Evaluation of speech unit modelling for HMM-based speech synthesis for Arabic

Amal Houidhek, Vincent Colotte, Zied Mnasri, Denis Jouvet

Evaluation of Speech Unit modelling for HMM-Based Speech Synthesis for Arabic

Amal Houidhek[1,2], Vincent Colotte[2], Zied Mnasri[1], Denis Jouvet[2]

[1]Electrical Engineering Department, Ecole Nationale d'Ingénieurs de Tunis, University Tunis El Manar, Tunisia
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{amal.houidhek, vincent.colotte, denis.jouvet}@loria.fr, zied.mnasri@enit.rnu.tn

Abstract

This paper investigates the use of hidden Markov models (HMM) for Modern Standard Arabic speech synthesis. HMM-based speech synthesis systems require a description of each speech unit with a set of contextual features that specifies phonetic, phonological and linguistic aspects. To apply this method to Arabic language, a study of its particularities was conducted to extract suitable contextual features. Two phenomena are highlighted: vowel quantity and gemination. This work focuses on how to model geminated consonants (resp. long vowels), either considering them as fully-fledged phonemes or as the same phonemes as their simple (resp. short) counterparts but with a different duration. Four modelling approaches have been proposed for this purpose. Results of subjective and objective evaluations show that there is no important difference between differentiating modelling units associated to geminated consonants (resp. long vowels) from modelling units associated to simple consonants (resp. short vowels) and merging them as long as gemination and vowel quantity information is included in the set of features.

Keywords: parametric speech synthesis, statistical modelling, Arabic language, speech unit modelling, vowel quantity, gemination

1. Introduction

Making allowance for the development of human machine interaction, speech synthesis is becoming more and more widespread and omnipresent in mobile applications (e.g., mail reading service) and information delivery applications. Speech synthesis consists in the automatic generation of a natural voice from a written text; the process is called Text To Speech (TTS) synthesis (Taylor, 2009). The automatic conversion of the text into a speech signal begins with a text analysis based on natural language processing. This corresponds to the front-end part, and consists in text segmentation into different levels (sentences, words...), text normalization, part of speech tagging, and conversion of the text into a sequence of phonemes. Then, a speech waveform is generated thanks to the concatenation of speech units corresponding to the desired sequence of phonemes or using speech parametric models. Developing a full speech synthesis system in a new language requires resources and knowledge. Active learning-based approaches have been proposed to reduce the need for language-specific expert knowledge (Watts et al., 2013). The work presented in this paper deals with the adaptation of the speech modelling part to the Arabic language, and does not investigate the front-end part.

Amongst the TTS techniques, statistical parametric speech synthesis (SPSS) (Tokuda et al., 2002; Black et al., 2007) has been widely employed. SPSS is based on hidden Markov models (HMM) which can be automatically trained using speech data. Models depend on a representation of the speech signal with a set of parameters (e.g., duration of sounds, spectrum, and fundamental frequency (F0)). Statistics (e.g., means and variances of probability density functions) are used to describe the distribution of the speech parameter values in the training corpus. Note that, before being applied to speech synthesis, HMMs have been widely and successfully used for automatic speech recognition, thanks to several toolkits such as HTK (Young, 1994).

HTS, the HMM-based speech synthesis system, is based on HTK, and presents the advantage of being trainable, making changing voice characteristics possible and produces a rather good quality speech signal. It has been applied to many languages, such as Japanese (Yoshimura et al., 1999), English (Tokuda et al., 2002), and German (Krstulovic et al., 2007). The performance of HMM speech synthesis system depends on the parameterisation of the speech signal and on the modelling of speech units. In HTS, context dependent phoneme models are employed. This requires a description of each speech unit with a set of contextual features that includes factors affecting the pronunciation of the corresponding sound. The set of contextual features comprises linguistic as well as prosodic and phonological information to describe all the characteristics of the speech unit.

In practise, the choice of contextual features is highly prominent because they are involved in the different parts of HTS: in the training part when building decision trees for parameter sharing and in the synthesis part where the context-dependent HMM are used to predict speech parameters. Thus, they have a considerable impact on the quality of the generated speech. Although a standard set of features was proposed for English in (Tokuda et al., 2002), a part of contextual features is language-dependent. Hence, previous adaptation of HTS for other languages went through ignoring or adding some features to cover particularities of the target language. Le Maguer et al. (2013) suggested a different set of features for French and evaluated the impact of adding new information on the modelling of acoustic parameters, whereas Silén et al. (2010) focused on evaluating the modelling of consonants and vowels durations for Finnish.

This paper investigates the adaptation of HTS, a widely used HMM-based speech synthesis system, for Modern Standard Arabic (MSA). The set of contextual features needs to take into account the specificities of Arabic language (Al-Ani, 1970) to generate natural Arabic speech. As claimed by studies of Arabic phonology, MSA presents particular phenomena that are not considered in the standard set of contextual features, such as gemination and vowel quantity (long vs. short vowels). As stated by Khouja and Zrigui (2005), a geminated consonant is twice as long as its simple counterpart is, and a long vowel is twice as long as its short counterpart is. This paper proposes and evaluates several choices for modelling the speech units that differentiate, or not, units associated to long vowels from those associated to short vowels, and/or units associated to geminated consonants from those associated to simple consonants. In the experiments, information about gemination (simple or geminated consonant) and vowel quantity (short or long vowel) is always included in the set of contextual features. Arabic utterances generated by the various modelling approaches are compared to each other based on an objective evaluation of the duration of the speech segments, and on perceptive tests. Crowdsourcing approaches have been proposed for conducting subjective evaluation tests, but the detection of cheating is critical (Buchholz & Latorre, 2011), and wrong decisions may bias the results. Thus, we have applied a more traditional evaluation procedure, where listeners were physically present for evaluations.

The rest of this paper is organised as follows. Section 2 presents a brief overview about speech synthesis and describes the HTS system. Section 3 describes the aspects of the Arabic speech to be taken into account in synthesis using HTS system, and introduces the proposed modelling approaches. Section 4 exposes the set of conducted experiments and discusses the results of objective and subjective evaluations.

2. Speech Synthesis

2.1. Brief overview

Previous works in TTS domain came up with several techniques to implement speech synthesis systems. Earlier methods relied on rule-based formant synthesis. They use phonetic units and produce speech based on rules of evolution of formants between phones (Klatt, 1980; Taylor et al., 1991). Afterwards, a variety of concatenation-based methods have been developed. They are based on the use of a corpus, a database of recorded utterances, and they consist in joining speech units (from the corpus) which can be phonemes, diphones or syllables to obtain the desired speech signal. Among concatenation variants, one particular method called unit selection, showed improvement in terms of quality and naturalness. It consists in selecting the best sequence of speech units among many candidate units from the speech database to produce the speech signal (Hunt and Black, 1996). This method considers both a target cost to measure similarity between selected unit and target characteristics, and a join cost to measure concatenation quality. (Schwarz et al., 2006) has proposed an approach based on k-nearest neighbours to speed up the unit selection process.

This paper focuses on a particular approach of TTS: Statistical Parametric Speech Synthesis (SPSS), which is based on HMMs. Figure 1 presents the main blocks of a SPSS system. The process starts with text processing, then corresponding speech parameters are predicted and finally processed using a synthesis filter to generate a speech signal.

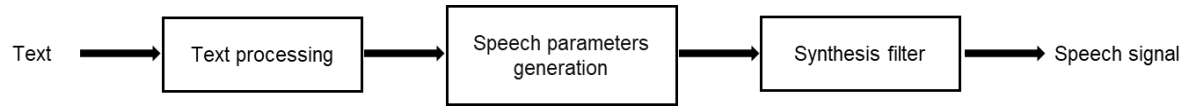

Figure 1: Overview of SPSS

In SPSS systems, the speech signal is characterized by a set of parameters extracted at regular time intervals using a sliding window; this representation combines spectral envelope parameters, fundamental frequency and aperiodic excitation parameters.

2.2. HMM-based speech synthesis system (HTS)

In HTS, speech parameters of each speech unit, i.e., spectrum features (e.g., Mel-cepstral coefficients and their dynamic features), fundamental frequency (F0), and phoneme duration are statistically modelled and generated based on the use of HMMs. First versions of HTS were using Mel-cepstral coefficients to represent the speech spectrum (Tokuda et al., 2002). Later, the MGC-LSP coefficients – Mel-Generalized-Cepstrum-based Line-Spectrum-Pair – (Koishida et al., 1997) have been used, and associated with the STRAIGHT vocoder (Kawahara et al., 1999), they led to improved speech quality (Zen et al., 2006). Since then, the STRAIGHT vocoder is used with HTS. With respect to the modelling, context-dependent models are used in HTS to handle the fact the speech parameters of a speech unit are dependent on contextual features. Contextual features include information on the preceding and following segments, on the position of the segment in the syllable and in the word, and a variety of other details (cf. Section 2.2.2).

2.2.1. General aspects of HTS

Figure 2 describes the HTS process (Zen et al., 2009). The mechanism includes two main blocks, i.e., training and synthesis. The input of the training process is a data corpus of natural speech signals. For each signal, acoustic parameters are extracted at a frame rate of 5 ms. Parameters include spectrum features (e.g., Mel-cepstral coefficients and their dynamic features), excitation (including log (F0) and its dynamic features) and aperiodicity parameters. Acoustic parameters are modelled using context dependent HMMs. Each speech segment is described with a set of contextual features. Once the training is achieved, three models are obtained: one for the duration of the sounds, one for the spectrum parameters and the last one for fundamental frequency and aperiodicity.

The synthesis process of HTS begins with converting the given text (to be synthesized) into a sequence of contextual features. The corresponding context-dependent HMMs are joined together to build the HMM of the utterance. State durations of the utterance HMM are determined in order to maximize the output probability. In HTS, state durations are predicted separately based on duration HMMs.

Spectrum and excitation parameters are generated from the HMMs using a speech parameter generation algorithm, i.e., MLPG (Maximum Likelihood Parameter Generation) that maximizes the output probabilities. Ultimately, the synthesis filter, i.e., MLSA (Mel-Log Spectrum Approximation) produces the speech waveform using the generated excitation and spectrum parameters.

The high number of contextual features in HTS (about 50 features in the standard set) increases the complexity to pick essential and useful context information. Therefore, decision trees (Jurafsky and Martin, 2009) are involved to group into clusters similar probability density functions. In HTS, the distributions for the spectrum, excitation and duration are clustered separately; in consequence, different phonetic decision trees are obtained for the modelling of spectrum, excitation and duration.

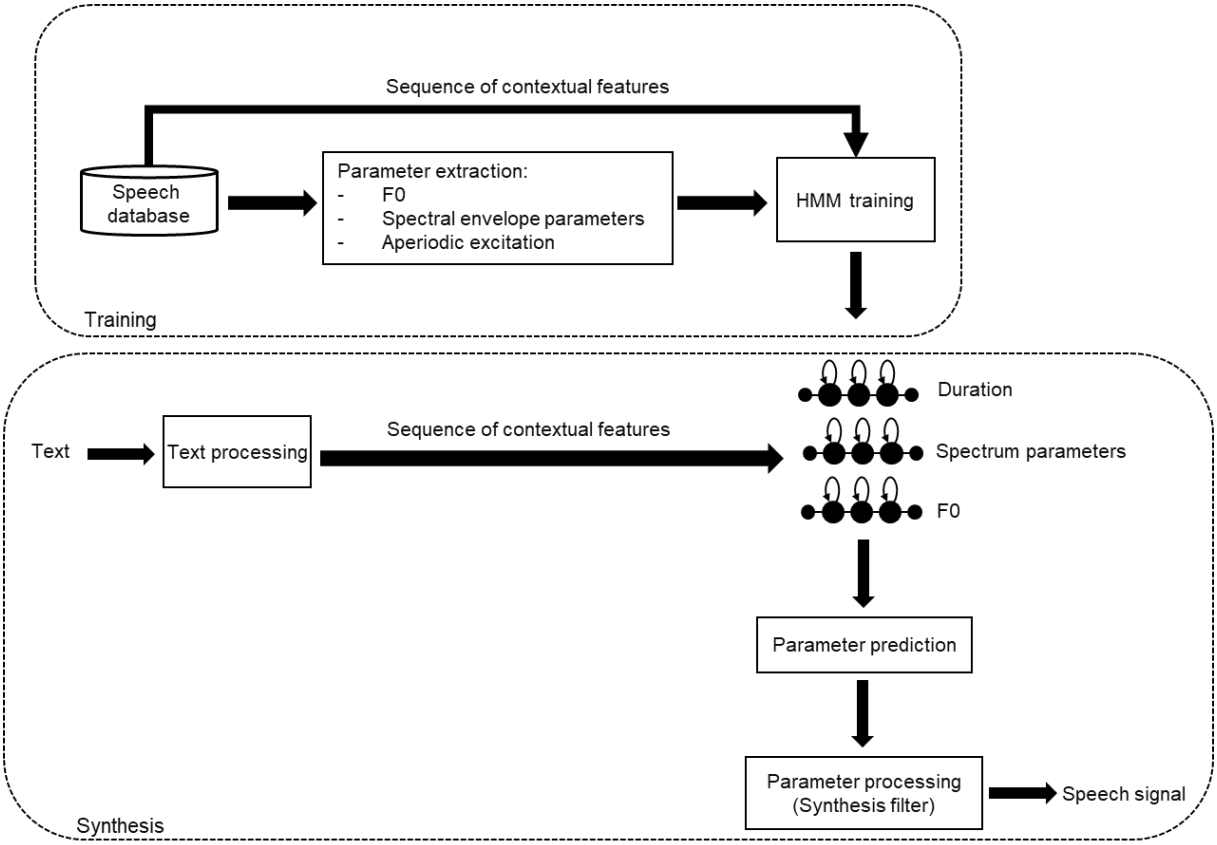

Figure 2: Overview of HTS

The vocoder STRAIGHT is used to extract acoustic parameters from speech signals of the training database and to process the predicted speech parameters to generate a speech signal.

2.2.2. Contextual features

The conversion of the text into a sequence of speech segments (phones and pauses) with associated contextual features is an essential step in HTS. The contextual features should capture all factors that can affect how a speech unit is pronounced in a particular context. Therefore, they include different types of information; either linguistic, phonologic or prosodic. The set of features comprises factors at different levels, within syllables, within words and within utterances. During the synthesis phase, decision trees and contextual features are employed to predict speech parameters. Thus, the choice of contextual features is crucial and affects the speech quality.

Tokuda et al. (2002) suggested a list of features that influence the pronunciation of speech units:

- Phoneme level features
    - Current phoneme.
    - Previous and next phonemes.
    - Position of current phoneme within the current syllable (forward and backward).

- Syllable level features
    - Number of phonemes within current, previous and next syllables.
    - Stress and accent information of current, previous and next syllables.
    - Vowel identity of current syllable.
- Word level features
    - Part Of Speech tagging of current, previous and next words.
    - Number of syllables of current, previous and next words.
    - Position of current word within current phrase (forward and backward).
- Phrase level features
    - Number of syllables of current, previous and next phrases.
    - TOBI end tone of current phrase (Silverman et al., 1992).
- Utterance level features
    - Number of syllables, words and phrases in utterance.

2.3. Arabic speech synthesis

Over the last three decades, the interest in Arabic speech synthesis has been steadily rising. Therefore, most of the TTS techniques have been adapted for Arabic language. Based on formants, i.e., maxima of the spectrogram, acoustic speech is generated through formants characteristics (i.e., bandwidth and amplitude) and rules of evolution of formants between phonemes, Rajouani et al. (1987) proposed a rule-based Arabic speech synthesis system.

Various corpus-based concatenative synthesis methods have been applied to Arabic language. These techniques consist in joining speech units of different sizes. Baloul (2003) proposed an automatic speech synthesis system for Arabic based on diphone concatenation. Diphone were chosen because they are considered a stable speech unit, i.e., a diphone starts from the middle of a phoneme (stable part) until the middle of the neighbouring phoneme, thus it covers the transition phase between the two adjacent phonemes (Moulines et al., 1990). Later Cheffour et al. (2000) brought to the fore larger speech units and came up with a concatenative speech system for Arabic based on di-syllables, i.e., units that starts from the middle of a syllable until the middle of the next syllable. Furthermore, triphones, i.e., sequences of three phonemes (Kishore and Black, 2003), were used in Ahmed (2004) to create Arabic speech signals. Abdelmalek and Mnasri (2016) applied unit selection technique for Arabic using phonemes and syllables as speech units. Halabi (2015) built an Arabic speech database and used Innoetics, which is considered a high quality speech synthesiser based on unit selection (Chalamandaris et al., 2013), to generate speech signals.

Besides, HTS has been also applied to MSA. Abdel-Hamid et al. (2006) suggested some modifications to improve speech quality. The scope of their work was on modifications of the speech parameters and on signal processing parts, a multi-band excitation model was applied and spectral parameters were extracted from spectral envelope. Khalil and Cherif (2013) have used the basic HMM-based speech synthesis system to produce Arabic speech relying on phonemes speech units and on the STRAIGHT vocoder.

3. Arabic HMM-based speech synthesis

To apply HTS to MSA, as for any other languages, the written text must be transformed into a sequence of contextual features. For Arabic, choosing the features requires the knowledge of particularities of MSA so that all factors that may affect the pronunciation of the speech units are captured.

3.1. Arabic phonology

MSA has a set of 28 consonants, which can exist in two forms: simple and geminated, and six vowels: three short and three long vowels (Newman, 1984). Previous studies of Arabic phonology have analysed the following phenomena for MSA: stress, emphatic consonants, vowel quantity and gemination.

3.1.1. Stress

It is a prominence given to a syllable within a word and usually called lexical stress or word stress (Black et., al 1998). In Arabic language, the position of the stress in a word can be predicted through predefined rules (Kouloughli, 1976; Halpern, 2009; Al-Ani, 1970).

3.1.2. Emphatic consonants

It consists in the pharyngealization of consonants in Arabic (Halabi, 2015). Consonants of MSA can be divided into three classes; always-emphatic consonants, consonants that can never be emphatic and two-state consonants, i.e., consonants than can be emphatic in particular contexts (Laufer and Baer, 1988).

3.1.3. Vowel quantity

MSA has short and long vowels (Selouani and Caelen, 1998). In spelling, long vowels are always indicated, unlike short vowels, by the following graphemes <و> /uu/, <ي> /ii/, and <ا> /aa/. When replacing a short vowel with its long counterpart, the meaning of the word changes, e.g., "هتف" /h**a**tafa/ means, "he shouted", whereas "هاتف" /h**aa**tafa/ means, "he telephoned".

3.1.4. Gemination

Consonants of MSA have two forms: simple and geminated. In spelling, a geminated consonant is distinguished from its simple counterpart by adding a diacritic sign (ّ) called *shadda* above the concerned consonant (Newman, 1984). The presence of a geminated consonant changes the meaning of the word, e.g., "درس" /da**r**asa/ means, "he studied", whereas "درّس" /da**rr**asa/ means, "he taught".

3.2. Speech unit modelling

This work investigates the choice of speech units and their modelling in an HMM-based speech synthesis system for Arabic language. Thus, particularities of the language are taken into account with a focus on gemination and on vowel quantity. For what concern emphatic consonants, they are distinguished from other consonants in the data we are using; hence, they correspond to specific modelling units. With respect to stress, stress information already belongs to the set of contextual features in the standard HTS system. Previous works on Arabic speech synthesis based on HMM did not mention any explicit interest on how to model geminated consonants and long vowels. Therefore, here, several choices of unit modelling are considered to investigate whether adding gemination and vowel quantity information into the set of contextual features is enough, or should geminated consonant (resp. long vowel) be considered as fully-fledged speech units. Thus, four possible speech unit-modelling approaches are proposed below. Note that for the four proposed approaches, the vowel quantity and the gemination information are always present in the set of contextual features. It is assumed that if such information is redundant for some choice of speech units, the decision trees that are built during the training process will simply ignore it.

3.2.1. Single model for simple and geminated consonants, and single model for short and long vowels (C1V1)

This is the most compact system, where a geminated consonant and its corresponding simple consonants are modelled with the same unit, and similarly, a long vowel and its corresponding short vowel are modelled with the same unit. This system in called C1V1.

3.2.2. Differentiating only short vs. long vowels (C1V2)

In this system, a single unit models both a geminated consonant (e.g., /dd/) and its simple counterpart (/d/). Whereas a long vowel (e.g., /aa/) and its short counterpart (e.g., /a/) are modelled by two different units. This system is named C1V2.

3.2.3. Differentiating only simple vs. geminated consonants (C2V1)

This system uses a single unit to model both a long vowel and its short counterpart. While for consonants, two units are used, one for the simple consonant and one for its geminated counterpart. This system is named C2V1.

3.2.4. Differentiating simple vs. geminated consonants and short vs. long vowels (C2V2)

This is the most detailed system, where two different units are used to model a simple consonant (e.g., /d/) and its geminated counterpart (e.g., /dd/). In addition, for vowels, two different units also are used to model a short vowel (e.g., /a/) and its long counterpart (/aa/). This model is named C2V2.

4. Experiments

4.1 Data

In order to develop and evaluate the different systems corresponding to the speech unit modelling variants mentioned above, a set of experiments was conducted. For this purpose, an Arabic corpus was used (Halabi and Wald, 2016). It contains 1806 utterances corresponding to a total of 4 hours of recorded speech; the corpus is a collection of news bulletins uttered in a neutral style by a male speaker (Halabi, 2015). The audio signals were recorded using a professional studio and speech signals were sampled at 48 kHz. The software Pro Tools 11 was used during recording. Besides, the Studio Microphone Neumann TLM 103 was employed, a choice justified by the fact that the device is successfully used to record human speech with high quality. When recording, the speaker was in a soundproof anechoic booth.

In the experiments reported below, a speaker-dependent modelling was used and a distinct HTS model was trained for each choice of speech unit modelling. A training corpus of 1565 Arabic utterances was considered and 30 other utterances were kept apart for the evaluation. Not all the corpus was used because there were some transcription issues, which could affect the quality of the training process; the corresponding utterances were thus ignored. To guarantee the best possible speech quality, the vocoder STRAIGHT was used.

4.2. Contextual features and modelling

The set of contextual features proposed for MSA speech synthesis was inspired from the standard set of contextual features defined for English (Tokuda et al., 2002). However, information related to tone accent and TOBI was ignored, as Arabic language does not present these characteristic rules (Kouloughli 1976; Al-Ani 1970). Moreover, no additional feature about emphatic aspect was added to the set of contextual features as emphatic consonants were already distinguished from other consonants in the transcription of the corpus.

However, two additional features were introduced to take into account specificities of the Arabic language. The first feature is used to indicate the gemination characteristic (possible values are ‘simple consonant’, ‘geminated consonant’, or ‘not a consonant’); the second one refers to vowel quantity (possible values are ‘short vowel’, ‘long vowel’, or ‘not a vowel’).

In the experiments, speech signals are generated with the four proposed systems using the same set of contextual features, which contain information about vowel quantity and consonant gemination. As mentioned, decision tree will simply ignore extra or redundant features. To measure the performance of a speech synthesis system and to compare them, subjective and objective evaluations are carried out.

4.3. Objective evaluation of phone duration

4.3.1 Evaluation protocol

An objective evaluation is conducted to evaluate the predicted duration of the speech units. For speech signals produced with each system (C1V1, C1V2, C2V1 and C2V2) the average, over the vowels, of the ratios between the mean duration of long vowels (LV) and the mean duration of corresponding short vowels (SV) is calculated, as well as the average ratio for geminated consonants (GC) vs. simple consonants (SC). Only phonemes with more than 10 occurrences for each class (simple/geminated consonants and short/long vowels) are considered. The calculated average ratios are compared to those obtained on natural speech.

4.3.2 Results and discussion

The obtained values, reported in Table 1, show that for the four systems, the duration ratios of long to short vowel and the duration ratios of geminated to simple consonant are similar to the ones calculated on natural speech.

Table 1. Duration ratios.

		LV / SV	GC / SC
	Number of occurrences	262 / 884	104 / 1315
Models	C1V1	1.8	2.2
	C1V2	1.9	2.2
	C2V1	1.8	2.1
	C2V2	1.8	2.0
	Natural	2.0	2.1

To better understand the similarity of the obtained duration ratios, the decision trees associated to the modelling of duration of the four systems have been analysed. For each system, there is only one single tree for all the units (phonemes and silence). Therefore, the four decision trees have been analysed, and Figure 3 represents the top part of the duration decision tree for the model (C2V2).

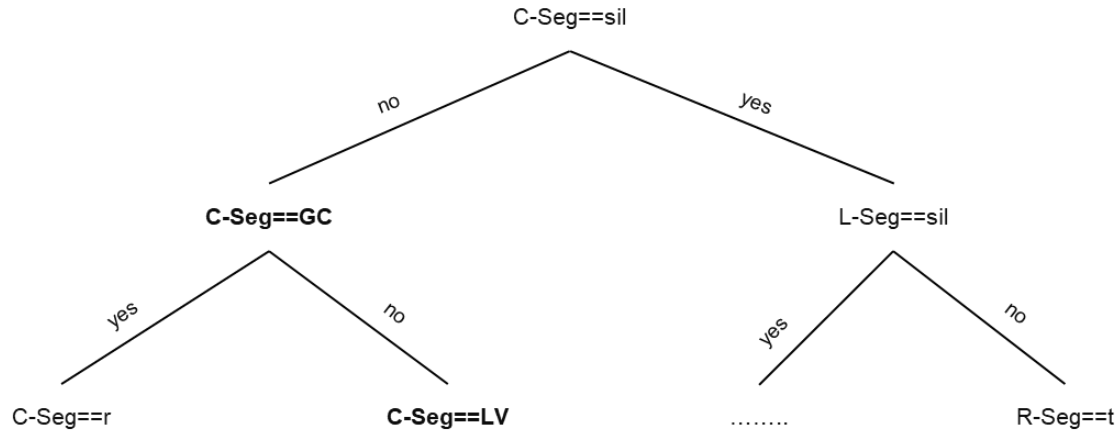

Figure 3: Duration decision tree of (C2V2) model (top part)

In the figure, "C-Seg" refers to the current segment, "L-Seg" to the previous (left) segment, and "R-Seg" to the next (right) segment; "sil" indicates a silence, “r” and “t” are the phonemes /r/ and /t/, “GC” indicates a geminated consonant, and “LV” indicates a long vowel. Figure 3 shows that questions about the nature of speech segment such as geminated consonant (C-Seg==GC) and long vowel (C-Seg==LV) are situated at the top of the tree. This behaviour was observed for the four models.

To further investigate the prediction of phone durations, root mean square error (RMSE) between natural duration and predicted durations with HTS was calculated on generated signals for the four systems C1V1, C1V2, C2V1 and C2V2 within different phoneme classes (simple and geminated consonants, and short and long vowels). The obtained RMSE are presented in Figure 4. Results show that for each phoneme class, there is no prominent difference between the RMSE measured for the four-speech unit modelling approaches.

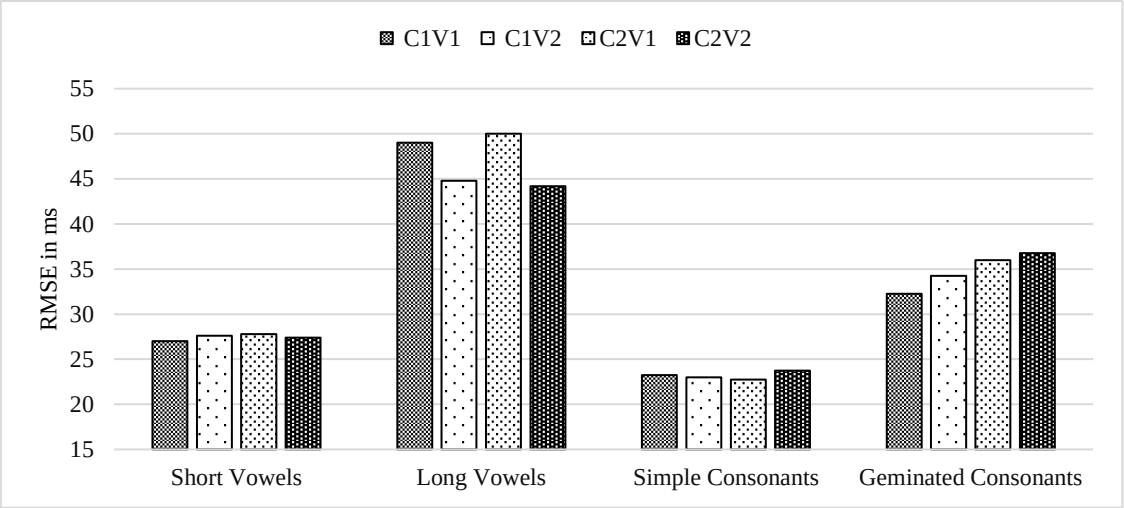

Figure 4: RMSE between natural durations and HTS generated durations

Figure 5 displays the normalized root mean square error (NRMSE) obtained by normalizing the RMSE by the mean duration values of the phoneme classes. Obtained NRMSE values are quite similar; around 25% for geminated consonants and 35% for the other phoneme classes (simple consonants, short vowels, and long vowels).

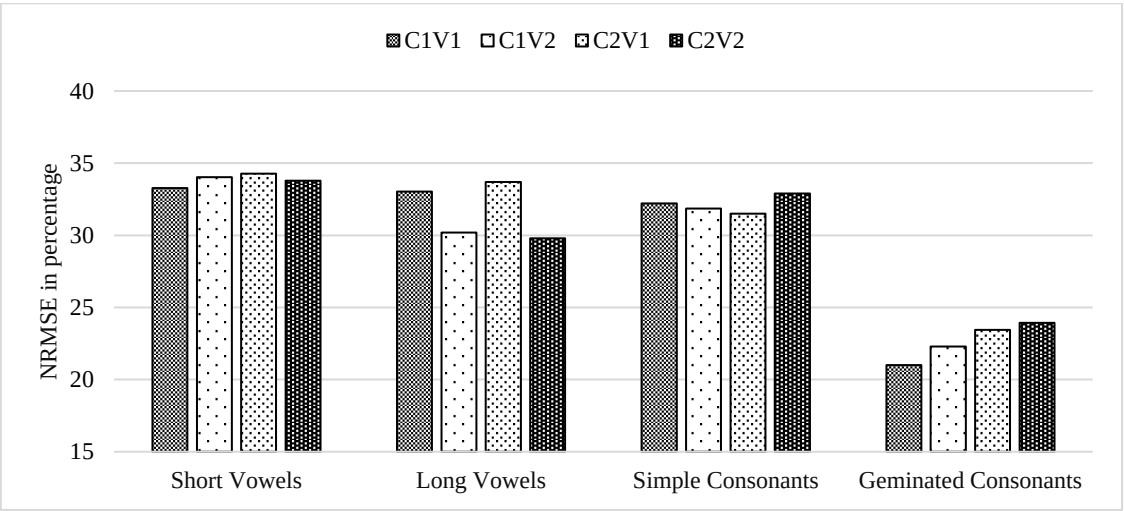

Figure 5: NRMSE between natural durations and HTS generated durations

4.4 Subjective evaluation of global quality and naturalness

4.4.1 Evaluation protocol

In the present work, MOS (Mean Opinion Score) tests (ITU, 1996) were carried out to evaluate the global quality of the speech signals generated by the four systems C1V1, C1V2, C2V1 and C2V2. Two factors contribute to the global quality: naturalness and overall signal quality. The overall signal quality refers to the quality of the produced acoustic signal. The naturalness is evaluated by referring to the intonation (whether the pitch's change is natural) and the rhythm (whether length of phonemes sounds natural too).

Nine Arabic native speakers took part in this evaluation. Each participant evaluated a set of 40 stimuli, i.e., 10 stimuli from each system (C1V1, C1V2, C2V1 and C2V2) and judged the corresponding overall quality and naturalness. Listeners were asked to answer the question "*How do you rate the overall quality and naturalness of what you have just heard compared to a natural speech (pronounced by a human being)?"* by giving a score ranging from 1 to 5 as follows:

- Very close to natural speech (5)

- Close to natural speech (4)
- A little bit far from natural speech (3)
- Far from natural speech (2)
- Very far from natural speech (1)

4.4.2 Results and discussion

The resulting MOS scores are presented in Figure 6 with their associated 95% confidence intervals. Results show that for the four modelling approaches, C1V1, C1V2, C2V1 and C2V2, HTS produces speech signals with similar perceived characteristics (overall quality, intonation and rhythm).

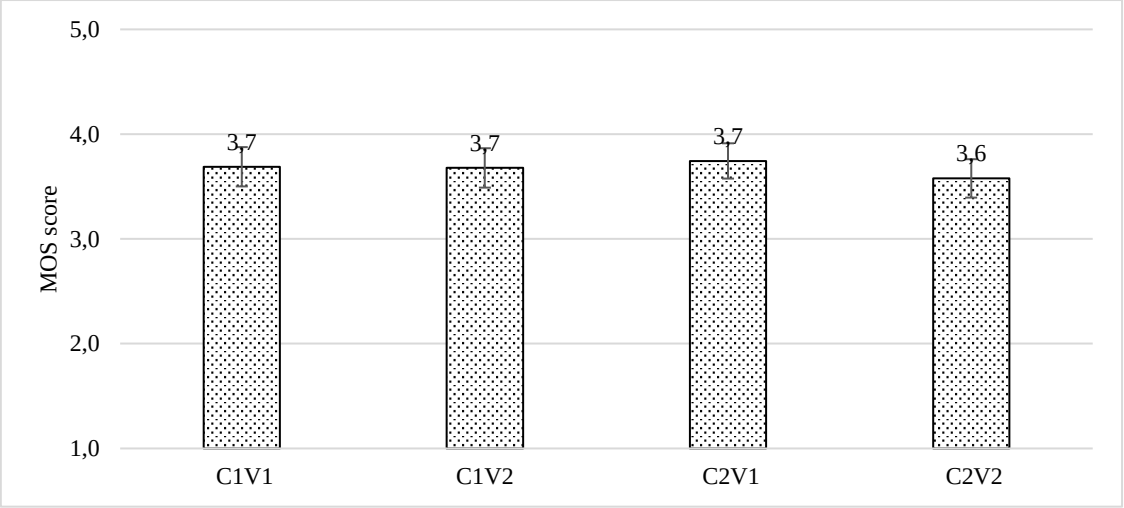

Figure 6: Results of global quality evaluation and naturalness

4.5 Subjective evaluation of degradation

4.5.1 Evaluation protocol

In this part, a DMOS (Differential Mean Opinion Score) test (ITU, 1996) was conducted to assess the degree of perceived degradation in the HTS generated speech signals, by comparing each speech signal produced by one of the four modelling approaches to the corresponding natural speech signal. For this evaluation, twelve listeners participated; each one evaluated a set of 20 pairs, each pair consists of the same utterance produced with one of the four systems and the corresponding natural signal. Signals were presented in a defined order; first, the reference (natural signal), followed by the signal produced by one of the systems (C1V1, C1V2, C2V1 and C2V2). Participants evaluated the degradation of signals by answering the question "*How do you judge the degradation of the second signal compared to the first one?*" using a five-point degradation category scale:

- Inaudible degradation (5)
- Audible but not annoying degradation (4)
- Degradation a little annoying (3)
- Annoying degradation (2)
- Very annoying degradation (1)

4.5.2 Results and discussion

The obtained DMOS are represented in Figure 7 with their associated 95% confidence intervals. The higher the score is, the lower the degradation is. Obtained DMOS scores show that signals produced with the four models C1V1, C1V2, C2V1 and C2V2 present a similar degree of degradation when compared to the natural speech. These results are consistent with the fact that the four systems provide a similar speech quality according to the results of global quality evaluation.

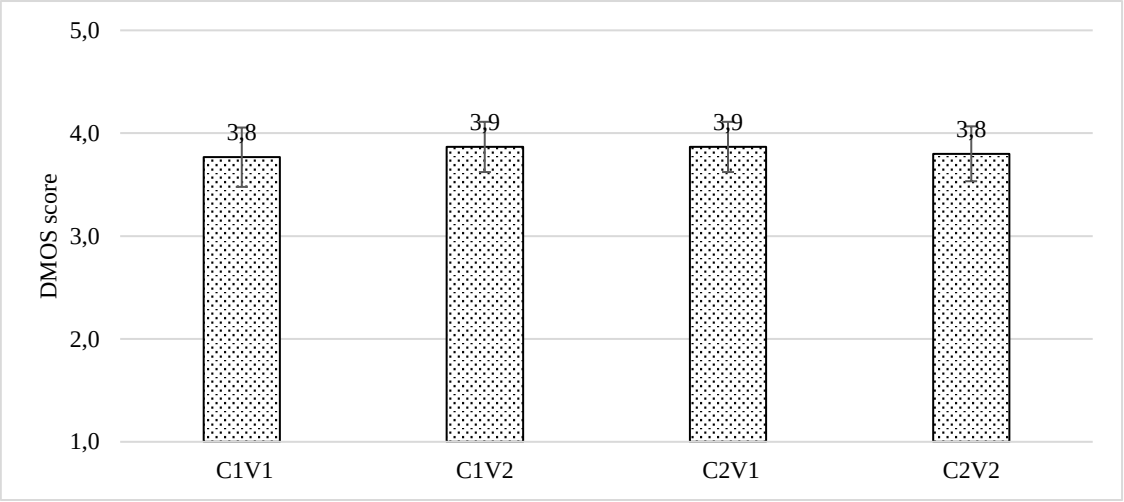

Figure 7: Results of degradation evaluation

4.6 Comparison of modelling approaches

4.6.1 Evaluation protocol

A preference test (ITU, 1996) was carried out to compare the four proposed approaches for speech unit modelling with respect to the quality of produced speech. Speech signals produced with the four models (C1V1, C1V2, C2V1 and C2V2) are compared to each other. Twenty-seven Arabic native speakers participated in this evaluation. Each one evaluated a set of 23 pairs of speech signals; each pair consists of the same utterance produced with two different systems. The order of presenting the speech signals is randomly chosen for each trial. During the evaluation, participants were asked to point to the preferred signal based on the global quality of produced speech, by answering the question "*How do you judge the quality of the second signal compared to the first one?*" and giving a score from 1 to 7 ranging from "*Much worse*" to "*Much better*". For the analysis of the results, the scores have been grouped into three categories: 'first preferred', 'no preference' and 'second preferred', as follows:

- Much better (7)
- Best (6)

} Second preferred

- A little better (5)
- About the same (4)
- A little worse (3)

} No preference

- Worst (2)
- Much worse (1)

} First preferred

4.6.2 Results and discussion

Results are presented in Figure 8. For example, on the figure, the bottom line corresponds to the comparison of C1V2 with C2V1. For this line, 12% of the answers give a preference to the first system (left side, i.e., C1V2), 6% of the answers give a preference to the second system (right side, i.e., C2V1), and the middle part shows that 82% of the answers do not express any preference. The one-to-one comparison of the four models shows that listeners had no clear preference for any particular system. Therefore differentiating geminated consonants (resp. long vowels) from simple consonants (resp. short vowels) or merging them when defining the speech units, lead to a similar speech synthesis quality.

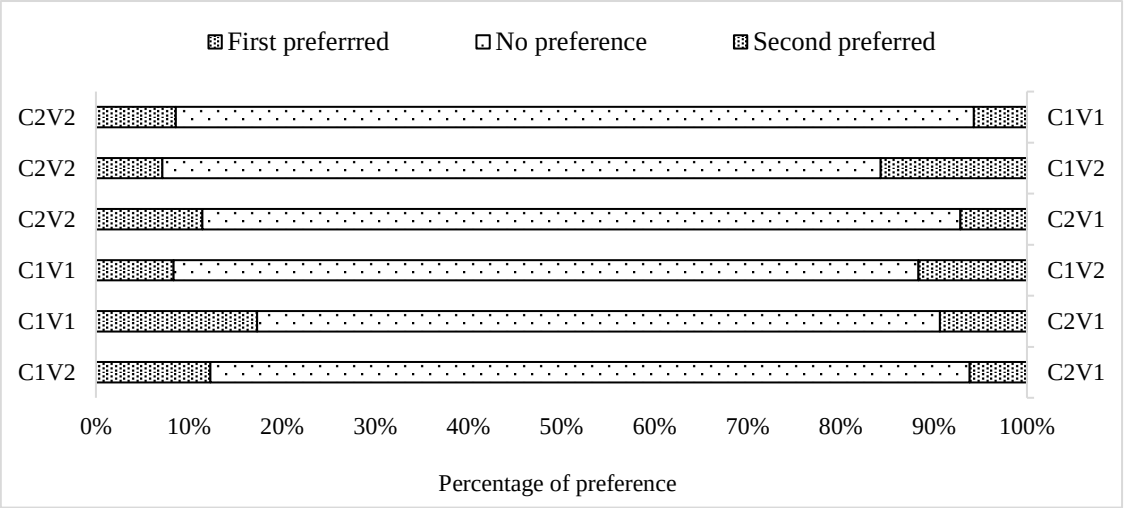

Figure 8: Results of preference test

5. Conclusions and perspectives

This paper studied different possible approaches of speech unit modelling for Arabic speech synthesis based on HMM (HTS). Both training and synthesis part of HTS system require a description of the text with a sequence of contextual features. These contextual features are used to build the decision trees during the training of the HMM models, and to predict the acoustic parameters to generate the waveform in the synthesis phase. Accordingly, the choice of the contextual features is decisive. Part of the contextual features are language dependent, and Arabic language presents two specific phenomena that have to be considered: gemination and vowel quantity.

These particularities of the Arabic language are not handled in the standard set of contextual features unlike stress. Consequently, two extra features were added to take into account these specificities. However, is it enough to add this information into the set of contextual features or should units be differentiated? In other words, should a geminated consonant (resp. a long vowel) be considered as a fully-fledged phoneme to be modelled with a specific unit, or are they the same (or very similar) phoneme as their simple (short) counterpart but with different duration (which can be modelled by the same unit)? To answer these questions, several possible modelling approaches of the speech segments have been investigated such as, the use of different units for modelling long vs. short vowels, and/or the use of different units for modelling simple vs. geminated consonants. These combinations were compared to another one, where a short vowel and its long counterpart are modelled with the same unit, and a geminated consonant and its simple counterpart are modelled with the same unit.

Objective measures show that segment durations generated with HTS for the four models are similar. This conclusion was validated by listening tests (MOS, DMOS and preference tests). Results showed that there is no prominent difference between the various modelling approaches. Thus, the identification of geminated consonants and long vowels as fully-fledged phonemes (hence modelled by specific units) is not compulsory when applying HTS for MSA, as long as this information exists in the set of contextual features.

Future work will explore Arabic speech synthesis based on deep learning approaches (Zen et al., 2013). Recently, different variants and architectures of Deep Neural Networks (DNN) have been introduced for speech synthesis (Zen & Sak, 2015; Wu et al., 2016), and the obtained results showed that DNN improve the quality of produced speech signals and their naturalness as well. Thus, it will be interesting to find out if DNN would, benefit from the explicit differentiation of geminated vs. simple consonants and long vs. simple vowels unlike HMM.

References

Abdel-Hamid, O., Abdou, S. M. & Rashwan, M. (2006). Improving Arabic HMM based speech synthesis quality. In Interpseech 2006, 9th Annual Conference of the International Speech Communication Association. Pittsburgh, Pennsylvania, USA.

Abdelmalek, R. & Mnasri, Z. (2016). High quality Arabic Text-to-speech synthesis using unit selection. In *SSD'2016, IEEE Conference on Signal, Systems and Devices*

Ahmed, B. (2004). Réalisation d'un système hybride de synthèse de la parole Arabe utilisant un dictionnaire de polyphones. In *JEP-TALN2004. Journées d'Etude sur la Parole*, Fès, Maroc.

Al-Ani, S. H. (1970). Arabic phonology: An acoustical and physiological investigation. In *ERIC*.

Black, A., Taylor, P., Caley, R., & Clark, R. (1998). The festival speech synthesis system.

Black, A. W., Zen, H. & Tokuda, K. (2007). Statistical parametric speech synthesis. In *ICASSP 2007, IEEE International Conference on Acoustics, Speech and signal Processing*. Honolulu, HI, USA, Vol. 4, pp. IV-1229.

Baloul, S. (2003). Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé. Doctoral dissertation, Le Mans.

Buchholz, S., & Latorre, J. (2011). Crowdsourcing preference tests, and how to detect cheating. In *INTERSPEECH'2011, Annual Conference of the International Speech Communication Association*.

Chalamandaris, A., Tsiakoulis, P., Karabetsos, S. & Raptis, S. (2013). The ILSP/INNOETICS Text-to-Speech System for the Blizzard Challenge 2013. In *The Blizzard Challenge 2013 workshop*. September 2013. Reykjavik, Iceland.

Cheffour, N., Benabbou, A. & Mouradi, A. (2000). Étude et Evaluation de la Di-Syllabe comme Unité Acoustique pour le Système de Synthèse Arabe PARADIS. In *LREC'2000, International Conference on Language Resources and Evaluation*, Athens, Greece.

Halabi, N. (2015). Modern standard Arabic speech corpus. Doctoral dissertation in University of Southampton.

Halabi, N. & Wald, W. (2016). Phonetic inventory for an Arabic speech corpus. In *LREC 2016, 10th International Conference on Language Resources and Evaluation*, Slovenia, pp. 734-738.

Halpern, J. (2009). Word stress and vowel neutralization in modern standard Arabic. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools.*

Hunt, A. J. & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP'1996, IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta Georgia, USA, Vol. 1, pp. 373-376.

ITU (1996). Recommendation P.800. Methods for subjective determination of transmission quality. International Telecommunication Union.

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.

Kawahara, H., Masuda-Katsuse, I. & De Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency- based F0 extraction: Possible role of a repetitive structure in sounds. In *Speech Communication*, vol. 27, pp. 187– 207.

Khalil, K. & Cherif, M.C. (2013). Arabic HMM-based speech synthesis. In *ICEESA'2013, International Conference on Electrical Engineering and software Applications*, Hammamet, Tunisia, pp. 1-5.

Khouja, M.K. & Zrigui, M. (2005). Durée des consonnes géminées en parole Arabe : mesures et comparaison. In *TALN-RECITAL 2005, Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Dourdan, France.

Kishore, S. P. & Black, A. W. (2003). Unit size in unit-selection speech synthesis. In *EUROSPEECH'2003, Eighth European Conference on Speech Communication and Technology*.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. In *The journal of the Acoustical Society of America*, vol. 67, pp. 971-995.

Koishida, K., Tokuda, K., Kobayashi, T. & Imai, S. (1997). Efficient encoding of mel-generalized cepstrum for CELP coders. In *ICASSP'1997, IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 1355-1358.

Kouloughli, D. (1976). Contribution à l'étude de l'accent en arabe littéraire. In *Annales de l'Université d'Abidjan Série H : Linguistique Abidjan*, volume 9, pp. 115-130.

Krstulovic, S., Hunecke, A. & Schroder, M. (2007). An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. In *EUROSPEECH'2007, European Conference on speech Communication and Technology*, vol. 7.

Laufer, A. & Baer, T. (1998). The emphatic and pharyngeal sounds in Hebrew and in Arabic. *Language and speech* 31 (2): 181-205.

Le Maguer, S., Barbot, N. & Boeffard, O. (2013). Evaluation of contextual descriptors for HMM-based speech synthesis in French. In *SSW'2013, ISCA Tutorial and Research Workshop on Speech Synthesis*, pp. 153-158, Barcelona, Spain.

Moulines, E., Emerard, F., Larreur, D., Le Saint Milon, J. L., Le Faucheur, L., Marty, F., Charpentier, F. & Sorin, C. (1990). A real-time French text-to-speech system generating high-quality synthetic speech. In *ICASSP'1990, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 309-312.

Newman, D. (1984). *The phonetics of Arabic. In Journal of the American Oriental Society*, vol. 46, pp. 1-6.

Rajouani, A., Najim, M., Chiadmi, D. & Zyoute, M. (1987). Synthesis-by-rule of Arabic language. In *EUROSPEECH'987, European Conference on Speech Technology*.

Schwarz, D., Beller, G., Verbrugghe, B., & Britton, S. (2006). Real-time corpus-based concatenative synthesis with catart. In *DAFx'2006, 9th International Conference on Digital Audio Effects*, pp. 279-282.

Selouani, S. A. & Caelen, J. (1998). Arabic phonetic features recognition using modular connectionist architectures. In *IVTTA'1998, IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 155-160, Torino, Italy.

Silén, H., Helander, E., Nurminen, J. & Gabbouj, M. (2010). Analysis of duration prediction accuracy in HMM-based speech synthesis. In *SP'2010, Speech Prosody*.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992). Tobi: A standard for labelling English prosody. In *ICSLP'1992, International Conference on Spoken Language Processing*, vol. 1, pp. 867–870.

Taylor, P. A., Nairn, I. A., Sutherland, A. M., Jack, M. A., Bagshaw, P. C., Renals, S. & Sutherland, A. M. (1991). A real time speech synthesis system. In *IEEE symposium*, pp. 101-106.

Taylor, P. (2009). *Text-to-speech synthesis*. In Cambridge University Press, Cambridge.

Tokuda, K., Zen, H. & Black, A. W. (2002). An HMM-based speech synthesis system applied to English. In *IEEE Speech Synthesis Workshop*, pp. 227-230.

Watts, O., Stan, A., Clark, R. A., Mamiya, Y., Giurgiu, M., Yamagishi, J., & King, S. (2013). Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found'data: evaluation and analysis. In *SSW'2013, 8th ISCA Speech Synthesis Workshop*.

Wu, Z., Watts, O. & King, S. (2016). Merlin: An open source neural network speech synthesis system. In *SSW'2016, 9th ISCA Speech Synthesis Workshop*, Sunnyvale, USA.

Young, S.J. (1994). The HTK hidden Markov model toolkit: Design and philosophy. Department of Engineering, Cambridge University, UK, Tech. Rep. TR.152.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *EUROSPEECH'1999, European Conference on Speech Communication and Technology*.

Zen, H., Toda, T., & Tokuda, K. (2006). The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. In *Proceedings Blizzard Challenge Workshop*.

Zen, H., Tokuda, K. & Black, A.W. (2009). Statistical parametric speech synthesis. In *Speech Communication*, vol. 51, no 11, pp. 1039-1064.

Zen, H. & Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *ICASSP'2015, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4470-4474.

Zen, H., Senior, A. & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *ICASSP'2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7962-7966.