IMPROVING SPEAKER RECOGNISABILITY IN PHONETIC VOCODERS

Carlos M. Ribeiro — *Isabel M. Trancoso*

INESC/ISEL-CEDET — *INESC/IST*

INESC, Rua Alves Redol, 9, 1000 Lisbon, Portugal.

E-mail: {cmr, Isabel.Trancoso}@inesc.pt Phone: +351 1 3100 314 Fax: +351 1 3145843.

ABSTRACT

Phonetic vocoding is one of the methods for coding speech below 1000 bit/s. The transmitter stage includes a phone recogniser whose index is transmitted together with prosodic information such as duration, energy and pitch variation. This type of coder does not transmit spectral speaker characteristics and speaker recognisability thus becomes a major problem. In our previous work, we adapted a speaker modification strategy to minimise this problem, modifying a codebook to match the spectral characteristics of the input speaker. This is done at the cost of transmitting the LSP averages computed for vowel and glide phones. This paper presents new codebook generation strategies, with gender dependence and interpolation frames, that lead to better speaker recognisability and speech quality. Relatively to our previous work, some effort was also devoted to deriving more efficient quantization methods for the speaker-specific information, that considerably reduced the average bit rate, without quality degradation. For the CD-ROM version, a set of audio files is also included.

1. INTRODUCTION

The goal of this work is to develop a very low bit rate coder (400-800 bits/s) for transmission and storage purpose, using a phonetic vocoding scheme. 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. In our vocoder, however, the LPC coefficients are fed into a HMM phone recogniser to derive a phone index. The transmitted information is now the phone index and duration, together with pitch, voicing and energy information, thus resulting in a variable bit rate scheme.

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 [6] 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-speaker codebooks [3]. Another is to adapt the codebook to a new speaker [8,9].

The latter type of approach is the one used in the present work. We have adapted a speaker modification strategy to solve this problem, which is based on changing formant frequencies and bandwidths. In the original method [10], statistics of the radius and angle of the poles associated with formants are collected, for each frame corresponding to the same vowel class, and for each speaker. In order to change the speech from a source speaker into a target speaker, each pole of the source speaker is moved towards the mean of the target speaker for the corresponding phone, by zscore normalisation.

Our adaptation strategy is slightly different [7], being based on the modification of LSP coefficients, instead of directly modifying formant frequencies and bandwidths. As discussed in [2], 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 further 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.

In order to implement this speaker adaptation strategy, we need to transmit the average values of each LSP coefficient, over the whole duration of the phone. This extra speaker specific information is only transmitted for vowel (and glide) phones, where the speaker characteristics are perceptually more important [6].

In the receiver stage, the phone index together with the previous and next indexes, is used to retrieve a set of LSP coefficients from a context-dependent codebook. Time scale modification is adopted to adjust the duration of the normalised phone in the codebook to the transmitted duration. For vowel and glide phones, the LSP coefficients are restored matching the transmitted input values, that carried speaker information, but relying in the dynamic characteristics stored in the codebook. The adaptation procedure is based on the following expression:

$$LSP_{i_{\mathrm{mod}}} = LSP_{i_{codebook}} - \overline{LSP}_{i_{codebook}} + \overline{LSP}_{i_{input}} \qquad (1)$$

The restored LSP coefficients are then used, together with energy and pitch information, to produce the synthetic speech, either using a conventional LPC vocoder scheme or, for instance, the magnitude-only sinusoidal synthesis scheme [5], with exactly the same input parameters, as the amplitude of the harmonics can be derived from the LPC filter.

Relatively to our previous work [7], new codebook generation strategies have been derived, which we will discuss in section 2. Automatic speech segmentation is described in section 3. Some effort was also devoted to deriving more efficient quantization methods for the speaker-specific information, which we will summarise in section 4. Section 5 presents our conclusions and discusses future work. The list of audio files included in the CD-ROM version is shown in section 6.

2. CODEBOOK

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. The codebook includes one normalised codeword for each different context-dependent phone. Each codeword is computed as the 10th order LSP centroid of the most likely duration (measured in number of frames, using a frame update period of 11.25 ms), for a given context-dependent phone. This averaging procedure avoids mixing phones with different duration.

Only 6089 different context-dependent phones were found in this training data. In European Portuguese, as in many other languages, the number of existent left and right context-dependent phones is much larger. When a context-dependent phone of the test data is not present in the codebook, it must be replaced by an existing context. However, replacing by a very different context (like replacing the /a/ in context /p-a+s/ by the same phone in the context /w-a+p/) can create artefacts in the output speech. Some phonetically based procedure was, therefore, implemented to search the codebook for the closest phone context

Some artefacts can arise from concatenation of phones, due to possible strong variations in the spectrum. This effect can be minimised by storing two additional frames for each codeword, one in the left and another in the right context, that allow for interpolation segment of two frames between adjacent phones. This procedure creates a reasonable smoothness between phones.

The proposed adaptation strategies 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 introduce artefacts in the speech output. To minimise this problem, duplicate codebooks were built, one for male and another for female speech. 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.

3. HMM SEGMENTATION

The segmentation procedure we have used is based on state-of-the-art HMM (Hidden Markov Models) phone recognition. The recogniser uses 3-state, 3-mixture-per-state models. Each input vector has 26 coefficients (12 cepstra, 12 delta-cepstra, energy and delta-energy).

The HMMs were trained with the same hand-labelled data used to create the codebook.. Two versions of the recogniser were implemented: a context-independent version and a context-dependent one.

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

	Recognition Error %
Context-independent models	43
Context-dependent models	32

Table 1: HMM recognition results

The amount of hand-labelled data we have is, in fact, too small, either for training HMM models or for creating LSP 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 assess the perceptual quality of the coder.

Given the better recognition performance, context-dependent models were used to segment the input speech in the phonetic vocoder. In spite of the 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.

4. QUANTIZATION

The set of parameters to be quantized and transmitted comprises: the index and duration of each phone; the LSP average values in vowel and glide regions; the RMS amplitude, the voiced/unvoiced decision (V/UV) and the pitch period. Frames of 11.25 ms are used and 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.

4.1. Phone index

The 53 phones are coded using the Shannon-Fano [1] memoryless source coding. The entropy computed on the basis of the phone probabilities for the training corpus was 5.23 bits per phone. Our quantizer achieves 5.27 bits per phone which corresponds to 99% efficiency. The most likely phones are coded with 4 bits and the least likely phones with 9 bits. For an average of 16 phones per second (excluding silences), this corresponds to about 86 bit/s.

Further bit rate savings could be achieved using language modelling (e.g. phone bigrams or trigrams), to explore the phonotatics of the language. However, some loss of robustness in the presence of phone recognition errors may be expected which prevented us from adopting this solution.

4.2. Duration

The phone duration is measured in terms of number of frames (11.25 ms long). The entropy computed for this parameter in the training corpus was 3.64 bits per phone. 68% of the phones in this corpus have a duration of 1 to 6 frames. These duration values are coded with 3 bits, and larger values are coded using Huffman tree coding. The average number of bits per phone achieved with this quantization scheme is 3.67, which corresponds to 99% efficiency. For an average rate of 16 phones per second (excluding the silences), this amounts to about 60 bit/s.

The phone duration, as the energy, is correlated with the type of phone, but it is difficult to take advantage of this dependence, as a speaker may have a higher or lower speaking rate than average, and the same can be said about the loudness of the sentence.

4.3. LSP differential coefficients

An estimate of the expected value of each average i^{th} LSP coefficient for the j^{th} phone is

$$E\left[\overline{LSP}_{ji}\right] = \frac{1}{N_j}\sum_n \overline{LSP}_{jni} \qquad (2)$$

where N_j is the number of occurrences of the j^{th} phone in the training corpus. The difference between the computed LSP average values and this expected value carries the speaker information, and its dynamic range is smaller than the one of the computed values. It can, therefore, be transmitted more efficiently. In the receiver, of course, these values must be added to the expected values in order to restore the average LSP values.

This procedure can be interpreted as a vector-scalar quantizer, with as many vectors as vowel and glide phones. The index to the vector part of this quantizer 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 LSP differential coefficient is quantized with 2 bits, using the LBG algorithm [4], resulting in a total of 20 bits for the scalar part of the quantizer. For an average rate of 6 vowel and glides per second (excluding silences), this corresponds to about 120 bit/s.

The total spectral information transmitted in this vocoder also includes the phone index and duration. These are transmitted using an average rate of 8.94 bits per phone, resulting in 28.94 bits per phone to transmit the spectral information in vowel and glide regions, and only 8.94 in the remaining regions.

4.4. Energy

An optimal quantization table for this parameter was also derived using the LBG algorithm. The RMS value is quantized every other frame, i.e., every 22.5 ms. Interpolation is used to reconstruct the frame by frame information. RMS information for the first and last frames of each phone are always transmitted, as amplitude variations are perceptually more important there. This results in an updating rate slightly higher than 44.4 double-frames per second. For a 4-bit quantizer, the average value achieved is around 196 bit/s.

The transmission of this parameter effectively consumes the biggest slot of the bit rate, as can be observed in Table 2. In order to reduce this slot, the average value of the RMS can be transmitted once per phone, and a procedure equivalent to the one used for LSP quantization can be used to restore the values on a frame by frame basis. For a 4 bit quantizer and an average rate of 16 phones per second (excluding silences), this corresponds to about 64 bit/s. This alternative quantization scheme introduced some artefacts in the synthetic signal, even using interpolation between phones. Hence, a better control of this parameter should be explored.

4.5. Voicing and pitch information

A differential scheme was also adopted for transmitting the 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 differentially coded using only 3 bits. However, the transmitted values are not selected consecutively, in order to increase the dynamic range. The 8 hypotheses correspond to the unvoiced code and differences of {-9, -6, -3, 0, 3, 6, 9}.

4.6. Total bit rate

The quantization scheme we have just described was tested with all the hand-labelled passages of the 10 speakers from the *few talkers* subset of the EUROM.1 corpus, and all the automatically aligned sentences of the *many talkers* subset corresponding to 60 speakers. This amounts to 60 minutes of speech, after removing silences.

Table 2 lists the average bit rate achieved for each quantizable parameter. An average total bit rate of 596 bit/s was obtained. The spectral information is coded with 266 bit/s corresponding to the phone index and duration and the LSP differential values in vowel or glide segments. As pointed out before, the RMS is the parameter that consumes the biggest slice of the total bit rate.

The minimum, average and maximum values of the total bit rate, phone rate and vowel rate obtained for this test corpus are listed in table 3. Adaptation was performed in 38% of the phones. The bit rate is always smaller than 800 bit/s and higher than 500 bit/s.

Parameter	**Average bit rate**
Phone index	86
Phone duration	60
LSP differential values	120
RMS	196
V/UV and pitch	134
Total	**596**

Table 2: Average values of bit rate for each quantizable parameter.

	Minimum	Average	Maximum
Bit/s	512	596	730
Phones/s	11	16	24
Vowel-glides/s	4	6	9

Table 3::Minimum, average and maximum values for bit rate, phone rate and vowel-glide rate.

5. CONCLUSIONS AND FUTURE WORK

The main goal of the worked described in this paper was the improvement of our previous variable rate phonetic vocoder, based on a speaker adaptation strategy. New codebook generation strategies were proposed, with gender dependence and interpolation frames, which lead to better speech quality and speaker recognisability. A set of quantization schemes for transmitting the speaker specific information was developed, that also considerably reduces the average bit rate, without degrading the synthetic speech quality. Further bit rate savings are envisaged, based on more sophisticated schemes for transmitting the average RMS value.

Quality assessment, so far, has been restricted to very informal listening tests. A more formal assessment using MOS or DRT tests is planned for the near future. We are currently testing new methods for improving the synthetic speech quality and the robustness of the overall coding scheme.

6. AUDIO DEMONSTRATION

The CDROM version of this paper includes an audio demonstration with the following set of 5 files:

Original (16 bit, 8 KHz sampling frequency)

[SOUND 0448_01.WAV]

Reference LPC Vocoder (unquantized)

[SOUND 0448_02.WAV]

Phonetic vocoder (unquantized, automatic phone alignment)

[SOUND 0448_03.WAV]

Phonetic vocoder (quantized, automatic phone alignment)

[SOUND 0448_04.WAV]

Phonetic vocoder (quantized, automatic phone recognition)

[SOUND 0448_05.WAV]

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues Drs. Céu Viana and Isabel Mascarenhas (CLUL), for their suggestions and their hard work in the manual correction of the EUROM.1 corpus labelling and also Prof. Luís Oliveira (INESC/IST), for many fruitful discussions.

7. REFERENCES

1. A. B. Carlson, “Communication System”, Mc Graw Hill, Book, Third Edition, 1886.

2. J. R. Crosmer, T. P. Barnwell, “A Low Bit Rate Segment Vocoder Based on Line Spectrum Pairs”, IEEE, Int. Conf. Acoust., Speech, Signal Processing, 1985.

3 P. Jeanrenaud, P. Peterson, “Segment Vocoder Based on Reconstruction With Natural Segments”, Int. Conf. Acoust., Speech, Signal Processing, 1991.

4. Y. Linde, A. Buzo, R.. Gray, “An Algorithm for Vector Quantization”, IEEE Trans, on Communications, 1980.

5. R. McAulay, T. Quartieri, “Magnitude-Only Reconstruction Using a Sinusoidal Speech Coder”, ", Int. Conf. Acoust., Speech, Signal Processing, 1984.

6. J. Picone, G. Doddington, “A Phonetic Vocoder”,- Proc. Int. Conf. Acoust., Speech, Signal Processing, 1989.

7. C. Ribeiro, I. Trancoso, “Phonetic Vocoding With Speaker Adaptation”, Proc. 5th European Conference on Speech Communication and Technology, 1997.

8. S. Roucos, A. Wilgus, “Speaker Normalization Algorithms For Very Low Speech Coding”, Proc. Int. Conf. Acoust., Speech, Signal Processing, 1984.

9. Y. Shiraky, M. Honda, “Speaker Adaptation Algorithms Based on Piece-wise Moving Adaptive Segment Quantization Method”, Int. Conf. Acoust., Speech, Signal Processing, 1990.

10. J. Slifka, T. Andersom, “Speaker Modification With LPC Pole Analysis”, Proc. Int. Conf. Acoust., Speech, Signal Processing,, 1995.