Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 1589 – 1596

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2016, 5-7 September 2016, York, United Kingdom

Learning HMM state sequences from phonemes for speech synthesis

Giorgio Biagetti[a], Paolo Crippa[a,*], Laura Falaschetti[a], Simone Orcioni[a], Claudio Turchetti[a]

[a] *DII – Department of Information Engineering, Università Politecnica delle Marche, via Brecce Bianche, 12, I-60131 Ancona, Italy*

Abstract

This paper presents a technique for learning hidden Markov model (HMM) state sequences from phonemes, that combined with modified discrete cosine transform (MDCT), is useful for speech synthesis. Mel-cepstral spectral parameters, currently adopted in the conventional methods as features for HMM acoustic modeling, do not ensure direct speech waveforms reconstruction. In contrast to these approaches, we use an analysis/synthesis technique based on MDCT that guarantees a perfect reconstruction of the signal frame feature vectors and allows for a 50% overlap between frames without increasing the data rate. Experimental results show that the spectrograms achieved with the suggested technique behave very closely to the original spectrograms, and the quality of synthesized speech is conveniently evaluated using the well known Itakura-Saito measure.

Peer-review under responsibility of KES International

Keywords: Learning, HMM, Speech synthesis, EM estimation, MDCT, MFCC ;

1. Introduction

Hidden Markov model (HMM) statistical parametric speech synthesis has proven to be a particularly flexible and robust framework to generate synthetic speech with various speaking styles and emotional expression[1,2]. Thanks to the ability in representing not only the phoneme sequences but also various contexts of the linguistic specification, HMM-based speech synthesis has recently been a major topic in speech research systems[3,4,5,6,7].

In conventional techniques based on the source-filter model assumption, phonetic and prosodic information are assumed to be conveyed primarily by the spectral envelope, fundamental frequency (F0), and the duration of individual phones[8]. However although these efforts have produced good performances, there are still limitations in this approach. In particular the modeling of F0 is difficult due to the discontinuity nature of F0 caused by the voice and unvoiced speech regions[9]. Moreover the spectral envelope defines a non-invertible transform so that the speech signal cannot be perfectly reconstructed from the feature sequence[10,11].

In this paper a novel HMM statistical parametric speech synthesis approach, based on learning HMM state sequences from phonemes and the modified discrete cosine transform (MDCT), which guarantees the perfect recon-

* Corresponding author. Tel.: +39-071-220-4541 ; fax: +39-071-220-4464.
E-mail address: p.crippa@univpm.it

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.206

struction of speech signal given the feature sequence and overcomes the main lacks of Mel-cepstral analysis/synthesis technique, is proposed.

2. Speech vector sequence generation

2.1. MDCT feature vector

Let us represent the sampled signal S as a sequence of $T+1$ blocks of D samples:

$$S = [s_1^T, s_2^T, \ldots, s_{T+1}^T]^T \in \mathbb{R}^{(T+1)D\times 1} \quad , \tag{1}$$

where

$$s_t \in \mathbb{R}^{D\times 1} \tag{2}$$

is the single block of length D.

In signal sampling with overlap, a sequence of frames

$$X = [x_1^T, x_2^T, \ldots, x_T^T]^T \in \mathbb{R}^{T(2D)\times 1} \quad , \tag{3}$$

is obtained, where

$$x_t = \begin{pmatrix} x_t^L \\ x_t^R \end{pmatrix} = \begin{pmatrix} s_t \\ s_{t+1} \end{pmatrix} \in \mathbb{R}^{2D\times 1}, \quad t = 1, \ldots, T \tag{4}$$

is the single frame corresponding to a window of length $2D$.

The sequences S, X, and the overlap regions are depicted in Fig. 1. As you can see the blocks x_t and x_{t+1} overlap for a length D, and the following condition holds:

$$x_t^R = x_{t+1}^L. \tag{5}$$

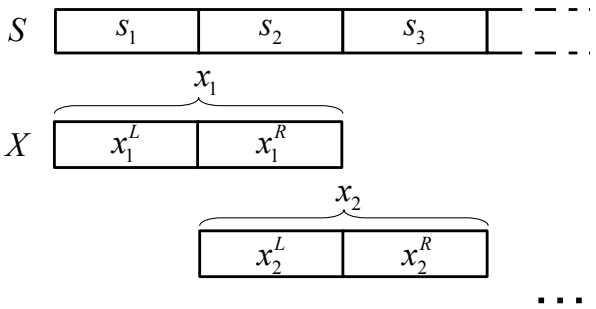

Fig. 1. The sequences S, X, and the overlap regions between different blocks.

The usually adopted model for speech parametrization is the source-filter model which leads to the extraction of parameters (features) such as linear predictive coding (LPC), Mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP) coefficients, etc. Among these, MFCCs are demonstrated the most successful due to their particular robustness to the environment and flexibility[12]. MFCC feature extraction corresponds to a transform F such that

$$\hat{o}_t = F x_t \tag{6}$$

where the vector $\hat{o}_t$ represents the so-called feature vector belonging to an appropriate subspace.

The main problem in speech synthesis is that, given the vector $\hat{o}_t$ from transcription, the frame signal x_t cannot be derived univocally from (6) because the transform F is not invertible. In order to face this problem we use an analysis/synthesis technique based on the MDCT that ensures a perfect reconstruction of the signal from feature vectors and allows for a 50% overlap between blocks without increasing the data rate.

Denoting with $A = (A_1 A_2) \in \mathbb{R}^{D\times 2D}$ the matrix that represents the MDCT [13], and with o_t the MDCT feature vector, it results

$$o_t = Ax_t = A \begin{pmatrix} s_t \\ s_{t+1} \end{pmatrix} = (A_1 A_2) \begin{pmatrix} s_t \\ s_{t+1} \end{pmatrix} = A_1 s_t + A_2 s_{t+1} \tag{7}$$

where A_1 , $A_2 \in \mathbb{R}^{D\times D}$. In matrix form we have

$$O = WS \tag{8}$$

with

$$W = \begin{pmatrix} A_1 & A_2 & \cdots & \cdots & 0 \\ 0 & A_1 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & A_1 & A_2 \end{pmatrix} \in \mathbb{R}^{TD\times(T+1)D} \tag{9}$$

and

$$O = [o_1^T , o_2^T , \ldots , o_T^T]^T \in \mathbb{R}^{TD\times 1} \tag{10}$$

is the MDCT feature vector corresponding to the signal S.

2.2. Learning HMM state sequences and maximum likelihood estimation

The speech synthesis algorithm we propose determines the sequence X of the synthetic signal, given the sequence O of features corresponding to the transcription (or sequence of phonemes) H to be synthesized.

In an HMM modeling we need to derive first the state sequence that generates the sequence O. To this end let

$$P(O, Q/\lambda) = \pi_{\theta_0} \prod_{t=1}^{T} a_{\theta_{t-1}\theta_t} b_{\theta_t}(o_t) \tag{11}$$

be the joint pdf of O and Q, given the model λ, where

$$Q = \{\theta_1, \theta_2, \ldots, \theta_T\} = \{(q_1, i_1), (q_2, i_2), \ldots, (q_T, i_T)\}, \tag{12}$$

being $\theta_t = (q_t, i_t)$ the substate associated to the Gaussian mixture i_t of the state q_t at the time instant t, that is

$$b_{\theta_t}(o_t) = (2\pi)^{-D/2} \left|U_{\theta_t}\right|^{-1/2} \cdot \exp\left\{-\frac{1}{2}(o_t - \mu_{\theta_t})^T U_{\theta_t}^{-1}(o_t - \mu_{\theta_t})\right\} \tag{13}$$

with $\mu_{\theta_t} \in \mathbb{R}^{D\times 1}$, $U_{\theta_t} \in \mathbb{R}^{D\times D}$. π_{θ_0} is the initial-state probability, and $a_{\theta_{t-1}\theta_t}$ is the state-transition probability.

Since $H = \{h_1, h_2, \ldots\}$ is a sequence of phonemes, we restrict the mathematical formulation to a single phoneme h alone. Given the phoneme h the sequences O and Q are chosen in such a way the joint pdf

$$P(O, Q/\lambda) = P(O/Q, \lambda)P(Q/\lambda), \tag{14}$$

which represents the likelihood of the set $\chi = \{O, Q\}$, is maximum. The sequence Q is obtained during learning phase as the one that satisfies $\max P(Q/\lambda)$. At the end of training to a given h corresponds a set $\{Q_1, Q_2, \ldots\}$ of substate sequences, thus we choose Q as the one that satisfies

$$Q = Q_{best} = \arg\max_i P(Q_i/\lambda). \tag{15}$$

Having derived Q, the sequence O is given by the maximum of the likelihood $\log P(O/Q, \lambda)$ which can be written as

$$\mathcal{L}(O) = \log P(O/Q, \lambda) = \sum_{t=1}^{T} \log b_{\theta_t}(o_t). \tag{16}$$

After some manipulations we have

$$\mathcal{L}(O) = -\frac{1}{2} O^T U^{-1} O + O^T U^{-1} M + k \tag{17}$$

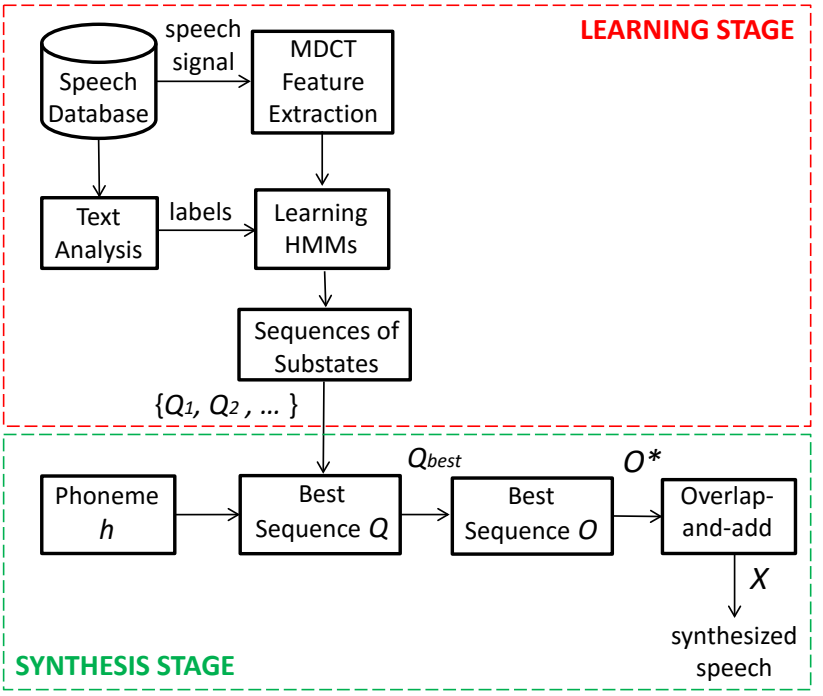

Fig. 2. Block diagram of the MDCT-based speech synthesis system.

where

$$U^{-1} = \operatorname{diag}\left[U^{-1}_{q_1,i_t}, U^{-1}_{q_2,i_t}, \ldots, U^{-1}_{q_T,i_t}\right] \in \mathbb{R}^{TD\times TD}, \quad M = \left[\mu^T_{q_1,i_t}, \mu^T_{q_2,i_t}, \ldots, \mu^T_{q_T,i_t}\right]^T \in \mathbb{R}^{TD\times 1} \tag{18}$$

and

$$k = k' + k'', \tag{19}$$

being

$$k' = \sum_{t=1}^{T} \log\left(2\pi\right)^{-D/2} \left|U_{\theta_t}\right|^{-1/2}, \quad k'' = \mu^T_{q_t,i_t} U^{-1}_{q_t,i_t} \mu_{q_t,i_t}. \tag{20}$$

The sequence O can be derived as the one that maximizes (17).

Having achieved the optimum sequence Q_{best} of substates for a given phoneme h, to such a sequence corresponds a set $\{O_1, O_2, \ldots\}$ of feature sequences and a set of likelihood values $\{\mathcal{L}(O_1), \mathcal{L}(O_2), \ldots\}$. In order to maximise the joint pdf (14), the sequence $O^* = \{O_1, O_2, \ldots\}$ such that $\mathcal{L}(O^*) = \max\{\mathcal{L}(O_1), \mathcal{L}(O_2), \ldots\}$ is chosen.

Finally, once the optimum sequence of feature vectors O^* is obtained, the sequence X of synthesized signal frames is derived by the overlap-and-add synthesis process.

An overview of the speech synthesis algorithm is shown in Fig. 2. The block diagram shows the two fundamental steps of the proposed approach: the *learning stage*, that is the off-line stage, and the *synthesis stage*, that is the on-line stage. The first step extracts from the input database (audio and text sources) the MDCT features and derives throw an HMM modeling the substates for all input sequences of phonemes; the second step, given an input text and on the basis of the classified sequences of substates, determines the best sequence of states Q_{best} and the corresponding best sequence of features O^* for every input phoneme. At the end, the overlap-and-add synthesis process returns the synthesized speech of the input text.

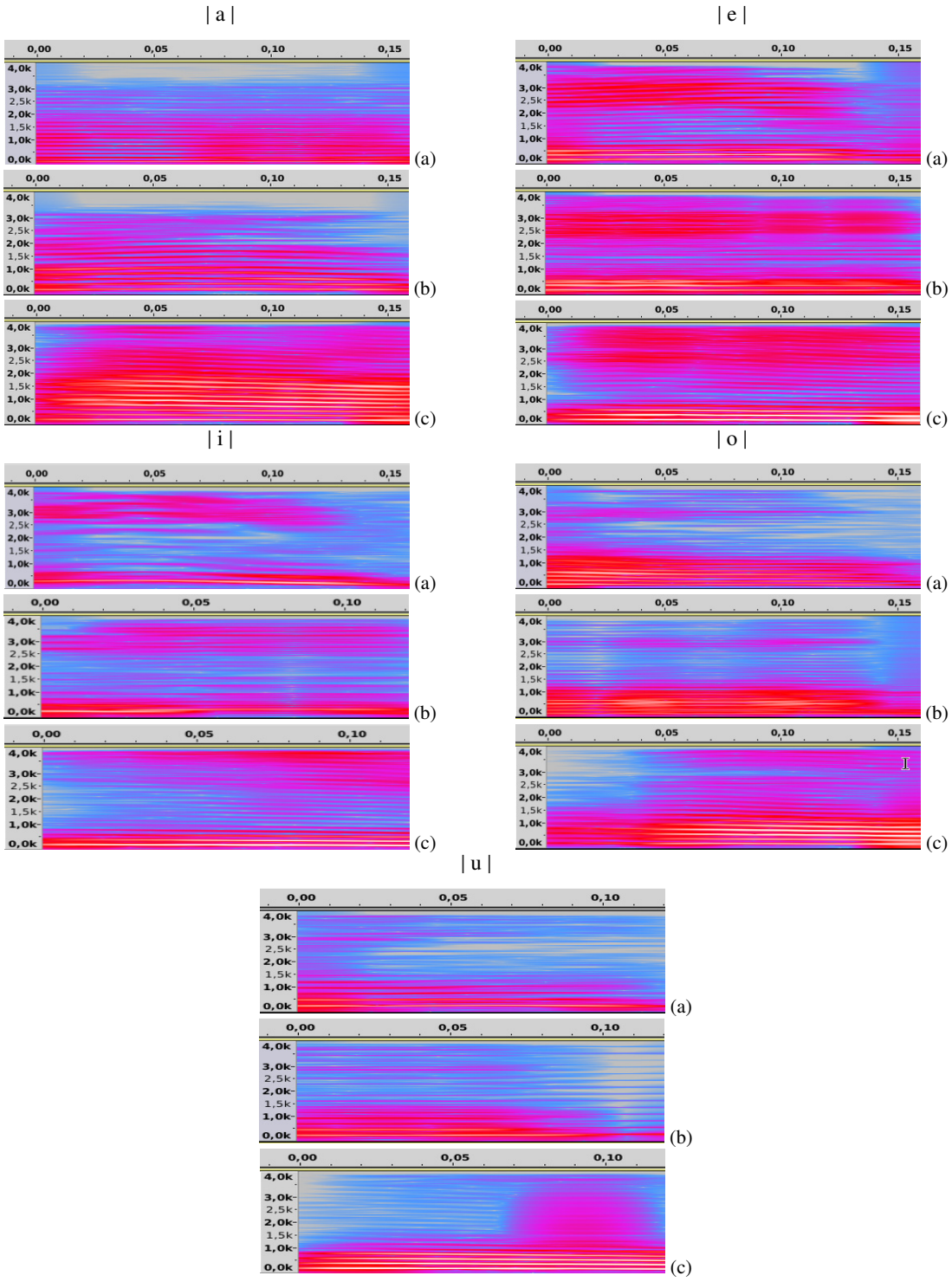

Fig. 3. Spectrograms of the Italian vowels |*a*|, |*e*|, |*i*|, |*o*|, |*u*| for the: (a) original signal, (b) signal synthesized by our technique, and (c) signal synthesized by diphones technique.

3. Experimental results

3.1. Acoustic model training

The first stage in the experiments we carried out to validate the proposed synthesis approach, was training the HMM acoustic model.

The material adopted for training was based on a 22 hours audio recording of a female speaker extracted from an Italian audiobook. The feature vector has been derived by applying the MDCT to the $2D = 20$ ms signal frame x_t, 50% overlapped with the successive frame. With a sampling rate of 8 kHz, a frame length of 80 samples (corresponding to the overlap length) is obtained.

The training was conducted with the Baum-Welch algorithm that performs an EM estimation of the audio modeling parameters.

To determine the most probable state sequences, we used the same training material and the Baum-Welch algorithm for the audio/text alignment at the HMM states level. In such a way, once the most probable state sequence for a given transcription is derived, the matrices in (18) can be computed.

3.2. Vowel synthesis

To validate the proposed speech synthesis technique the above scheme was used to synthesize the five Italian vowels, once the best substate sequences are given.

For comparison the same phonemes were synthesized using the "eSpeak" software[14] and the MBROLA (it-4) female recording audio extracted from ITC-irst data base[15]. MBROLA is an diphone-based algorithm[16] for speech synthesis. The MBROLA project web page provides diphone databases for a large number of spoken languages. "eSpeak" is a compact open source software speech synthesizer that can be used as a front-end to MBROLA diphone voices.

Figure 3 reports the spectrograms of the five Italian vowels $|a|$, $|e|$, $|i|$, $|o|$, $|u|$, as achieved by a 20 ms, 50% overlapped window. The first spectrogram in each figure depicts the behavior of the original audio signal, while the second and third spectrograms are related to the signal synthesized with our approach and the diphone (i.e. the second half of one phone plus the first half of the following) technique , respectively. As you can see, the spectrograms achieved with the suggested technique behave very closely to the original spectrograms. Diphones instead give spectrograms that are quite different from those expected.

3.3. Word synthesis

To further validate the proposed technique several Italian words have been synthesized.

Figure 4 reports the spectrograms of the three Italian word *topo* (| t o p o |), *casa* (| k a z a |), *Alice* (| a l i ʧe |), as achieved by a 20 ms, 50% overlapped window. The first spectrogram in each figure depicts the behavior of the original audio signal, while the second spectrogram is related to the signal synthesized with this approach. As you can see, the spectrograms achieved with the suggested technique behave very closely to the original spectrograms.

In addition Table 1 shows for the same three words *topo* (| t o p o |), *casa* (| k a z a |), *Alice* (| a l i ʧe |), and for the additional two *voce* (| v o ʧe |) and *troppo* (| t r o p p o |), the Itakura-Saito measure (ISM)[17,18] both for the synthesized words and a population of observations extracted from the original database adopted for training, with respect to the most likely (the target) realizations of such words. As you can see, the values of ISM for the three synthesized words are inside the ranges achieved for the population of original words, thus confirming the quality of synthesized speech.

4. Conclusion

This paper has derived a new HMM-based framework for speech synthesis. This framework combines an MDCT representation that guarantees a perfect reconstruction of the signal from feature vectors, a technique for learning HMM state sequences from phonemes. In the paper the rigorous mathematical apparatus, which the technique is founded on, has been reported together with some experimental results showing the validity of the approach.

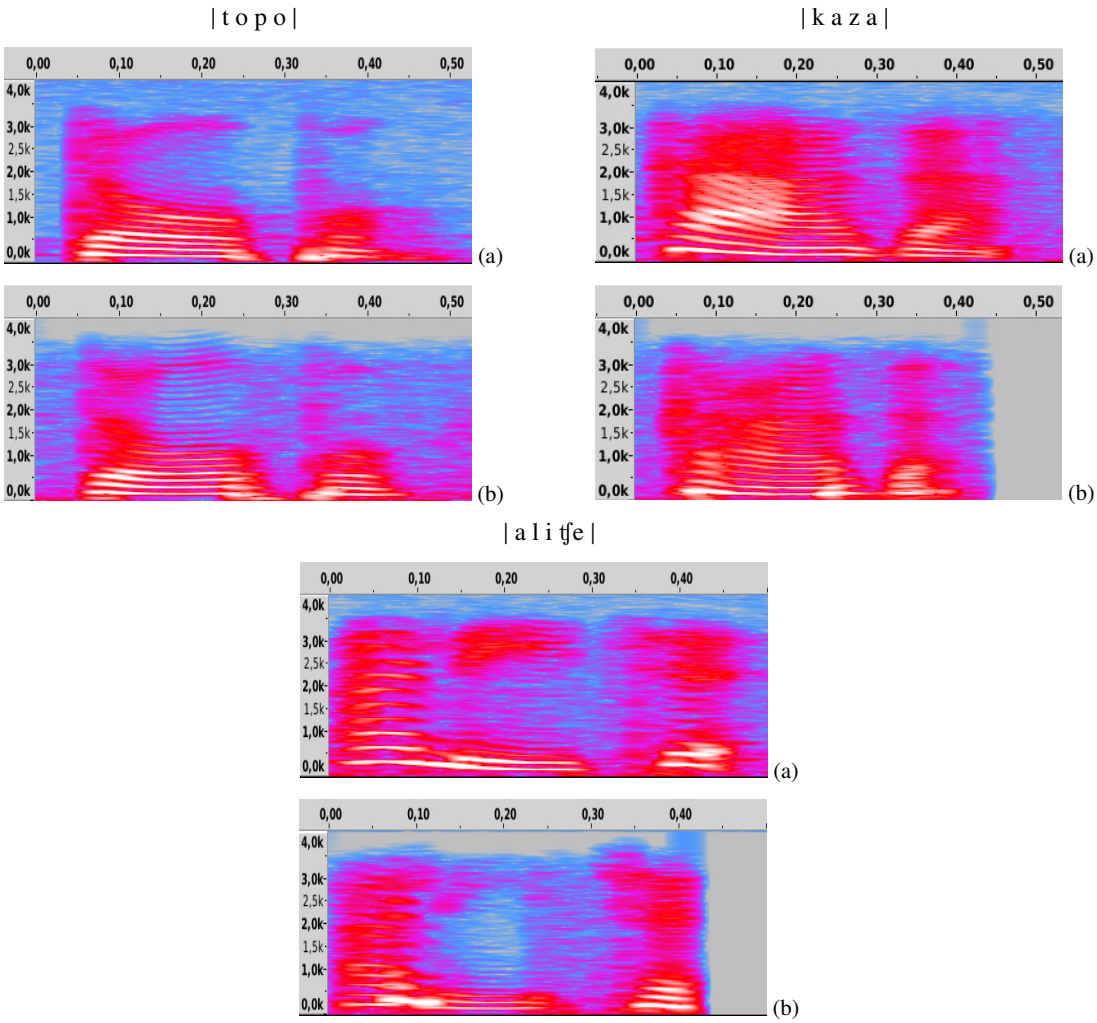

Fig. 4. Spectrograms of the Italian words *topo* (| t o p o |), *casa* (| k a z a |), *Alice* (| a l i ʧe |) for the: (a) original signal, (b) signal synthesized by our technique.

Table 1. Itakura-Saito measure for a population of observations and the synthesized Italian words.

Word	**Original Words**		**Synthesized Word**
	min	max	
| t o p o |	4.2371	18.1802	7.1187
| k a z a |	4.5560	31.8218	14.7179
| a l i ʧe |	8.8605	28.5589	11.1542
| v o ʧe |	10.2970	27.6711	20.8455
| t r o p p o |	1.2301	23.8528	5.2783
| t ɛ r r a |	7.5787	21.3814	16.5928

References

1. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE* 2013;**101**(5):1234–1252.

2. Donovan, R.E., Woodland, P.C.. A hidden Markov-model-based trainable speech synthesizer. *Computer Speech & Language* 1999;**13**(3):223 – 241.
3. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.. Speaker interpolation in HMM-based speech synthesis system. In: *EUROSPEECH*. 1997, .
4. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.. Speech parameter generation algorithms for HMM-based speech synthesis. In: *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, (ICASSP'00)*; vol. 3. 2000, p. 1315–1318.
5. Toda, T., Tokuda, K.. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In: *9th European Conf. Speech Communication and Technology*. 2005, p. 2801–2804.
6. Yoshimura, T.. *Simultaneous Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-Based Text-to-Speech Systems*. Ph.D. thesis; Nagoya Institute of Technology; 2002.
7. Yamagishi, J., Nose, T., Zen, H., Ling, Z.H., Toda, T., Tokuda, K., et al. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans Audio, Speech, and Language Processing* 2009;**17**(6):1208–1230.
8. Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., et al. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Trans Audio, Speech, and Language Processing* 2011;**19**(1):153–165.
9. Yu, K., Young, S.. Continuous F0 modeling for HMM based statistical parametric speech synthesis. *IEEE Trans Audio, Speech, and Language Processing* 2011;**19**(5):1071–1079.
10. Ling, Z.H., Deng, L., Yu, D.. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Trans Audio, Speech, and Language Processing* 2013;**21**(10):2129–2139.
11. Cabral, J.P., Richmond, K., Yamagishi, J., Renals, S.. Glottal spectral separation for speech synthesis. *IEEE Journal of Selected Topics in Signal Processing* 2014;**8**(2):195–208.
12. Dobrowolski, A.P., Majda, E.. Cepstral analysis in the speakers recognition systems. In: *Proc. Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference (SPA)*. 2011, p. 1–6.
13. Bosi, M., Goldberg, R.E.. *Introduction to digital audio coding and standards*. Springer; 2003.
14. eSpeak text to speech. 2007. `http://espeak.sourgeforge.net`.
15. The MBROLA project. 2006. `http://tcts.fpms.ac.be/synthesis/mbrola`.
16. Moulines, E., Charpentier, F.. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 1990;**9**(5–6):453 – 467.
17. Itakura, F., Saito, S.. Analysis synthesis telephony based on the maximum likelihood method. In: *Proceedings of the 6th International Congress on Acoustics*; vol. 17. pp. C17–C20; 1968, p. C17–C20.
18. Chen, G., Koh, S.N., Soon, I.Y.. Enhanced Itakura measure incorporating masking properties of human auditory system. *Signal Processing* 2003;**83**(7):1445–1456.