USING LINGUISTIC KNOWLEDGE TO IMPROVE THE DESIGN OF LOW-BIT RATE LSF QUANTISATION

J.J. Parry, I.S. Burnett, J.F. Chicharo

TITR Whisper Laboratories,
University of Wollongong,
NSW, Australia

ABSTRACT

In this paper we investigate an alternative approach to the design of low-bit rate (LBR) quantisation. This approach incorporates phonetic information into the structure of Line Spectral Frequency (LSF) codebooks. In prior work vector quantisation (VQ) has been used to quantise stochastic processes. Speech signals can, however, be described in terms of phonetic segments and linguistic rules. A trained LSF codebook, like the phonetic inventory of a language, is a static description of spectral behaviour of speech. As clear relationships exist between phonetic segments and LSFs the structure of an LSF codebook can be analysed in terms of the phonetic segments. The investigation leads to the conclusion that phonetic information can be usefully employed in codebook training in terms of perceptual performance and bit-rate reductions.

1. INTRODUCTION

The quantised spectral envelope of speech represents an important part of the bit allocation in speech coding. Low bit-rate approaches to spectral envelope quantisation utilize Linear Prediction (LP) techniques to exploit the redundancies offered through the quasi-periodic structure of speech. The efficient representation of LP coefficients (or LPCs) can be achieved using reflection coefficients, the arcsine of reflection coefficients, log-area ratios, intermittence spectral frequency pairs and line-spectral frequency-pairs (LSFs). LSFs are a very popular representation due to their stability and advantages in efficiency and error correction.

Reducing rate-distortion levels while maintaining speech transparency is increasingly difficult at low-bit rates. The success of several LBR quantisation techniques is due to the utilisation of speech structure in quantiser design. Paliwal and Atal [1] report performance improvements by placing emphasis on formant peaks improving the representation of vowel structure, a factor considered to be perceptually important in speech. It has also been shown that humans have a reduced perceptual resolution in the higher frequency bands of speech [2]. Listeners were found to have difficulty distinguishing unvoiced speech from power-matched gaussian white noise.

Varying the levels of information content required for different speech classes has also been explored. In some work on variable rate coding [3,4] speech was divided into general categories based on silence, voiced and unvoiced speech and voice-onset information. While this and other work [5] has claimed to use phonetic segmentation, there has been no attempt to incorporate actual phonetic information into the design of the quantiser.

The main motivation for the work presented in this paper is to investigate the role of phonetic structure in the quantisation of low-bit rate speech coding parameters. Prior work [6] shows that inter-language phonetic differences are not reflected in the structure of vector quantisers designed using a standard mean squared error (MSE) measure. When quantising speech, not pertaining to the language of the codebook training set, quantitative cross-language performance tests yielded significant type 2 outliers. The three criterion for transparent speech are 1) an average spectral distortion of 1dB, 2) less than 2% of outliers between 2dB and 4dB (type 1 outliers) and 3) no outliers greater than 4dB (type 2 outliers). The globally minimal solution of a MSE approach provides a robust quantiser design but information theory [7] suggests that a much lower entropy solution could be achieved through the analysis and exploitation of redundancies in the phonetic structure of language. Fundamental work in information theory [8] suggested that the minimum entropy of speech is based in part on the phonetic constituents of language. It is therefore reasonable to suggest that quantisers design based on phonetic structure will provide improvements in rate-distortion ratios.

The organization of the paper is as follows. Section 2 presents a phonetic analysis of the LSF domain and explains how the various phonetic components contribute to the overall codebook structure. Section 3 explains how structural phonetic information can be used to effectively design LSF codebooks with comparable subjective and objective quality to standard codebook design approaches.

2. PHONETIC ANALYSIS OF THE LSF DOMAIN

LSFs, unlike LPCs, provide us with a perceptually meaningful representation of a section of speech. The frequency values of the LSFs directly correspond to the speech spectrum and their behaviour over time can be directly related to evolutionary characteristics of that spectrum e.g. the growth and dissipation of formant activity. Further, the analysis of LSFs across individual phonemes yields important information about the distinct structure of a given speech segment in the LSF domain. LSFs (as opposed to other representations of the LP parameters) are particularly useful since they exhibit a localized spectral sensitivity property [9]. This facilitates an isolated investigation into the different phonetic components of language.

2.1 Voiced phonemes in the LSF domain

The relatively high peaks in the LPC power spectrum are indicative of the presence of voiced speech and, for a correctly dimensioned LP analysis, will correspond to the formant activity. In the presence of formants, LSFs have a tendency to cluster around the angular positions corresponding to the roots of the LPC filter when they are close to the unit circle [10]. Additionally, LSFs have characteristic regions of activity inside which their clustering contributes to the representation of formant behaviour (Figure 1).

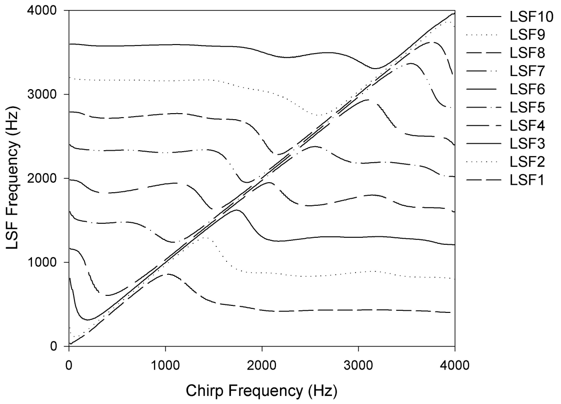

Figure 1: The regions of activity of LSFs. By converting a 0 to 4kHz simulated formant frequency sweep into LSFs, distinct regions inside which LSFs contribute to the representation of formants can be seen.

In the LSF domain the regions of activity of formants can be clearly seen. This region is bounded by a lower diagonal asymptote corresponding to the true formant frequency (Figure 2). A standard vowel is characterised by a set of formants that remain constant for a distinct period of time and, typically, two or three formants are required to specify the vowel. The synthetic experimental results shown in Figure 1 indicate that for any one formant present in a particular region of LSF space, the local (and successive) LSFs tend to cluster towards each other. This results in the lower asymptotic diagonal that can be seen across all LSFs. The clear diagonal provides a continuous mapping of the vowels, from lower frequency back vowels at the lower left to front vowels at the upper right of the graph. The characteristic formant location tends to result in the first formant lying within LSFs 1 to 3, the second formant within LSFs 4 to 6 and the higher formants occurring within the remaining higher order LSFs. This observation suggests that for split VQ, a 3/3/4 configuration is preferable as it maintains the integrity of formants. Thus for a particular language, the LSF space for vocoids can be described with the known characteristic boundaries of successive formants in a region about the diagonal.

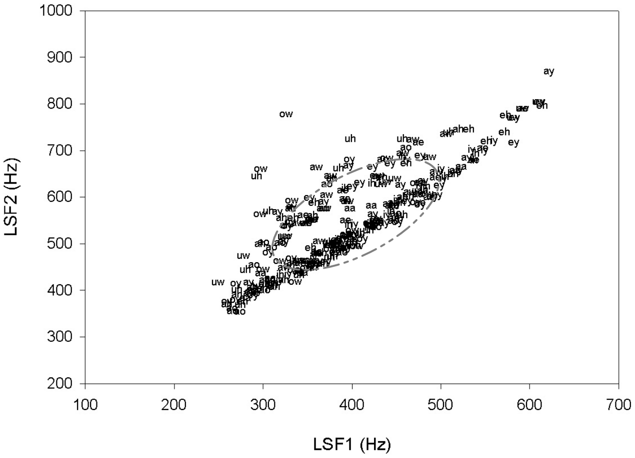

Figure 2: Scatter plot displaying the distribution of voiced phonemes (vowels, glides and semi-vowels) in a phonetically structured 30-bit LSF codebook

2.2 Consonants in the LSF Domain

As mentioned in section 1, it is well known that humans have a reduced level of perceptual frequency resolution in the upper end of the speech spectrum and that in this region it is difficult to distinguish between the complex structure of unvoiced speech and simple Gaussian noise. It can be seen in Figure 3 that the density of LSF vectors is considerably less dense for the unvoiced consonants.

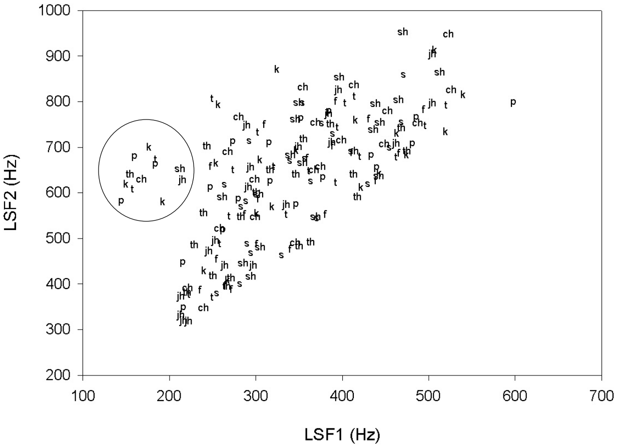

Figure 3: Scatter plots displaying the distribution of unvoiced pulmonic consonants (plosives, fricatives, affricates and laterals) in a phonetically structured 30-bit LSF codebook

Consonants can be completely unvoiced or consist of a mix of voiced and unvoiced speech. The differing vector distributions of voiced and unvoiced consonants are shown in Figure 3 and 4. The voiced diagonal aspect of the consonants is clearly seen in Figure 4. The common unvoiced aspects of each voiced and unvoiced consonant occupy a common region of the LSF space. The circled area in Figure 3 and 4 indicates a clear region pertinent to voiced and unvoiced plosives. Similarly it can be shown that in a pair-wise manner, the unvoiced content of voiced and unvoiced consonants map onto similar regions of LSF space.

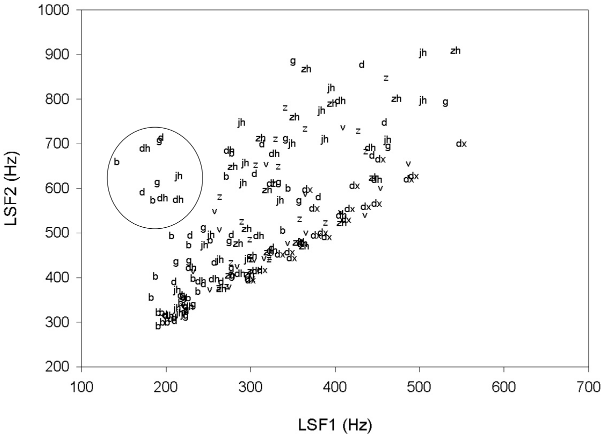

Figure 4: Scatter plot displaying the distribution of voiced pulmonic consonants (plosives, fricatives, affricates and laterals) in a phonetically structured 30-bit LSF codebook

3. EXPERIMENTS

To further investigate the nature of the phonetic structures described in section 2 a series of codebooks were generated in a process whereby a phonetic constraint was used in the training. The same set of speech was used to train a set of standard MSE split VQ codebooks. A "control" set of codebooks was also generated having an arbitrary constraint to ascertain that the role of phonetic structure was the contributing factor in performance.

3.1 Speech Database

The speech data used in this investigation were extracted from the TIMIT database. The data used in training comprised of speech from 460 speakers (female and male). The speech was preprocessed with a sampling frequency of 8kHz. The data was then phonetically segmented using the TIMIT phonetic transcription corresponding to the training speech (there are 51 specified phonetic groups used in the TIMIT transcription alphabet).

3.2 Phonetic codebook design

From the information gathered in section 2 the following guidelines can be made with regards to the phonetic design of LSF codebooks.

1) From [4] a bit allocation ratio of ~ 3:1 between voiced and unvoiced frames can be used to achieve subjective transparency at 24 bits/frame.

2) Using the TIMIT phonetic labeling, 30 voiced phonemes comprise the voiced phonetic space of English, which after [4] corresponds to 9 vectors per voiced phone.

3) Due the mapping property of pulmonic consonants, it suffices to train the consonant codebook structure using only the unvoiced consonants (if a voiced-unvoiced pair exists in the language being modeled).

Each vector in the training sequence was allotted to separate phone-specific LSF codebooks. On completion of training the codebooks were reintegrated into a complete set of 3 split VQ sub-codebooks. This was repeated for 5 different codebook sizes having the specified ratio of vectors for each phoneme type.

3.3 Quantisation Performance

Objective Tests

Codebook quantisation distortion between the original LPC spectra and the phonetically quantised LPC spectra was measured using the standard SD measure (in dB) where for a particular frame i the SD is measured as:

$$SD_i = \frac{1}{F_s}\int_0^{F_s}\left[10\log_{10}(P_i(f)) - 10\log_{10}(\hat{P}_i(f))\right]^2 df$$

where Fs is the sampling frequency and $P_i(f)$ and $\hat{P}_i(f)$ are the LPC power spectrum. The set of TIMIT "test" sentences was used in objective testing across a range of codebook sizes. Figure 5 compares the SD quantisation performance of the phonetically structured quantiser with the standard MSE quantiser across the set of test sentences. As would be reasonably expected the MSE trained codebook outperforms the phonetically trained codebook in terms of the square error based SD. However as is well established coder performance is best assessed using subjective tests. In this case a series of simple pair-wise comparisons were performed across a listener base.

Subjective Tests

A set of six sentences (balanced between male and female) were selected from the "test" section of the TIMIT database. The sentences were encoded using both a standard MSE codebook and a phonetically structured codebook for a range of codebook sizes. The sentences were played to a listener base of ten adult speakers who were asked to indicate which sentence of the pair they preferred. The results of this pair-wise comparison are presented in Figure 6.

For equivalent quality the MSE and phonetic codebooks would score equal rankings in the subjective pair-wise testing. Alternatively listeners would express no preference. The results show that in general the perceptual phonetic codebook (which is substantially non-optimised) compares favourably with the standard MSE trained codebook. Further a dramatic difference in performance between the phonetic and control codebooks is demonstrated. It is clear that the use of phonetic segmentation techniques can produce codebooks of similar performance to existing training techniques. Further refinement of the techniques for building phonetically segmented codebooks should lead to improved overall performance at lower bit rates. It is important to note the contrast in these subjective results compared to the objective results discussed previously. While this is to be expected from the well-known differences in subjective and objective measures of speech quality, the differences also highlight more complex perceptual effects. For example investigation of the objective performance

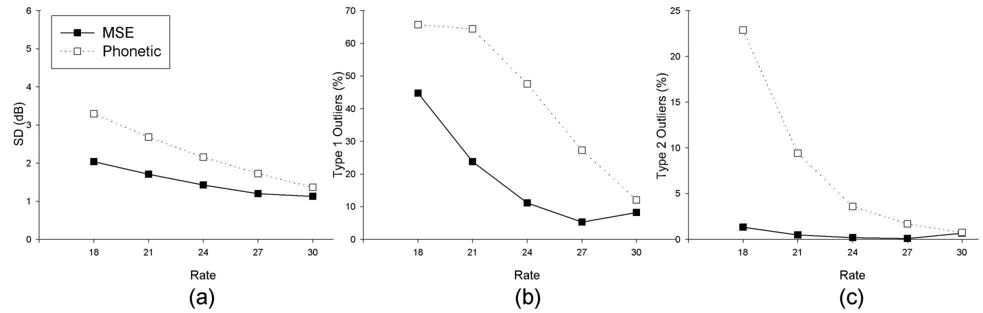

Figure 5: SD Objective performance tests; MSE cw. phonetic segmentation (a) Mean SD (b) Type 1 outliers (c) Type 2 outliers

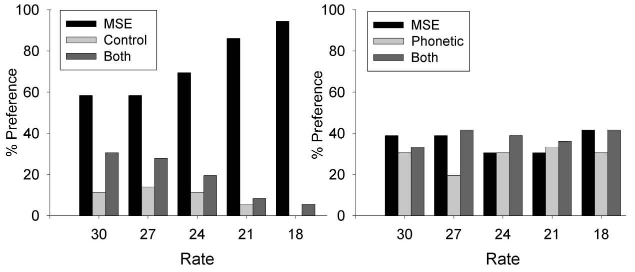

Figure 6: Pair-wise subjective performance tests comparing standard MSE codebooks with phonetically segmented codebooks

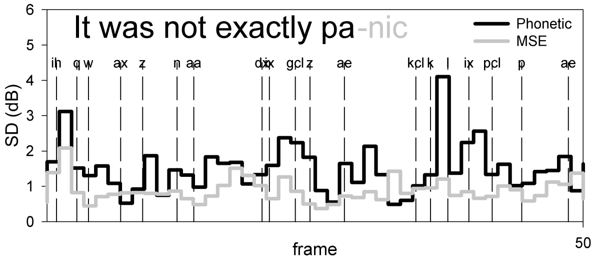

Figure 7: Phoneme-by-phoneme SD distortion measures comparing standard MSE codebooks with phonetically segmented codebooks (30 bit codebook)

demonstrates substantially increased type 1 and type 2 outliers for the phonetic codebook. However these subjective results indicate that, and checks confirm that the outliers correspond to perceptually unimportant phonetic segments. A simple indication of this phenomenon is shown in Figure 7 where phonetically labeled plots of SD for both standard and phonetic codebooks are shown.

4. CONCLUDING REMARKS

This investigation into the use of phonetic structure in the design of LSF codebooks has illustrated through subjective and objective tests that significant redundancies are present in the standard MSE style approach to quantiser design. Similar performance levels to the MSE were reported when using a quantiser that was designed using training stimulus limited to individual phoneme categories. This approach further refines other approaches to phonetic segmentation [4] by restricting the spectral representation to only the pertinent phonetic regions. This provides a framework for further reductions of rate-distortion levels as a function of language.

5. ACKNOWLEDGEMENTS

J.J. Parry is funded under an Australian Postgraduate research scholarship and is the recipient of a Motorola (Australia) Partnerships in Research Grant.

6. REFERENCES

1. K.K. Paliwal and B.S. Atal, “Efficient vector quantization of LPC parameters at 24 bits/frame”, IEEE Trans. Of Speech and Audio Process, pp.3-7, Jan., 1993
2. G. Kubin, B.S. Atal, and W.B. Kleijn, “Performance of noise excitation for unvoiced speech,” *Proc. IEEE Workshop Speech Coding for Telecom.*, pp. 35-36, October 1993.
3. S. Wang and A. Gersho, “Phonetically-based vector excitation coding of speech at 3.6 kbit/s,” *Proc.IEEE Int. Conf. On Acoust., Speech, and Sig. Proc.*, May 1989, pp.I-49-52.
4. R. Hagen et al. “Variable rate spectral quantization for phonetically classified CELP coding,” *Proc. IEEE Int. Conf. On Acoust., Speech, and Sig. Proc.*, 1995, pp.748-751.
5. T.M. Liu and H. Hoege, “Phonetically-based LPC vector quantization of high quality speech,” *Proc. European Conf. Speech Technology,* September 1989.
6. J.J. Parry et al. “Language-specific phonetic structure and the quantisation of the spectral envelope of speech” *paper in preparation*.
7. W.B. Kleijn and K.K. Paliwal, “Quantisation of LPC Parameters”, in: W.B. Kleijn and K.K. Paliwal, *Speech Coding and Synthesis* (Elsevier Science), pp. 433-466, 1995.
8. C.E. Shannon, “A mathematical theory of communication.” *Bell Syst. Tech. J.*, Vol. 27, pp. 379-423, pp. 623-656, 1948.
9. F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals”, *J.Ac.Soc.Am.*, Vol.57 ,1975.
10. G.S. Kang and L.J. Fransen, “Low-bit rate speech coders based on line spectral frequencies (LSFs)”, *Naval Research Laboratories Report 8857*