Phonetic Inventory for an Arabic Speech Corpus

Nawar Halabi, Mike Wald
School of Electronics and Computer Science
University of Southampton
E-mail: nh2f13@ecs.soton.ac.uk, mw@ecs.soton.ac.uk

Abstract

Corpus design for speech synthesis is a well-researched topic in languages such as English compared to Modern Standard Arabic, and there is a tendency to focus on methods to automatically generate the orthographic transcript to be recorded (usually greedy methods). In this work, a study of Modern Standard Arabic (MSA) phonetics and phonology is conducted in order to create criteria for a greedy meth-od to create a speech corpus transcript for recording. The size of the dataset is reduced a number of times using these optimisation methods with different parameters to yield a much smaller dataset with identical phonetic coverage than before the reduction, and this output transcript is chosen for recording. This is part of a larger work to create a completely annotated and segmented speech corpus for MSA.

Keywords: Phonology, Corpus Design, Corpus Evaluation

1. Introduction

Speech corpora for speech synthesisers (namely Unit Selection synthesisers) in different languages have increased in number and vary in size. The size of these corpora is rarely justified in the literature as there is no consensus on what is the minimum length for a speech corpus required to build a synthesiser with a natural voice (Bonafonte et al., 2008; B. Bozkurt, Dutoit, Prudon, D'Alessandro, & Pagel, 2002; Clark, Richmond, & King, 2007; Kominek & Black, 2003; Tao, Liu, Zhang, & Jia, 2008). Bonafonte et al. 2008 produced 10 hours each of male and female Catalan speech targeted at speech synthesis. They claim, in their review, that they have found speech corpora ranging from 3 to 12 hours of speech. Oliveira et al., 2008 recorded 13 hours of speech from each of their 4 subjects, also targeting speech synthesis. They claim that unit selection usually requires 3 to 10 hours of speech, without taking a specific language into consideration. Tao et al., 2008 used 7 hours of recording for the Blizzard Challenge 2008 to build a synthesiser for Mandarin. Some speech corpora are shorter like the "awb" voice produced by Kominek and Black (Kominek & Black, 2003) using the ARCTIC database which consists of 1.4 hours of speech. There is even an attempt to build a speech synthesiser using only an hour of speech for Portuguese (Parlikar & Black, 2012). Many more examples of corpora exist with different lengths which do not cater for redundancy. But what matters more than the speech recording length is the transcript that has been chosen to produce it. The transcript needs to have good phonetic and prosodic coverage of the target language such as the examples of Modern Standard Arabic (MSA) used in this work.

2. Phonetic and Prosodic Coverage

Phonetic coverage is the ability of the speech corpus to be used to generate as natural synthetic speech that does not occur in the corpus. The main speech synthesis methods that determine this coverage in this work are concatenative, unit selection synthesis methods (A. W. Black, 2002). In quantifiable terms, the phonetic coverage of a speech corpus transcript is the subset of all target units' set (phones, diphones or triphones…) with a frequency of occurrence associated with each unit in this subset. In this work the target units' set is called the "phonetic vocabulary".

Prosodic coverage loosely means the extent to which a speech corpus could be used to generate new sentences that sound natural with changes in pitch and intensity that resemble human speech with pauses positioned correctly to convey the meaning. A more rigorous definition of prosodic coverage is the ability to synthesise speech from the speech corpus with as many possible prosodic states (declarative intonation or interrogative intonation for example). However, to define prosodic coverage more rigorously, one should define prosody. Prosody is defined here as the changes in pitch (intonation), rhythm, pausing (sometimes classed under rhythm) and stress (which relates to intensity and intonation) to convey the speaker's state or the features of the speaker's utterance (Fernández & Cairns, 2011).

It is important to define the scope of the corpus to be produced. The terms "domain specific" and "open domain" (sometimes called "general purpose"). This work falls under the "open domain" category.

Note that in this paper, the International Phonetic Alphabet (IPA) along with the Arabic letters is used to represent phones if not otherwise stated.

3. MSA Phonetics

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

3.1 Stress (Lexical Stress)

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 included in many works reviewed. It was used as a feature of segments in speech corpus for both optimising the phonetic coverage before the recording and to help with choosing the best unit for concatenation in speech synthesis (Barros & Möbius, 2011; Kominek & Black, 2014). In the case of coverage optimisation, stress is used usually as a feature of vowel phones as stress mainly affects vowels (Biadsy, Hirschberg, & Habash, 2009; de Jong & Zawaydeh, 1999) (pitch is altered and vowel length is changed). So a stressed vowel is considered a different phone compared to 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 algorithm for determining stressed syllables in a text transcript is based on a set of rules presented by Halpern (Halpern, 2009) in a recent stress study where the target language is MSA. Halpern illustrates 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. The steps taken being a series of conditional statements as follows:

- If last syllable is super-heavy, stress falls on it, or else
- If word is monosyllabic, it is stressed, or else
- If word is disyllabic, the penultimate is stressed, else
- If word has more than two syllables and the penultimate is heavy, stress falls on the penultimate syllable, or else
- Stress falls on the antepenultimate syllable

It is important to note that when analysing the stress of a word, all prefixes must be ignored according to source of these stress rules.

3.2 Pausing

For pausing, every phone in the phonetic vocabulary has been 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 recording the data was instructed to utter some of the short word-final consonants followed by a "sokoon" with a short pause after. All vowels were included before phrase boundaries, so the instruction was not repeated for vowels. The talent was also instructed to be consistent in pausing in case the corpus is to be used for prosody modelling. However, it was noticed after the recording that some of the pauses were not semantically placed but occurred due to breathe.

3.3 Sentence Stress

Sentence stress is sometimes referred to as 'contrastive stress', 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). In this work, this phenomenon was considered too strong emotionally and context sensitive and so the talent was instructed not to provide emphasis on any word in the utterances in the transcript.

3.4 Intonation

It was agreed that the more automatic changes to the pitch and speed (duration) of natural human speech recording, the more unnatural it becomes (Baris Bozkurt, Dutoit, & Pagel, 2002; Clark et al., 2007; Maia, Toda, Zen, Nankaku, & Tokuda, 2007). So the talent was required to speak in a consistent, declarative and non-emotional manner (especially regarding pitch).
To estimate the pitch range that the talent should stay within, this study reviews research into speech synthesis where the authors have 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 was to see if it was possible to find a threshold of the ratio of change in f0, above which, any ratio of change in f0 would cause the segment to become unnatural and/or incomprehensible. Kawai et al (Kawai, Yamamoto, Higuchi, & Shimizu, 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 was considered acceptable. Kawanami et al. (2002) based their work on that of 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 talent had to be instructed not to change their pitch and speed for every recording.

3.5 Gemination

Gemination or "shadda" ("tashdeed") in MSA and in Arabic in general is described as the doubling of a consonant so that the resulting segment is double the length of its non-geminated counterpart (Selouani & Caelen, 1998). Gemination as a term is used in different ways in the literature but in this work, a geminate consonant is defined phonetically as an elongated consonant that is phonetically and 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".
In practice, gemination is not realised simply as a doubling of a consonant. Gemination is realised by increasing the duration of the articulation of the consonant, and depending on the type of consonant, the realisation differs. For plosives (stops), the length (duration) of low energy region before the explosion is

increased. This region is sometimes called the "plosive closure" and is adopted in this paper. For all other consonants (fricatives, nasals, approximates…), the length of articulation of the spectrally stable section of the phone is increased (Essa, 1998; Selouani & Caelen, 1998). The geminated consonants then are not merely a repetition of their corresponding consonants rather they are new phones to be added to the vocabulary that is considered for optimising the phonetic coverage of the authors' corpus. This adds to the 28 consonant phones another 28 geminated consonant phones for the phonetic vocabulary.

3.6 Emphasis and Nasalisation

4. Optimisation (Corpus Reduction)

The initial transcript was optioned by scraping the Aljazeera Learn website (Aljazeera, 2015) because it contains fully diacritised text which is hard to obtain. Most MSA text on the web is written without diacritisation. The initial script contained 23,531 words.

All the works reviewed for corpus optimisation for speech synthesis used greedy methods (Bonafonte et al., 2008; François & Boëffard, 2002; Kawai et al., 2000; Kawanami et al., 2002; Tao et al., 2008). Greedy methods as explained in the "National Institute of Standards and Technology" (P. E. 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 heuristic/s were different between works in the literature. Also the unit of choice for optimisation (triphone, diphone, phone…) varies. It should be noted that greedy methods do not guarantee the production of a globally optimal solution, as the corpus selection problem is NP hard (François & Boëffard, 2002) which requires a brute force search to find the optimal solution. This requires astronomical processing power. 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 greater than 10^{16}. François & Boëffard, 2002 classified greedy algorithms into three categories:

- Greedy: The initial solution is the empty set, and then sentences that increase coverage the most (relative to solution at iteration) are added to the solution until certain target coverage is achieved.
- Spitting: The initial solution is the whole sentence set and then sentences that are contributing least to coverage are removed iteratively until a sentence removal would damage coverage in some way.
- Exchange: Starting from a specific solution exchange one of the solution's sentences with one of the sentences excluded from the solution if this exchange increases coverage until no increase in coverage is possible maintaining a static set size.

The criteria for the three different approaches above were simple. They used unit counts from each sentence to give a score. "Useful units" in a sentence being 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) and "useless units" being the units that are redundant as the set that already has a number of units with the same identity that equals or is higher than the limit (3 in their case). The authors have used unit counts with the sentence cost (length) in different ways which were then compared. It was shown that by using "Spitting" after "Greedy" methods coverage cost improves (number of chosen sentences and their average length) but the method does not necessarily increase coverage. The way they combined the two methods was by running "Greedy" and then running "Spitting" restricting its choice of sentences to the output of "Greedy".

Since in this work the primary concern is coverage and not necessarily length of corpus, but the length of the generated speech (2 hours maximum for proper utterances), the "Greedy" method was chosen. To choose criteria for iteratively choosing sentences, a simple count was adopted where each sentence was scored by the following formula:

$$SS(S,C) = \sum_{k=0}^{n} \frac{SUF_k(S)}{CUF_k(C)} \quad if\ CUF_k(C) > SUF_k(S)\ \ for\ all\ k \quad (1)$$

$$= -1 \qquad otherwise$$

Where $SS(S,C)$ is the "Sentence Score" of the sentence S relative to corpus C, $SUF(S)$ is the "Sentence Unit Frequency" which is the number of times a specific unit indexed by k appears in the sentence S and $SUF(S)$ is the "Corpus Unit Frequency" which is the number of times a specific unit indexed by k appears in the corpus C. $CUF_k(C)$ 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.

A subset of diphones in Arabic was used for optimisation. The reason for using diphones, as the unit of choice, was based on the fact that this subset was the most used method in the literature reviewed (Barros & Möbius, 2011; Bonafonte et al., 2008; Kelly, Berthelsen, Campbell, Ni Chasaide, & Gobl, 2006; Kominek & Black, 2003; Matoušek & Romportl, 2007). The choice of this subset is informed by the study carried out so far and will be further elaborated on in section 3.1.

4.1 Phonetic Vocabulary

The diphones (see Table 1) only cover "short syllable diphones" and "half syllable diphones" which were the only diphones included in the optimisation. Both of these terms are used in this work for convenience and are not defined elsewhere. In this work, a short syllable is a syllable starting with a consonant (could be geminated) and ending with a vowel (no consonant coda), and a half syllable is the second part of a syllable ending with a consonant (a vowel followed by a strictly non-geminated consonant).

Table 1 shows the phonetic vocabulary used in this work for optimisation (the table does not include geminated consonants which are just represented by doubling the

letter or adding a colon after the consonant letter in all representations). Text in blue is vowels and diphthongs (at the bottom). Text in black is consonants. Text in green is foreign phones found in the corpus. A slightly modified Buckwalter transliteration was used here (Buckwalter, 2002) just to illustrate the phonetic nature of the phones. Brackets indicate emphaticness (pharyngealisation). Out of the complete set of theoretically possible diphones ($67^2 = 4489$ including geminated consonants diphones and excluding diphones containing a foreign origin phoneme), most were excluded for the following reasons:

- Emphatic consonants cannot be followed or preceded by a non-emphatic vowel /a/ or /a:/. This excludes $14 * 2 + 14 * 2 = 56$ diphones of this form.
- Consonant clusters (referred to here as "cc") were excluded because some of them do not occur in MSA. As for the rest, (Yi, 2003) has shown how certain concatenation points between specific types of phones 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 phenomena on a different scale (Tench, 2015; Yi, 2003). This makes it possible to construct those consonant clusters from smaller units by concatenating at the low amplitude region before the consonant. This excludes $56 * 56 = 3136$ diphones.
- "vv" clusters were excluded as they do not occur in MSA. This excludes 100 diphones.
- Diphones of the form "vc" (vowel-consonant) were reduced by unifying the identity of a "vc" diphone for long and short vowels of the same kind. This is assuming that the length of the vowel can be increased by the preceding unit in unit selection speech synthesisers. This excludes a further 210 diphones.

$$Diphones\ left = 4489 - 56 - 3136 - 210 - 100 - 1 = \mathbf{986}$$

Phonemes (Left: Arabic. Middle: IPA. Right: Buckwalter) except for last section where there is no IPA available														
أ	ʔ	ʾ	ر	r	r	غ	ɣ	g	ي	j	y	-(i)	i0
ب	b	b	ز	z	z	ف	f	F	ڤ	v	v	[ـُ]	u1
ت	t	t	س	s	s	ق	q	q	پ	p	p	[ـِ]	i1
ث	θ	^	ش	ʃ	$	ك	k	K	ج	d͡ʒ	J	[ا]	A:
ج	ʒ	j	ص	sˤ	S	ل	l	l	ا	a:	a:	[ـَ]	A
ح	ħ	H	ض	dˤ	D	م	m	m	و	u:	u0:	[و]	u1:
خ	x	x	ط	tˤ	T	ن	n	n	ي	i:	i0:	[ي]	i1:
د	d	d	ظ	ðˤ	Z	ه	h	h	ـَ	a	a	sil	sil
ذ	ð	*	ع	ʕ	E	و	w	w	ـُ	u	u0		
Diphthongs for general knowledge (Left: Arabic. Right: IPA)														
ـَي	/aj/	ـَو	/aw/	و(-)	/(a)w/	ي(-)	/(a)j/							

Table 1: Phonemes from which diphones were created. (sil) stands for pause and it is considered a phone. Square brackets represent emphatic (pharyngealised vowels).

The validity of these exclusions was only theoretical and based on rules before the recording, but was 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.

After running the optimisation script, 884 utterances were left in the data set out of the complete 2092 (for a threshold of 3). The optimisation process took place several times with the threshold for the minimum number of diphone occurrences changed every time. Table 2 shows the results. The threshold 3 was chosen because of resource limitations (10 hours recording studio time and talent time) and more utterances from the bigger sets were planned for recording in case studio time was more than sufficient.

Threshold	# of words	# 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

Table 2: Optimisation results. The row in blue was chosen based on resources available.

5. Discussion

The main contributions of this work follow:

- Conduct a study of MSA phonetics with a speech synthesis application in mind.
- Create a phonetic vocabulary with a set of di-phones that should exist in a transcript in MSA for it to be suitable for speech synthesis. This is based on analysis of classical Arabic and MSA phonetics and phonology and also optimisation for a certain speech synthesis method namely unit selection.
- Design and build a greedy algorithm to reduce the initial transcript keeping coverage optimal while reducing effort required.

The recording, segmentation and annotation of this corpus has not been covered in this paper and is intended as future work that will be conducted as part of this project.

6. Conclusion

It has been shown in this paper how important a complete study of phonetic and phonology of a language is for using greedy methods for reduction. It is particularly important where there is limited research into the appropriateness of the methods chosen for the language being studied namely Modern Standard Arabic. This work showed the theory behind MSA speech synthesis creation but further analysis of this is suggested as future work using subjective listening tests which is eventually be part of this project. This is to justify the choices made from a practical point of view.

7. Acknowledgements

The ECS accessibility team would like to thank the speech talent, linguistic experts and the sound engineer for their amazing effort.

8. Bibliographical References

Aljazeera. (2015). Aljazeera Learn. Retrieved February 15, 2015, from http://learning.aljazeera.net/arabic

Barros, M., & Möbius, B. (2011). *Human Language Technology. Challenges for Computer Science and Linguistics*. (Z. Vetulani, Ed.) (Vol. 6562). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-20095-3

Biadsy, F., Hirschberg, J., & Habash, N. (2009). Spoken Arabic dialect identification using phonotactic modeling, 53–61. Retrieved from http://dl.acm.org/citation.cfm?id=1621774.1621784

Black, A. W. (2002). Perfect Synthesis For All Of The People All Of The Time. In *IEEE 2002 Workshop on Speech Synthesis*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.8631

Black, P. E. (2005). Greedy algorithms. Retrieved March 15, 2015, from http://www.nist.gov/dads/HTML/greedyalgo.html

Bonafonte, A., Adell, J., Esquerra, I., Gallego, S., Moreno, A., & Pérez, J. (2008). Corpus and Voices for Catalan Speech Synthesis. In *LREC 2008*. Retrieved from http://aclweb.org/anthology/L08-1517

Bozkurt, B., Dutoit, T., & Pagel, V. (2002). Re-Defining Intonation From Selected Units For Non-Uniform Units Based Speech Synthesis. In *Proc. SPS-IEEE Benelux Signal Process. Symp* (pp. 141–144). Leuven, Belgium. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.2935

Bozkurt, B., Dutoit, T., Prudon, R., D'Alessandro, C., & Pagel, V. (2002). Improving quality of MBROLA synthesis for non-uniform units synthesis. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.* (pp. 7–10). IEEE. doi:10.1109/WSS.2002.1224360

Buckwalter, T. (2002). Buckwalter Arabic Transliteration. Retrieved June 27, 2013, from http://www.qamus.org/transliteration.htm

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

de Jong, K., & Zawaydeh, B. A. (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics*, *27*(1), 3–22. doi:10.1006/jpho.1998.0088

Essa, O. (1998). Using Prosody in Automatic Segmentation of Speech. In *Proceedings of the ACM 36 th Annual Southeast Conference* (pp. 44–49). New York, New York, USA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.4359

Fernández, E. M., & Cairns, H. S. (2011). *Fundamentals of Psycholinguistics* (1st ed.).

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'02)*. European Language Resources Association (ELRA). Retrieved from http://aclweb.org/anthology/L02-1265

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

Kawai, H., Yamamoto, S., Higuchi, N., & Shimizu, T. (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)* (pp. 420–425). Beijing, China.

Kawanami, H., Masuda, T., Toda, T., & Shikano, K. (2002). Designing speech database with prosodic variety for expressive TTS system. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)* (pp. 2039–2042). European Language Resources Association (ELRA). Retrieved from http://aclweb.org/anthology/L02-1337

Kelly, A. C., Berthelsen, H., Campbell, N., Ni Chasaide, A., & Gobl, C. (2006). Speech Technology for Minority Languages: the Case of Irish (Gaelic). In *INTERSPEECH*. Pittsburgh, Pennsylvania. Retrieved from http://www.tara.tcd.ie/handle/2262/39404

Kenworthy, J. (1987). *Teaching English Pronunciation*.

Kominek, J., & Black, A. W. (2003). CMU Arctic Databases for Speech Synthesis. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.64.8827

Kominek, J., & Black, A. W. (2014). THE CMU ARCTIC SPEECH DATABASES. In *5th ISCA Speech Synthesis Workshop* (pp. 223–224). Pittsburgh, Pennsylvania.

Maia, R., Toda, T., Zen, H., Nankaku, Y., & Tokuda, K. (2007). An Excitation Model for HMM-Based Speech Synthesis Based on Residual Modeling. In *6th ISCA Workshop on Speech Synthesis* (pp. 131–136). Bonn, Germany: International Speech Communication Association. Retrieved from http://library.naist.jp/dspace/handle/10061/8269

Matoušek, J., & Romportl, J. (2007). 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* (pp. 326–333). Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=1776334.1776380

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

Oliveira, L., Paulo, S., Figueira, L., Mendes, C., Nunes, A., & Godinho, J. (2008). Methodologies for Designing and Recording Speech Databases for Corpus Based Synthesis. In *LREC 2008*. Retrieved from http://aclweb.org/anthology/L08-1484

Parlikar, A., & Black, A. W. (2012). Data-driven phrasing for speech synthesis in low-resource languages. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4013–4016). IEEE. doi:10.1109/ICASSP.2012.6288798

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. IVTTA '98 (Cat. No.98TH8376)* (pp. 155–160). IEEE. doi:10.1109/IVTTA.1998.727712

Tao, J., Liu, F., Zhang, M., & Jia, H. (2008). Design of Speech Corpus for Mandarin Text to Speech. In *The Blizzard Challenge 2008 workshop*. Brisbane, Australia.

Tench, P. (2015). Consonants. Retrieved March 15, 2015, from http://www.cardiff.ac.uk/encap/contactsandpeople/academic/tench/consonants.html

Yi, J. R.-W. (2003). Corpus-based unit selection for natural-sounding speech synthesis. Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/16944