Hidden-Markov-Model based speech synthesis in Hungarian

BÁLINT TÓTH, GÉZA NÉMETH

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics

{toth.b, nemeth}@tmit.bme.hu

Keywords: speech synthesis, Text-To-Speech (TTS), Hidden-Markov-Modell (HMM)

This paper describes the first application of the Hidden-Markov-Model based text-to-speech synthesis method to the Hungarian language. The HMM synthesis has numerous favorable features: it can produce high quality speech from a small database, theoretically it allows us to change characteristics and style of the speaker and emotion expression may be trained with the system as well.

1. Introduction

There are several speech synthesis methods: articulatory and formant synthesis (trying to model the speech production mechanism), diphone and triphone based concatenative synthesis and corpus-based unit selection synthesis (based on recordings from a single speaker). Currently the best quality is produced by the unit selection method, although the size of the corpus database is quite big (it may be in the gigabyte range), voice characteristics are defined by the speech corpus and these features cannot be changed without additional recordings, labelling and processing which increase the cost of generating several voices.

Text-to-speech (TTS) systems based on the Hidden-Markov-Model (HMM) are categorized as a kind of unit selection speech synthesis, although in this case the units are not waveform samples, but spectral and prosody parameters extracted from the waveform. HMMs are responsible for selecting those parameters which most precisely represent the text to be read. A vocoder generates the synthesized voice from these parameters. HMM-based text-to-speech systems are becoming quite popular nowadays because of their advantages: they are able to produce intelligible, naturally sounding voice in good quality and the size of the database is small (1,5-2 Mbytes). It is also possible to adapt the voice characteristics to different speakers with short recordings (5-8 minutes) [1-4], and emotions can also be expressed with HMM TTS [5].

The current paper gives an overview about the architecture of HMM based speech synthesis, investigates the first adaptation of an open-source HMM-based TTS to Hungarian, and describes the steps of the adaptation process. The results of a MOS-like test are also introduced and future plans of the authors are mentioned as well.

2. The basics of Hidden-Markov-Models

Hidden-Markov-Models are mainly used for speech recognition [6] in speech research, although in the last decade or so they have also been applied to speech synthesis. The current section briefly introduces the basics of HMMs, a detailed description can be found in [7].

A HMM $\lambda(A,B,\pi)$ is defined by its parameters: A is the state transition probability, B is the output probability and π is the initial state probability. In case of text-to-speech let us assume that λ is a group of HMMs, which represent quintphones (a quintphone is five phones in a sequence) in sequence (*Fig. 1*). The series of quintphones define the word, which we would like to generate. The goal is to find the most probable state sequence of state feature vectors **X**, which will be used to generate the synthetic speech (see Section 3 for more details).

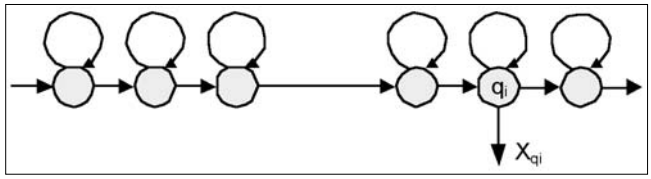

Figure 1. Concatenated HMM chain in q_i state, at i^{th} time, the output is X_{qi}.

The X_{qi} output is an M-dimensional feature vector at state q_i of model λ:

$$X_{q_i} = (x_1^{(q_i)}, x_2^{(q_i)}, x_3^{(q_i)} \ldots x_M^{(q_i)})^T$$

The aim is to define the $\underline{x} = (X_{q_1}, X_{q_2} \ldots X_{q_L})$ output feature vector (of length is L, which is the number of sounds/phonemes in an utterence) from model λ, which maximizes the overall likelihood $P(x|\lambda)$:

$$\underline{x} = \arg\max_x \{P(x|\lambda)\} = \arg\max_x \left\{ \sum_Q P(x|q,\lambda) P(q|\lambda) \right\},$$

Where $Q = (q_1, q_2 \ldots q_L)$ is the state sequence of model λ. The $P(x|\lambda)$ overall likelihood can be computed by adding the product of joint output probability $P(x|q,\lambda)$ and state sequence probability $P(q|\lambda)$ over all possible Q state-sequence.

To be able to compute the result within moderate time, the Viterbi-approximation is used:

$$\underline{x} \approx \arg\max_x \{P(x|q,\lambda,L) P(q|\lambda,L)\}$$

The $\underline{q}$ state sequence of model λ can be maximized independently of $\underline{x}$:

$$\underline{q} = \arg\max_{q} \{P(q|\lambda, L)\}$$

Let us assume that the output probability distribution of each state q_i is a Gaussian density function with μ_i mean value and Σ_i covariance matrix. The model λ is the set of all mean values and covariance matrices for all N states:

$$\lambda = (\mu_1, \Sigma_1, \mu_2, \Sigma_2 \ldots \mu_N, \Sigma_N)$$

Consequently the log-likelihood function is:

$$\ln\{P(x|q,\lambda)\} = -\frac{LM}{2}\ln\{2\pi\} - \frac{1}{2}\sum_{t=1}^{L}\ln\{|\Sigma_{q_t}|\} - \frac{1}{2}\sum_{t=1}^{L}(x_t - \mu_{q_t})^T \Sigma_{q_t}^{-1}(x_t - \mu_{q_t})$$

If we maximize x then the solution is trivial: the output feature vector equals to the states' mean values

$$\underline{x} = (\mu_{q_1}, \mu_{q_2} \ldots \mu_{q_L})$$

This solution is not representing the speech well because of discontinuities at the state boundaries. The feature vectors must be extended by the delta and delta-delta coefficients (first and second derivatives):

$$\underline{x} = ((x_{q_i})^T, (\Delta x_{q_i})^T, (\Delta^2 x_{q_i})^T)$$

3. HMM-based speech synthesis

HMM-based speech synthesis consists of two main phases: the training phase *(Fig. 2)* and the speech generation phase *(Fig. 3)*. During training the HMMs "learn" the spectral and prosodic features of the speech corpus, during speech generation the most likely parameters of the text to be synthesized are extracted from the HMMs.

For training a rather large speech corpus, the phonetic transcription of it and the precise position of the phoneme boundaries are required. The mel cepstrum, its first and second derivatives, the pitch, its first and second derivatives are extracted from the waveform. Then the phonetic transcription should be extended by context dependent labelling (see Subsection 4.2). When all these data are prepared, training can be started. During the training phase the HMMs are "learning" the spectral and excitation parameters according to the context dependent labels. To be able to model parameters with varying dimensions (e.g. $\log\{F_0\}$ in case of unvoiced regions) multidimensional probability density functions are used. Each HMM has a state duration density function to model the rhythm of the speech.

There are two standard ways to train the HMMs: (1) with a 2-4 hour long speech corpus from one speaker or (2) with speech corpora from more speakers and then adapt it to a speaker's voice characteristics with a 5-8 minute long speech corpus [1,2]. This way new voice characteristics can be easily prepared with a rather small speech corpus. According to [1,2] the adaptive training technique produces better quality than training from one speech corpus only. Furthermore there are numerous methods to change the voice characteristics [3,4].

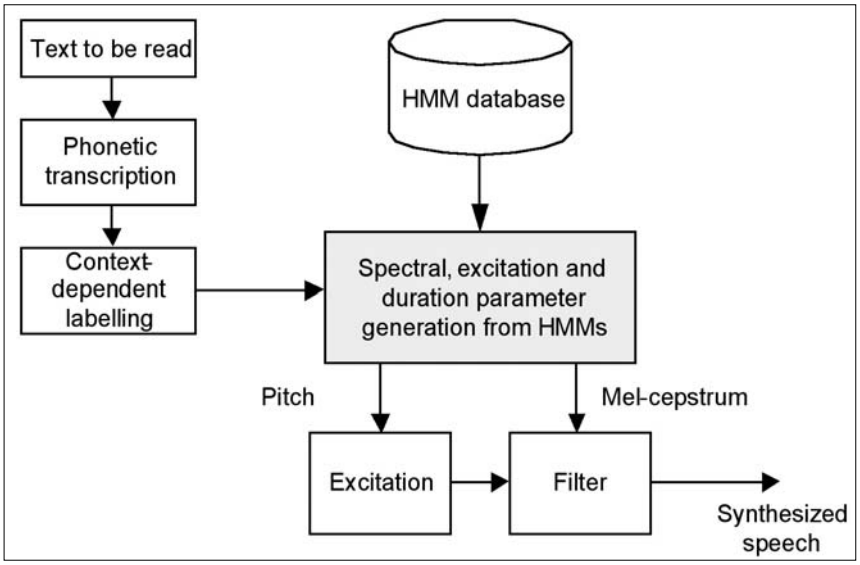

Figure 3. Speech generation with HMMs

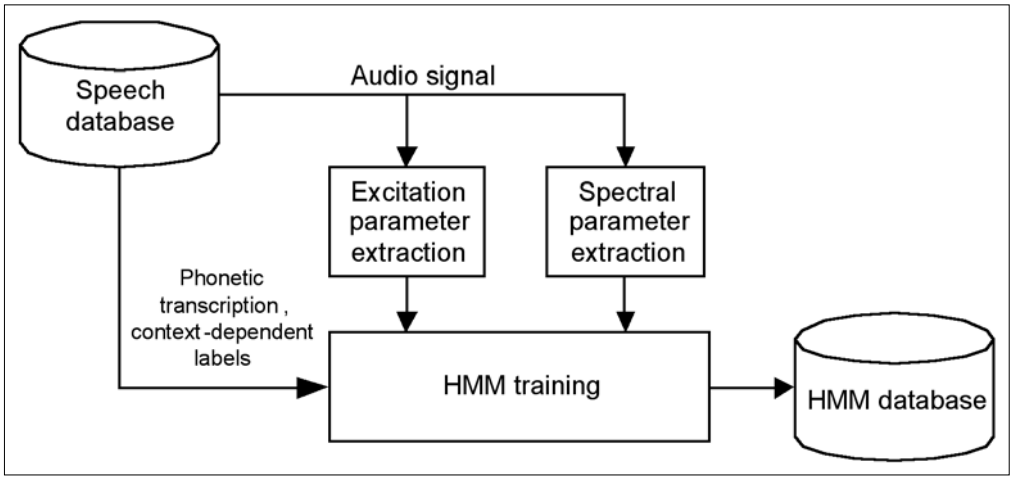

Figure 2. Training of the HMMs

Sounds	• The two previous and the two following sounds/phonemes (quintphones). Pauses are also marked.
Syllables	• Mark if the actual/previous/next syllable is accented. • The number of phonemes in the current/previous/next syllable. • The number of syllables from/to the previous/next accented syllable. • The vowel of the current syllable.
Word	• The number of syllables in the current/previous/next word. • The position of the current word in the current phrase (forward and backward).
Phrase	• The number of syllables in the current/previous/next phrase. • The position of the current phrase in the sentence (forward and backward).
Sentence	• The number of syllables in the current sentence. • The number of words in the current sentence. • The number of phrases in the current sentence.

Table 1.
The prosodic features used for context dependent labelling

To generate speech after training is done, the phonetic transcription and the context dependent labelling (see Subsection 4.2) of the text to be read must be created. Then the phones' duration is extracted from the state duration density functions and the most likely spectral and excitation parameters are calculated by the HMMs. With these parameters the synthetic speech can be generated: from the pitch value the excitation signal of voiced sections is generated and then it is filtered, typically with a mel log spectrum approximation (MLSA) filter [8]. Simple vocoders were used earlier, lately mixed excitation models are applied in order to achieve better quality [9].

4. Adapting HMM-based TTS to Hungarian

The authors conducted the experiments with the HTS framework [10]. For the Hungarian adaptation a speech database, the phonetic transcription of it, the context-dependent labelling and language specific questions for the decision trees were necessary. In the following the most important steps will be described.

4.1 Preparation of the speech corpus

The authors used 600 sentences to train the HMMs. All the sentences were recorded from professional speakers, it was resampled at 16.000 Hz with 16 bits resolution. The content of the sentences is weather forecast and the total length is about 2 hours. The authors prepared the phonetic transcription of the sentences, the letter and word boundaries were determined by automatic methods, which are described in [11].

4.2 Context-dependent labelling

To be able to select to most likely units, a number of phonetic features should be defined. These features are calculated for every sound.

Table 1 summarizes the most important features.

Labelling is done automatically, which may include small errors (e.g. defining the accented syllables), although it does not influence the quality much, as the same algorithm is used during speech generation, thus the parameters will be chosen by the HMMs consistently (even in case of small errors in the labelling).

4.3 Decision trees

In Subsection 4.2 context-dependent labelling was introduced. The combination of all the context-dependent features is a very large number. If we take into account the possible variations of quintphones only, even that is over 160 million and this number increases exponentially if other context dependent features are included as well. Consequently, it is impossible to design a natural speech corpus, where all the combinations of context-dependent features are included. To overcome this problem, decision tree based clustering [12,13] is used.

As different features influence the spectral parameters, the pitch values and the state durations, decision trees must be handled separately. *Table 2* shows which features were used in case of Hungarian for the decision trees [14].

For example, if the decision tree question regarding the length of the consonants is excluded, then the HMMs will mostly select short consonants even for long ones, as these are not clustered separately and there are much more short consonants in the database.

4.4 Results

In order to be able to define the quality of Hungarian HMM-based text-to-speech synthesis objectively, the authors conducted a MOS (Mean Opinion Score) like listening test. Three synthesis engines were included in the test: a triphone-based, a unit selection system and a HMM-based speech synthesis engine.

Table 2.
Features used for building the decision trees

Phonemes	• vowel / consonant • short / long • stop / fricative / affricative / liquid / nasal • front / central / back vowel • high / medium / low vowel • rounded / unrounded vowel	
Syllable	• stressed / not stressed • numeric parameters	*(see Table 1.)*
Word	• numeric parameters	*(see Table 1.)*
Phrase	• numeric parameters	*(see Table 1.)*
Sentence	• numeric parameters	*(see Table 1.)*

At the beginning of the test 3-3 sentences randomly generated from each system were played, these sentences were not scored by the subjects. The reason for doing this is to have the subjects used to synthetic voice and to show them what kind of qualities they can expect.

At the next step 29 sentences generated by each system were randomly played in different sequences in order to avoid 'memory effects' [15]. The content of the test sentences were from the weather forecast domain. The triphone based system is the ProfiVox domain independent speech synthesizer, the HMM based TTS was trained with weather forecast sentences (cca. 600 sentences) and the unit selection system had a large weather forecast speech corpus (including cca. 7000 sentences). The same 29 sentences were generated by all systems, but none of these sentences were present in the speech corpora. The subjects scored the sentences from 1 to 5 (1 was the worst, 5 was the best).

12 subjects were included in the test. The results are shown in *Fig. 4*.

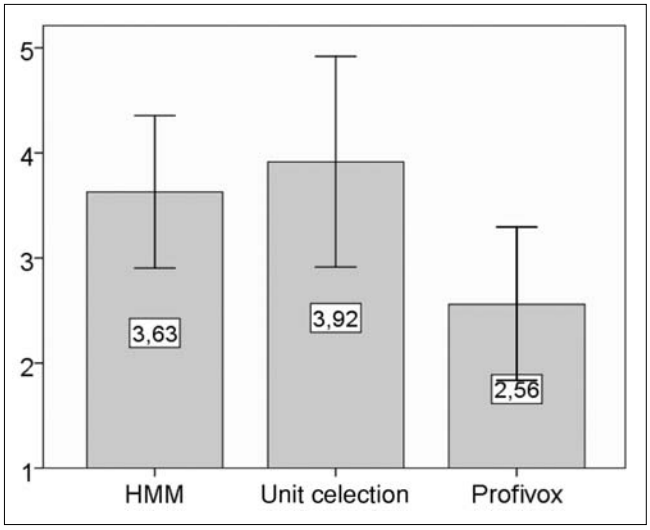

Figure 4.
The results of the MOS like test

The triphone-based system scored 2.56, the HMM-based TTS scored 3.63 and the unit selection one scored 3.9 on average (mean value). The standard deviation was in the same order 0.73, 1 and 0.73. Although the unit selection system was considered better then the HMM-based TTS, it can read only domain-specific sentences with the same quality, while the HMM-based TTS can read any sentence with rather good quality. Furthermore the database of the unit selection system includes more then 11 hours of recordings, while only 1.5 hours of recordings was enough to train the HMMs. The size of the HMM database is under 2 Mbytes, while the unit selection system's database is over 1 GByte.

The triphone based system was designed for any text, domain specific information is not included in the engine. This can be one reason for the lower score. The absolute value of the results is not so important, rather the ratio of them contains the most relevant information.

5. Future plans

The current paper introduced the first version of the Hungarian HMM-based text-to-speech system. As the next step the authors would like to record additional speech corpora in order to test adaptive training, to achieve more natural voice and to be able to create new voice characteristics and emotional speech with small (5-8 minutes long) databases.

Because of the small footprint and the good voice quality, the authors would like apply the system on mobile devices as well. To be able to port the hts_engine to embedded devices, it may occur, that the core engine must be optimized to achieve real-time response on mobile devices.

6. Summary

In this paper the basics of Hidden-Markov-Models were introduced, the main steps of creating the first Hungarian HMM-based text-to-speech synthesis system were described and a MOS-like test was presented. The main advantage of HMM-based TTS systems is that they can produce natural sounding voice from a small database, it is possible to change the voice characteristics and to express emotions.

Acknowledgements

We would like to thank the test subjects for their participation. We acknowledge the support of Mátyás Bartalis for creating the Web-based MOS-like test environment and the help of Péter Mihajlik in using the Hungarian speech recognition tools. The research was partly supported by the NKTH in the framework of the NAP project (OMFB-00736/2005).

Authors

BÁLINT PÁL TÓTH graduated in 2005 and received a diploma with honours from the Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics. He continued his research on speech synthesis and multimodal user interfaces as Ph.D. student right after the diploma. His main research topics are Hidden Markov-Model based speech synthesis and multimodal user interfaces on mobile devices.

GÉZA NÉMETH graduated from the Budapest University of Technology and Economics, Faculty of Electrical Engineering, in 1983, and obtained an engineering specialization diploma in 1985. He worked as a development engineer in BEAG Electroacoustic Factory between 1985 and 1987. Dr. Németh has been with the Department of Telecommunications and Media Informatics at Budapest University of Technology and Economics since 1987 where has been teaching measurement technologies, communication systems, telecommunications, signal processing, telecommunication management, speech information systems. He is also leading the Speech Technology Laboratory. Dr. Németh has been instrumental in transferring speech research results into practice, several applications have been developed under his leadership.

References

[1] T. Masuko, K. Tokuda, T. Kobayashi, S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," Proceedings of ICASSP, 1997, pp.1611–1614.

[2] M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," Proceedings of ICASSP, 2001, pp.805–808.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," Proceedings of Eurospeech, 1997, pp.2523–2526.

[4] M. Tachibana, J. Yamagishi, T. Masuko, T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," IEICE Trans. Inf. & Syst., Vol. E88-D, No.11, 2005. pp.2484–2491.

[5] S. Krstulovic, A. Hunecke, M. Schroeder, "An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements," Proceedings of Interspeech, 2007.

[6] Mihajlik P., Fegyó T., Németh B., Tüske Z., Trón V., "Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages: Hungarian ASR for the MALACH Project," In: Matousek V., Mautner P. (eds.), Text, Speech and Dialogue: 10th International Conf., TSD 2007, Pilsen, Czech Republic, September 2007. Proceedings, Berlin; Heidelberg: Springer, Lectures Notes in Computer Sciences, pp.342–350. (Lecture Notes in Artificial Intelligence, p.4629.)

[7] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, 77 (2), Febr. 1989, pp.257–286.

[8] S. Imai, "Cepstral analysis synthesis on the mel frequency scale" Proceedings of ICASSP, 1983, pp.93–96.

[9] R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," Proceedings of Interspeech, Aug. 2007, pp.1909–1912.

[10] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, "The HMM-based speech synthesis system v.2.0," Proceedings of ISCA SSW6, Bonn, Germany, Aug. 2007.

[11] Mihajlik, P., Révész, T., Tatai, P., "Phonetic transcription in automatic speech recognition," Acta Linguistica Hungarica, 2003, Vol. 49., No.3/4, pp.407–425.

[12] J. J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD dissertation, Cambridge University, 1995.

[13] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), Vol. 21., No.2, 2000. pp.79–86.

[14] Gósy M., Fonetika, a beszéd tudománya, Budapest, Osiris Kiadó, 2004.

[15] Jan P. H. van Santen, Perceptual experiments for diagnostic testing of text-to-speech systems, Computer Speech and Language, 1993, pp.49–100.