16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

A PHONETIC VOCODER WITH ADAPTATION TO SELECTABLE SPEAKER CODEBOOKS

Israel Halaly and Yuval Bistritz

Department of Electrical Engineering, Tel Aviv University
Tel Aviv 69978, Israel
halaly@bezeqint.net, bistritz@eng.tau.ac.il

ABSTRACT

The paper presents a very low bit rate phonetic vocoder based on speech recognition and synthesized speech with speaker adaptation using a set of speaker phoneme codebooks (SPCBs). The vocoder incorporates a well designed set of speaker phonemes codebooks that are available to both the encoder and decoder. The encoder performs periodically 'analysis by synthesis' that compares the incoming speech to speech that the decoder could synthesize from the output stream of the phoneme recognizer and the quantized pitch data per each SPCB and adapts it to the incoming speech by spectral warping. The index of the best performing SPCB and its adaptation parameter are transmitted to the decoder, together with the pitch and recognizer output bit streams, to synthesize speech that resembles better the speaker. In experiments held at a typical low bit rate of phonetic vocoders (below 300 bps), the incorporated adaptation reduced the average spectral distortion and increased speaker recognizability as judged by listeners.

1. INTRODUCTION

Speech vocoders are in demand for secure military applications because they meet well the need for voice encryption and channels with narrow bandwidth. Several low bit rate speech coders were adopted as standards for such applications, such as NATO STANAG 4591 triple-rate for 2400, 1200 and 600 bps, based on MELP algorithm [1]. However, for communication channels that operate at very narrow bandwidths the bit rates of the available vocoders are still too high. There is therefore need for further squeezing of the bit rate with willingness to sacrifice speech quality as long as the message is transferred reliably, the speech is intelligible and the speaker's identity is preserved.

Two encouraging approaches to achieve further bit rate reduction are phonetic and segment vocoders [2]-[6]. One problem with these vocoders is that they typically produce speech with poor speaker recognizability. This happens because the reproduced speech relies on the decoder's speech units (phonetic units or acoustically derived segment units), regardless of the encoded input speaker features. Some of the proposed coders proposed to alleviate this difficulty by adding speaker adaptation schemes but paying for it by a significant increase in the overall bit rate. A very low bit rate phonetic speech coder with speaker adaption was proposed in [2]. The adaption is performed by transmitting a vector to adjust the mismatch between the input speech and the Hidden Markov Model (HMM). This vocoder achieves bit rate of around 300 bps with reasonable speech quality. Another phonetic vocoder proposed in [3] uses HMM segmentation by transmitting average values of LSP coefficients for each phone. Speaker adaption is done by modifying the coefficients of the LPC synthesis filter. The achieved bit rate is 840 bps. In another approach proposed in [4], the speech is synthesized by concatenating the waveforms of units selected from a large database. The coder produces a natural sounding speech at bit rate of 833.4 bps.

This paper considers a phonetic vocoder that consists of a speech recognition unit, a synthesis unit, and an adaptation scheme that involves speaker phoneme codebooks (SPCBs), which are adapted by spectral warping. A special algorithm is devised to optimize the choice of a desirable number of SPCBs from a large database. The codebooks are available to both the encoder and the decoder. The speech synthesized for a specific SPCB is adapted to the input speaker by spectral warping using a technique called vocal tract normalization or vocal tract length normalization (VTLN) [7] [8]. The encoder performs a kind of 'analysis by synthesis' at the end of which, it picks one SPCB and its warping factor (WF) that produces the most similar synthesized speech to the input speech. The decoder uses the chosen SPCB and WF to produce speech with improved speaker features.

The next section 2 describes the design of the phonetic vocoder and the speaker adaptation scheme. The subsequent section 3 brings experimental results and their evaluation. The conclusion also brings some points that deserve further study.

2. PHONETIC VOCODER

The vocoder comprises an HMM phonetic recognizer and a speech synthesizer. The block diagram of the proposed speech encoder is illustrated in Figure 1. The encoder models the input speech spectra by a set of Mel-Cepstral Coefficients (MCCs), which are extracted by mel-cepstral analysis technique [9]. An HMM-based phoneme recognizer uses the MCCs to extract phoneme indexes and state durations. The pitch contour is also extracted from the input speech. The output of the HMM recognizer and the pitch contour are sent to the decoder. The encoder contains some more units (shown in the dashed line part) that performs the speaker adaptation to be described in detail below.

The block diagram of the proposed decoder is illustrated in Figure 2. Phoneme HMMs are concatenated according to the phoneme indexes, and the transmitted state durations. The HMM parameter generation derives the MCCs directly from the HMM [10]. This system uses dynamic features, i.e. delta and delta-delta mel-cepstral coefficients, as its feature vectors. The sequence of MCCs is obtained by maximizing the likelihood of the feature vector with respect to the concatenated HMM model. This inclusion of static and dynamic features admits smooth and natural sounding synthe-

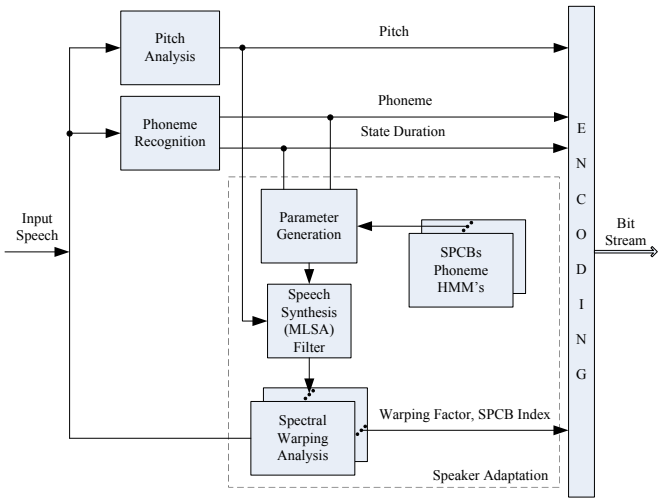

Figure 1: Block diagram of the speech encoder

sized speech. The speech signal in then constructed by using Mel Log Spectrum Approximation (MLSA) filter, directly from the MCCs [9]. The MLSA filter is an IIR stable filter that can approximate the modeled speech spectrum with sufficient accuracy. The excitation of the MLSA filter combines pulse train or white Gaussian noise, for voiced or un-voiced frames, respectively. The remaining units in the decoder perform speaker adaptation as explained next.

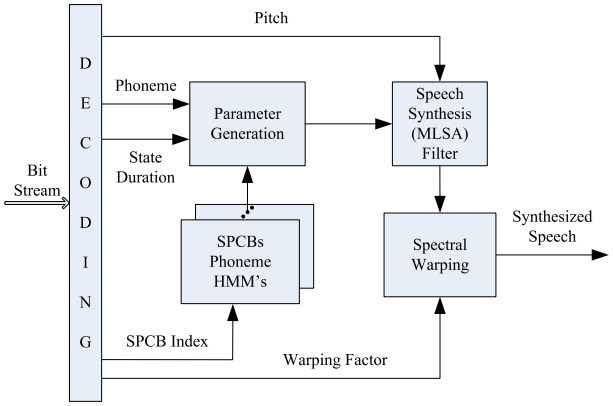

Figure 2: Block diagram of the speech decoder

The vocoder described so far suffers from poor speaker recognizability, a drawback typical to also other phonetic vocoder. The decoder produces speech from the transmitted phonemes and state durations using HMMs within its codebook, regardless of distinguishing qualities of the input speaker. To sooth this difficulty some mechanisms were proposed that adapt continuously the synthesized speech to the input speaker [2] [3]. However, the adaptation scheme tends to spoil the most attractive feature of the phonetic vocoder - natural sounding speech at very low bit rate. The method proposed in this paper suggests an adaptation scheme to the speaker with only a small (10 bps) increase in the bit rate. It consist of three components; (i) A set of trained speakers phoneme codebooks (SPCB) were added to the encoder and the decoder (subsection 2.1); (ii) An adaptation technique based on spectral warping that is defined by a single parameter (subsection 2.2); and (iii) The encoder performs an adjustment that selects a codebook and its warping factor pair, that attain the best similarity to the input speaker characteristics (subsection 2.3).

2.1 Codebook Design

The codebook design algorithm combines clustering and log spectral distortion (LSD) measures, iteratively. Given a set of P training speakers, the goal is to produce $N(< P)$ HMM SPCBs for N of the speakers that best present the training database within the adaptation scheme described below. The codebook design consists of two stages. The first is the selection of the best N speakers out of P speakers in a database. The second is the creation of HMM for the phonemes of the chosen speakers.

The selection starts by choosing randomly N speakers. The spectra of each of the remaining P-N speakers is warped towards the spectra of the chosen N speakers (as described below), and is associated to the group of the speaker to which it is closest by the LSD measure. Among speakers assigned to a same group, each speaker spectra is warped toward all the others, and the speaker that achieves the minimal LSD to the rest of speaker in the group, becomes the new representer of the group. In the next iteration, each of the new P-N outsider speakers is associated to a closest group and then, again, a new representer for the group is chosen. After several iterations, the algorithm converges to N speakers that can best represent the training data within our intended admission of adaptation by spectral warping.

In the second stage, the SPCB for each of the N chosen speakers is obtained by training HMM models for the phonemes from his available speech.

2.2 Spectral Warping

The underlying assumption in using spectral warping to improve speaker recognizability, is that differences among speakers depend strongly on the individual formants location of their phonemes. The spectral warping brings the formants structure of the synthesized phonemes, perceptually closer to the input speaker. The chosen VTLN technique was used in speech recognition [7] and was applied successfully to voice conversion [8]. We refer to warping function as $\tilde{\omega}(\omega)$, where ω is the source frequency and $\tilde{\omega}$ is the warped frequency. The applied warping function was chosen to be symmetric piece-wise linear function with two segments, whose slope α denotes the WF values [7]:

$$\tilde{\omega}_\alpha(\omega) = \begin{cases} \alpha\omega & ,\omega \le \omega_0 \\ \alpha\omega_0 + \frac{\pi-\alpha\omega_0}{\pi-\omega_0}(\omega-\omega_0) & ,\omega \ge \omega_0 \end{cases} \quad (1)$$

where ω_0 is,

$$\omega_0 = \begin{cases} \frac{7}{8}\pi & ,\alpha \le 1 \\ \frac{7}{8\alpha}\pi & ,\alpha \ge 1 \end{cases} \quad (2)$$

When $\alpha = 1$, the decoder outcome is synthesized from a selected SPCB without adaptation. A value close to $\alpha = 1$ occurs when one of the codebooks matches well the input

speaker. For an input speaker without a trained phoneme codebook, the frequency axis of a selected SPCB will be compressed ($\alpha < 1$) or stretched ($\alpha > 1$). This warping degrades the naturalness of the synthesized speech and wide deviation of α from its center at $\alpha = 1$ should be avoided. Following some subjective tests on the naturalness of the warped speech, we limited its admissible values to $0.5 \leq \alpha \leq 1.5$. Figure 3 illustrates a piece-wise linear warping function for different values of α.

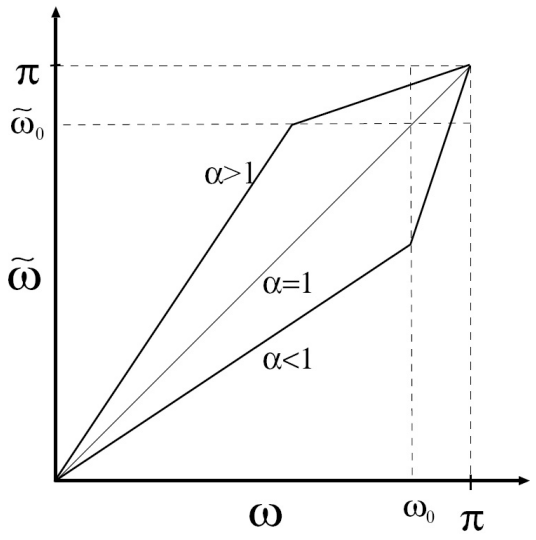

Figure 3: Piece-wise linear warping function

2.3 The Adaptation Scheme

Now we are ready to attend to the adaptation units in the encoder (Figure 1) and the decoder (Figure 2). Based on the extracted phoneme indexes and state durations, the encoder performs an analysis by synthesis task in order to select the codebook and its spectral warping factor that produces the synthesized speech that is closest to the input speech. The decoder reproduces its synthesized speech according to the received SPCB index and WF.

The units involved in the adaptation scheme at the encoder are marked by dashed line in Figure 1. The distance between the input speech X_s and the speech X_i produced by the i-th SPCB , $d(X_s, X_i)$, was measured by LSD. This measure is particularly adequate for this task because, as is well known, it is in correlation with perceptual conception. The WF α_i is chosen for each SPCB such that $d(X_s, \hat{X}_i)$ is minimized over the range of admissible values of the warping factor, where $\hat{X}_i$ denotes the adaptation of the speech synthesized by the i-th SPCB. The WF performs well its expected task when there is a good alignment between the input speech and the speech synthesized from the phonemes sequence and state durations. The validity of this condition depends on decent performance of the phonetic recognizer. It is also concurrent with finding a SPCB that produces the closest speech under the LSD criteria. Dynamical allocation of SPCBs helps to maintain this requirement and therefore improves the coder performance. Let

$$k = \arg\min_i d\left(X_s, \hat{X}_i\right) , \tag{3}$$

over the $i = 1, \ldots, N$ SPCBs using their best adapted synthesized speech $\hat{X}_i$. Then k is the selected SPCB index the α_k that created $\hat{X}_k$ is the chosen warping factor.

Figure 4 illustrates the adaptation performance improvement as the number of SPCBs increases.

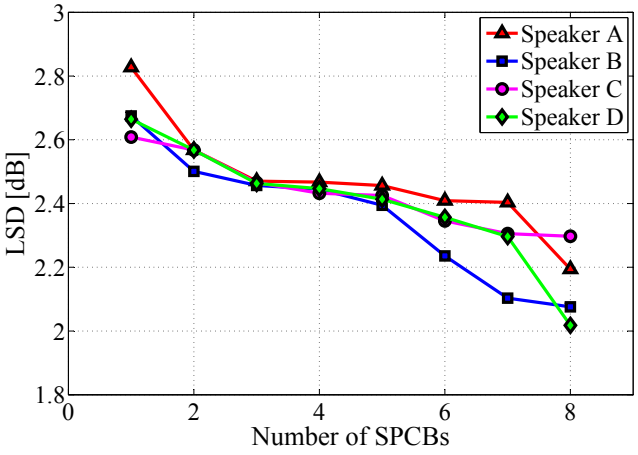

Figure 4: LSD score behavior with different SPCBs

The SPCB and WF selection can be done at the beginning of the session or periodically. The periodical adaptation updates and transmits the two parameters on a long term basis (e.g. every one second). Unrestrained periodical adaptation may create jumps in the characteristics of the synthesized speech. Thus, the adaptation algorithm allows changes in the WF values within a bounds of ± 10 percent from the chosen SPCB's WF. As long as the WF value varies within the bounds, the SPCB index is left unchanged. Otherwise, the SPCB index is changed to improve adaptation. This long-term adaptation scheme helps the convergence of the spectral warping algorithm and can face even change of the speaker.

3. EXPERIMENTS

For speech recognition we used a continuous HMM speech recognizer. The HMM was a 5-state left to right triphone model with no skips. Each state was modeled using a single Gaussian distribution with diagonal covariance. 44 phoneme models and one model for silence were used. Decision tree based model clustering was applied to each set of triphones models. The resulting set of tied triphone models has 1690 distributions including silence model. The speech recognizer models were developed from the training partition of the TIMIT corpus. Speech signals were sampled at 16kHz and windowed by 25msec Blackman window with a 5msec shift. The speech spectrum was modeled by 25 MCCs, including the 0-th coefficient. Since there were some phonemes which occurred more than others, we applied Huffman coding based on the occurrence probability distribution of phonemes. The resulting bit rate for the phonemes information is 50 bps.

The state durations lengths of each model were treated as three-dimensional vectors and vector quantization was applied to them [2]. A codebook of 1024 entries was designed using the LBG algorithm. The codebook design used the trained state durations lengths, which were obtained by the speech recognizer. Huffman coding was applied on the VQ

index, resulting in a bit rate of 100 bps.

For better bit reduction, the pitch information was quantized on a contour basis and not per frame. Following [4], the pitch contour quantization used linear approximation based on rate distortion criteria, by finding a set of points that approximates the contour, under a rate distortion constraint. The achieved bit rate for the pitch information is 130 bps.

The WF value is calculated according the chosen SPCB index for the input speaker, and both are transmitted every second.

We used for the search of α a step size of 0.01 that requires 7 bits. We used 5 male and 3 female SPCBs. Thus, the index of the chosen codebook requires 3 bits.

The overall vocoder's bit rate, assuming a mean phoneme duration of 100ms, is summarized in Table 1. They bits assignment can be divided into three groups. The first layer consists of phoneme indexes and it requires about 50 bps. At this rate, it is possible to held speech communication based on transmission of only the phonetic transcript of the message. The second layer adds speech prosody information that consists of the state durations and pitch contour bits. It utilizes about 230 bps. The third layer adds speaker specific information and is intended to improve the similarity of the produced speech to the speaker. This layer requires an extra of only 10 bps.

Table 1: Mean bit rate of the vocoder

Parameters	**Bits/Sec**
Phoneme indexes	50
State durations	100
Pitch	130
SPCB index	3
WF	7
Total	**290**

Quantitative measurements for the performance of the proposed vocoder are presented in Table 2. They were conducted under three different scenarios as follows: (i) Speaker Dependent (SD) - one of the SPCB belongs to the speaker; (ii) Speaker Independent (SI) without adaptation - the speaker does not have an own codebook and no adaptation is applied; (iii) SI with adaptation - the speaker does not have an own codebook and adaptation is applied. The measurements were carried out using LSD measures over a set of 8 SPCBs, which were chosen from a 20 speakers database. The SD scenario achieved the highest score, as expected. The SI without adaptation case showed a noticeable degradation of 1.37 dB that is due to the dissimilarity between the speaker and the selected phoneme codebook. Adding adaptation to the SI resulted in improvement of 0.74 dB. The performance of the vocoder in the SI with adaptation scenario is below that of the SD scenario. However, the proposed adaptation improved the vocoder's overall performance in comparison with the un-adapted SI situation.

In the remaining of this section we report several subjective tests that were carried out to further evaluate the proposed vocoder: speech quality (§3.1), speaker recognizability (§3.2) and speech intelligibility tests (§3.3). The tests follow the methodology that was used in the selection of the MELP 2400 bps US DoD standard [11][12]. As a reference coder, the MELP 2400 bps was also assessed.

Table 2: LSD scores for vocoder performance

Scenario	**Mean LSD [dB]**
(i) SD	0.92
(ii) SI w/o Adaptation	2.29
(iii) SI with Adaptation	1.55

3.1 Speech Quality Test

Speech quality assessment was performed using a Mean Opinion Score (MOS) subjective listening test following [12]. The test data was 10 sentences that were not included in the training. 10 untrained listeners were requested to rate the speech quality on a five-point scale. The quality test results are illustrated in Table 3. The SI without adaptation scenario achieved a relatively high score (at this point no assessment of speaker recognizability was required). Adding adaptation to the SI scenario reduced the grade by 0.19 points. This is attributed to the degradation in the speech naturalness that was introduced by the spectral warping.

Table 3: Speech quality test results

Scenario	**MOS score**
MELP	3.20
SI w/o Adaptation	2.13
SI with Adaptation	1.94

3.2 Speaker Recognizability Test

Speaker recognizability tests were performed as suggested in [11]. Pairs of utterances were presented to a listener and he was requested to judge if they were spoken by a *same* or a *different* speaker. Two sets of experiments were conducted. In the first experiment, denoted by U-P (Unprocessed-Processed), we examined to what extent the coder preserves the speaker identity. The first sentence was not coded and the second was coded. The second experiment, denoted by P-P (Processed-Processed), was carried out to assess how well each coder preserves information necessary to distinguish one speaker from another. This time both sentences were coded. The tests were conducted under two scenarios: SI without adaptation and SI with adaptation. We used 20 speakers, 10 female and 10 male. 10 untrained listeners participated in the experiments.

The percentage of correct responses for *same* and *different* pairs and their average values are presented in Figure 5. The apparent trend common for both the processed and unprocessed speech is that it is easier to distinguish between *different* speakers than to identify a *same* speaker situation.

The SI without adaptation achieved a very low percentage of correct answers (40%) for the *same* speakers in the U-P test. In this case the synthesized speech results from an un-adapted codebook speech that conveys no useful characteristics to identify the speaker. The score for the distinction between *different* speakers tends to be higher also because they are often assigned a different SPCB. The improvement

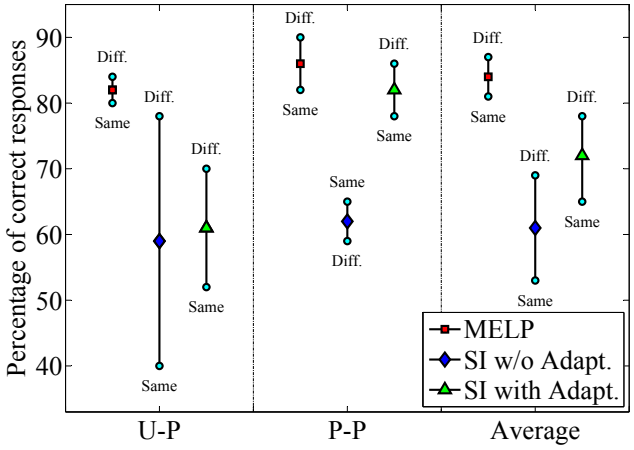

Figure 5: Speaker recognizability test results

introduced by the speaker adaptation algorithm is noticeable in all the test conditions.

3.3 Speech Intelligibility Test

Intelligibility tests were performed as suggested in [12] using Diagnostic Rhyme Test (DRT). The DRT intelligibility test is a regular assessment method for low bit rate speech coders. It uses pairs of rhyming word that differ only in the first consonant. First, a pair of words is shown to the listener. Then, he is aurally exposed to one of the two words and has to decide which of the two words he heard. The experiment was conducted in two environments: quiet and office conditions, and in two scenarios: SI without adaptation and SI with adaptation. 10 untrained listeners were asked to listen to 30 pairs of words.

The percentage of correctly chosen words are presented in Figure 6. The SI without adaptation achieved a very high percentage of correct answers (91%) in quiet environment since in this case the synthesized speech suffered no degradation in naturalness. The office environment caused some phoneme recognition errors. The degradation in SI with adaptation increased the number of errors.

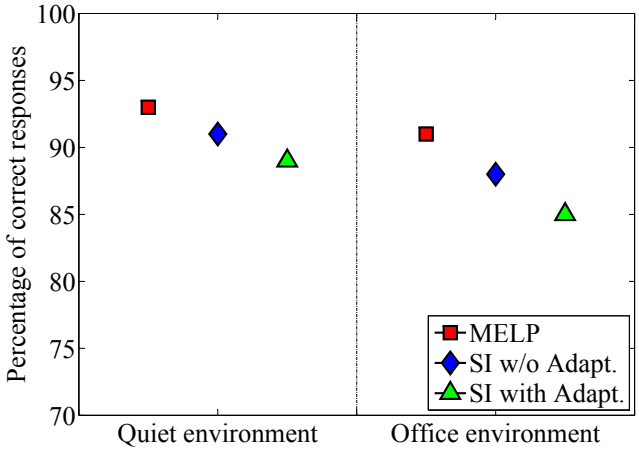

Figure 6: Speech intelligibility test results

4. CONCLUSION

The paper considered a very low bit rate vocoder based on phonemes recognition and synthesized speech with adaptation to the input speaker. The synthesized speech is adapted to a well trained selection of speaker codebooks by means of a spectral warping algorithm. The adaptation improves the speaker recognizability and requires only a very small increase in the overall bit rate. The new approach admits an interesting tradeoff between the performance level and the number of SPCBs that deserves further study. It is also interesting to study the impact of refined warping schemes on the performance of the vocoder.

5. ACKNOWLEDGMENT

We would like to thank Dr. David Sündermann for his voice conversion software.

REFERENCES

[1] A. McCree and T. P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, July 1995.

[2] T. Masuko, K. Tokuda and T. Kobayashi, "A Very Low Bit Rate Speech Coder using HMM with Speaker Adaptation," in *Proc. ICSLP 1998*, pp. 507–510.

[3] C. M. Ribeiro and I. M. Trancoso, "Phonetic Vocoding with Speaker Adaptation," in *Proc. EUROSPEECH 1997*, pp. 1291–1294.

[4] K.-S. Lee and R. V. Cox, "A Very Low Bit Rate Speech Coder based on a Recognition/Synthesis Paradigm," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 482–491, July 2001.

[5] S. Roucos and A.M. Wilgus, "The Waveform Segment Vocoder: A New Approach for Very-Low-Rate Speech Coding," in *Proc. ICASSP 1985*, pp. 236–239.

[6] J. Černocký, G. Baudoin and G. Chollet, "Segmental Vocoder-going beyond the Phonetic Approach," in *Proc. ICASSP 1998*, pp. 605–608.

[7] L. F. Uebel and P. C. Woodland, "An Investigation into Vocal Tract Length Normalization," in *Proc. EUROSPEECH 1999*, pp. 2527–2530.

[8] D. Sündermann, H. Ney and H. Höge, "VTLN-Based Cross-Language Voice Conversion," in *Proc. ASRU 2003*, pp. 676–681.

[9] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," in *Proc. ICASSP 1992*, pp. 137–140.

[10] K. Tokuda, T. Kobayashi and S. Imai, "Speech Parameter Generation from HMM using Dynamic Features," in *Proc. ICASSP 1995*, pp. 660–663.

[11] A. Schmidt-Nielsen and D.P. Brock, "Speaker Recognizability Testing for Voice Coders," in *Proc. ICASSP 1996*, pp. 1149–1152.

[12] J.D. Tardelli and E.W. Kreamer, "Vocoder Intelligibility and Quality Test Methods," in *Proc. ICASSP 1996*, pp. 1145–1148.