APPLICATION OF SPEAKER MODIFICATION TECHNIQUES TO PHONETIC VOCODING

Carlos M. Ribeiro

INESC/ISEL-CEDET
cmr@inesc.pt

Isabel M. Trancoso

INESC/IST
Isabel.Trancoso@inesc.pt

INESC - Instituto de Engenharia e Sistemas de Computadores
Rua Alves Redol, Nº 9, 1000 Lisbon.

ABSTRACT

The goal of the work described in this paper is to develop a very low bit rate vocoding scheme. The vocoder is a typical LPC vocoder, whose parameters are post-processed on a phone-by-phone basis, resulting in a variable bit rate segment vocoder.

Given the well known speaker recognizability problems presented by vocoders at such low bit rates, we have attempted to integrate a speaker modification method based on altering the formant frequencies and bandwidths of vowel segments. This is done by transmitting the mean value and standard deviation of the radius and angle of the poles corresponding to formant frequencies for each phone. In the decoder stage, the phone index is used to retrieve a set of normalized values from a codebook of 'typical' phones. This set is speaker adapted to preserve the static characteristics (average and standard deviation) but relies in the typical phone to represent the dynamic characteristics such as formant trajectories.

1. INTRODUCTION

Segment or Phonetic vocoders are one of the most frequently proposed methods to code speech at rates below 1000 bit/s [4, 6]. This type of coder attempts to decompose speech into a sequence of segments that are compared to a codebook of pre-stored segments, which, depending on the vocoder, may be phones, transitions between phones or arbitrary sequences of sounds. Since the speech segments that the vocoder tries to recognize have varying duration, this type of vocoder works in a variable frame rate environment.

The transmitter stage usually includes a recognizer of the HMM (Hidden Markov Model) or DTW (Dynamic Time Warping) type whose task is to recognize and segment the input speech signal. Typically, the stored codebook includes LPC-based spectral parameters. In the receiver, a synthesizer reconstructs a speech segment, based on the transmitted information: codeword index, RMS, pitch, voicing and duration.

Speaker recognizability is one of the main problems faced by vocoders at the lowest bit rates, given the need to reduce speaker specific information. Hence, phonetic vocoders are very suitable to speaker dependent coding. For speaker independent coding, some type of speaker adaptation may be performed. One possible method is to choose the best codebook from a set of multiple speakers codebooks [3]. Another is to adapt the codebook to a new speaker [5]. This latter type of approach is the one used in the present work.

The adaptation strategy we have followed is based on the speaker modification work described in [7]. The authors introduced a method of altering the formant frequencies of vowel segments using LPC analysis/synthesis. The pole location modification is based on statistical references and provides individual control over formant frequencies and bandwidths. The method is based on collecting statistics of the radius and angle of the poles associated with formants, for each frame corresponding to the same vowel class, and for each speaker. LPC analysis is performed on the utterance from the source speaker which we want to modify, in order to make it sound as spoken by the target speaker. Each pole of the linear prediction polynomial (expressed in its polar form $r\,e^{j\theta}$) is then moved toward the mean of the target speaker for that particular class by zscore normalization:

$$r' = (r - \bar{r}_{source}) / \sigma_{rsource} \quad (1a)$$

$$\theta' = (\theta - \bar{\theta}_{source}) / \sigma_{\theta source} \quad (1b)$$

where $\bar{r}_{source}$ and $\bar{\theta}_{source}$ are the mean values of the radius and angle, respectively, for the source speaker, and $\sigma_{rsource}$ and $\sigma_{\theta source}$ are the corresponding standard deviations. The pole modification is achieved by introducing the target speaker statistics:

$$r_{\mathrm{mod}} = r' \times \sigma_{rtarg} + \bar{r}_{targ} \quad (2a)$$

$$\theta_{\mathrm{mod}} = \theta' \times \sigma_{\theta targ} + \bar{\theta}_{targ} \quad (2b)$$

After reconstructing the modified linear prediction polynomial, LPC synthesis is performed using a modified residual.

We have tried to adapt the same type of strategy to the framework of phonetic vocoding. Besides using this strategy based on modifying pole locations, we have also experimented with a similar one based on the modification of Line Spectrum Pair (LSP) coefficients. This work is still in its earlier stages and it remains to be determined whether real bit savings can be derived from these strategies.

The organization of this paper is as follows: the codebook normalization and speaker adaptation strategies are described in Section 2. Section 3 presents the overall coder and decoder. The training and test corpora and the corresponding results are described in section 4. Finally, section 5 presents our conclusions and discusses future work.

2. CODEBOOK NORMALIZATION AND SPEAKER ADAPTATION

The speaker modification method introduced in [7] was adapted to the framework of phonetic vocoding as follows: for all the frames corresponding to the duration of a single phone, the mean value and standard deviation of the radius and angle of each pole (corresponding to a formant frequency) are computed and transmitted, together with the phone index, pitch, voicing and duration information. In the decoder, the phone index is used to retrieve a set of normalized codewords from a codebook of 'typical' phones.

The codebook includes one normalized codeword for each vowel class, of dimension $L^x(p+1)$, where L is the duration of the stored vowel in terms of number of frames, and the additional term is the RMS per frame. The stored values are normalized with zero mean and unit standard deviation, according to (1).

This set of values is then speaker adapted, as in (2), in order to preserve the static characteristics of the speaker for the uttered phone (average and standard deviation), while relying on the typical phone stored in the codebook to represent the dynamic characteristics such as formant and RMS trajectories inside the phone.

We have used basically the same procedure to modify the set of LSP coefficients and RMS corresponding to a phone, and adapting them to the input speaker. The Line Spectrum Pair transformation of LPC prediction coefficients has been introduced by Itakura in 1975 [2] and some of its properties were later studied by Soong and Juang [8]. By definition, the LSP coefficients are the angles of the roots of two polynomials, P(z) and Q(z), derived from the forward and backward predictor polynomials:

$$P(z) = A(z) + B(z) \tag{3a}$$

$$Q(z) = A(z) - B(z) \tag{3b}$$

where

$$B(z) = Z^{p} A(z^{-1}) \tag{4}$$

and p is the order of the LPC analysis. It is well known that P(z) and Q(z) are symmetric and antisymmetric, respectively, with a root at z = 1 (for P(z)) or z = -1 (for Q(z)), and the remaining roots all lying on the unit circle. The LSP transformation is equivalent to replacing the impedance of the glottal source is an acoustic tube model with either a closed (P(z)), or open (Q(z)) tube section. For a stable vocal tract filter, the roots of the closed and open glottis are interleaved on the unit circle.

As discussed in [1], the closer two consecutive LSP coefficients are together, the narrower the bandwidth of the corresponding pole of the vocal tract filter. Hence, formants are marked by two close LSP coefficients, whereas spectral tilt is primarily marked by LSP coefficients which are farther apart. The roots of P(z) have been named as position coefficients, because the closed glottis model is the best approximation for a lossless approximation of the vocal tract filter. Hence, whenever formants are present, one can find a correspondence between the roots of P(z) and the locations of the formant frequencies. The roots of Q(z), on the other hand, have been called difference coefficients, because of their role in marking the presence and absence of a formant by their closeness to a position coefficient.

The statistical relationship between the locations and bandwidths of the speech formants and the position and difference coefficients was explored in [1], in order to code each coefficient with a different granularity, taking into account perceptual considerations. In the present work, we will explore this relationship to extend the speaker modification method which was originally formulated in terms of the radius and angle of the poles of the predictor filter.

The codebook is now a set of normalized codewords, one for each class of phones, of dimension $L^x(p+1)$ each. The stored values are normalized, with zero mean value and unit standard deviation, according to:

$$LSP_i^{'} = (LSP_i - \overline{LSP_i}) / \sigma_{LSPi} \tag{5a}$$

$$E' = (E - \overline{E}) / \sigma_E \tag{5b}$$

where LSP_i is the i^{th} LSP coefficient, E is the RMS, and ´-´ and σ denote, as usual, mean value and standard deviation of the corresponding values. Speaker adaptation is done in the receiver stage, on a frame by frame basis within each phone, by performing the inverse normalization procedure, in order to match the transmitted mean value and standard deviation.

$$LSP_{i\,\mathrm{mod}} = LSP_i' \times \sigma_{LSP_{itarg}} + \overline{LSP_{i\,\mathrm{targ}}} \tag{6a}$$

$$E_{\mathrm{mod}} = E' \times \sigma_{E\mathrm{targ}} + \overline{E}_{\mathrm{targ}} \tag{6b}$$

The modified values of the LSP coefficients are monitored to avoid instability of the LPC filter.

The phone characteristics in the middle of the phone are well defined, but boundaries are strongly dependent from nearest phones. Better performance can be obtained by deriving codewords for each phone in different left and right contexts, in spite of increasing the processing delay and the codebook storage.

Time warping is adopted to adjust the duration of the normalized phone in the codebook to the transmitted phone duration. In order to decrease the distortion due to this warping procedure, one can store the same phone with different duration. We have not tested this extended storage scheme, but, once again, potential improvements may be achieved with the trade-off of larger codebook storage requirements.

3. PHONETIC VOCODER

A block diagram of the overall phonetic coder and decoder structure is shown in Figure 1. The input speech is fed through an LPC vocoder analysis stage (p=10), that computes the LPC coefficients, RMS, voiced/unvoiced decision and pitch, on a frame by frame basis. Depending on the type of speaker adaptation strategy, the LPC coefficients are then used to derive either the poles corresponding to formant frequencies or LSP coefficients.

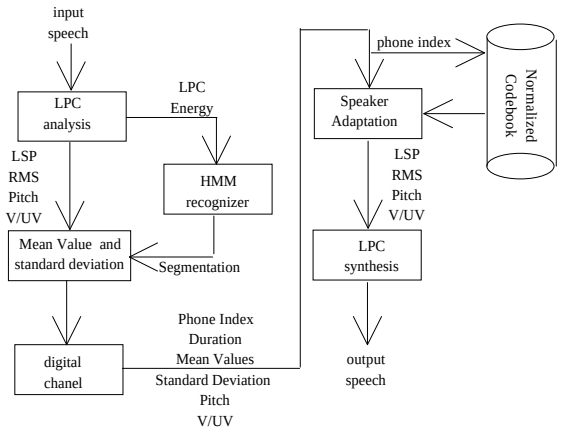

Figure 1: Block diagram of the phonetic vocoder with speaker adaptation (LSP modification).

The LPC coefficients and energy are used as input to an HMM phone recognizer based in the HTK software, in order to segment the speech signal. The recognizer uses 3-state, 3-mixture-per-state models. Each input vector has 26 coefficients (12 cepstra, 12 delta-cepstra, energy and delta-energy). The phone index and duration are then computed and transmitted once per phone, together with the mean values and standard deviations of the derived parameters.

Voicing decision is corrected for phones that a priori are either always voiced or always unvoiced, and transmitted with the pitch estimation on a frame-by-frame basis and not segment-by-segment. Nevertheless, for phones which are a priori always unvoiced, no information about voicing decision or pitch estimation is transmitted. For phones with a duration of only one or two frames, the direct values of the LSP coefficients and RMS are transmitted, as no further bit rate reduction is possible.

In the receiver stage, the phone index and the neighboring phone indexes are used to retrieve the corresponding context dependent codeword, as described in section 2. The phones chosen as 'typical' are just, in this early stage of the work, the corresponding phone in the corpus with the largest duration. RMS and LSP coefficients are restored frame-by-frame, adapting the codeword to the input speaker by matching the mean value and standard deviation. Time warping is used to adapt the codeword length to the transmitted duration. Finally, a normal LPC vocoder synthesizer is used to reproduce the speech.

4. CORPUS AND RESULTS

The corpus used in this work is a subset of the EUROM.1_P corpus (Portuguese version of the EUROM.1 corpus, collected in the SAM_A ESPRIT project). This subset has 10 speakers (Few Talkers corpus), each of them with 15 passages (5 sentences each), amounting to approximately 53 minutes of speech.

The set of tests described in the paper have used this hand-labeled corpus, in order to avoid recognizer errors and thus better identify any problem derived from our approach. It is worth noticing, however, that the subjective speech quality may not be seriously degraded if a recognizer error results in a sufficiently good acoustic matching.

The corpus labeling was done in a semi-automatic way, using the HTK software. For the Portuguese phones which were close to English ones, initial phone models were derived from the labeled TIMIT corpus (N.Y. dialect). For the remaining ones, the manually labeled CVC subset of EUROM.1_P was used. The training was done using Baum-Welch reestimation. These initial models were used to automatically align the above mentioned subcorpus of passages, using the Viterbi algorithm. This aligned corpus was then used to retrain new phone models and realign the phonetic transcriptions, in a boot-strap procedure which was iterated a couple of times. This automatic time alignment was then manually corrected to create the final label files. Finally, triphone models were also trained.

Figure 2 plots the phoneme duration histogram in 20 ms frames. The average value is 7 phones per second, and the largest phone duration corresponds to 25 frames or 500 ms. 9% of the phones have less then 2 frames duration, resulting in no bit rate reduction.

In order to evaluate the subjective quality of the phonetic coder, an A-B test was performed, comparing the well known FS1015 LPC-10 2400 bit/s vocoder with the phonetic coder. Since we are still in the preliminary stages of this work and the phonetic coder

does not involve yet any quantization of the mean values and standard deviations, we have also compared it with an unquantized version of the LPC vocoder. In very informal listening tests, the quality of the synthetic speech produced by the phonetic vocoder was judged as only slightly inferior to the one of the unquantized LPC vocoder, specially in some non-vowel segments, where some artifacts can be found. Both present the same level of speaker recognizability.

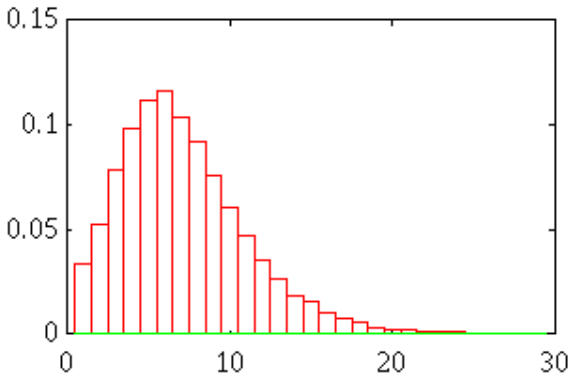

Figure 2: Histogram of phone duration measured as the number of frames (20 ms), for a subset of the EUROM.1_P FewTalkers corpus.

5. CONCLUSIONS AND FUTURE WORK

We have presented a new variable rate segmental vocoder, based on a speaker adaptation strategy. At such an early stage of the work, the list of possible improvements is rather large. Alternative adaptation schemes are currently being studied, as well as codebook design strategies and quantization techniques for the mean values and standard deviations. The current synthesis method can also be significantly improved, specially in what concerns the rather simple excitation model.

ACKNOWLEDGEMENTS

We want to thank our colleagues Drs. Céu Viana and Isabel Mascarenhas (CLUL), for their suggestions and their hard work in the manual correction of the EUROM.1 corpus labeling.

6. REFERENCES

1. J. R. Crosmer, T. P. Barnwell, "A Low Bit Rate Segment Vocoder Based on Line Spectrum Pairs", IEEE, *Int. Conf. Acoust., Speech, Signal Processing*, pp 240-243, 1985.

2. F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals", J. Acoust. Soc. Amer., vol 57, S35, 1975.

3. P. Jeanrenaud, P. Peterson, "Segment Vocoder Based on Reconstruction With Natural Segments", IEEE, *Int. Conf. Acoust., Speech, Signal Processing*, pp 605-608, 1991.

4. J. Picone, G. Doddington, A Phonetic Vocoder, IEEE, *Acoust., Speech, Signal Processing*, pp580-583, 1989.

5. Yoshinao Shiraky, Masaaki Honda, "Speaker Adaptation Algorithms Based on Piece-wise Moving Adaptive Segment Quantization Method", IEEE, *Int. Conf. Acoust., Speech, Signal Processing*, pp 657-660, 1990.

6 Yoshinao Shiraky, Masaaki Honda, "LPC Speech Coding Based on Variable-Length Segment Quantization", IEEE *Trans. On Acoust., Speech, and Signal Processing*, Vol. 36, Nº 9, pp 1437-1444.

7. Janet Slifka, Timothy R. Andersom, "Speaker Modification with LPC pole analysis", *Int. Conf. Acoust., Speech, Signal Processing*, pp 645-647, 1995.

8. F. Soong, B. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression", IEEE, *Int. Conf. Acoust., Speech, Signal Processing*, 1.10.1-1.10.4, 1984.