Statistical modelling of speech units in HMM-based speech synthesis for Arabic

Amal Houidhek, Vincent Colotte, Zied Mnasri, Denis Jouvet, Imene Zangar

Statistical modelling of speech units in HMM-based speech synthesis for Arabic

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

[1]Electrical Engineering Departement, Ecole Nationale d'Ingénieurs de Tunis, University Tunis El Manar, Tunisia
[2]Inria, Villers-lès-Nancy, F-54600, France
[2]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[2]CNRS, LORIA, UMR 750, Villers-lès-Nancy, F-54600, France
{amal.houidhek,vincent.colotte,denis.jouvet}@loria.fr, zied.mnasri@enit.rnu.tn

Abstract

This paper investigates statistical parametric speech synthesis of Modern Standard Arabic (MSA). Hidden Markov Models (HMM)-based speech synthesis system relies on a description of speech segments corresponding to phonemes, with a large set of features that represent phonetic, phonologic, linguistic and contextual aspects. When applied to MSA two specific phenomena have to be taken in account, the vowel lengthening and the consonant gemination. This paper studies thoroughly the modeling of these phenomena through various approaches: as for example, the use of different units for modeling short vs. long vowels and the use of different units for modeling simple vs. geminated consonants. These approaches are compared to another one which merges short and long variants of a vowel into a single unit and, simple and geminated variants of a consonant into a single unit (these characteristics being handled through the features associated to the sound). Results of subjective evaluation show that there is no significant difference between using the same unit for simple and geminated consonant (as well as for short and long vowels) and using different units for simple vs. geminated consonants (as well for short vs. long vowels).

Keywords: speech synthesis, statistical modeling, Arabic language, speech unit modeling

1. Introduction

During the last decade, statistical parametric speech synthesis (SPSS) systems, such as HMM-based system called HTS (Black et al., 2007), have attracted a lot of interest compared to those based on unit selection (Hunt and Black, 1996). Actually, HTS system is able to produce a smooth and good quality speech. Besides, it presents the advantages of being trainable, having a small footprint and changing voice characteristics. This approach has been already applied to many languages, such as Japanese (Zen et al., 2009), English (Tokuda et al., 2002) and French (Le Maguer et al., 2013). HTS system is split into two main parts corresponding to training and synthesis (Zen et al., 2009). The training part consists of modeling spectrum features (e.g., Mel-cepstral coefficients and their dynamic features) and excitation features (including log (F0) and its dynamic features) from a speech database. In this stage, both linguistic and prosodic contexts are taken into account. Then, speech features are modeled by context-dependent HMMs. The synthesis part is a multi-step process: first an HMM representing the sentence is built by concatenating the context dependent phone HMMs, then, the HMMs' state durations are calculated through maximizing their probability, and speech level features (e.g., Mel-Cepstral coefficients, log(F0)…etc) are predicted. Finally, these features are processed by the synthesis filter, i.e., MLSA (Mel-Log Spectrum Approximation) to generate a waveform. In HTS, a speech unit is described with a set of contextual features at different levels: phoneme, syllable, word, phrase and utterance. These features are related to different aspects of a speech unit, whether linguistic, phonetic, phonologic or prosodic. These contextual features, which describe the characteristics of the speech segments associated to the phone units, are used during the HTS training process to build decision trees for parameter sharing. The relevance of the contextual features have a considerable effect on the HMM models and thus on synthesis quality. Part of the contextual features is language dependent and thus some descriptors may be added or neglected according to the language. In (Tokuda et al., 2002) a set of specific descriptors has been presented for English speech synthesis using HTS, whereas in (Le Maguer et al., 2013) a different set of descriptors was introduced for French. However, previous studies focused only on evaluating the impact of a chosen subset of descriptors on acoustic parameter modeling, like in (Le Maguer et al., 2013), and on modeling the duration of consonants and vowels (Silen et al., 2010).

The scope of this paper is to adapt the HTS system to the Arabic language through the introduction and the study of the relevance of some specific features. The originality of this work consists in evaluating the effects of many specific features combinations on the quality of HTS-synthesized Arabic speech. To extract relevant descriptors, a comprehensive review of Arabic language specifications (Al-Ani and Salman, 1970) was referred to.

In fact, MSA presents different specificities like stress (lexical stress), gemination and vowel lengthening. Long vowels are almost twice as long as short vowels and geminated consonants are twice as long as simple consonants (Khouja and Zrigui, 2005). Whereas stress was introduced in the original labels list of HTS (Tokuda et al., 2002) specific descriptors for gemination and vowel lengthening have not been so far considered in the standard set of descriptors for HTS. Thus, this paper expands and compares several choices of units that differentiate, or not, long vowels from short vowels, and/or geminated consonants from simple consonants. Using these different models of speech unit, Arabic utterances were produced for each combinations. Then, they were evaluated based on perceptive tests and on an objective evaluation of the duration of the phones.

The paper is organized as follows; section 2 presents different aspects of speech unit modelling in Arabic speech synthesis, particularly in SPSS, and then details the four proposed systems, developed for evaluating unit modelling in Arabic speech synthesis using HTS. Section 3 describes the evaluation protocol and data used in the experiments, then presents and discusses the results of the evaluation.

2. Speech-unit modeling for Arabic speech synthesis

Speech synthesis has been in continuous progress since the very early modern speech synthesizers (Klatt, 1980) and (Taylor et al., 1995) until latest techniques, like SPSS. Depending on approaches and languages, different speech units are used. A speech unit corresponds to a phonetic or a phonological segment, which could be a phone, a diphone (Moulines et al., 1990), a triphone or a syllable (Kishore and Black, 2003). The phoneme is considered as the smallest speech entity, any modification of a phoneme in a word may change its meaning. According to (Taylor, 2009) a phone is defined as a realization of a phoneme in a certain context.

2.1. Arabic speech synthesis

Arabic speech has always been catching up with each novel TTS (Text To Speech) technique, starting from articulatory speech synthesis, formant-based, concatenative and more recently statistical parametric speech synthesis. In (Rajouani et al., 1987), synthesis by rules is used to produce Arabic speech. Based on formants, i.e., maxima of the spectrogram, acoustic speech is generated through formants characteristics (i.e., bandwidth and amplitude) in addition to the rules of evolution of formants between phonemes. Concatenation-based methods consist in putting together a dynamically selected set of natural speech units to produce acoustic signals. Diphone units are speech signal segments going from the middle of one speech sound (phoneme) to the middle of the next one; diphone concatenation-based approaches lead to a better speech quality, compared to previous methods. In (Baloul, 2003) diphone concatenation is used to produce Arabic speech. Further researches, used larger speech units in concatenation-based methods, such as di-syllable in (Cheffour et al., 2000) and triphones in (Ahmed, 2004). A particular approach of concatenation method, and maybe the most successful one in terms of quality and naturalness, is based on unit selection. It consists in selecting speech units of different sizes, to be concatenated, in order to produce speech. Applied to Arabic language in (Abdelmalek and Mnasri, 2016), it was based on phonemes and syllables.

2.2. Statistical parametric speech synthesis (SPSS)

HTS was adapted to many languages, and different set of descriptors have been used corresponding to the characteristics of each targeted language. For French (Le Maguer et al., 2013), English (Tokuda et al., 2002) and German (Krstulovic, 2007) the phoneme was the speech unit. However, Japanese (Zen, 2006) Swedish (Lundgren, 2005) and Portuguese HTS-based speech synthesis systems were based on the syllable as a speech unit. Previous works on HTS system to produce Arabic speech used phonemes as speech units. In (Khalil and Cherif, 2013), the basic HTS system was adapted to MSA without explicitly considering the gemination or the vowel lengthening phenomena. (Abdel-Hamid et al., 2006) focused on some modification of the features and of the signal generation processing to improve the quality of produced speech.

2.3. Modern Standard Arabic phonology

MSA has 28 consonants (each consonant has a simple and a geminated version) and 3 vowels (they exist in both short and long version). Studies of Arabic phonology present specific characteristics of MSA (Newman, 1984), in particular gemination and vowel lengthening.

Gemination ("shadda") (Newman, 1984): In spelling, it is represented by adding a diacritic sign (ّ) above a consonant followed by a short vowel. Acoustically, it corresponds to making the duration of the geminated consonant twice as long as that of the simple consonant. A geminated consonant is phonetically and phonologically different from the same non-geminated one. This feature is considered as a distinctive factor, for example, a geminated /r/ changes the meaning of the word /darasa/ (he studied) to /darrasa/ (he taught).

Vowel lengthening ("madd") (Selouani and Caelen, 1998): It consists of lengthening the duration of the vowels. Whereas short vowels are not usually written, long vowels are always indicated by the following graphemes و /uu/, ي /ii/, ا/aa/. In addition, vowel lengthening changes the meaning of words, for example, if the initial vowel /a/ in the word /hatafa/ (he shouted) is lengthened, the resulting word /haatafa/ means (he telephoned).

2.4. Proposed unit modelling systems

To investigate the modeling of speech units in Arabic speech synthesis (normal vs. geminated consonant, short vowel vs. long vowel), a study of different choices of speech units is conducted. The aim is to investigate whether it is better to consider geminated consonants (resp. long vowels) as fully fledged phonemes (as implied by the Arabic phonology) or if both gemination and vowel lengthening can only be considered as specific linguistic features among the others?
Below, four sets of units corresponding to four possible combinations are presented:

1) Differentiating simple vs. geminated consonants and short vs. long vowels (C*2__V*2): In this first system, a simple consonant (e.g., /d/) and its geminated counterpart (e.g., /dd/) are modelled with different units. Similarly, for vowels, a short vowel (e.g., /a/) and a long one (e.g., /aa/) are represented by two distinct units. This leads to the most expanded system of speech unit modelling relying on two units per consonant and two units per vowel.

2) Differentiating simple vs. geminated consonants and merging long and short vowels (C*2__V*1): In this approach, a long vowel and its short counterpart are modelled with the same unit (for example the same model

for /aa/ and /a/). On the other hand, the model of a geminated consonant is kept different from the model of the corresponding simple consonant, thus leading to two units per consonant and one unit per vowel.

3) Merging simple and geminated consonants and differentiating short vs. long vowels (C*1__V*2): The third system separates long vowels from short ones; so for example, the long vowel /aa/ and the short vowel /a/ are modelled by two different units. However, simple consonants and their geminated counterparts are modelled by the same units. This leads to one unit per consonant and two units per vowel.

4) Merging simple and geminated consonant and merging short and long vowels (C*1__V*1): It is the most compact system. Here a simple consonant (e.g., /d/) and the corresponding geminated consonant (/dd/) are represented by the same unit. The same for vowels, a long vowel (e.g., /aa/) and the corresponding short one (/a/) are modelled by the same unit. This leads to one unit per consonant and one unit per vowel.

In addition, linguistic features are computed over the phoneme sequence of each stimuli in the training corpus. The set of contextual features for defining the phone labels is inspired from the one defined for English (Tokuda et al., 2002). Each segment is described with a set of features corresponding to different levels (phoneme, syllable, word, phrase, utterance). However, features related to the accent information and to TOBI (Silverman et al., 1992) are not employed in the proposed set. Indeed, the novelty is adding two additional features to take into account specificities of the Arabic language. Whereas the first feature is used to indicate the possible geminated characteristic (possible values are simple consonant, geminated consonant, or not a consonant), the second one refers to vowel lengthening (possible values are short vowel, long vowel, or not a vowel). Finally, it should be noted that the four systems use the same set of descriptors.

3. Evaluation

3.1. Data

For the development and the evaluation of the systems described above, a Modern Standard Arabic corpus is used (Halabi and Wald, 2016). It consists of 1595 separate utterances recorded from a male-speaker reading news bulletin, in a neutral style. The signal is sampled at 48 kHz. For speech synthesis, a speaker-dependent modelling is used, and one HTS model is developed (trained) for each of the four choices of units. For each system, 1565 Arabic utterances from the corpus are used for training the model parameters and 30 utterances are kept for evaluation purpose (test set). Signals are produced for each system using HTS with STRAIGHT vocoder (Kawahara et al., 1999).

For evaluating and comparing the different approaches, the 30 utterances of the test set were synthesized with each system, thus providing 30 speech stimuli for each system.

3.2. Results and discussions

In this experiment, seven native Arabic listeners, who neither are speech specialists nor involved in speech synthesis, participated in the tests.

3.2.1. Global quality and naturalness evaluation

Listeners were asked to assess two features. The first one is the overall quality which refers to the general quality of produced signals and if it was easy to listen to. The second one is the naturalness of the synthesized speech by focusing on the evaluation of the intonation (whether the pitch's evolution is natural) and of the rhythm, (whether the length of phonemes sounds natural too). Signals produced with the four proposed systems are assessed by giving a score on a scale from 1 to 5 ranging from very bad to excellent. Fig. 1 and Fig. 2 display the mean values of MOS listening tests with 95% confidence intervals, with respect to quality and naturalness.

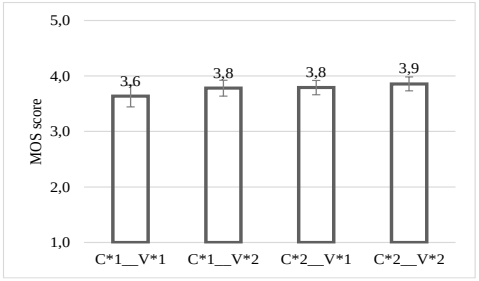

Fig.1 . Global quality results with 95% confidence intervals.

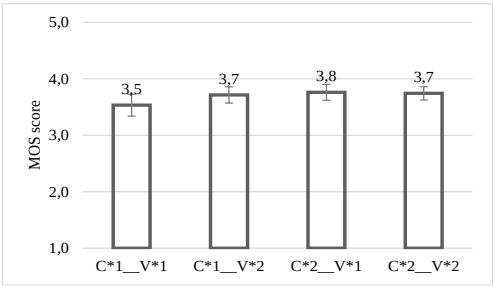

Fig.2 . Naturalness results with 95% confidence intervals.

According to the MOS scores, the four models lead to rather similar quality of speech signals. The similarity between quality and naturalness results can be explained by the fact that listeners, who neither are specialists in phonetics nor accustomed to speech evaluation, had not seen the difference between the global quality and naturalness questions even though each question was preceded with an introductory explanation.

3.2.2. Comparison test

Speech signals produced with the four approaches of speech unit modeling A (C*2__V*2), B (C*1__V*1), C (C*1__V*2) and D (C*2__V*1) are compared to each other.

Each pair consists of the same utterance produced with two different systems. During the test, the order of presenting the speech signals is randomly chosen for each trial. Listeners were asked to compare the second signal to the first one and to give a score from 1 to 7 ranging from much worse to much better.

The one-to-one comparison of the four systems in Fig. 3 shows that there is no clear preference for a particular system.

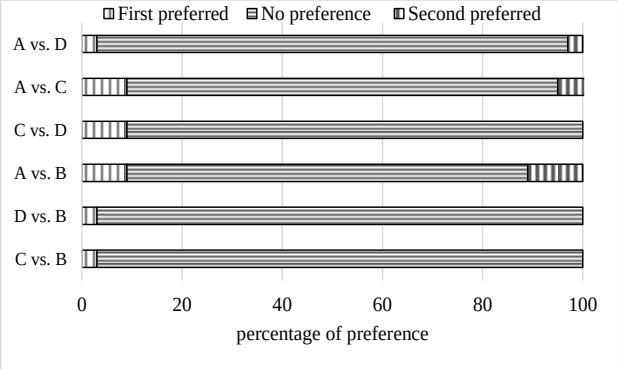

Fig. 3. Results of comparisons tests (percentage of preference)

3.2.4. Evaluation of duration

For each system, 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). These average ratios are then compared to those obtained for natural speech. Only phonemes with a number of occurrences greater or equal to ten are considered.

	Number of occurrences	LV / SV		GC / SC	
		262	884	104	1315
Models	C*1__V*1	1.8		2.2	
	C*1__V*2	1.9		2.2	
	C*2__V*1	1.8		2.1	
	C*2__V*2	1.8		2.0	
	Natural	2.0		2.1	

Table 1. Duration ratios

Table 1 shows that ratios of long vowel durations to short vowel durations, as well as ratios of geminated consonant durations to simple consonant durations, computed for the different synthesis systems are similar to the ratios calculated on natural speech.

To investigate the reasons of the similarity of ratios values of the four models, the duration decision trees were analysed for the four models. Note that there is only a single tree for all phonemes in each model. Fig. 4, which displays the top of the decision tree, shows that questions about the length of the current segment (i.e., geminated consonant or not: "C-Cur-Gem-Cons" and long vowel or not: "C-Seg ==vl") are placed on the top of the tree; this is observed for the four models.

Note that, "C-"refers to the current phoneme, "L-"to the left phoneme and "sil" to silence.

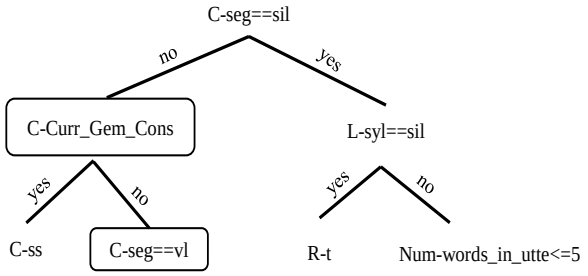

Fig. 4. Duration decision tree of (C*2__V*2) model (top part).

Root mean square error (RMSE) between natural durations and HTS generated durations (for each model) has been calculated for each phoneme in the test corpus.

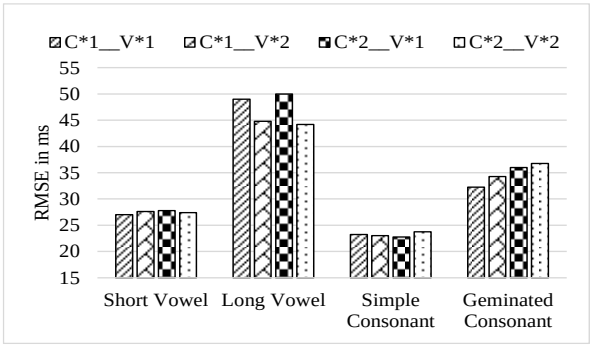

Fig. 5. RMSE between natural durations and HTS generated durations.

Fig. 5 displays the values of RMSE for four phoneme classes (short vowels, long vowels, simple consonants and geminated consonants). Results show that for each class, there is no important difference between RMSE values of the four models. Besides, when considering the mean duration values of each class, NRMSE (normalized root mean square error) values are similar, around 25% for geminated consonant and 35% for the 3 other classes.

4. Conclusion

In this paper, HMM based speech synthesis (HTS) was applied to Modern Standard Arabic. In Arabic language, two specific phenomena need to be considered: the consonant gemination and the vowel lengthening. Thus, two new features were included in the HTS descriptors to take into account these specificities.

Moreover, several modelling approaches have been investigated with respect to the choice of modelling units, as for example, 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 approaches were compared to another one, which merges short and long variants of a vowel into a single unit, and simple and geminated variants of a consonant into a single unit. Subjective listening evaluation tests (MOS and preference tests) showed that there is no significant difference between the various modelling systems.

When using HMM-based speech synthesis system for MSA, an identification of geminated consonants and long vowels as fully-fledged phonemes is not necessary, as long as this information exists in the set of the specific phonetic feature.

As DNN are now being used for speech synthesis, it will be interesting to study if, unlike for HMM speech synthesis, neural networks would benefit from the explicit differentiation of geminated vs. simple consonants and long vs. simple vowels.

5. Acknowledgements

This research work was conducted under PHC-Utique Program in the framework of CMCU (Comité Mixte de Coopération Universitaire) grant N°15G1405.

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, Leipzig, Germany.

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

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

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

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

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, Athènes, Greece.

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

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, Vol. 1, pp. 373-376. Atlanta Georgia, USA.

Kawahara, H., Masuda-katsuse, I., De Cheveign, 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 1. 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, pp. 1-5. Hammamet, Tunisia.

Khouja, M.K., Zrigui, M. (2005). Durée des consonnes géminées en parole arabe : mesures et comparaison. In TALN-RECITAL 2005, Dourdan, France.

Kishore, S. P., Black, A. W. (2003). Unit size in unit selection speech synthesis. In 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.

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 Proceedings of the European Conference on speech Communication and technology (Eurospeech) Citeseer, vol. 7.

Le Maguer, S., Barbot, N., Boeffard, O. (2013). Evaluation of contextual descriptors for HMM-based speech synthesis in French. In SSW 2013, pp. 153-158. Barcelona, Spain.

Moulines, E., Emerard, F., Larreur, D., Le Saint Milon, J. L., Le Faucheur, L., Marty, F. et al. (1990). A real-time French text-to-speech system generating high-quality synthetic speech. In ICASSP 1990, 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 European Conference on speech Technology.

RSHþHN, I. (1997). Syllable based speech synthesis. In WHP, vol. 267, no 1.

Selouani, S. A., Caelen, J. (1998). Arabic phonetic features recognition using modular connectionist architectures. In IVITA 1998, 4th Proceedings 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 Proceedings of speech prosody.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. (1992). Tobi: A standard for labeling English prosody. In CSLP 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 I {EEE} symposium Vol. 19, 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.

Zen, H. (2006). An example of context-dependent label format for HMM-based speech synthesis in English. The HTS CMUARCTIC demo, vol. 133. Lundgren, A. (2005). An HMM-based Text-to-speech system applied to Swedish. Master's thesis. Royal Institute of Technology, KTH, Sweden.

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