PHONETIC VOCODING WITH SPEAKER ADAPTATION

Carlos M. Ribeiro and Isabel M. Trancoso
INESC/ISEL-CEDET INESC/IST
cmr@inesc.pt Isabel.Trancoso@inesc.pt

INESC, Rua Alves Redol, 9, 1000 Lisbon, Portugal
Phone: +351 1 3100314; FAX: +351 1 3145843

ABSTRACT

This paper describes a phonetic vocoding scheme which relies on speaker adaptation to capture important speaker characteristics. These are typically lost in phonetic vocoders which transmit only information about the phones which are recognized, together with some prosodic information. In our scheme, however, additional speaker characteristics are transmitted in vowel regions (average values of LSP coefficients for each phone). This additional information yielded potentially good speaker recognizability results, in informal listening tests, while still achieving a rather low average bit rate, suitable for many transmission and storage applications. This work extends our previous phonetic vocoding scheme described in [5]. The vocoder is now fully quantized and the number of transmitted parameters had been significantly reduced.

1. INTRODUCTION

Segment or Phonetic vocoders are one of the most frequently proposed methods to code speech at rates below 1000 bit/s [4, 7]. 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 to be recognized 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 the speech segment, based on the transmitted information: codeword index, duration, RMS, pitch and voicing decision.

Speaker recognizability is one of the main problems faced by vocoders at these 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 codebooks [3]. Another, is to adapt the codebook to a new speaker [6]. This latter type of approach is the one adopted in our work. The basic scheme was described in a earlier paper [5] and a modified version is now summarized in section 2.

Section 3 presents the overall coder and decoder and the use of speaker adaptation in the context of phonetic coding.

In our previous work, hand labeled speech was used and no quantization was performed. The present paper describes the automatic segmentation procedure we have adopted (based on HMM) in section 4, and the quantization scheme in section 5. Finally, section 6 presents our conclusions and discusses future work.

2. SPEAKER ADAPTATION

The adaptation strategy we have followed is based on the speaker modification work described in [8]. 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:

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

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

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

We have adopted the same type of strategy to the framework of phonetic vocoding. Besides using this strategy based on modifying pole locations, we have also tested a similar one based on the modification of Line Spectrum Pair (LSP) coefficients [2, 9]. The procedure adopted was the following: for all the frames corresponding to the duration of a single vowel, the mean value and standard deviation of the LSP coefficients are computed and transmitted, together with the phone index and duration, RMS, pitch and voicing information.

In the decoder stage, 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, where L is the duration of the stored vowel in terms of number of frames, and p is the LPC analysis order. The stored values are normalized with zero mean and unit standard deviation, according to

$$LSP_i' = (LSP_i - \overline{LSP}_i) / \sigma_{LSPi} \qquad (3)$$

where LSP_i is the i^{th} LSP coefficient, and $\bar{\ }$ 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

$$LSP_{i\,\mathrm{mod}} = LSP_i' \times \sigma_{LSP_{itarg}} + \overline{LSP}_{i\,\mathrm{targ}} \qquad (4)$$

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

The characteristics in the middle of the phone are well defined, but boundaries are strongly dependent on the 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. This is the procedure adopted.

After codeword desnormalization, time warping is performed to adjust the duration 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 yet tested this scheme, but once again, potential improvements may be achieved with the trade-off of larger codebook storage requirements.

For non vowel phones, we have used the codeword without speaker adaptation. Speaker adaptation can be used to adapt non vowel segments, but, as expected, just a very slight improvement was obtained despite the increase of transmission parameters. We have also tried to extend this procedure to the RMS values, and transmit the average value for each phone, instead of transmitting RMS values on a frame by frame basis. The bit rate was reduced but also some artifacts were introduced in the speech signal.

The main information about the speaker characteristics is contained in the average values and not in the standard deviations. Using the standard deviations of the codewords and transmitting only the average values yields little speech degradation but strongly reduces the parameters to be transmitted.

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 estimate the pitch, on a frame by frame basis. LPC coefficients are then used to compute LSP coefficients.

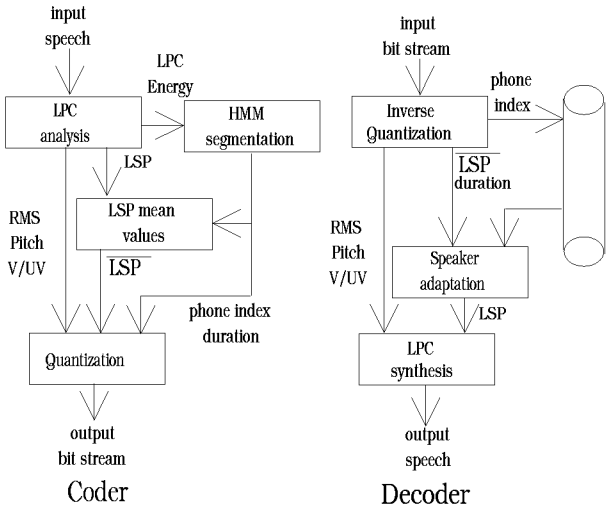

Figure 1
Phonetic vocoder with speaker adaptation

The LPC coefficients and energy are used as input to an HMM phone recognizer, in order to segment the speech signal. The phone index and duration are then computed and transmitted once per phone, together with the LSP mean values.

In the decoder stage, the phone index and the neighboring phone indexes are used to retrieve the corresponding context dependent codeword, as described in section 2. The phone chosen as 'typical' is just, in this stage of the work, the corresponding phone in the corpus with the largest duration. LSP coefficients are restored frame-by-frame, adapting the codeword to the input speaker by matching the LSP mean values. 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 signal.

4. HMM SEGMENTATION

The segmentation procedure we have used is based on state-of-the-art HMM phone recognition. 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).

At the time when we started this research work, no phonetically labeled spoken Portuguese corpus was available. We have thus trained our initial models using the TIMIT database. These models were used to align a subset of the EUROM.1 Portuguese corpus (the passages from the *few talkers* subset), by establishing some correspondences between phones in American English and in European Portuguese. The alignment is done by creating a net that recognizes the exact input sequence of phones. The recognizer outputs the same sequence, giving information about its time alignment. This first set of labels was used to train new models directly from the EUROM.1 Portuguese corpus, and the procedure was iterated until no changes in the segmentation were observed.

The resulting labels were then manually corrected, but this process took considerably less than if no prior labels were available. In many cases, the labeling was correct or slight adjustments of the boundaries were needed.

The corrected phone labels were used to retrain the final models, using the passages from 8 of the 10 speakers of the *few talkers* subset. This corresponds to approximately 32 minutes of speech, after removing silences. Two versions of the recognizer were implemented. One with 53 phones (including silences), and another with these 53 phones within different left and right contexts, amounting to a total of 5972 models.

The 2 remaining speakers of the *few talkers* subset (totaling 8 minutes of speech) were used for testing the recognizer. The results are listed in Table 1.

	Recognizer	Aligned
Context-independent phones	57%	100%
Context dependent phones	68%	100%

Table 1
HMM recognition results

The alignment results show potential for a boot strap procedure, in order to create more training data for the HMM recognizer.

In spite of the recognizer errors, a sufficiently good acoustic matching is generally obtained. This type of errors does not seriously degrade the subjective speech quality of phonetic vocoders. Perceptual results based in the comparison between synthetic speech produced using hand labeled and automatically segmentated speech (context dependent phones) have confirmed this statement.

5. QUANTIZATION

The resulting set of parameters to be quantized and transmitted comprises: the index and duration of each phone; the LSP average values in vowel regions; the RMS, voiced/unvoiced decision (V/UV) and pitch. The LSP coefficients are updated every 10 ms. The RMS values, V/UV decision and pitch information are transmitted every other frame. In the receiver, these values are interpolated to reconstruct the frame by frame information.

5.1 Phone index and duration

The 53 phones are coded using 6 bits and the duration using 5 bits, which allows a duration range of 320 ms (32 quantization levels x 10 ms). For an average of 17 phones per second (excluding silences), this corresponds to about 187 bit/s.

5.2. LSP average values

The LSP average values in vowel regions are quantized using the quantization table proposed by the FS-1016 4.8 kbit/s CELP Coder [1]. This is a non-differential quantization scheme of every LSP coefficient, which amounts to 34 bits. For an average of 6 vowels per second, this correspond to 204 bit/s.

5.3. RMS quantization

The RMS values are quantized using 5 bits, according to the FS-1015 RMS table [10]. As the transmission rate is 50 times per second, this results in 250 bit/s. If the duration of the phone is an odd number of 10-ms frames, the last RMS value is not transmitted and is reconstructed by interpolation in the receiver. In the case of phone durations of just one frame, however, the RMS value is always transmitted. The resulting bit rate is always smaller than 250 bit/s.

5.4. Voicing and pitch information

If the previous frame is unvoiced, 1 bit is transmitted to inform if the present frame is voiced or unvoiced. If the present frame is voiced, 7 more bits are transmitted with the estimated pitch value.

If the previous frame is voiced, the current pitch estimate is differential coded using only 5 bits. In order to reduce error propagation, the first voiced frame of each phone is never differentially encoded.

Further reduction may also be obtained if no voicing decision is transmitted for phones that a priori are either always voiced or always unvoiced. In the presence of recognizer errors, however, this procedure may impose significant speech degradation and was, therefore, discarded.

5.5 Quantization results

The quantization scheme we have just described was tested with all the passages of the 10 speakers of the *few talkers* subset of the EUROM.1 Portuguese corpus. This corresponds to 40 minutes of speech, after removing silences. The minimum, average and maximum values of the bit rate, phone rate and vowel rate that we have obtained for this subset are listed in table 2.

	Minimum	Average	Maximum
Bit/s	709	840	970
Phone/s	12	17	21
Vowel/s	4	6	8

Table 2
Minimum, average and maximum values for bit rate, phone rate and vowel rate.

Table 3 lists the average bit rate for each quantizable parameter.

Parameter	bit/s
phone	91
phone duration	84
LSP average values	200
RMS	236
V/UV and pitch	226
Total	**840**

Table 3
Average values of bit rate for each quantizable parameter

6. CONCLUSIONS

The phonetic vocoder described in this paper presents significant improvements over our earlier work: the coder is now fully quantized, the HMM segmentation is implemented and considerable bit savings were achieved by not transmitting the standard deviations of the LSP coefficients.

Very informal listening tests have yielded fair results in terms of speaker recognizability, which indicates the feasibility of our speaker adaptation scheme.

Further research should concentrate on deriving better codebooks and improving the very basic synthesis model we have adopted in our earlier experiments.

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.

REFERENCES

1. J. Campbell, V. Welch and T. Tremain, "*The DoD 4.8 Kbps Standard*", Advances in Speech Coding, Kluwer Academic Publishers, 1990.
2. F. Itakura, "*Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals*", J. Acoust. Soc. Amer., vol. 57, S35, 1975.
3. P. Jeanrenaud and P. Peterson, "*Segment Vocoder Based on Reconstruction With Natural Segments*", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 605-608, 1991.
4. J. Picone and G. Doddington, "*A Phonetic Vocoder*", Proc. IEEE Acoust., Speech, Signal Processing, pp. 580-583, 1989.
5. C. Ribeiro and I. Trancoso, "*Application of Speaker Modification Techniques to Phonetic Vocoding*", Proc. ICSLP´ 1996.
6. Y. Shiraky and M. Honda, "*Speaker Adaptation Algorithms Based on Piece-wise Moving Adaptive Segment Quantization Method*", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 657-660, 1990.
7 Y. Shiraky and M. Honda, "*LPC Speech Coding Based on Variable-Length Segment Quantization*", IEEE Trans. On Acoust., Speech, and Signal Processing, Vol. 36, N. 9, pp. 1437-1444.
8. J. Slifka and T. Andersom, "*Speaker Modification with LPC pole analysis*", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 645-647, 1995.
9. F. Soong and B. Juang, "*Line Spectrum Pair (LSP) and Speech Data Compression*", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 1.10.1-1.10.4, 1984.
10. T. Tremain, "*The Government Standard Linear Predictive Coding Algorithm: LPC-10*", Speech Technology, April 1982.