HMM-Based Speech Synthesis for the Greek Language

Sotiris Karabetsos, Pirros Tsiakoulis, Aimilios Chalamandaris, and Spyros Raptis

Institute for Language and Speech Processing (ILSP) / R.C. "Athena",
Voice and Sound Technology Department
Artemidos 6 & Epidavrou, Maroussi, GR 15125, Athens, Greece
{sotoskar,ptsiak,achalam,spy}@ilsp.gr
http://www.ilsp.gr

Abstract. The success and the dominance of Hidden Markov Models (HMM) in the field of speech recognition, tends to extend also in the area of speech synthesis, since HMM provide a generalized statistical framework for efficient parametric speech modeling and generation. In this work, we describe the adaption, the implementation and the evaluation of the HMM speech synthesis framework for the case of the Greek language. Specifically, we detail on both the development of the training speech databases and the implementation issues relative to the particular characteristics of the Greek language. Experimental evaluation depicts that the developed text-to-speech system is capable of producing adequately natural speech in terms of intelligibility and intonation.

Keywords: HMM, Speech Synthesis, Text to Speech, Greek Language, Statistical Parametric Speech Synthesis, Hidden Markov Model.

1 Introduction

Nowadays, the most common approach for achieving high quality near-natural speech synthesis is the corpus-based unit selection technique. In principle, this method presumes no adoption of any specific speech model and simply relies on runtime selection and concatenation of speech units from a large speech database using explicit matching criteria [1]. However, despite the dominance of corpus-based unit selection text to speech systems, there is an increased interest for speech synthesis based on an efficient model-based parametric framework. The capability of producing synthetic speech of high naturalness and intelligibility through a proper manipulation of the parameters of the model, offers significant advantages not only in controlling the synthesis procedure but also in easily adapting the technology in different languages, contexts and applications [2].

In parametric speech synthesis, the most common model is the, so-called, source-filter model of speech production [3]. Early developed speech synthesis techniques (e.g. formant synthesis) adopted this model in a rule-based framework so as to achieve general purpose speech synthesis. Recently, the same model has been efficiently adopted in a data-driven approach utilizing the statistical framework of the Hidden Markov Models (HMM) leading to the HMM-based speech synthesis [4,5]. Although the HMM-based speech synthesis framework is still not superior to the corpus-based approach, recent

P. Sojka et al. (Eds.): TSD 2008, LNAI 5246, pp. 349–356, 2008.

results have established it as one of the most successful parametric speech synthesis techniques [6,7].

Consequently, the HMM-based speech synthesis framework has been successfully applied in different languages where the emphasis is mostly put on investigating the particular language requirements in order to efficiently model the contextual information. Besides the Japanese language, examples of text to speech systems in other languages using HMM-based speech synthesis include, English [8], German [9], Chinese [10], Slovenian [11], Portuguese [12], Spanish [13], Korean [14], Arabic [15], to name a few. In most of these cases, the resulting text to speech system was found to perform well in terms of achieved quality. In this work, we describe the adaption, the implementation and the evaluation of the HMM speech synthesis framework for the case of the Greek language. Specifically, we report on both the development of the training speech databases and the implementation issues relative to the particular characteristics of the Greek language. The developed text to speech system is assessed by comparative subjective tests using both diphone and corpus-based unit selection speech synthesis systems. To our knowledge, this is the first report on the exploitation of the HMM speech synthesis framework for the Greek language.

The rest of the paper is organized as follows: in Section 2 the HMM framework for speech synthesis is briefly discussed while Section 3 describes its adaption for the case of the Greek language. Section 4, provides the experimental assessment and Section 5 concludes the work.

2 The HMM Speech Synthesis Framework

The schematic representation of the HMM-based speech synthesis framework is illustrated in Figure 1. As mentioned, the data-driven methodology is followed since the model is trained on a pre-recorded and annotated speech database. Thus, the framework consists of both a training part and a synthesis part.

The responsibility of the training part is to perform the statistical modeling of speech, based on characteristics extracted from the speech database. In particular, the extracted characteristics relate not only to prosodic (source) and vocal tract (spectral representation) information, but also extend on contextual and durational modeling [16,17]. More specifically, the HMM-based speech synthesis framework performs simultaneous modeling of pitch and spectrum taking into account the dynamics of both quantities as well. Spectral representation utilizes Mel-based cepstral coefficients while prosody is represented as logF0. Multi Space probability Distribution (MSD) modeling is performed to alleviate the problem of non continuous pitch values in unvoiced regions. Moreover, context clustering is performed using decision trees so as to fully exploit the contextual information in lexical and syntactic level [17]. Duration models are also constructed according to the method described in [16].

At the synthesis part, the input text is analyzed and converted to a context-dependent phoneme sequence. Next, the synthesis utterance is represented by the concatenation of the individual context-dependent HMMs. The speech signal is synthesized (reconstructed) from spectral coefficients and pitch values using the MLSA filter. Speech generation is guided in a way that the output probability for the HMMs is maximized [5,8].

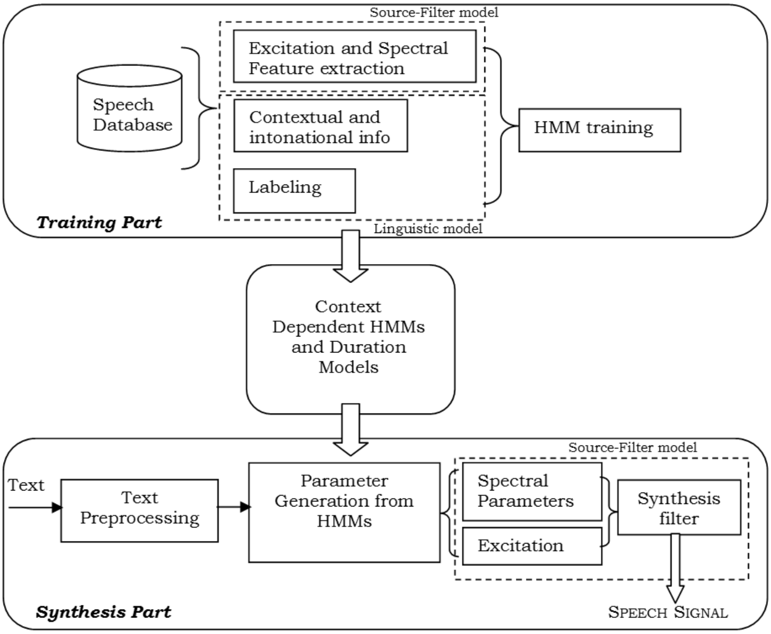

Fig. 1. The training and synthesis parts of the HMM speech synthesis framework

The inclusion of both spectral and prosodic dynamics (delta and acceleration coefficients) ensures a smooth and pragmatic speech generation. Since the source-filter model of speech production and reconstruction is employed, various types of excitation signals may be utilized which may result "buzzy" ([8]) and high quality speech [18].

3 HMM Speech Synthesis for the Greek Language

The HMM-based speech synthesis system for the Greek language follows the framework depicted in Figure 1. Its adaption considers the particular characteristics of the Greek language and mainly focuses on the analysis and the contextual modeling since contextual information is language dependent. However, the HMM framework provides a general setup for sufficient context modeling that can be easily adopted for different languages.

For the phonemic representation of Greek phonemes, a set of 37 phonemes was adopted that are divided in 26 consonants and allophones, 5 unstressed vowels, 5 stressed vowels and one for the silence (sentence beginning, sentence ending and pause). The set of the 37 phonemes define nine classes as follows:

- Unvoiced fricatives: /f/, /x/, /X/, /T/, /s/;
- Voiced fricatives: /v/, /J/, /j/, /D/, /z/;
- Liquids: /l/, /r/, /L/;
- Nasals: /m/, /M/, /n/, /N/, /h/;

- Unvoiced stops: /p/, /t/, /k/, /c/;
- Voiced stops: /b/, /d/, /g/, /G/;
- Silence: /-/;
- Stressed Vowels: /a'/, /e'/, /i'/, /o'/, /u'/;
- Unstressed Vowels: /a/, /e/, /i/, /o/, /u/.

An advantage of the Greek language is that the stress is either defined in free running text or can be extracted by the grapheme to phoneme rules, thus stressed vowels are represented uniquely by different phonetic symbols.

The contextual information that has been considered for the developed system accounts for both phonetic and linguistic information. As mentioned previously, since stressed vowels are represented uniquely, Tones and Break Indices (ToBI) analysis and annotation was partly involved. The list of the contextual factors that has been extracted and taken into account is the following:

- **Phonetic level**
  - The Phoneme identity
  - The identities of both the previous and the next phonemes and the phonemes before the previous and after the next phoneme
  - The position of the phoneme in the current syllable both forward and backward
- **Syllable level**
  - Determination whether the current, next and previous syllable is stressed
  - Number of phonemes in current, next and previous syllable
  - Position of the syllable both in phrase and word
  - Number of vowels in the syllable
  - Number of stressed syllables in the current phrase before and after the current syllable, and the number of syllables between the current and the previous and between the current and the next stressed syllables.
- **Word level**
  - The number of syllables in the current, next and previous word
  - The position of the word in the phrase (both forward and backward) and the sentence
- **Phrase level**
  - The number of syllables in the current, next and previous phrase
  - The position of the phrase in the sentence (both forward and backward)
- **Sentence level**
  - The number of phrases in the sentence
  - The number of words in the sentence
  - The number of syllables in the sentence

For the training of the HMMs, two speech databases were considered: one using a female and the other using a male speaker. Since both of the produced systems performed equivalently in terms of achieved quality, only the results for the female speaker are shown in the next section. The training speech database consisted of 1200 phonetically balanced reading-style utterances. The sampling frequency was 16KHz with 16 bits resolution. The segmentation and the labelling of the databases was performed automatically (by using HMMs) and manually refined. Notice that the same speech databases were utilized for the construction of the relevant corpus-based unit selection text to speech system. For the HMM synthesis, parameter extraction was carried out using a 25msec Blackman-windowed frame length at a 5msec frame rate. The spectral representation considers mel-based cepstral coefficients together with their delta and delta-delta values including the zeroth coefficient. Pitch is represented by using logF0 along with delta and delta-delta values as well. The HMM topology utilized was a 5-state left-to-right with no skip. Based on the database and the contextual factors considered, the HMM speech synthesis system resulted a total number of full context models of 116432, with 1293 questions for the construction of the state clustering decision trees.

4 Experimental Evaluation

The assessment of the produced HMM-based speech synthesis system was based on conducting small scale listening tests. The system was compared against two speech synthesis systems namely, a diphone-based and a corpus-based unit selection system. The diphone-based system has only one instance per diphone. Evaluation is based on the Mean Opinion Score (MOS) concerning the naturalness and the intelligibility (clearness) on a scale of one to five where one stands for "bad" and five stands for "excellent". A total of 15 no-domain specific sentences were synthesized by the three systems. Each sentence was 5 to 10 words long and was not included in the training database. A group of 10 listeners, comprised by both speech and non-speech experts were asked to express their opinion for each sentence in a MOS scale. For every sentence, the three versions (produced by each system) were presented to each listener in random order each time. Every group could be heard by the listeners more than once. The results of the test are summarized in Table 1, where the MOS is shown for both naturalness and intelligibility The results depict that the HMM-based system performs slightly better than the diphone-based one, in terms of overall quality. On the other hand, the corpus-based unit selection system outperforms both systems in overall quality. However, the HMM-based system achieves a good score in the achieved intelligibility and a fair score in naturalness. Nevertheless, the MOS result on naturalness, depends mostly on the deterioration due to vocoder-like speech generation and secondly on the resulting intonation. Thus, the adoption of a more sophisticated manipulation for the source-filter model (e.g. as in [18]) and the enhancement of prosodic information, could lead to a high quality HMM-based speech synthesis system. An example of HMM-based synthesis is shown in Figure 2, where both the spectrogram and the pitch contour are illustrated for the Greek utterance "kata ti Jno'mi mu eksi'su

Table 1. MOS-based comparative evaluation of the HMM-based speech synthesis system. The system is compared against both a diphone-based and corpus-based general purpose text to speech systems.

SYSTEM	NATURALNESS	INTELLIGIBILITY
HMM-based	3.9	4.2
Diphone-based	3.7	3.8
Corpus-based	4.5	4.6

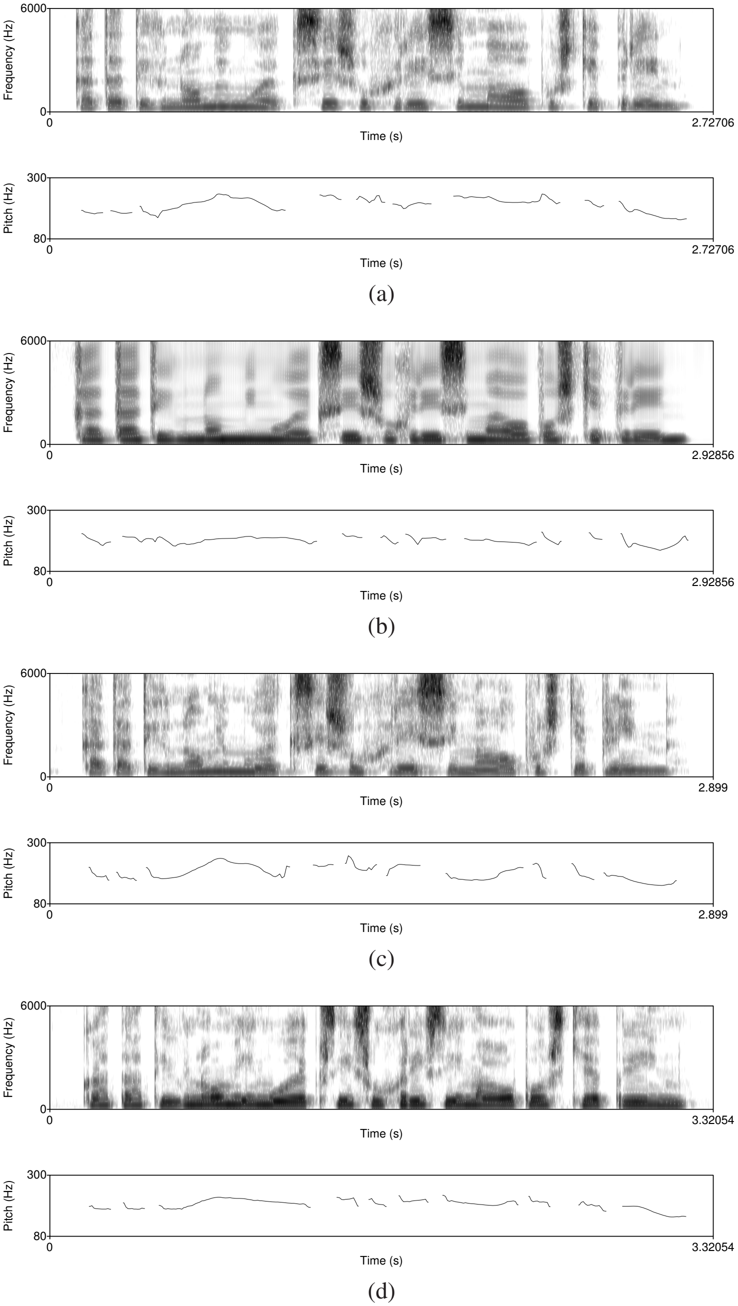

Fig. 2. Example of spectrogram and pitch contour of a synthesized utterance: a) natural speech signal, b) HMM-based, c) corpus-based and d) diphone-based

xri'sima o'pos ce pri'n". The figure also shows the original signal along with the synthesized version using the diphone-based and the corpus-based system. It is seen that HMM-based synthesis produces smooth speech signal but tends to follow a monotone pitch contour similarly to the diphone-based case. The opposite situation is true for the corpus-based approach. This result was expected since, in the training part, prosodic information was taken into account for stressed vowels only. Some examples of synthetic utterances produced by the three text-to-speech systems are available at the following URL address: `http://speech.ilsp.gr/TSD2008/samples.htm`

5 Conclusions and Further Work

HMM-based speech synthesis provides an efficient model-based parametric method for speech synthesis that is based on a statistical framework. In this work, the HMM-based speech synthesis framework was adapted for the development of text to speech system for the Greek language. The performance of the system has been evaluated through comparative listening tests, against a diphone-based and a corpus-based unit selection system. The MOS results have shown that the HMM system performs better than the diphone-based system but worse then the corpus-based one. This is mainly due to the reconstruction process for speech generation which leads to a vocoder-like style of speech. However, these results are encouraging since this is the first HMM speech synthesis system built for the Greek language and many improvements are possible. Further work entails the adoption of a more sophisticated model for speech reconstruction and the inclusion of more prosodic properties in order to increase the naturalness of the produced speech.

Acknowledgments. This work has been partially supported through Community and National funding under the "Information Society" programme of the Greek General Secretariat of Research and Technology, and through the Operational Programme Interreg III Á / PHARE CBC Greece Bulgaria 2000–2006. We would also like to thank all the persons involved in the listening tests and contributed to this work.

References

1. Hunt, A., Black, A.: Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: ICASSP 1996, Atlanta, pp. 373–376 (1996)
2. Black, A., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Hawaii, pp. 1229–1232 (2007)
3. Quatieri, T., F.: Discrete Time Speech Signal Processing, Principles and Practice. Prentice Hall, Upper Saddle River (2002)
4. Tokuda, K., Masuko, T., Yamada, T.: An Algorithm for Speech Parameter Generation from Continuous mixture HMMs with Dynamic Features. In: Proc. of Eurospeech (1995)
5. Tokuda, K., Yoshimura, K., Masuko, T., Kobayashi, T., Kitamura, T.: Speech Parameter Generation Algorithms for HMM-based Speech Synthesis. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000), pp. 1315–1318 (June 2000)

6. Zen, H., Toda, T.: An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In: Proc. of Interspeech 2005, Lisbon, pp. 93–96 (2005)
7. Zen, H., Toda, T., Tokuda, K.: The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. In: Proc. Blizzard Challenge 2006 (2006)
8. Tokuda, K., Zen, H., Black, A.: An HMM-based speech synthesis system applied to English. In: Proc. of IEEE Speech Synthesis Workshop 2002 (IEEE SSW 2002) (September 2002)
9. Krstulovic, S., Hunecke, A., Schroeder, M.: An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements. In: Proc. of Interspeech 2007, Antwerp (2007)
10. Qian, Y., Soong, F., Chen, Y., Chu, M.: An HMM-based Mandarin Chinese text-to-speech system. In: Huo, Q., Ma, B., Chng, E.-S., Li, H. (eds.) ISCSLP 2006. LNCS (LNAI), vol. 4274, pp. 223–232. Springer, Heidelberg (2006)
11. Vesnicer, B., Mihelic, F.: Evaluation of the Slovenian HMM-based speech synthesis system. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 513–520. Springer, Heidelberg (2004)
12. Maia, R., Zen, H., Tokuda, K., Kitamura, T., Resende, F., G, Jr.: Towards the development of a Brazilian Portuguese text-to-speech system based on HMM. In: Proc. of Eurospeech 2003, pp.2465–2468, Geneva (2003)
13. Gonzalvo, X., Iriondo, I., Socor, J., Alas, F., Monzo, C.: HMM-based Spanish speech synthesis using CBR as F0 estimator. In: ISCA Tutorial and Research Workshop on Non Linear Speech Processing - NOLISP 2007 (2007)
14. Kim, S.-J., Kim, J.-J., Hahn, M.-S.: HMM-based Korean speech synthesis system for handheld devices. IEEE Trans. Consumer Electronics 52(4), 1384–1390 (2006)
15. Abdel-Hamid, O., Abdou, S., Rashwan, M.: Improving Arabic HMM based speech synthesis quality. In: Proc. of Interspeech 2006, Pittsburg, pp. 1332–1335 (2006)
16. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. of Eurospeech 1999, pp. 2347–2350 (September 1999)
17. Yamagishi, J., Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: A context clustering technique for average voice models. IEICE Trans. Inf. & Syst. E86-D(3), 534–542 (2003)
18. Yamagishi, J., Zen, H., Toda, T., Tokuda, K.: Speaker-Independent HMM-based Speech Synthesis System – HTS-2007 System for the Blizzard Challenge 2007. In: Proc. of Blizzard Challenge 2007 workshop, Bonn, pp. 1–6 (2007)