised (participants rated stimuli from "System 1" and then "System 2" but they were not informed about the existence of different systems in MOS and DMOS tests beforehand). This might allow the effect of habituation (Müller,

2007) to influence the results as “System 2” was always presented last, which may have led to higher scores given to “System 2”. This means that the MOS and DMOS results are most useful for comparisons with different works rather than between the two systems.

7.7 Summary

In this chapter, several important contributions have been made.

- A review of works has been conducted on subjective evaluations of speech synthesis systems and speech corpora. This review was to enable the evaluation method of the speech corpus to be chosen.
- Based on this review, it was decided that listening tests should be conducted to evaluate the speech corpus and show the usefulness of the stress features extracted orthographically (Halpern, 2009).
- Having conducted the listening tests, the results showed that the listeners preferred system 2, which uses the orthographically extracted stress features (Halpern, 2009). This was in terms of both naturalness and overall impression.
- The results also showed that the MOS average from system 2 are similar to the MOS average of the best system in the Blizzard Challenge 2013 (Chalamandaris et al., 2013).
- This chapter also highlighted the limitations of MOS tests. Both works in the literature and expert advice disagreed regarding the reliability of MOS results, especially when comparing systems (Clark, Podsiadło, et al., 2007; Huckvale et al., 2012). To overcome this problem, the results of the MOS tests were triangulated with those from the preferences tests.

Chapter 8 Conclusion and Future Work

This chapter lists the contributions achieved, and discusses how the research questions have been answered. The limitations of this work are discussed leading to suggestions for future work.

8.1 Comparison with other Arabic single speaker Speech Corpora

Almosallam et al. (2013) is the only published work for Arabic single speaker speech corpora to date. This section compares their work and this highlighting our contributions.

8.1.1 Phoneme Set

Almosallam et al. (2013) did not include several phonemes presented in this work which are essential in MSA. The phoneme set used in the newly-developed corpus included all of theirs and added phonemes that appeared to have been missed. An example is the 'leaned' vowels /u1/ and /i1/, which were not mentioned in their work. Another example is ignoring the discussion of geminated consonants.

In addition, the phonemic vocabulary of MSA was formalised, and the speech corpus created using the outcome of this formalisation, making this the first work to create a phoneme set for MSA in the context of speech synthesis.

8.1.2 Phonetisation Rules

Almosallam et al. (2013) did not describe how they converted diacritised MSA text to phoneme sequences before alignment. This was comprehesivly discussed in this work and a full set of phonetisation rules has been included. This is the first time a full set of phonetisation rules for MSA has been published, with a software implementation of these rules.

8.1.3 Corpus Evaluation

Understandably, because of the size and nature of their publication (5 page paper), Almosallam et al. (2013) did not include an evaluation as comprehensive as the one introduced in this work. They generated an HTS voice and conducted a small-scale MOS test for intelligibility and naturalness with only 10 participants with 5 stimuli each (50 samples), without conducting an *a*

priori test to calculate the number of the required samples. Table 8-1 shows general statistics about this work's corpus relative to Almosallam et al.'s corpus. The table shows there is no way to directly compare this work's corpus to theirs because of the different metrics used, but it is felt that the small number of participants they used, and the lack of a background study into listening tests in their work, makes this work a worthy source of knowledge for MSA corpus design.

Table 8-1. Corpus content comparison with Almosallam et al. (2013)

	Almosallam et al. (2013)	This work's corpus
Length	7 hours	3.7 hours
Electroglottograph signal (EGG)?	Yes	No
Aligned with phoneme sequence	Yes	Yes
MOS for overall impression	N/A	3.8494 (System 2)
MOS for naturalness	3.58	N/A
MOS for intelligibility (Not Recommended) (Huckvale et al., 2012)	3.9	N/A
DMOS for overall impression	N/A	3.7583 (System 2)
DMOS for naturalness	N/A	3.8208 (System 2)

Using MOS tests for intelligibility has been criticised in the literature (Huckvale et al., 2012), and performance-based tests are recommended as carried out in the Blizzard Challenge (SynSIG, 2016). Finally, the choice of naturalness as a factor by Almosallam et al. (2013) as a single test for his speech corpus could be seen as an incomplete and unreliable assessment of all aspects of the speech synthesis. However, it is recognised that time constraints may have been an issue. In the time allowed for this study it has been possible to offer a more comprehensive phonological study of MSA and an evaluation using methods supported by an increased amount of research.

8.2 Research Contributions

This work's contributions are listed below.

1. A set of phonemes of MSA and phonetisation rules for converting MSA text to phoneme sequences. These were not evaluated directly, but their quality was assumed to be inferred from the success or failure of evaluations of tools which use them (see contributions 3 and

7). These rules helped in building a complete MSA phonetiser as part of this work which could be used later as part of a fully functional TTS system.

2. A 3.7 hour, single speaker, MSA recorded speech corpus targeted at parametric and unit selection speech synthesis. This first-of-a-kind corpus is a seed for future work in Arabic speech synthesis.
3. An analysis of data collected objectively from the speech corpus to assess its length and coverage. An analysis of data collected subjectively from a set of listening tests to evaluate the usefulness of orthographically extracted stress features, as completed by Halpern (2009), in synthesising MSA speech, and to decide whether the speech corpus is useable for generating natural Arabic speech.
4. A critique of the way researchers in the literature often fail to mention their procedures when setting up listening tests and their evaluation methodologies. This could help future researchers in making better decisions when conducting listing tests and analysing the data.
5. A set of sequentially ordered steps for creating a speech corpus, describing the human and technical resources required for each step. These steps were not evaluated but are a result of a traditional literature review. Their validity was inferred from the quality of the resulting speech corpus. See Section 2.2 for more detail.
6. A software tool for converting Arabic (MSA specifically) text to a phoneme sequence (phonetic or phonemic transcript). This is usually called a phonetiser and is based on the phonetisation rules resulting from contribution 1.
7. A software tool for segmenting recorded MSA speech and aligning it with its phonetic transcript. The segmentation accuracy was presented as an evaluation of this tool.

Here is how some of the contributions answered all of our proposed research questions (some contributions did not correspond to a research question directly, but were a by-product of this work)

Question 1: “What is the phonemic inventory for MSA? Put in a different way: Which phonemes occur in connected MSA speech?” Answered by contribution 1.

Question 2: “What are the rules which govern how the orthographic transcript is to be converted to the phonetic transcript to reflect how the utterances are going to be pronounced during the recordings?” Answered by contribution 1.

Question 3: “How can an accurate segmentation and alignment system be achieved, that has been trained on the proposed corpus? How can this be answered comparatively and take into

account all the possible parameters that can be changed in that system?" Answered by contributions 2 and 7.

Question 4: "How much coverage does the reduced orthographic transcript achieve for MSA phonetics and phonology?" Answered by contribution 3.

Question 5: "How would a speech synthesis system based on unit selection and built with the proposed speech corpus perform when subjectively evaluated?" Answered by contribution 3.

Question 6: "Would adding stress markings to the speech corpus – based on the orthographically generated word stress markings from the work of Halpern (2009) – improve the results of the subjective tests conducted to answer question 5?" Answered by contribution 3 and 6.

This work has shown, by using the phonetic rules, tools and guidelines introduced throughout, that it is possible to produce a high quality MSA speech corpus, and that the stress features introduced by Halpern (2009) can be used to improve the speech quality and naturalness. However, as a result of this work, it is suggested that the evaluation methodologies for speech synthesisers and corpora need to be revised and standardised.

8.3 Future Work

This section suggests future work to complement the contributions of this work.

8.3.1 Modernising Arabic Phonetics and Phonology

The word "Arabic" was chosen as this section is not solely focused on MSA.

The results obtained in this work can be improved further to cover different dialects and phonetic aspects of the language (Primary stress, secondary stress, vowel duration,…).

The phonetical study was limited to the purposes of this research, which was acceptable mainly because of the permissive nature of modern speech synthesis systems like Innoetics (Chalamandaris et al., 2013), which contain advanced methods to prune and correct the database and model prosody and duration, giving room for errors when constructing a speech corpus or using incomplete phonetic rules.

Better research on Arabic phonetics and phonology would be useful for speech synthesis and it is necessary to know how Arabic – in all dialects – is actually spoken ,with a full phonetic analysis rather than coping with explanations and a limited range of examples from early works on the

subject. Old works (prior to 2008) had to be consulted in this work due to a lack of modern works (Thelwall & Sa'Adeddin 2009; de Jong & Zawaydeh 1999; Halpern 2009), and at times this hampered research into the complexities of the subject and failed to offer in-depth support that may be readily available for other languages.

This work is a seed for future works on Arabic phonetics and phonology in different dialects for speech synthesis.

8.3.2 Written and recorded sources for the Arabic Language

Before recording the prompts, all the text gathered to generated the prompts was scraped from the Aljazeera learn website (Al Jazeera, 2015). At the time of scraping, the complete text obtained was around 24000 words long. This work has proven that such a short text was enough to produce a high quality corpus, but it would be interesting to find out how the quality of the corpus would improve if the transcript size was increased. The corpus produced is far shorter than any of the corpora reviewed in the literature, which usually start with hundreds of thousands, or millions, of words before reducing the script for the prompts, and the sources of the texts are usually more diverse than just news as in this work (Umbert et al., 2006b; Bonafonte et al., 2008).

It is suggested that more research into using other text sources for Arabic in different dialects should be carried out. The Arabic Treebank is an example of a larger, syntactically annotated, fully diacritised orthographic corpus for MSA (Maamouri et al., 2005). It was not used in this work due to resource limitations, namely the purchasing of a licence and the fact that it was copyrighted. Using this Treebank would restrict the license of the corpus generated by this work. It was felt important to make our content publicly available.

The diversity of the sources should be wide enough to cover different topics and styles (declarative versus interrogative). This is not easy to obtain in a fully diacritised form, as most digital Arabic written text is undiacritised. For MSA, this problem can be solved by using automatic diacritisers, which – although not 100% accurate – can then be post-corrected using the recorded acoustic information by the segmentation systems (Torres and Gurlekian, 2008; Young et al., 1997). But the problem of text sources for Arabic dialects other than MSA remains unsolved.

For speech recognition and synthesis, recorded speech corpora for Arabic language remain scarce and lack any form of evaluation as well as being restricted in the number of dialects covered. In this work, only a basic survey of the existing corpora was presented, because most of the corpora were targeted at speech recognition rather than synthesis. A full, more detailed survey is suggested as future work. There are proprietary MSA corpora and voices on the market (Ocean,

2016; Acapela Group, 2015). Details about these voices remain unavailable to researchers, but an attempt to contact the owners of these voices to add the details of their corpora to the survey is suggested.

8.3.3 Corpus Segmentation

Although the aligned corpus was sufficient for producing a high quality speech corpus, the method used for segmenting the corpus in this work is not state of the art, and there is potential for improving the accuracy of the alignments.

This work used a basic HMM segmentation system (see Chapter 5) and compared it to a more accurate ANN/HMM system (Hosom, 2009). Even the more accurate ANN/HMM system does not claim to be state of the art.

As this work did not aim to improve automatic speech segmentation and did not include a comprehensive review of all the many speech segmentation methods, it is suggested that a more complete survey of recent improvements should be undertaken. It is suggested that the combination of this work's corpus along with more accurate segmentation and alignment would improve the quality of synthesis.

However, within the constraints placed on this research work, the segmentations generated were sufficient to produce high quality speech synthesis.

8.3.4 Subjective Corpus Evaluation

Chapter 7 included a comprehensive subjective evaluation of our speech corpus, mainly based on previous works (Dall et al., 2014; Wester et al., 2015; SynSIG, 2016; Chalamandaris et al., 2013). It also included the limitations and the doubts about the reliability of MOS scores and the different statistical significance tests used in the literature and repeated in this work. No claim was made here about the superiority of the different evaluation methodologies used, but it can be said that the Blizzard Challenge's methodology is the most established (Clark, Podsiadło, et al., 2007) and hence recommended for future works.

All aspects of the Blizzard Challenge's methodology were followed here except for the number of participants. The different Blizzard Challenges conducted so far have a statistically significant, higher number of participants relative to other published works (Wester et al., 2015). This is mainly due to the difficulty and high cost of hiring hundreds of paid participants and controlling each listening test. For future work, it is suggested that the Blizzard Challenge's listening test software be shared (the researcher was not able to acquire it) as it allows for randomisation and

helps standardise the testing procedure. This helps overcome the restrictions that other survey systems have, such as the one employed in this work (University of Southampton, 2016).

The costs and difficulties in hiring more participants may be solved by the use of crowdsourcing techniques. Crowdsourcing has already been reported in the literature (Buchholz and Latorre, 2011) and has been shown to correlate to some extent with controlled test results. The reason it was not used in this work is the continuing preference for controlled listening tests.

The evaluation of our speech corpus did not include a full comparative analysis against other works in the literature in Arabic or other languages. This is due to the lack of established criteria and lack of other speech corpora, mainly in Arabic. The survey discussed in Section 1.2 can be further enriched by conducting a comparative analysis of the different corpora in Arabic and other languages available.

The evaluation factors used in this work (naturalness and overall impression) are two of many used in previous work. It is also suggested that other factors be used in the survey if listening tests were to be included in that survey.

8.3.5 An Arabic Text-To-Speech front-end

This is the most important future work suggested here, as it complements the back-end (synthesiser) built using this work's corpus. There is so far no complete **front-end** stack for Arabic TTS available, but tools are available that could help in building a complete front-end. eSpeak (Multiple Owners, 2016) is an open source speech synthesiser which allows the addition of transcription rules for numbers and dates, converting them to text. This does not include named entity recognition, or recognising words with irregular pronunciation or abbreviations. The ICU project (Multiple Owners, 2016) is a set of C/C++ and Java libraries for conducting linguistic tasks in Unicode that are locale specific. The project includes converting Arabic numbers and dates to undiacritised text.

Automatic prediction of diacritics for Arabic Script is a well-researched subject in the literature (Haertel et al., 2010; Habash and Rambow, 2007; Ananthakrishnan and Avenue, 2005). A number of high quality systems are available with different licenses. The most accurate one is MADAMIRA (Pasha et al., 2014) based on MADA (Habash and Rambow, 2012), but this is only freely available with a research license. It can be obtained with other license with commercial agreement. Other less accurate ones are available for free, such as Mishkal (Zerrouki, 2014).

Phonetisation is another component needed for an Arabic front-end, which is not available. As a by-product of this work a phonetiser was created (see Section 5.1). In summary, it is suggested

that future work is conducted to create a full Arabic (MSA specifically) front-end using the components suggested here. Further research is needed into improving the individual components like phonetisers and diacritisers.

8.4 Summary

The contributions made as part of this work have an important impact on the future of Arabic speech synthesis.

First is our speech corpus generated in this work, which is available online for downloading. At the time of writing it is considered to be the first freely available, fully annotated, single speaker, MSA speech corpus built as a result of scientific research. The hurdle of obtaining data for creating synthesisers for MSA is now overcome and future research can focus on other factors related to Arabic TTS, such as front-ends or speech synthesis.

Second, the phoneme set and the rules generated in this work resulted in the building of a MSA phonetiser which can be used as part of front-end in a TTS system for MSA. These rules can also be adapted to suit future research related to other Arabic dialects.

Third, the subjective listening tests have shown that using the orthographically extracted stress features allow them to be used to improve quality of synthesis in an MSA TTS. This is the first method of stress annotation proven to be suitable for MSA TTS in scientific research.

Fourth, the methodology chapter of this work included a full discussion of the stages and methods used to build the speech corpus. These methods can be adapted to other Arabic dialects to speed up the process of building other Arabic speech corpora.

At the time of writing, Innoetics have used our speech corpus, phonetiser, number-to-word converter produced in this work, along with MADAMIRA diacritiser (Pasha et al., 2014), to produce a high-quality, MSA TTS system viable for commercial purposes. In addition, two academic projects have been launched using our speech corpus as part of their work.

In summary, this work has succeeded in producing novel contributions which have resulted in direct impact on both academic and commercial work on MSA TTS. Not only has a new corpus been built, but research contributions have been made which could facilitate and expedite future work in other Arabic dialects.

Appendix A Acoustic Features

No speech recognition and segmentation systems perform inference directly on the speech frames. There is always a layer that transforms the raw speech data to a sequence of feature vectors that are calculated from a window with a certain width and shifted by a certain amount to calculate the next frame. The window size and shift (offset) are always measured in milliseconds and typical they are 20-25 ms and 5-10 ms respectively (Young et al., 1997).

The majority of the methods reviewed use mel frequency cepstral coefficients (MFCC) as acoustic features and often in combination with other features. These are extracted from the acoustic signal before any training or inference is done.

MFCC (Jurafsky and Martin, 2009) are coefficients that have been found to have strong correlation with the human vocal tract physical state and from this physical state it is possible to determine which phone is being uttered. This justifies the choice of MFCC as features because they enable the classification of phones from the vocal tract's physical state to then be classified from a correlated parameter that is MFCC.

MFCC are a representation of the speech signal that tries to ignore the unwanted information such as speaker identity and the loudness of speech. In tasks like speech recognition, it is not of interest to know whether the speaker is male or female (unless performing speaker recognition) or how loud they are speaking; rather to know which phone they are uttering, so two speakers with different sound characteristics and possibly gender should generate similar MFCC when uttering the same phones.

There were attempts to improve MFCC precision in speech segmentation by using it alongside other features. Hosom (2009) tried adding additional features related to bursts or sudden increase in loudness and intensity of speech which could indicate occurrences of events such as phone boundaries. Even though MFCC does have loudness and intensity change detection characteristics, it has been argued that these additions make the system more sensitive to those changes. The claim is that these feature additions have improved boundary detection for most boundary types.

Perceptual Linear prediction (PLP) (Hermansky, 1990) shares similarities with MFCC. It is also inspired by the human auditory system. The main difference in PLP (Hönig et al., 2005) is that it performs Linear Predictive Coding (LPC) to the pre-emphasised Bark scale transformed power spectrum, which generates an approximation of this spectrum. All this is before moving to the cepstral coefficients similar to MFCC. From the literature, no claim has been found that this linear

coding step simulates any stage of hearing in the auditory system and it appears to be just a dimension reduction method.

Other feature extraction techniques have been encountered such as Discrete Cosine Transform Coefficients (DCTC) and LPC coefficients. The former is strongly related to MFCC and PLP as it estimates the cepstrum of the speech. The latter is strongly related to the human vocal tract state. Karnjanadecha and Zahorian (2012) conducted a comparison between different types, MFCC, PLP and DCTC, and showed that MFCC is best when using 39 cepstral and energy coefficients, but they also showed that DCTC with 78 cepstral and energy coefficients outperformed all the other methods. They did not test for more coefficients for MFCC and PLP.

Appendix B Listening Test Briefing

Following are the points which the participants are briefed about before the listening tests.

1. Please sit in a relatively quiet place (a room that does not have windows on main roads, for example). This is in case you are doing the experiment at a later date than the briefing.
2. Please use good quality headphones if they are available to you, and set the volume to a slightly higher value than your comfort level so that you hear clearly.
3. Please read the first page very carefully (instructions page), as it contains the following instructions and more.
4. In summary, there are (after the first page which is just demographics) 50 questions divided into 10-20-20 (there are no section dividers).
5. You are encouraged to only listen to each stimulus ONCE. Don't listen again unless you were distracted.
6. Please pay close attention to two important things:
7. First, the meaning of naturalness. By naturalness I'm focusing on Intonation (pitch changes, is it natural) and rhythm (length of phonemes relative to each other, is it natural). Rhythm should not be confused with speed.
8. Second, the meaning of the 1-5 scale values that you will be scoring questions 11 to 50. In questions 11 to 30, 1 means "total/complete degradation" and 5 means "no degradation". In questions 31 to 50, 1 means "bad" and 5 means "excellent". It is very important that these terms are taken literally, as consistency is very important.

Please ignore some of the linguistic errors that are found in the speech stimuli. They are not to be considered as part of the test.

Appendix C Test Instructions

The following is the content of the first page of the listening test which contains instruction and consent information.

Hello,

Thank you for agreeing to participate in this survey.

This survey has been designed to conduct Mean Opinion Score and Preferences tests for three different Text To Speech systems (systems which generate sound from input text) using three different versions of an Arabic Speech Database to assess the efficacy of some features and certain parts of the database. This is part of a work by Nawar Halabi at the University of Southampton.

You have been chosen to participate in this survey because you are a native Arabic Speaker. No special knowledge of the Arabic language is required to complete the survey, but you will be asked to give a rating of your Standard Arabic proficiency. You will be asked a set of 50 questions. 10 in the first section, 20 in the second, and 20 in the last. Before each question you will listen to two audio files. There are three types of question:

1. Preference questions (10 questions): Please choose which prompt you thought was most "Natural" and had better "Overall Impression" out of a pair (each separately).
2. DMOS questions (20 questions): Please assess how close a synthesised audio file is to another prompt which contains natural speech.
3. MOS questions (20 questions): Please assess the quality of a synthesised audio file relative to your own expectation.

In the survey, you will come across two terms frequently: "Naturalness" and "Overall Impression".

- Naturalness: Refers to how natural a spoken prompt is in terms of intonation and rhythm (does the pitch sound natural and do the lengths of the vowels and consonants sound natural).
- Overall Impression: Refers to the general quality of the synthesised speech and how easy it was to listen to.

This survey may take up to 45 minutes to complete, so I am grateful for your support and time.

Please sit in a quiet place with headphones set to a comfortable volume level (possibly slightly higher for clear listening) and try to listen to the stimulus only once (you are free to listen more times).

Your data will be kept anonymous and confidential, only being used for research purposes. You can withdraw yourself from participating in this study at any time and for any reason.

You can find here a link to a document containing the participant information. Please read carefully before participating in the survey. And please do not hesitate to contact the researcher if you have any queries.

All the best,

Mr. Nawar Halabi

Postgraduate Researcher in Computer Science and Arabic Corpus Design in the Web And Internet Science group.

University of Southampton

If you would like to contact the researcher directly please do so on nh2f13@soton.ac.uk

By ticking the check box below you are agreeing to the following:

- I have understood the Participant information sheet (17/11/2015 version 1) and had the opportunity to ask questions about it.
- I agree to take part in this research project and agree for my data to be used for the purpose of this study.
- I understand that information collected about myself during my participation in this study will be stored on a password protected computer and that this information will only be used for the purpose of this study. All files containing any personal data will be made anonymous.

☐ Please tick this box to indicate that you consent to taking part in this survey.

Appendix D Survey Questions Screenshots

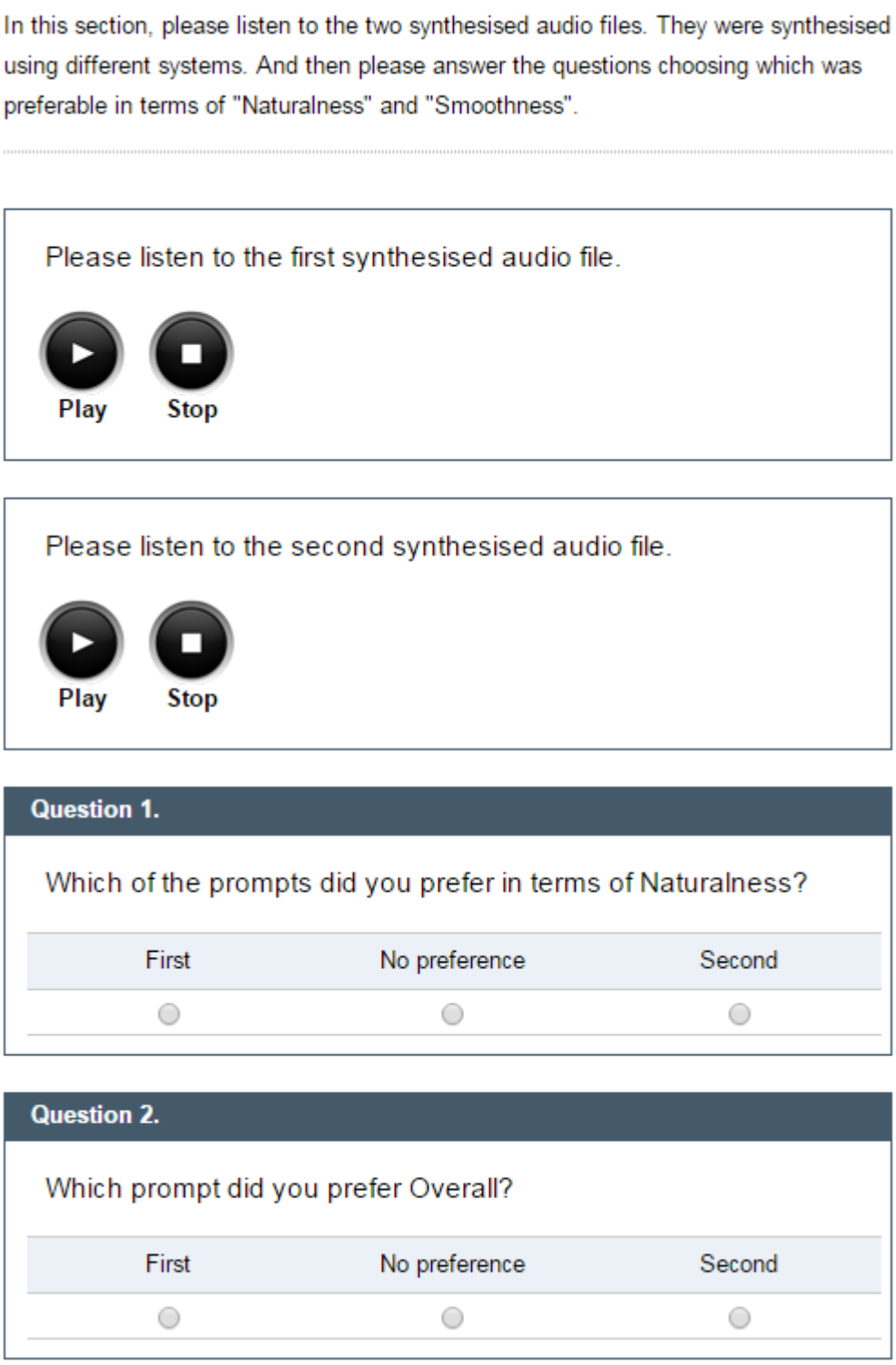

Figure 8-1. Screenshot of Preference question

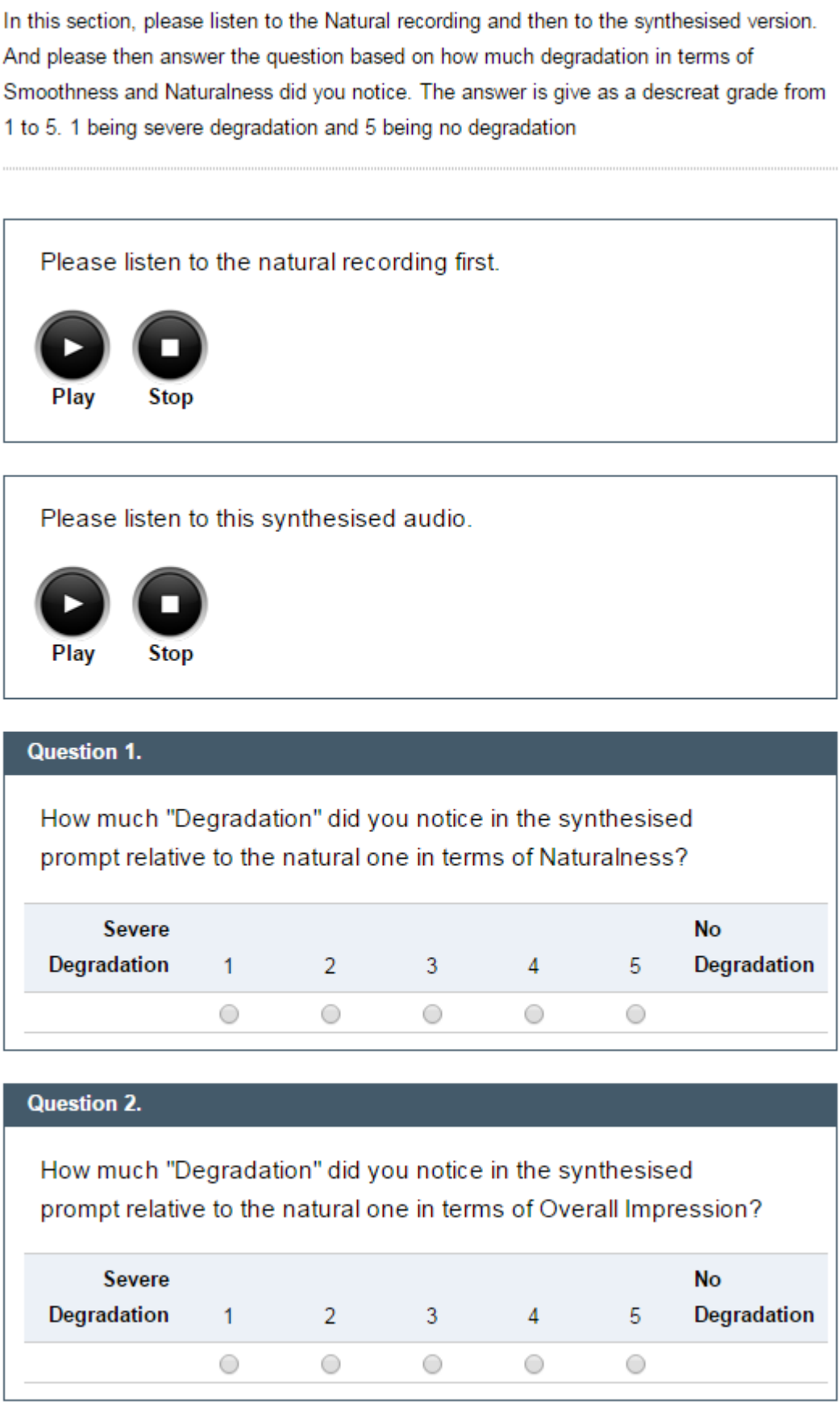

Figure 8-2. Screenshot of DMOS question

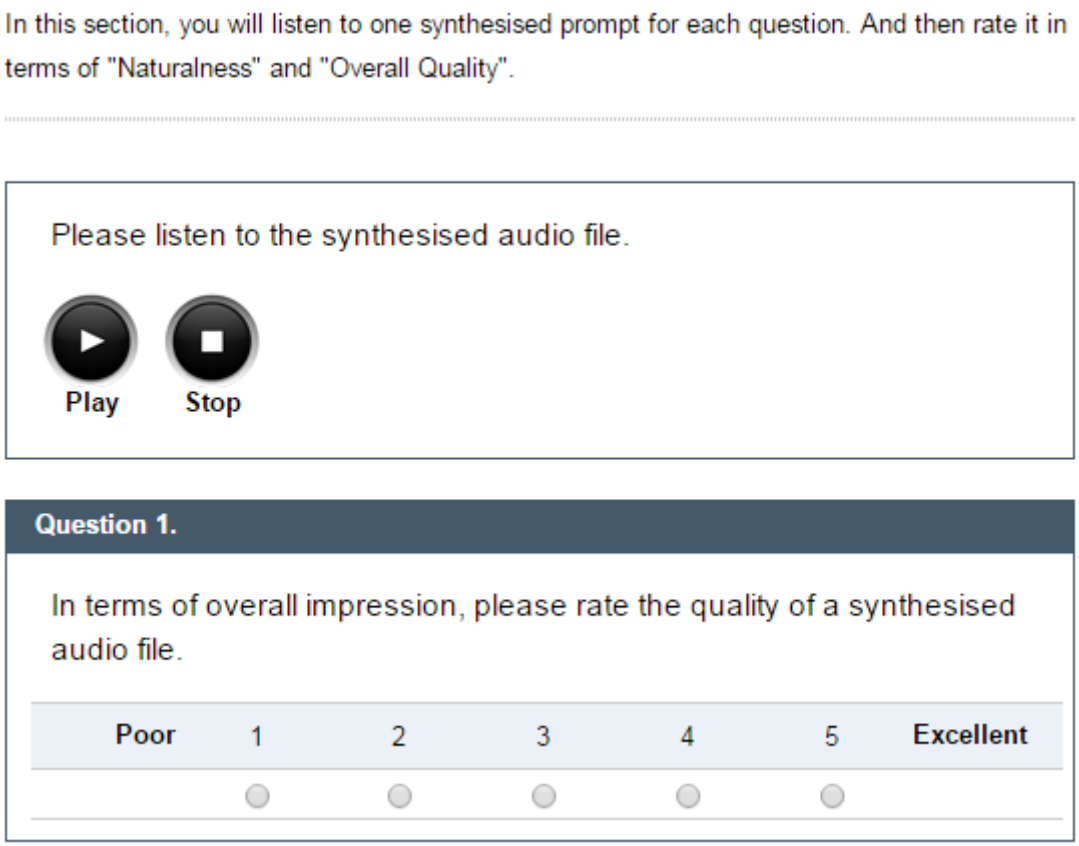

Figure 8-3. Screenshot of MOS question

Appendix E MOS and DMOS results

Table 8-2. DMOS and MOS test statistics

Test		Mean	STD	95% intervals +/–
System 1	DMOS for naturalness	3.5583	1.0901	0.1379
	DMOS for overall impression	3.5000	1.1475	0.1452
	MOS for overall impression	3.7071	1.0328	0.1316
System 2	DMOS for naturalness	3.8208	1.0199	0.1289
	DMOS for overall impression	3.7583	1.1398	0.1433
	MOS for overall impression	3.8494	1.0299	0.1303

Appendix F Instructions for the voice talent

1. Please do not omit vowels from end of words (Apocope).
2. Please do not apply “Tajweed”.
3. Please do not geminate consonants (“Shadda”) too hard. Please keep t subtle but still different from the non-geminate consonants.
4. There may be some unexpected Sokoons. Please pronounce those with the Sokoon even if wrong grammatically.
5. (-) and (“) mean short pause.
6. Consistency in Loudness. (Please listen to previous day’s recording for reference). Consistency in pitch. (Please listen to previous day’s recording for reference). Please speak around your comfortable pitch with no sudden changes in pitch. No excess emotion of any kind that might affect the pitch. Please create pitch boundaries around your comfortable pitch and try not to leave this range. And in the first iteration of recording we will be in a “declarative” or “pitch-descending” mood and the second as questions or “pitch-ascending”. (This instruction was discarded but could be included in case corpus was insufficient in size).
7. Consistency in speed (words per minute). (Please listen to previous day’s recording for reference). For this we recommend using another TTS (Text to Speech or a Speech Synthesiser) just to get an idea.
8. Consistency in pronunciation. When there is an error, the project leader will fix it and give feedback. Some errors are intentional.
9. No emphasis on specific words please. This causes a lot of spectral and intensity variation.
10. Consistency in recording environment (distance/angle from the microphone every time).
11. Ask about the voxforge recommendation for distortion as we cannot tolerate noise. We need to keep it within 0.5 dbs and export to a wav with 16 bit or higher pcm.

Bibliography

Acapela Group (2015) *Text to Speech and Voice Solutions* [online]. Available from: http://www.acapela-group.com/ (Accessed 3 September 2015).

Alderete, J. & Frisch, S. A. (2009) 'Phonotactic Learning without a Priori Constraints: A Connectionist Analysis of Arabic Cooccurrence Restrictions', in *Proceedings of the 48th annual meeting of the Chicago Linguistics Society*. 2009 Chicago, Illinois: . [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.387.5358 (Accessed 9 February 2015).

Alghamdi, M. (2003) 'KACST Arabic Phonetics Database', in *The Fifteenth International Congress of Phonetics Science, Barcelona*. 1 January 2003 Barcelona, Spain: . pp. 3109–3112. [online]. Available from: http://www.researchgate.net/publication/229049878_KACST_Arabic_Phonetics_Database (Accessed 3 November 2014).

Ali, H. K. & Ali, H. S. (2011) Epenthesis in English and Arabic. A Contrastive Study. *Journal of Tikrit University for the Humanities*. 18 (6), 648–660.

Almeman, K. et al. (2013) 'Multi dialect Arabic speech parallel corpora', in *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. [Online]. February 2013 Sharjah, United Arab Emirates: IEEE. pp. 1–6. [online]. Available from: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6487288 (Accessed 5 November 2014).

Almosallam, I. et al. (2013) 'SASSC: A Standard Arabic Single Speaker Corpus', in *proceedings of 8th ISCA Speech Synthesis Workshop*. 2013 Barcelona, Spain: .

Alsulaiman, M. et al. (2011) 'Building a Rich Arabic Speech Database', in *2011 Fifth Asia Modelling Symposium*. [Online]. May 2011 Manila, Philippines: IEEE. pp. 100–105. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5961222 (Accessed 2 January 2015).

Alsulaiman, M. et al. (2013) 'KSU Speech Database: Text Selection, Recording and Verification', in *2013 European Modelling Symposium*. [Online]. November 2013 Manchester, UK: IEEE. pp. 237–242. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6779852 (Accessed 31

December 2014).

Amith, J. (2012) 'Assessing agreement level between forced alignment models with data from endangered language documentation corpora', in *13th Annual Conference of the International Speech Communication Association*. 2012 Portland, Oregon: ISCA.

Ananthakrishnan, S. & Avenue, P. (2005) 'Automatic Diacritization of Arabic Transcripts for Automatic Speech Recognition', in *Proceedings of International Conference on Natural Language Processing (ICON)*. 2005 Kanpur, India: .

Van Bael, C. et al. (2007) Automatic phonetic transcription of large speech corpora. *Journal of Computer Speech & Language*. [Online] 21 (4), 652–668. [online]. Available from: http://www.sciencedirect.com/science/article/pii/S0885230807000228 (Accessed 30 May 2014).

Barros, M. & Möbius, B. (2011) 1436 sentence corpus made by manually constucting sentences. Each sentence focuses on a diphone and convays it in differnet contexts. *Human Language Technology. Challenges for Computer Science and Linguistics*. Lecture Notes in Computer Science. Zygmunt Vetulani (ed.). Vol. 6562. [Online]. Berlin, Heidelberg: Springer Berlin Heidelberg. [online]. Available from: http://www.springerlink.com/index/10.1007/978-3-642-20095-3 (Accessed 1 January 2015).

Benoît, C. et al. (1996) The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Journal of Speech Communication*. [Online] 18 (4), 381–392. [online]. Available from: http://www.sciencedirect.com/science/article/pii/016763939600026X (Accessed 29 September 2015).

Bertrán, A. P. (1999) Prosodic Typology: On the Dichotomy between Stress-Timed and Syllable-Timed Languages. *Journal of Language Design*. (2), 103–130.

Biadsy, F. et al. (2009) 'Spoken Arabic dialect identification using phonotactic modeling', in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*. 31 March 2009 Athens, Greece: Association for Computational Linguistics. pp. 53–61. [online]. Available from: http://dl.acm.org/citation.cfm?id=1621774.1621784 (Accessed 20 January 2015).

Biadsy, F. & Hirschberg, J. B. (2009) 'Using Prosody and Phonotactics in Arabic Dialect Identification', in *10th Annual Conference of the International Speech Communication Association*. 2009 Brighton, UK: International Speech Communication Association.

[online]. Available from: http://academiccommons.columbia.edu/catalog/ac:159968 (Accessed 20 January 2015).

Black, A. W. (2006) 'CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling', in *7th Annual Conference of the International Speech Communication Association*. 2006 Pittsburgh, Pennsylvania: International Speech Communication Association. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.71.8604 (Accessed 11 September 2013).

Black, A. W. et al. (2008) 'CMU Blizzard Challenge: Optimally using a large database for unit selection synthesis.', in *Blizzard Challenge 2008*. 2008 Brisbane, Australia: . [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.151.3270 (Accessed 19 August 2013).

Black, A. W. (2002) 'Perfect Synthesis For All Of The People All Of The Time', in *IEEE 2002 Workshop on Speech Synthesis*. 2002 Santa Monica, CA: . [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.8631 (Accessed 6 June 2014).

Black, A. W. (2003) 'Unit Selection and Emotional Speech', in *8th European Conference on Speech Communication and Technology*. 2003 Geneve, Switzerland: . pp. 1649–1652. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.8457 (Accessed 20 August 2013).

Black, P. E. (2005) *Greedy algorithms* [online]. Available from: http://www.nist.gov/dads/HTML/greedyalgo.html (Accessed 15 March 2015).

Boersma, P. & Weenink, D. (2015) *Praat Software* [online]. Available from: http://www.praat.org/ (Accessed 10 August 2015).

Du Bois, J. W. et al. (1992) Definition of intonaiton unit. *Discourse Transcription*. John W. Du Bois (ed.). Santa Barbara, California: Department of Linguistics, University of California.

Bonafonte, A. et al. (2008) 'Corpus and Voices for Catalan Speech Synthesis', in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2008 Valletta, Malta: . [online]. Available from: http://aclweb.org/anthology/L08-1517.

Boros, T. et al. (2014) 'RSS-TOBI - A Prosodically Enhanced Romanian Speech Corpus', in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2014 Reykjavik, Iceland: . pp. 316–320. [online]. Available from: http://www.lrec-

conf.org/proceedings/lrec2014/summaries/727.html.

Bozkurt, B. et al. (2002) 'Re-Defining Intonation From Selected Units For Non-Uniform Units Based Speech Synthesis', in *Proceedings of SPS-IEEE Benelux Signal Process. Symp*. 2002 Leuven, Belgium: . pp. 141–144. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.2935 (Accessed 7 January 2015).

Braunschweiler, N. (2006) 'The Prosodizer - Automatic Prosodic Annotations of Speech Synthesis Databases', in *Proceedings of Speech Prosody*. 2006 Dresden, Germany: . pp. PS5–PS27 – 76.

Brognaux, S. et al. (2012) 'Train & align: A new online tool for automatic phonetic alignment', in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. [Online]. December 2012 Miami, Florida: . pp. 416–421.

Bryman, A. (2006) *Social Research Methods*. 3rd edition. Oxford University Press.

Buchholz, S. & Latorre, J. (2011) 'Crowdsourcing Preference Tests, and How to Detect Cheating.', in *12th Annual Conference of the International Speech Communication Association*. 2011 Florence, Italy: ISCA. pp. 3053–3056. [online]. Available from: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2011.html#BuchholzL11.

Buckwalter, T. (2002) *Buckwalter Arabic Transliteration* [online]. Available from: http://www.qamus.org/transliteration.htm (Accessed 27 June 2013).

Cambridge University Press (2014) *Internation Phonetic Alphabet* [online]. Available from: https://www.internationalphoneticassociation.org/.

Cerňak, M. & Rusko, M. (2005) 'An evaluation of a synthetic speech using the PESQ measure', in *Proceedings of Forum Acusticum 2005*. 1 September 2005 Budapest, Hungary: . [online]. Available from: http://www.researchgate.net/publication/233843303_An_evaluation_of_a_synthetic_speech_using_the_PESQ_measure (Accessed 23 September 2015).

Chalamandaris, A. et al. (2013) 'The ILSP/INNOETICS Text-to-Speech System for the Blizzard Challenge 2013', in *The Blizzard Challenge 2013 workshop*. September 2013 Reykjavik, Iceland: .

Chevelu, J. et al. (2015) 'How to Compare TTS Systems: A New Subjective Evaluation

Methodology Focused on Differences', in *16th Annual Conference of the International Speech Communication Association*. 2015 Dresden, Germany: .

Clark, R. A. J., Richmond, K., et al. (2007) Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Journal of Speech Communication*. 49 (4), 317–330.

Clark, R. A. J., Podsiadło, M., et al. (2007) 'Statistical analysis of the Blizzard Challenge 2007 listening test results', in *6th ISCA Workshop on Speech Synthesis*. 2007 Bonn, Germany: . pp. 1–6.

CSTR (2016) *CSTR's subjective evaluation guide* [online]. Available from: https://wiki.inf.ed.ac.uk/twiki/pub/CSTR/Speak14To15/evaluation.pdf (Accessed 1 January 2016).

Dall, R. et al. (2014) 'Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation', in *Proceedings of Speech Prosody*. 2014 Dublin, Ireland: . [online]. Available from: http://www.research.ed.ac.uk/portal/en/publications/rating-naturalness-in-speech-synthesis-the-effect-of-style-and-expectation(a63232ab-d099-43b0-8200-00410b7544a9).html (Accessed 16 October 2015).

Education, C. R. & (2016) *Conference Rankings* [online]. Available from: http://portal.core.edu.au/conf-ranks/ (Accessed 6 May 2016).

Elshafei, M. A. (1991) Toward an Arabic Text-to-speech System. *The Arabian Journal for Science and Engineering*. 16 (4b), 565–583.

Essa, O. (1998) 'Using Prosody in Automatic Segmentation of Speech', in *Proceedings of the 36th annual Southeast regional conference*. 1998 New York, New York, USA: . pp. 44–49. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.4359 (Accessed 11 June 2014).

Field, A. (2009) *Discovering Statistics using SPSS*. 3rd edition. SAGE Publications Ltd. [online]. Available from: http://doi.wiley.com/10.1002/bjs.7040.

François, H. & Boëffard, O. (2002) 'The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database.', in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. 2002 Las Palmas, Canary Islands, Spain: European Language Resources Association (ELRA). [online]. Available from: http://aclweb.org/anthology/L02-1265.

Gadoua, A. H. (2000) Consonant Clusters In Quranic Arabic. *Cahiers Linguistiques d'Ottawa/Ottawa Papers in Linguistics (Journal)*. 2859–85.

Ghahramani, Z. (2001) An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*. 1 (25), 9–42. [online]. Available from: http://dl.acm.org/citation.cfm?id=505741.505743 (Accessed 12 August 2015).

Gilbert, L. (2015) *Expert Opinion*.

Glass, R. L. et al. (2004) An analysis of research in computing disciplines. *Communications of the ACM (Journal)*. [Online] 47 (6), 89–94. [online]. Available from: http://dl.acm.org/ft_gateway.cfm?id=990686&type=html (Accessed 2 February 2016).

Gorman, K. et al. (2011) Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics [online]. Available from: http://jcaa.caa-aca.ca/index.php/jcaa/article/view/2476 (Accessed 7 December 2014). 39 (3) p.192–193. [online]. Available from: http://jcaa.caa-aca.ca/index.php/jcaa/article/view/2476 (Accessed 7 December 2014).

Green, L. (2016) *Expert Opinion*.

Habash, N. (2010) Introduction to Arabic Natural Language Processing. *Synthesis Lectures on Human Language Technologies*. [Online] 3 (1), 1–187. [online]. Available from: http://www.morganclaypool.com/doi/abs/10.2200/S00277ED1V01Y201008HLT010 (Accessed 13 January 2015).

Habash, N. & Rambow, O. (2007) 'Arabic Diacritization through Full Morphological Tagging', in *The Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2007 Rochester, NY, USA: Association for Computational Linguistics. pp. 53–56. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.6049 (Accessed 11 June 2013).

Habash, N. & Rambow, O. (2012) *MADA+TOKAN Manual*.

Haertel, R. A. et al. (2010) 'Automatic diacritization for low-resource languages using a hybrid word and consonant CMM', in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2 June 2010 Stroudsburg, PA, USA: Association for Computational Linguistics.

pp. 519–527. [online]. Available from: http://dl.acm.org/citation.cfm?id=1857999.1858075 (Accessed 13 June 2013).

Halabi, N. (2015) *MPhil Code and Results* [online]. Available from: https://www.dropbox.com/s/0tjdj4coy1n094g/MPHIL FILES.rar?dl=0.

Halpern, J. (2009) 'Word Stress and Vowel Neutralization in Modern Standard Arabic', in *2nd International Conference on Arabic Language Resources and Tools*. 2009 Cairo, Egypt: . pp. 1–7.

Hermansky, H. (1990) Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*. 87 (4), 1738–1752. [online]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2341679 (Accessed 8 June 2016).

Hirose, K. & Tao, J. (eds.) (2015) t-test for MOS. *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Prosody, Phonology and Phonetics. [Online]. Berlin, Heidelberg: Springer. [online]. Available from: http://link.springer.com/10.1007/978-3-662-45258-5.

Hoffmann, S. & Pfister, B. (2010) 'Fully automatic segmentation for prosodic speech corpora.', in *11th Annual Conference of the International Speech Communication Association*. 2010 Makuhari, Chiba, Japan: . pp. 1389–1392.

Holz, H. J. et al. (2006) Research methods in computing. *The ACM Special Interest Group on Computer Science Education Bulletin*. [Online] 38 (4), 96. [online]. Available from: http://dl.acm.org/citation.cfm?id=1189136.1189180 (Accessed 9 February 2016).

Hönig, F. et al. (2005) 'Revising Perceptual Linear Prediction (PLP)', in *9th European Conference on Speech Communication and Technology*. 2005 Lisbon, Portugal: . pp. 2997–3000.

Hosom, J.-P. (2009) Speaker-Independent Phoneme Alignment Using Transition-Dependent States. *Journal of Speech communication*. [Online] 51 (4), 352–368. [online]. Available from: http://dl.acm.org/citation.cfm?id=1507768.1507931 (Accessed 30 September 2013).

Hu, Q. et al. (2014) 'An investigation of the application of dynamic sinusoidal models to statistical parametric speech synthesis', in *Interspeech 2014*. pp. 780–784.

Huckvale, M. et al. (2012) Performance-Based Measurement of Speech Quality with an Audio Proof-Reading Task. *Journal of the Audio Engineering Society*. 60 (6), . [online]. Available

from: https://www.researchgate.net/publication/264880199_Performance-Based_Measurement_of_Speech_Quality_with_an_Audio_Proof-Reading_Task (Accessed 28 January 2016).

Inai, T. et al. (2015) 'Sub-Band Text-to-Speech Combining Sample-Based Spectrum with Statistically Generated Spectrum', in *16th Annual Conference of the International Speech Communication Association*. 2015 Dresden, Germany: .

Indumathi, A. & Chandra, D. E. (2012) Survey on Speech Synthesis. *Signal Processing: An International Journal (SPIJ)*. 6 (5), 140–145.

ITU-T (2015) *Method for the subjective assessment of intermediate quality level of audio systems (BS.1534-3) (Standard)*.

ITU-T (1996) *Methods for objective and subjective assessment of quality (Rec P.800) (A Standard)*.

Jakovljević, N. et al. (2012) Automatic Phonetic Segmentation for a Speech Corpus of Hebrew. *INFOTEH-JAHORINA*. 11742–745.

Jarifi, S. et al. (2008) A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. *Journal of Speech Communication*. [Online] 50 (1), 67–80. [online]. Available from: http://www.sciencedirect.com/science/article/pii/S0167639307001215 (Accessed 21 August 2014).

Al Jazeera (2015) *Aljazeera Learn* [online]. Available from: http://learning.aljazeera.net/arabic (Accessed 15 February 2015).

Jesson, J. et al. (2011) For traditional and systematic literature reviews. *Doing Your Literature Review: Traditional and Systematic Techniques*. SAGE Publications. [online]. Available from: https://books.google.com/books?hl=en&lr=&id=LUhdBAAAQBAJ&pgis=1 (Accessed 7 April 2016).

de Jong, K. & Zawaydeh, B. A. (1999) Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics*. [Online] 27 (1), 3–22. [online]. Available from: http://www.sciencedirect.com/science/article/pii/S0095447098900882 (Accessed 4 January 2015).

Jurafsky, D. & Martin, J. H. (2009) *Speech and Language Processing*. Second Edi. New Jersey:

Pearson Education Inc.

Kain, E. et al. (2007) 'Spectral control in concatenative speech synthesis', in *6th ISCA Workshop on Speech Synthesis*. 2007 Bonn, Germany: . pp. 11–16. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.384.9067 (Accessed 7 January 2015).

Karnjanadecha, M. & Zahorian, S. A. (2012) 'Toward an Optimum Feature Set and HMM Model Parameters for Automatic Phonetic Alignment of Spontaneous Speech.', in *INTERSPEECH*. 2012 ISCA. [online]. Available from: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html#KarnjanadechaZ12.

Karrer, T. et al. (2006) 'PhaVoRIT: A Phase Vocoder for Real-Time Interactive Time-Stretching', in *Proceedings of the International Computer Music Conference*. 2006 New Orleans, USA: . pp. 708–715.

Kato, T. et al. (2011) 'Large-Scale Subjective Evaluations of Speech Rate Control Methods for HMM-Based Speech Synthesizers.', in *12th Annual Conference of the International Speech Communication Association*. 2011 Florence, Italy: . pp. 1845–1848.

Kawai, H. et al. (2000) 'A Design Method of Speech Corpus for Text-To-Speech Synthesis Taking Account of Prosody', in *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*. 2000 Beijing, China: . pp. 420–425.

Kawanami, H. et al. (2002) 'Designing speech database with prosodic variety for expressive TTS system.', in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. 2002 Las Palmas, Canary Islands, Spain: European Language Resources Association (ELRA). pp. 2039–2042. [online]. Available from: http://aclweb.org/anthology/L02-1337.

Kelly, A. C. et al. (2006) 'Speech Technology for Minority Languages: the Case of Irish (Gaelic)', in *7th Annual Conference of the International Speech Communication Association*. 2006 Pittsburgh, Pennsylvania: . [online]. Available from: http://www.tara.tcd.ie/handle/2262/39404 (Accessed 1 January 2015).

Kenworthy, J. (1987) *Teaching English Pronunciation*. New York, New York, USA: Longman Inc.

Kim, S. et al. (2006) HMM-based Korean speech synthesis system for hand-held devices. *IEEE Transactions on Consumer Electronics*. [Online] 52 (4), 1384–1390. [online]. Available

from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4050071 (Accessed 19 August 2013).

King, S. (2013) Measuring a decade of progress in Text-to-Speech. *Loquens (Journal)*. [Online] 1 (1), e006. [online]. Available from: http://loquens.revistas.csic.es/index.php/loquens/article/view/6/12 (Accessed 3 November 2014).

Kominek, J. & Black, A. W. (2003) *CMU Arctic Databases for Speech Synthesis (Manuel)*. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.8827 (Accessed 14 August 2013). [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.8827 (Accessed 14 August 2013).

Kominek, J. & Black, A. W. (2014) 'The CMU Arctic Speech Database', in *5th ISCA Speech Synthesis Workshop*. 2014 Pittsburgh, Pennsylvania: . pp. 223–224.

Krul, R. et al. (2007) 'Adaptive Database Reduction for Domain Specific Speech Synthesis', in *6th ISCA Workshop on Speech Synthesis*. 2007 Bonn, Germany: . pp. 217–222.

Kumar, R. et al. (2007) 'Building a Better Indian English Voice using "More Data"', in *6th ISCA Workshop on Speech Synthesis*. 2007 Bonn, Germany: International Speech Communication Association. pp. 90–94. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.62.4885 (Accessed 20 August 2013).

Lamere, P. et al. (2003) 'The CMU SPHINX-4 Speech Recognition System', in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2003 Hong Kong: . [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=F39A30BB1E9AF47B64D45B2AB0E3C27C?doi=10.1.1.406.8962 (Accessed 11 June 2015).

Latorre, J. et al. (2014) 'Speech intonation for TTS: Study on evaluation methodology', in *15th Annual Conference of the International Speech Communication Association*. 14 September 2014 Singapore: . [online]. Available from: http://www.researchgate.net/publication/264129463_Speech_intonation_for_TTS_Study_on_evaluation_methodology (Accessed 23 October 2015).

Laufer, A. & Baer, T. (1988) The emphatic and pharyngeal sounds in Hebrew and in Arabic. *Journal of Language and speech*. 31 (Pt 2)181–205. [online]. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/3256772 (Accessed 13 February 2015).

Lenzo, K. A. & Black, A. W. (2000) 'Diphone Collection and Synthesis', in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. [Online]. 2000 Beijing, China: . [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.2556 (Accessed 20 August 2013).

Lindstrom, A. et al. (1996) 'Prosody Generation in Text-to-Speech Conversion Using Dependency Graphs', in *Proceedings of Prosody Generation in Text-to-Speech Conversion Using Dependency Graphs*. 1996 Philadelphia, USA: . pp. 1341–1344.

Lowry, R. (2007) *Concepts & Applications of Inferential Statistics*. Online Boo. Vassar College. [online]. Available from: https://scout.wisc.edu/report/nsdl/met/2003/0214.

Lu, H. et al. (2011) 'Building HMM based unit-selection speech synthesis system using synthetic speech naturalness evaluation score', in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [Online]. May 2011 Prague, Czech Republic: IEEE. pp. 5352–5355. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5947567 (Accessed 19 August 2013).

Lu, H. et al. (2015) 'Pruning Redundant Synthesis Units Based on Static and Delta Unit Appearance Frequency', in *16th Annual Conference of the International Speech Communication Association*. 2015 Dresden, Germany: . pp. 269–273.

Maamouri, M. et al. (2005) *Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis)* [online]. Available from: https://catalog.ldc.upenn.edu/LDC2005T20.

Maia, R. et al. (2007) 'An Excitation Model for HMM-Based Speech Synthesis Based on Residual Modeling', in *6th ISCA Workshop on Speech Synthesis*. 1 August 2007 Bonn, Germany: International Speech Communication Association. pp. 131–136. [online]. Available from: http://library.naist.jp/dspace/handle/10061/8269 (Accessed 20 August 2013).

Malfrère, F. et al. (1998) 'Automatic Prosody Generation Using Suprasegmental Unit Selection', in *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*. 1998 Jenolan Caves House, Blue Mountains, NSW, Australia: . pp. 323–328. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.9166 (Accessed 5 January 2015).

Malfrère, F. et al. (2003) Phonetic alignment: speech synthesis-based vs. Viterbi-based. *Journal of Speech Communication*. [Online] 40 (4), 503–515. [online]. Available from: http://www.sciencedirect.com/science/article/pii/S0167639302001310 (Accessed 29 May 2014).

Matoušek, J. R. & Psutka, J. (2001) 'Design of Speech Corpus for Text-to-Speech Synthesis', in *7th European Conference on Speech Communication and Technology*. 2001 Aalborg, Denmark: . pp. 2047–2050. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.9719 (Accessed 1 January 2015).

Matoušek, J. R. & Romportl, J. (2007a) 'Recording and annotation of speech corpus for Czech unit selection speech synthesis', in *TSD'07 Proceedings of the 10th international conference on Text, speech and dialogue*. 3 September 2007 Springer-Verlag. pp. 326–333. [online]. Available from: http://dl.acm.org/citation.cfm?id=1776334.1776380 (Accessed 19 August 2013).

Matoušek, J. R. & Romportl, J. (2007b) *Text, Speech and Dialogue*. Lecture Notes in Computer Science. Václav Matoušek & Pavel Mautner (eds.). Vol. 4629. [Online]. Berlin, Heidelberg: Springer. [online]. Available from: http://www.springerlink.com/index/10.1007/978-3-540-74628-7 (Accessed 31 December 2014).

Michael, R. S. (2001) *Crosstabulation & Chi Square* [online]. Available from: http://www.indiana.edu/~educy520/sec5982/week_12/chi_sq_summary011020.pdf (Accessed 28 January 2016).

Mohammadi, A. et al. (2014) Eigenvoice Speaker Adaptation with Minimal Data for Statistical Speech Synthesis Systems Using a MAP Approach and Nearest-Neighbors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. [Online] 22 (12), 2146–2157. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6918404 (Accessed 20 January 2016).

Möller, S. et al. (2010) 'Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems', in *11th Annual Conference of the International Speech Communication Association*. 2010 Makuhari, Chiba, Japan: ISCA. pp. 1325–1328.

Moore, J. E. et al. (1997) robotic voice for blind. *Foundations of Rehabilitation Counseling with Persons Who Are Blind or Visually Impaired*. American Foundation for the Blind.

Moreno, P. J. et al. (1998) 'A recursive algorithm for the forced alignment of very long audio segments.', in *5th International Conference on Spoken Language Processing*. 1998 Sydney, Australia: ISCA. [online]. Available from: http://dblp.uni-trier.de/db/conf/interspeech/icslp1998.html#MorenoJTG98.

Mporas, I. et al. (2009) 'Using Hybrid HMM-Based Speech Segmentation to Improve Synthetic Speech Quality', in *13th Panhellenic Conference on Informatics*. [Online]. September 2009 Corfu Island, Greece: . pp. 118–122.

Müller, C. (ed.) (2007) Habituation. *Speaker Classification II*. Lecture Notes in Computer Science. Vol. 4441. [Online]. Berlin, Heidelberg: Springer. [online]. Available from: http://link.springer.com/10.1007/978-3-540-74122-0.

Multiple Owners (2016) *ICU Project* [online]. Available from: http://site.icu-project.org/ (Accessed 12 April 2016).

Murphy, K. P. (2012) *Machine Learning. A Probabilistic Perspective*. 1st edition. Cambridge, Massachusetts: MIT Press.

Muthukumar, P. K. & Black, A. W. (2014) A Deep Learning Approach to Data-driven Parameterizations for Statistical Parametric Speech Synthesis. *arXiv*. [online]. Available from: http://arxiv.org/abs/1409.8558 (Accessed 28 May 2015).

Newman, D. (1986) The Phonetics of arabic. *Journal of the American Oriental Society*. 1–6.

Niekerk, D. R. van (2014) *Tone Realisation For Speech Synthesis of Yoruba*. Vaal Triangle Campus of the North-West University.

Niekerk, D. R. van & Barnard, E. (2009) 'Phonetic alignement for speech synthesis in under-resourced languages', in *10th Annual Conference of the International Speech Communication Association*. 2009 Brighton, UK: ISCA. pp. 880–883. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.384.1749 (Accessed 10 September 2014).

Nunnally, J. C. & Bernstein, I. H. (1994) for 100 scale points or 5. *Psychometric Theory*. 3rd editio. McGraw-Hill.

Ocean, S. (2016) *Arabic Speech Database* [online]. Available from: http://www.speechocean.com/ (Accessed 4 February 2016).

Oliveira, L. C. et al. (2008) 'Methodologies for Designing and Recording Speech Databases for

Corpus Based Synthesis', in *6th edition of the Language Resources and Evaluation Conference (LREC)*. 2008 Marrakech, Morocco: . [online]. Available from: http://aclweb.org/anthology/L08-1484.

Pasha, A. et al. (2014) 'MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic', in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. May 2014 Reykjavik, Iceland: European Language Resources Association (ELRA).

Peddinti, V. & Prahallad, K. (2011) 'Exploiting Phone-Class Specific Landmarks for Refinement of Segment Boundaries in TTS Databases.', in *12th Annual Conference of the International Speech Communication Association*. 2011 Florence, Italy: ISCA. pp. 429–432. [online]. Available from: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2011.html#PeddintiP11.

Pereira, F. C. N. & Riley, M. D. (1996) Speech Recognition by Composition of Weighted Finite Automata. arXiv [online]. Available from: http://arxiv.org/abs/cmp-lg/9603001. p.24. [online]. Available from: http://arxiv.org/abs/cmp-lg/9603001.

Polkosky, M. D. & James, R. L. (2003) Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology*. 161–182. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.535.5372 (Accessed 16 October 2015).

Prahallad, K. (2010) *Automatic building of synthetic voices from audio books*. Carnegie Mellon University. [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.169.7981 (Accessed 14 October 2013).

Prahallad, K. et al. (2007) 'Automatic building of synthetic voices from large multi-paragraph speech databases', in *8th Annual Conference of the International Speech Communication Association*. [Online]. 2007 pp. 2901–2904.

Qian, Y. et al. (2008) 'HMM-Based Mixed-Language (Mandarin-English) Speech Synthesis', in *2008 6th International Symposium on Chinese Spoken Language Processing*. [Online]. December 2008 Kunming, China: IEEE. pp. 1–4. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4730269 (Accessed 20 August 2013).

Radermacher, H. L. (2006) *Participatory Action Research With People With Disabilities: Exploring Experiences Of Participation*. Victoria University. [online]. Available from: http://vuir.vu.edu.au/484/3/484whole.pdf (Accessed 7 April 2016).

Raitio, T. et al. (2015) 'Phase Perception of the Glottal Excitation of Vocoded Speech', in *16th Annual Conference of the International Speech Communication Association*. 2015 Dresden, Germany: . pp. 254–258.

Rodero, E. (2012) 'A comparative analysis of speech rate and perception in radio bulletins', in *Proceedings of Text & Talk*. 2012 pp. 391–411.

Romportl, J. (2010) 'Automatic Prosodic Phrase Annotation in a Corpus for Speech Synthesis', in *Proceedings of Speech Prosody*. 11 May 2010 Chicago, Illinois: . pp. 1–4. [online]. Available from: http://www.kky.zcu.cz/cs/publications/JanRomportl_2010_AutomaticProsodic (Accessed 10 March 2015).

Rothauser, E. H. et al. (1969) IEEE Recommnded Pratice for Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*. [Online] 17 (3), 225–246. [online]. Available from: http://www.researchgate.net/publication/265414192_IEEE_Recommended_Practice_for_Speech_Quality_Measurements (Accessed 29 September 2015).

Sainz, I. et al. (2008) 'Subjective Evaluation of an Emotional Speech Database for Basque', in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2008 Marrakech, Morocco: . [online]. Available from: http://www.lrec-conf.org/proceedings/lrec2008/summaries/437.html.

Sainz, I. et al. (2012) 'Versatile Speech Databases for High Quality Synthesis for Basque', in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*. May 2012 Istanbul, Turkey: European Language Resources Association (ELRA).

Scimago Lab (2016) *The SCImago Journal & Country Rank* [online]. Available from: http://www.scimagojr.com/index.php (Accessed 6 May 2016).

Selouani, S. A. & Caelen, J. (1998) 'Arabic phonetic features recognition using modular connectionist architectures', in *Proceedings 1998 IEEE 4th Workshop Interactive Voice Technology for Telecommunications Applications.* [Online]. 1998 Torino, Italy: IEEE. pp. 155–160. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=727712 (Accessed 3

November 2014).

Shannon, M. & Byrne, W. (2009) 'Autoregressive HMMs for speech synthesis', in *10th Annual Conference of the International Speech Communication Association*. 2009 Brighton, UK: ISCA (International Speech Communication Association).

Silverman, K. (2012) *Video Lecture: Speech Synthesis*. [online]. Available from: https://www.youtube.com/watch?v=7mjh0PSUv0M&spfreload=10. [online]. Available from: https://www.youtube.com/watch?v=7mjh0PSUv0M&spfreload=10.

Stolcke, A. et al. (2014) 'Highly Accurate Phonetic Segmentation Using Boundary Correction Models and System Fusion', in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2014 Florence, Italy: . [online]. Available from: http://research.microsoft.com/apps/pubs/default.aspx?id=209007.

Strom, V. et al. (2006) 'Expressive Prosody for Unit-selection Speech Synthesis', in *7th Annual Conference of the International Speech Communication Association*. 2006 Pittsburgh, Pennsylvania: .

SynSIG (2016) *Blizzard Challenge* [online]. Available from: http://www.synsig.org/index.php/Blizzard_Challenge (Accessed 14 January 2016).

Szaszák, G. et al. (2015) 'Using Automatic Stress Extraction from Audio for Improved Prosody Modelling in Speech Synthesis', in *16th Annual Conference of the International Speech Communication Association*. 2015 Dresden, Germany: . pp. 2227–2231.

Tao, J. et al. (2008) 'Design of Speech Corpus for Mandarin Text to Speech', in *The Blizzard Challenge 2008 workshop*. 2008 Brisbane, Australia: .

Taylor, P. (2009) *Text-To-Speech Synthesis*. Cambridge University Press.

Tench, P. (2015) *Consonants* [online]. Available from: http://www.cardiff.ac.uk/encap/contactsandpeople/academic/tench/consonants.html (Accessed 15 March 2015).

Thelwall, R. & Sa'Adeddin, M. A. (2009) Arabic. *Journal of the International Phonetic Association*. [Online] 20 (02), 37. [online]. Available from: http://journals.cambridge.org/abstract_S0025100300004266 (Accessed 3 November 2014).

Torres, H. M. & Gurlekian, J. A. (2008) Acoustic speech unit segmentation for concatenative synthesis. *Journal of Computer Speech & Language*. [Online] 22 (2), 196–206. [online].

Available from: http://www.sciencedirect.com/science/article/pii/S0885230807000484 (Accessed 15 August 2014).

Umbert, M. et al. (2006a) 'Spanish Synthesis Corpora', in *Proceedings of the International Conference of Language Resources and Evaluation (LREC)*. 2006 Genoa, Italy: . pp. 2102–2105.

Umbert, M. et al. (2006b) 'Spanish Synthesis Corpora', in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2006 Genoa, Italy: . pp. 2102–2105.

University of London (2016) *Research Ethics Guidebook* [online]. Available from: http://www.ethicsguidebook.ac.uk/ (Accessed 7 April 2016).

University of Southampton (2016) *iSurvey* [online]. Available from: isurvey.soton.ac.uk (Accessed 29 January 2016).

Vetulani, Z. (2009) 'Human Language Technology. Challenges for Computer Science and Linguistics', in *4th Language and Technology Conference, LTC 2009*. [Online]. 2009 Roznan, Poland: . pp. 42–43. [online]. Available from: http://books.google.com/books?id=l5FZxDY3yi0C.

van Vuuren, V. Z. et al. (2013) 'A dynamic programming framework for neural network-based automatic speech segmentation.', in Frédéric Bimbot et al. (eds.) *INTERSPEECH*. 2013 ISCA. pp. 2287–2291. [online]. Available from: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2013.html#VuurenBN13.

Watson, J. C. E. (2007) *The Phonology and Morphology of Arabic*. Oxford: Oxford University Press. [online]. Available from: http://books.google.com/books?hl=de&lr=&id=4RDIoDAF1e8C&pgis=1 (Accessed 3 November 2014).

Wester, M. et al. (2015) 'Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations', in *16th Annual Conference of the International Speech Communication Association*. 2015 Dresden, Germany: . [online]. Available from: http://www.research.ed.ac.uk/portal/en/publications/are-we-using-enough-listeners-no-an-empiricallysupported-critique-of-interspeech-2014-tts-evaluations(cf7b29ea-3f77-4785-96a3-2b5df768284c).html (Accessed 15 October 2015).

Wolters, M. K. et al. (2010) 'Evaluating Speech Synthesis Intelligibility using Amazon

Mechanical Turk', in *Proc. of 7th Speech Synthesis Workshop (SSW7)*. 2010 Kyoto, Japan: . pp. 136–141.

Xijun Ma et al. (2004) 'Probability based prosody model for unit selection', in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. [Online]. 2004 Montreal, Quebec, Canada: IEEE. pp. I – 649–652. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1326069 (Accessed 21 October 2014).

Yi, J. R.-W. (2003) *Corpus-based unit selection for natural-sounding speech synthesis*. Massachusetts Institute of Technology. [online]. Available from: http://dspace.mit.edu/handle/1721.1/16944 (Accessed 30 December 2014).

Young, S. et al. (1997) *The HTK book*. Vol. 2. Cambridge, Massachusetts: Cambridge University.

Yuan, J. et al. (2013) 'Automatic phonetic segmentation using boundary models.', in Frédéric Bimbot et al. (eds.) *14th Annual Conference of the International Speech Communication Association*. 2013 Lyon, France: ISCA. pp. 2306–2310. [online]. Available from: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2013.html#YuanRLSMW13.

Zen, H. et al. (2009) 'Recent development of the HMM-based speech synthesis system (HTS)', in *Proc. 2009 Asia-Pacific Signal and Information Processing Association*. 2009 Sapporo, Japan: . [online]. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.7846 (Accessed 19 August 2013).

Zen, H. et al. (2012) Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization. *IEEE Transactions on Audio, Speech, and Language Processing*. [Online] 20 (6), 1713–1724. [online]. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6148263 (Accessed 26 January 2016).

Zen, H. et al. (2013) 'Statistical Parametric Speech Synthesis Using Deep Neural Networks', in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2013 Vancouver, BC, Canada: . pp. 7962–7966.

Zen, H. et al. (2007) 'The HMM-based speech synthesis system (HTS) version 2.0', in *6th ISCA Workshop on Speech Synthesis*. 2007 Bonn, Germany: International Speech Communication Association. pp. 294–299. [online]. Available from: http://www.researchgate.net/publication/228365542_The_HMM-

based_speech_synthesis_system_(HTS)_version_2.0 (Accessed 20 August 2013).

Zerrouki, T. (2014) *Mishkal Diacritiser* [online]. Available from: http://www.tahadz.com/mishkal/.

Zue, V. W. & Seneff, S. (1996) 'Transcription and Alignment of the TIMIT Database', in *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*. [Online]. Elsevier. pp. 515–525. [online]. Available from: http://linkinghub.elsevier.com/retrieve/pii/B9780444816078500888.