SPEAKER ADAPTATION IN A PHONETIC VOCODING ENVIRONMENT

Carlos M. Ribeiro *Isabel M. Trancoso*
INESC/ISEL INESC/IST
INESC, Rua Alves Redol 9,
1000 Lisbon, Portugal

ABSTRACT

The coder proposed in this paper falls in the class of segmental vocoders known as phonetic vocoders. Speaker recognisability 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, and can achieve bit rates as low as 250 bit/s. For speaker independent coding, a speaker adaptation methodology is adopted, although resulting in higher bit rates to transmit the speaker specific information. In order to further reduce the corresponding bit rate, a new method is proposed that explores the intra-speaker correlation for the same phone.

1. INTRODUCTION

Phonetic vocoders can reach bit rates as low as 250 bit/s at the cost, however, of significant degradation of synthetic speech quality, and severe speaker recognisability problems. In fact, phonetic vocoders are very suitable to speaker dependent coding [6], given the need to reduce speaker specific information. For speaker independent coding, some type of speaker adaptation may be performed. Gender dependent synthesis based on average pitch is a bit free option, but insufficient to characterise the speaker.

One possible method to overcome this problem is to choose the best codebook from a set of multiple-speaker codebooks [1]. Another one is to adapt the codebook to the input speaker [9]. The latter type of approach is the one used in the present work.

Section 2 will describe our baseline phonetic vocoder without speaker adaptation. Section 3 will be devoted to the speaker adaptation methodology. Section 4 will present a new method to explore the speaker specific information that has been previously transmitted for the same phone. Finally, Section 5 will present the conclusions and further work.

2. PHONETIC VOCODER WITHOUT SPEAKER ADAPTATION

In the proposed vocoder, like in other basic LPC vocoders, the transmitter stage performs LPC analysis, and estimates pitch, voicing and energy parameters, on a frame-by-frame basis. LPC coefficients are then fed into an HMM phone recogniser which segments the speech signal and produces a phone index. This index and the corresponding phone duration now replace the LPC information transmitted in conventional LPC vocoders and, together with pitch, voicing and energy information, constitute the parameters to be transmitted.

In the receiver stage, the phone index and the previous and next indexes are used to retrieve a correspondent codeword of LSF coefficients (Line Spectrum Frequencies) from a codebook. Each context dependent codeword is a matrix of L_j x p coefficients, where L_j is the duration of the stored j^{th} phone in terms of number of frames, and p is the LSF order. Time scale modification is adopted to adjust the duration of the normalized phone in the codebook to the transmitted duration. The restored LSF coefficients are then used to produce the synthetic speech, together with energy and pitch information.

We have implemented both conventional LPC vocoder synthesis and magnitude-only harmonic synthesis. Both can be used in this framework, as interoperability [5] between the two schemes may be obtained by deriving the amplitude of the harmonics from the LPC filter.

The codebook was generated using speech data from 8 of the 10 speakers of the *few talkers* subset of the EUROM.1 *corpus* for European Portuguese, collected in the SAM_A ESPRIT project. This data (15 passages from each speaker) was hand-labelled using 53 phone labels, and corresponds to about 32 minutes of speech, after removing silences. Each context-dependent codeword is computed as the 10^{th} order LSF centroid of the most likely duration (measured in number of frames, using a frame update period of 11.25 ms) of the corresponding context-dependent phone. This averaging procedure avoids mixing phones with different durations.

The set of parameters to be quantised and transmitted comprises the index and duration of each phone; the energy, the voiced/unvoiced decision (V/UV) and the pitch period. The quantisation schemes for the transmitted parameters were tested with a different subset of the EUROM.1 *corpus*, comprising the automatically aligned filler sentences of the *many talkers* subset, corresponding to 50 speakers (originally 60 but 10 were common to the *few talkers* subset used for training the codebook). This amounts to 18 minutes of speech after removing silences. The average phone rate is 13 phone/s. The quantisation methods are explained in detail in [8], and can be summarized as follows:

- ***Phone Index*** **-** The 53 different phone indexes are encoded using the Shannon-Fano memoryless source coding. In the test *corpus,* this corresponds to an average rate of 76 bit/s.
- ***Phone Duration*** **-** Phone duration is encoded using Huffman tree coding. In the test *corpus,* this corresponds to 53 bit/s, on average.
- ***Energy*** **-** Energy can be transmitted once per frame, resulting in a bit rate of 178 bit/s (4 bit x 44.4 frames/s),

or on a phone-by-phone basis, resulting in 65 bit/s (5 bit x 13 phones/s), on average.

- ***V/UV and Pitch*** – Likewise, pitch and voicing information can be transmitted on a frame-by-frame basis, resulting in 311 bit/s (7 bit x 44.4 frames/s); with differential quantisation, this value can be reduced to 125 bit/s; if, on the other hand, one adopts quantisation on a phone-by-phone basis, the resulting bit rate can be as low as 73 bit/s.

Hence, the lowest possible bit rate achieved with our baseline phonetic vocoder was 267 bit/s, using a full phone-by-phone quantisation scheme. Frame-by-frame updating of energy, V/UV and pitch can be used for higher quality, resulting in a higher bit rate of 618 bit/s. Between these two extremes, one may consider intermediate schemes, which provide a good trade-off between speech quality, on one hand, and bit rate on the other.

One of the weakest points of phonetic vocoders in general, according to many authors, is their dependence on the accuracy of the speech recognition system. We have used a state-of-the-art HMM phone recogniser with 3-state, 3-mixture-per-state models. Each input vector has 30 coefficients (14 mel-cepstra, 14 delta-mel-cepstra, energy and delta-energy). 53 HMM context-independent models were trained with the same hand-labelled data used to create the codebook. The 2 remaining speakers of the *few talkers* subset (totalling 8 minutes of speech) were used for testing the recogniser, yielding 37% recognition errors. The amount of hand-labelled data we have is, in fact, too small, either for training HMM models or for creating LSF codebooks, indexed by triphones. Hence, a bootstrap procedure can be adopted to automatically align more data, using the context-independent models. This data can then be used to retrain models, create new codebooks or to assess the perceptual quality of the coder. In spite of the 37% recognition errors, a sufficiently good acoustic matching is generally obtained. This type of errors does not seriously degrade the subjective speech quality of phonetic vocoders [6]. This idea was confirmed by a perceptual comparison between synthetic speech produced using hand-labelled and automatically segmented speech. Nevertheless, the use of different parameters in recognition (mel-cepstra) and synthesis (LSF) may not be as efficient as an "unified" method [2]. Another approach we have tried which significantly reduced the recognition errors is the use of context-dependent phone models. The use of the right context-dependent phone, however, may be prohibitive in terms of delay. This approach yielded a 25% reduction in recognition errors, in experiments using cepstral coefficients [7], and is currently under test using mel-ceptral coefficients.

These results have been obtained using ergodic models. A continuous speech recogniser trained for European Portuguese with corresponding pronunciation lexica and language models would improve the above mentioned recognition score, but has not been implemented yet.

3. SPEAKER ADAPTATION

The adaptation strategy we have followed is based on the speaker modification work described in [10]. The authors introduced a method of altering the formant frequencies of vowel segments using LPC analysis/synthesis. Our strategy for speaker adaptation is slightly different, being based on the modification of LSF coefficients, instead. The retrieved codeword is first time-modified to the transmitted duration Lj' and then adapted to minimize the mean squared error between the time-modified *kth* order LSF vector (of dimension Lj') and the corresponding LSF input vector:

$$LSF_{k\,\mathrm{mod}} = \alpha_k + \beta_k\, LSF_k \tag{1}$$

where α_κ and β_κ can be shown to satisfy:

$$\beta_k = \frac{C_{k\,in,cw}}{C_{\mathrm{k}\,cw}} \tag{2a}$$

$$\alpha_k = \overline{LSP}_{k\,in} - \beta_k\, \overline{LSP}_{k\,cw} \tag{2b}$$

where $C_{k\ in,cw}$ corresponds to the covariance between the *kth* order LSF input vector and the corresponding LSF vector in the selected codeword, and $C_{k\ cw}$ corresponds to the autocovariance of the *kth* order LSF vector in the codebook.

We found that speaker characteristics are essentially preserved by matching the LSF average values, and that the scale factor β has a minor perceptually effect in the output speech, and can be set to one, reducing the transmitted parameters and consequently the bit rate [7]. Hence, in order to implement this speaker adaptation strategy, it is necessary to transmit only information about the average values of each LSF coefficient over the whole duration of the phone, that is, the set of α_κ. The speaker-adapted vector will have the same average as the input vector, while relying in the dynamic characteristics stored in the codebook. This extra speaker specific information is only transmitted for vowel and glide phones, where the speaker characteristics are perceptually more important, thus reducing the bit rate. For consonants, the retrieved codeword is used without speaker adaptation. One of the problems that may arise as a result of this speaker adaptation strategy is instability in the resulting LPC filter. This is avoided by checking for instability and forcing the filter to be stable.

The proposed adaptation strategy proved to be effective even for male to female voice adaptation or vice-versa. Nevertheless, strong prosodic variations due to changes in fundamental frequency can also introduce artifacts in the speech output signal. To minimize this problem, duplicate codebooks were built one for male and another for female speakers. The procedure for choosing the gender-dependent codebook is based on the average values of the pitch frequency, computed for each of the last three voiced phones. An average value greater than a given threshold is considered to be from a female voice, and smaller than that is from a male voice. A majority rule is then used for the decision. The threshold value (150 Hz) was computed based on the statistics of the training data. With this heuristic decision, the transmission of information about the speaker gender is therefore avoided.

This speaker adaptation procedure may be interpreted as a vector-scalar quantiser, with as many vectors as vowel and glide phones. The index to the vector part of this quantiser is the phone index that must be transmitted anyway. The scalar part, that is,

the variation relatively to the expected values, contains the speaker information needed to adapt the codeword. Each α_κ value is quantised with 2 bits, using the LBG algorithm [4], resulting in a total of 20 bits for the scalar part of the quantiser. The corresponding bit rate obtained with the test *corpus* was 92 bit/s.

4. ON LINE SPEAKER ADAPTATION

In order to further reduce the bit rate imposed by speaker adaptation, one may attempt to explore the speaker specific information that has been previously transmitted for the same phone. A very simple way of doing this is just to transmit the speaker information for the first occurrence of each phone and take these values for the next occurrences. This simple approach has two drawbacks. First, intra-speaker diversity is not well represented using just one instance of every phone, and some type of averaging must be implemented. Second, a reset procedure must be introduced to handle new speakers in the system. Both type of problems are solved with a step-by-step first order adaptation of the form:

$$\overline{LSF}_k = (1-\mu)\overline{LSF}_k + \mu\overline{LSF}_{k\,in} \quad 0 \le \mu \le 1 \quad (3)$$

where μ controls the adaptation speed. For μ close to one, rapid adaptation is obtained but less memory for the past values is kept. For μ close to zero, the adaptation is slower but more memory for the past values is held. In order to reduce the bit rate, the set of α_κ must be transmitted only if the spectral distance between the reconstructed LSF averages and the present LSF values exceeds a certain threshold (Th). This procedure requires one more bit per vowel or glide in order to code the presence or absence of adaptation. The distance we used is the weighted Euclidean distance:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{k=1}^{p} w_k \left(x_k - y_k\right)^2} \quad (4)$$

where the weights w_k are computed as proposed in [3]:

$$w_k = \frac{1}{LSF_k - LSF_{k-1}} + \frac{1}{LSF_{k+1} - LSF_k} \quad k = 1,.....p \quad (5)$$

This weighting is based on the property that the closer two consecutive LSF coefficients are, the narrower the bandwidths of the corresponding pole of the vocal tract filter. These regions are more sensitive to LSF changes, and the corresponding LSF should have higher weights. Another necessary condition for transmitting the adaptation parameters is that the spectral distance between the reconstructed vector and the input vector exceeds the spectral distance between the present vector and the input vector.

The higher the value chosen for Th, the less frequently the LSF values are updated, with the corresponding bit rate savings and speech recognisability degradation. For low values of Th, the LSF coefficients will be updated more often and the speech recognisability will be as good as with full speaker adaptation, but with an increasing bit rate.

The parameters μ and Th were fine tuned in order to produce a 50% reduction in bit rate when transmitting the set of α_κ values, comparing to the coder with full speaker adaptation. In this condition, speech recognisability was judged close to that with full speaker adaptation.

5. CONCLUSIONS AND FURTHER WORK

This paper presented a phonetic vocoder based on different strategies of speaker adaptation with a set of quantization schemes which allows a trade-off between quality and bit rate, in a range of 250 to 600 bit/s.

One of the main limitations that we faced was the reduced size of the hand-labelled spoken material, namely for codebook design and coder assessment. As better alignment methods are presently being developed in the scope of other projects, this limitation will soon become less relevant.

We are currently testing methods to better control the LPC filter instability that can arise from the speaker adaptation procedure. Another area of work is the use of the same parameters in the recognition and synthesis stages, in order to improve robustness in the presence of recognition errors. Finally, a set of formal assessment tests including speaker recognisability is also planned for the near future.

6. REFERENCES

[1] P. Jeanrenaud, P. Peterson, “Segment Vocoder Based on Reconstruction With Natural Segments”, Int. Conf. Acoust., Speech, Signal Proc., pp 605-608, 1991.

[2] W. J. Holmes, “Towards a Unified Model for Low Bit-Rate Speech Coding Using a Recognition-Synthesis Aproach”, Proc. ICSLP’98.

[3] R. Laroia, N. Phamdo, N. Farvadin, "Roboust and efficient quantization of speech LSP parameters using two-dimensional differential coding", Proc. Int. Conf. Acoust., Speech, Signal Processing, pp 641-644, 1991.

[4] Y. Linde, A. Buzo, R.. Gray, “An Algorithm for Vector Quantization”, IEEE Trans. Comm., pp 84-95, 1980.

[5] R. J. McAuley, T. Champion, “Improved Interoperable 2.4 kb/s LPC using Sinusoidal Transform Coder Techniques”, Proc. Int. Conf. Acoust., Speech, Signal Proc., pp. 641-643, 1990.

[6] J. Picone, G. Doddington, “A Phonetic Vocoder”,- Proc. Int. Conf. Acoust., Speech, Signal Proc., pp 5809-583, 1989.

[7] C. Ribeiro, I. Trancoso, Phonetic Vocoding with Speaker Adaptation", ", Proc. 5th European Conference on Speech Communication and Technology, 1997.

[8] C. Ribeiro, I. Trancoso, "Phonetic Vocoding", Proc. 2th conference on Telecommunication, 1999.

[9] S. Roucos, A. Wilgus, “Speaker Normalisation Algorithms For Very Low Speech Coding”, Proc. Int. Conf. Acoust., Speech, Signal Processing, pp 1.1.1-1.1.4, 1984.

[10] J. Slifka, T. Anderson, “Speaker Modification With LPC Pole Analysis”, Proc. Int. Conf. Acoust., Speech, Signal Processing, pp 644-646, 1995.