HIDDEN MARKOV MODEL BASED SPEECH SYNTHESIS SYSTEM IN SLOVAK LANGUAGE WITH SPEAKER INTERPOLATION

Martin SULÍR, Jozef JUHÁR
Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic
martin.sulir@tuke.sk, jozef.juhar@tuke.sk

ABSTRACT

This paper describes the first experiments with speaker interpolation in Slovak speech synthesis system. The interpolation provides an approach to voice characteristic conversion for hidden Markov model based text-to-speech synthesis system. The main idea of this technique is to synthesize an artificial speech with unseen and untrained output speech characteristic by interpolating of the existing sets of the pretrained models. The use of this technique allows to create new voices without the need to add additional data into training procedure. This is a major advantage especially for the low resources languages, such as Slovak language, where it is often difficult to obtain the necessary amount of data. The obtained results shows that the characteristics of the synthesized speech is changed from one male speaker to another one with the help of the interpolation ratio by which the Slovak voices with the new characteristics may be crated.

Keywords: *Hidden Markov models, models interpolation, statistical parametric speech synthesis, text-to-speech*

1. INTRODUCTION

The speech synthesis systems are today represented mainly by the computer systems which are able to convert the input text into output audio file which represents the speech. The use of this type of systems in practise is often faced with the reluctance of the users to communicate with the device, whereas the synthetic speech acts unnatural and it creates a barrier in communication [1]. That is why research in this field is oriented on the development of new advanced methods and their improvement in order to make final output speech from these systems as similar to the human interpretation as possible. The hidden Markov model (HMM) based speech synthesis method represents one of the most progressive approach how to convert written text into speech [2]. The progressiveness of this method is particularly involved in its high flexibility, where it allows to quite easily convert the voices with the help of an adaptation [3], interpolation [4] or, for example, using the technique of eigenvoice [5]. The utilization of these techniques arise from the using of hidden Markov models which can be properly mathematically modified in order to obtain their desired modified versions.

The interpolation allows to synthesized speech with unseen and untrained speaker's characteristics by modifying the HMM parameters among some pretrained speaker's HMM sets. It is possible to gradually change the characteristics of synthesized speech from one to another speaker with the help of the interpolation ration between the sets of the HMM models. The major advantage of this approach is that no further data are necessary what is an effective way how to create new voices especially for the low resources languages, such as Slovak language, where it is often difficult to obtain the necessary amount of data.

The major contribution of this work is an extension of the existing sets of the models for the Slovak HMM-based speech synthesis. It is very important to have as much models as it is possible particularly in the case of low resource languages such as Slovak language.

This paper is organized as follows: In Section 2 the HMM-based text-to-speech system with interpolation is described. Section 3 describes the speaker interpolation. In section 4, the experiments and results are presented and the conclusions are listed at the end of this paper.

2. HMM-BASED SPEECH SYNTHESIS SYSTEM WITH INTERPOLATION

A block diagram of the HMM-based speech synthesis system is shown in Figure 1. The system consists of three parts, namely there is the training, the interpolation and the synthesis part [4].

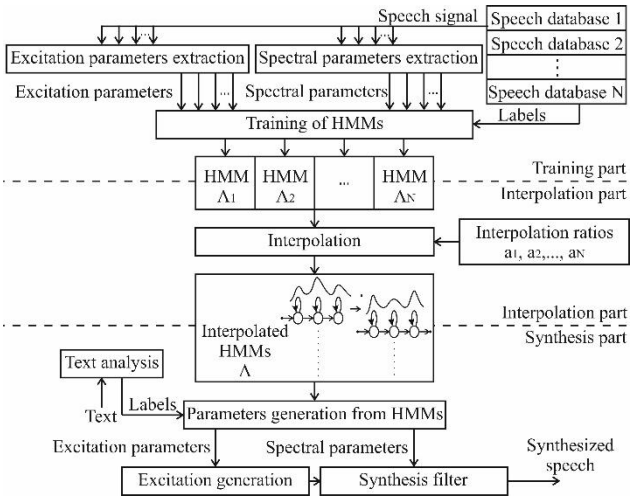

Fig. 1 Block diagram of speech synthesis system with interpolation

The training part consists of the training of the sets of HMM models [6]. Its main task is an extraction of the spectral and the excitation parameters from the speech databases as well as an implementation of the HMMs training. In case of the interpolation based training, the speech database consists of the multiple sub-databases,

where each of them was recorded by one certain speaker. The training procedure is carried out for each sub database separately and this process results in the set of individual HMM models. In the HMM-based speech synthesis method, each of the HMMs correspond to a left-to-right model where each output vector is composed of two components. It consists of the spectrum part, represented by the mel-cepstral coefficients and their delta and delta-delta coefficients and the excitation part which is represented by the excitation parameters and their corresponding delta and delta-delta dynamic features.

The second part of the system is interpolation step. The main task of this stage is interpolating representative HMM sets. It is necessary to generate a new HMM set by interpolating between the representative speaker's HMM sets with an arbitrary interpolation ratio. The following chapter describes the interpolation process in more detail.

The synthesis part consists of two main components. The first one is represented by the text analyser, which converts a given text into contextual label sequence. The second component is represented by the several blocks which are responsible for the parameter and the duration generation from the HMMs and excitation generation based on the generated excitation parameters. The vectors of mel-cepstral coefficients and logarithmic values of the fundamental frequency are generated based on the obtained HMM sequence and the speech waveform is synthesized from these vectors by using the speech synthesis filter, which represents the last block of the synthesis stage. A conventional system works as a mel-cepstral vocoder with a simple impulse train as the excitation signal, where a sequence of the periodic pulses and white noise together with a mel-log spectrum approximation (MLSA) filter are used [6]. Recently, a several high-quality vocoders with a more advanced excitation were implemented into system. Such methods include, e.g., MELP (Mixed Excitation Linear Prediction) method [7], HSM (Harmonic/Stochastic Model) model [8], excitation model based on modeling of residues [9], STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of Weighted Spectrum) [10] or AHOcoder [11].

Recently, the HMM-based speech synthesis technique has been reported for many languages, such as for large languages including the Mandarin Chinese [12] or Spanish [13], but the flexibility in the development of those systems also enabled the integration of small languages, such as Thai [14], Slovenian [15] and also Slovak.

3. SPEAKER INTERPOLATION

If the specific speakers who enter to the training procedure are marked as $S_1, S_2, \ldots, S_N$ and the HMM models pertaining to them are marked as $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$ and if the target speaker S is represented by the set of models Λ. Then the distance between the target speaker S and each of the specific speakers S_N can be measured by *Kullback* information measure between Λ and Λ_k as follows [16]:

$$I(\Lambda, \Lambda_k) = E_O\left[P(\mathrm{O}\,|\,\Lambda) \log \frac{P(\mathrm{O}\,|\,\Lambda)}{P(\mathrm{O}\,|\,\Lambda_k)} \right]. \quad (1)$$

When we consider interpolating between N HMM sets $\Lambda_1, \Lambda_2, \ldots, \Lambda_N$ with the weights $a_1, a_2, \ldots, a_N$, it is possible to determine interpolated HMM set of models Λ in a way that Λ maximize the cost function:

$$\varepsilon = \sum_{k=1}^{N} a_k I(\Lambda, \Lambda_k). \quad (2)$$

If we consider that each HMM state has a single Gaussian output probability density then it is necessary only to interpolate the N Gaussian probability density functions (pdf),

$$p_k(o) = \mathrm{N}(o; \mu_k, U_k), \quad (3)$$

where $k = 1,2,\ldots,N$ and μ_k together with U_k stand for mean vector and covariance matrix. The interpolated pdf is then determined by minimizing the function:

$$\varepsilon = \sum_{k=1}^{N} a_k I(p, p_k). \quad (4)$$

Subsequently, the *Kullback* information formula can be rewritten as follows:

$$I(p, p_k) = E_o\left[\mathrm{N}(o; \mu, U) \log \frac{\mathrm{N}(o; \mu, U)}{\mathrm{N}(\mu_k, U_k)} \right]. \quad (5)$$

Figure 2 shows the spatial view on the interpolation between the specific speakers and the target speaker.

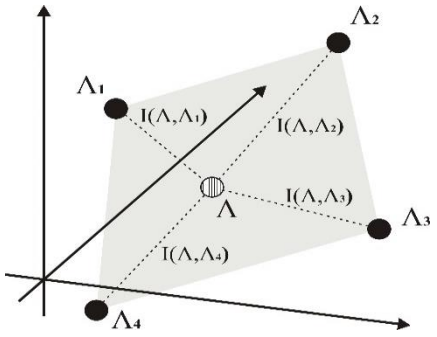

Fig. 2 The spatial view on the interpolation HMM models

The use of interpolation in case of the HMM-based speech synthesis thus enables the smooth change of the speaker characteristics with the help of the interpolation ratio between the speakers [17]. It is also possible to smoothly change from one to another speaker or change the level of the emotion in case of the emotional speech synthesis.

4. EXPERIMENTS

Together, two previously trained HMM-based voices, marked as *MSM* and *ADM*, were used and evaluated for the Slovak language using speaker interpolation technique [18]. These system use previously developed modules for Slovak text analysis together with the proposed language

dependent context clustering. For these experiments, the 5 state left-to-right models together with the conventional MLSA filter approach were used. Five different types of synthesized speech were generated with the help of the two HMM pretrained HMM sets and three newly created. The interpolation ratio was set as follows: $(MSM, ADM) = (1,0), (0.75,0.25), (0.5,0.5), (0.25,0.75), (1,1)$. Table 1. shows the description of *MSM* and *ADM* voices.

Table 1 Description of the initial HMM-based Slovak voices

Speech synthesis voice	Vocoder	Input corpus	Parameters
Male voice (MSM)	Mel – generalized cepstrum	3667 phonetically balanced sentences, 4 hrs 19 mins	MGC: 34 + f0: 1, Pulse plus noise excitation
Male voice (ADM)	Mel – generalized cepstrum	330 phonetically balanced sentences, 38 mins	MGC: 34 + f0: 1, Pulse plus noise excitation

4.1. ABX listening test

The evaluation of newly created interpolated samples was performed by the subjective listening tests. The ABX listening assessment was used for this purpose. In these test, five new sentences were evaluated where they were different from the training data. The basic idea of the test is that the three speech samples are used in the evaluation procedure. The speech samples A and B correspond to the synthesized speech samples from the original HMM sets of voices without interpolation. In our case, the samples A and B correspond to the generated sentences from the system *MSM* and *ADM*. This two samples represent some kind of the reference in the evaluation procedure. The assessed sample, marked as X, is represented by one of the interpolated synthesized speech sample with the particular interpolation ratio. The role of the participants in this evaluation is to consider, if the sample X is closer to the sample A or B. In our experiments, we had together 13 participants who evaluated the five submitted samples with different interpolation ratio. Figure 3 shows the experimental results of the ABX listening test.

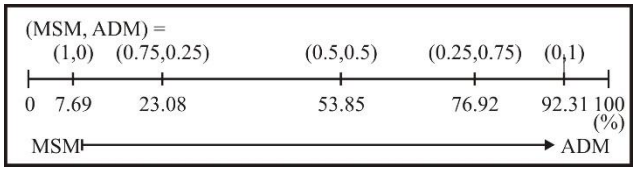

Fig. 3 Results of ABX listening test

The horizontal axis represents the rate that speech samples from the newly created interpolated HMM models were evaluated to be closer to the HMM models of the ADM speaker, and vice versa. It is evident that the interpolated synthesized samples faithfully represent the characteristics of the particular interpolation ratio. As can be seen, the largest deviation occurs when the interpolation ratio is equal to $(1,0)$, respectively $(0,1)$. The deviation still can be considered as a small, which could be caused, for example, by inattention of the participants and so on. The results for other interpolation ratios can be considered as very good.

4.2. Comparison of the generated fundamental frequency

Figure 4. shows the generated fundamental frequency of the interpolated speech samples. The five plots from top to bottom represent the interpolation ratio between the previously described two sets of the HMM models. The same Slovak sentence "*Syntéza slovenčiny*" was synthesized in each sample. As can be seen, the overall decrease of the frequency is evident when the interpolation ratio is changing from one speaker to another. The significant difference is noticeable especially at the beginning of the frequency contour where the two gaps were formed. In general, the fundamental frequency interpolation can be considered as appropriate.

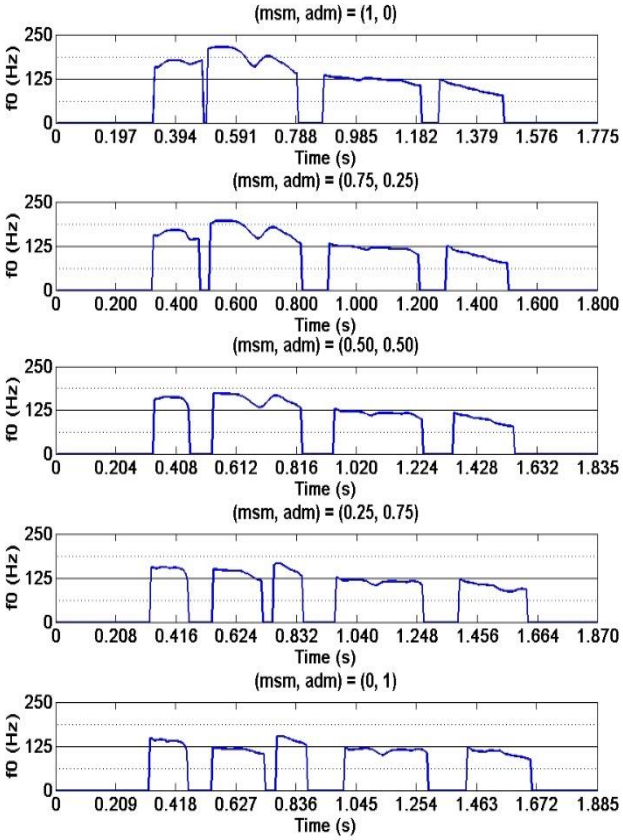

Fig. 4 Generated fundamental frequency for a sentence "Syntéza reči"

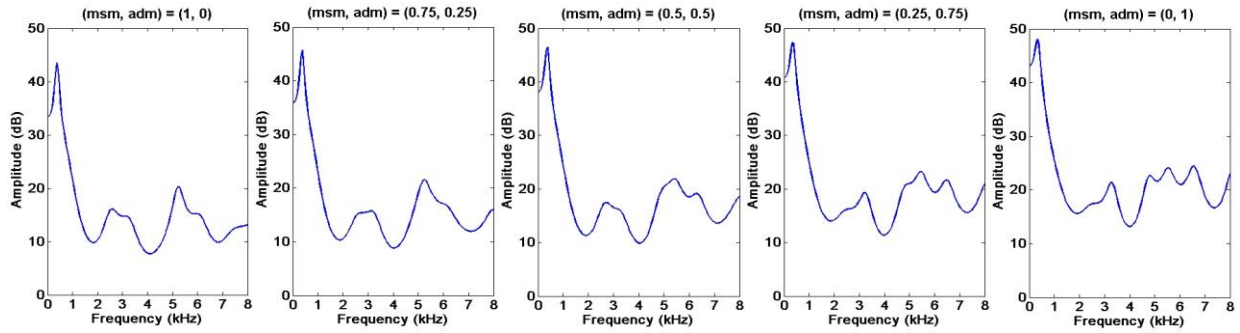

Fig. 5 Overall spectrum of a sentence "Syntéza slovenčiny"

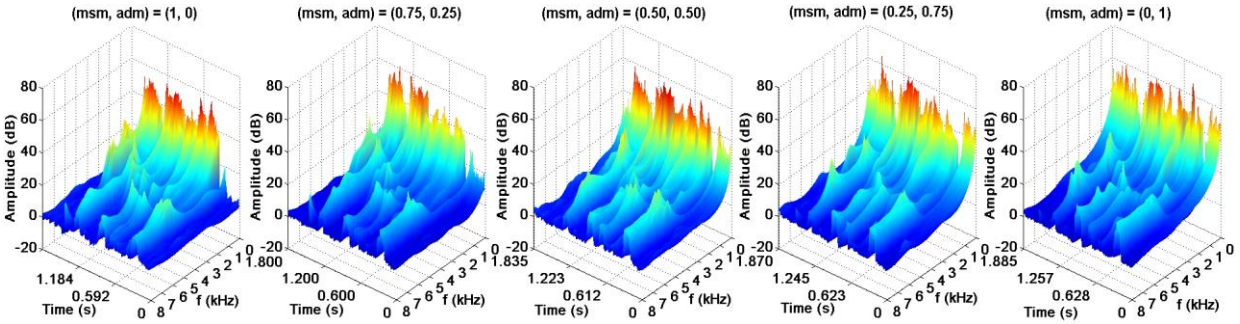

Fig. 6 Spectrum change in time of a sentence "Syntéza slovenčiny"

4.3. Comparison of the generated spectra

The generated spectrum of the synthesized interpolated speech is shown in Figure 5 and 6. The same sentence as in the case of the fundamental frequency was used for this purpose and only the frequency range from 0 to 8 kHz is shown because only there was a significant difference between the speech samples.

Figure 5 shows the overall spectrum of the synthesized samples where it is evident that the amplitude is smoothly changing from one speaker to another.

Figure 6 shows the spectrum change in time where the spectrum in each frame is demonstrated.

5. CONCLUSIONS

This paper presents the first experiments with the speaker interpolation for Slovak HMM-based speech synthesis system. Together, three new sets of models have been created with different interpolation ratio between the two previously trained models of the synthesized speech. The obtained results showed the potential of new voice creation with this technology. The interpolation provides an efficient technology how to create new Slovak voices with the combination of the pretrained HMM sets which is a great advantage in the developing and expanding of the Slovak text-to-speech systems.

ACKNOWLEDGMENTS

This publication is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF.

REFERENCES

[1] TAYLOR, P.: Text-to-speech synthesis. Cambridge: Cambridge University Press, 2009.

[2] KING, S.: Measuring a decade of progress in text-to-speech. Loquens, Vol.1, No.1, 2014.

[3] YAMAGISHI, J. – KOBAYASHI, T.: Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training. In: Journal IEICE - Transactions on Information and Systems archive, Vol.E90-D, No.2, pp. 533-543, 2007.

[4] YOSHIMURA, T. – MASUKO, T. – TOKUDA, K. – KOBAYASHI, T. – KITAMURA, T.: Speaker interpolation in HMM-based speech synthesis system. In: Proc. of Eurospeech, pp. 2523-2526, 1997.

[5] SHICHIRI, K. – SAWABE, A. – TOKUDA, K. – MASUKO, T. – KOBAYASHI, T. – KITAMURA, T.: Eigenvoices for HMM-based speech synthesis. In: Proc. of ICSLP, pp. 1269-1272, 2002.

[6] TOKUDA, K. – NANKAKU, Y. – TODA, T. – ZEN, H. – YAMAGISHI, J. – OURA, K.: Speech Synthesis Based on Hidden Markov Models. In: Proc. of the IEEE, Vol.101, No.5, pp. 1234-1252, 2013.

[7] YOSHIMURA, T. – TOKUDA, K. – MASUKO, T. – KOBAYASHI, T. – KITAMURA, T.: Mixed excitation for HMM-based speech synthesis. In: Proc. of the Eurospeech, pp. 2259-2262, 2001.

[8] ERRO, D. – MORENO, A. – BONAFONTE, A.: Flexible harmonic/stochastic speech synthesis. In: Proc. of the 6th ISCA Workshop on Speech Synthesis, pp. 194-199, 2007.

[9] MAIA R.S. – TODA, T. – ZEN, H. – NANKAKU, Y. – TOKUDA, K.: An excitation model for HMM-based speech synthesis based on residual modeling. In: Proc. of the 6th ISCA Workshop on Speech Synthesis, pp. 131-136, 2007.

[10] KAWAHARA, H.: Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. In: Acoustical science and technology, Vol.27, No.6, pp. 349-353, 2006.

[11] ERRO, D. – SAINZ, I. – NAVAS, E. – HERNAEZ, I.: Harmonics plus noise model based vocoder for statistical parametric speech synthesis. In: IEEE Journal of Selected Topics in Signal Processing, Vol.8, No.2, pp. 184-194, 2014.

[12] QIAN, Y. – SOONG, F. – CHEN, Y. – CHU, M.: An HMM-based Mandarin Chinesen text-to-speech system. In: Proceedings of the ISCSLP 2006, pp. 223-232, 2006.

[13] GONZALVO, X. – IRIONDO, I. – SOCORO, J.C. – ALIAS, F. – MONZO, C.: HMM-based Spanish speech synthesis using CBR as F0 estimator. In: Proceedings of the ISCA Tutorial and Research Workshop on Non Linear Speech Processing NoLISP 2007, pp. 7-10, 2007.

[14] CHOMPHAN, S. – KOBAYASHI, T.: Implementation and Evaluation of an HMM-based Thai Speech Synthesis System. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association, pp. 2849-2852, 2007.

[15] VESNICER, B. – MIHELIC, F.: Evaluation of the Slovenian HMM-based speech synthesis system. In: Proceedings of the TSD 2004, pp. 513-520, 2004.

[16] YOSHIMURA, T. – MASUKO, T. – TOKUDA, K. – KOBAYASHI, T. – KITAMURA, T.: Speaker interpolation for HMM-based speech synthesis system. In: Journal of Acoustical Society of Japan, Vol.21, No.4, pp. 199-206, 2000.

[17] TACHIBANA, M. – YAMAGISHI, J. – MASUKO, T. – KOBAYASHI, T.: Speech synthesis with various emotional expressions and speaking styles by style Interpolation and morphing. In: Journal IEICE - Transactions on Information and Systems archive, vol.E88-D, no.11, pp. 2484-2491, 2005.

[18] SULÍR, M. – JUHÁR, J.: Speaker adaptation for Slovak statistical parametric speech synthesis based on hidden Markov models. In: Radioelektronika 2015. - Piscataway : IEEE, pp. 137-140, 2015.

Received October 2, 2015 , accepted February 9, 2016

BIOGRAPHIES

Martin Sulír was born in Poprad, Slovakia in 1988. He received his M.Sc. (Ing.) degree in the field of Telecommunications in 2012 at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice. He is currently PhD student at the Department of Electronics and Multimedia Communications at the Technical University of Košice. His research interests include text to speech synthesis systems.

Jozef Juhár was born in Poproč, Slovakia in 1956. He graduated from the Technical university of Košice in 1980. He received Ph.D. degree in Radioelectronics from Technical university of Košice in 1991, where he works as a full professor at the Department of electronics and multimedia communications. He is author and co-author of more than 200 scientific papers. His research interest includes digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.