LINGUISTIC MAPPING IN LSF SPACE FOR LOW-BIT RATE CODING

J.J. Parry, I.S. Burnett, J.F. Chicharo
TITR Whisper Laboratories,
University of Wollongong,
NSW, Australia

ABSTRACT

In this paper we investigate the spectral density of Line Spectral Frequency (LSF) content in languages. The results show that the phonetic variation of languages is reflected in the LSF space. This leads to an alternative approach to the design of LSF quantisers. 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 new approach incorporates phonetic information into the structure of LSF codebooks through combining individual phonetic codebooks. 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 illustrates how the composition and density of the phonetic makeup of speech can vary across languages. Section 3 presents a phonetic analysis of the LSF domain and explains how the various phonetic components contribute to the overall codebook structure. Section 4 shows how structural phonetic information can be used to effectively design LSF codebooks with comparable subjective and objective quality to standard codebook design approaches.

2. STATISTICAL PROPERTIES OF LANGUAGE SPECTRA

As languages have a distinct phonetic make-up it follows that they will also have characteristic spectral structures. Figure 1 shows the bi-variate spectral histogram of the LSF (1-2) spectra of a statistically large sample (144,000 vectors) of Mandarin speech (taken from the OGI-multi-language speech database [9]). The magnitude of the bi-variate spectral histogram of a language provides information about the relative frequency of occurrence of sounds of a language. Figure 2 shows the bi-variate spectral histogram of a similar sample size of Vietnamese. It is clear that while Mandarin and Vietnamese occupy a similar LSF space, the characteristic distribution is quite different. Examination of the spectral structure of a large number of languages (reported in full detail in [10]) shows that the LSF spectra of languages generally occupy the same regions of LSF space but the distribution of vectors can be highly varied. Consequently it is important that

these variations are catered for when designing the structure of LSF codebooks.

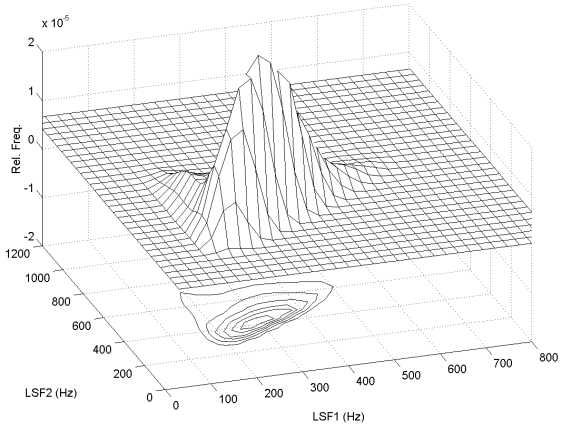

Figure 1: The cumulative bi-variate spectral density of the LSF (1-2) spectrum of OGI-11 MANDARIN (144,000 vectors)

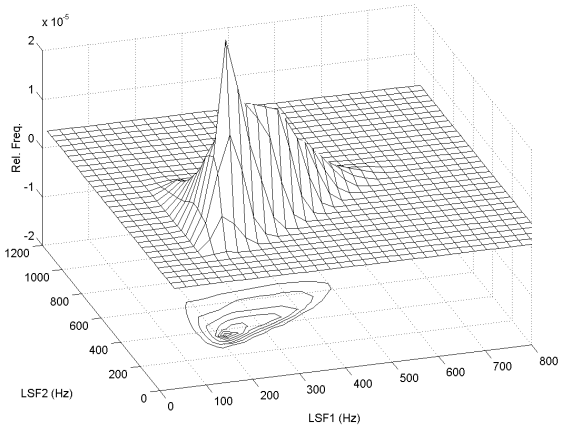

Figure 2: The cumulative bi-variate spectral density of the LSF (1-2) spectrum of OGI-11 VIETNAMESE (144,000 vectors).

3. MULTI-LANGUAGE SPECTRAL QUANTISATION

Work assessing multi-language quantisation performance [11,12,13] have reported that robust cross-language quantisation is achieved using both split and multi-stage VQ techniques. It is evident, from cumulative LSF density spectra across languages (a full range of languages is shown in [10]), that the claims of similar quantisation performance are valid within the scope of the presented work (A group of Indo-European languages, English, German, Italian and Norwegian). It is however, unwise to extrapolate these results to an assumption that quantisation performance is similar across all languages. In the training of LSF codebooks a Mean Squared Error (MSE) algorithm allocates codebook vectors as a function of the relative presence of each sound in the training stimulus. The associated repercussions of multi-language LSF quantisation with MSE designed codebooks can be seen most clearly in the nature of quantisation outliers in particular, type 2 outliers. In prior assessments of LSF quantisation [6] it was shown that the presence of type 2 outliers was quite varied across languages. Figure 3 gives an example of how type 2 outliers can vary across languages (see [10] for a more comprehensive study of all the 22 languages of the OGI database). Comparing Figures 3 (a) and (b) it is clear that the use of Vietnamese as a training language has resulted in a much higher presence of type 2 outliers than that of Mandarin. The visible differences seen in the cumulative LSF density spectra are clearly reflected in the structure of the codebooks and consequently in the relative magnitude of type 2 outliers (It is interesting to note that in [11] the behaviour of type 2 outliers was not considered).

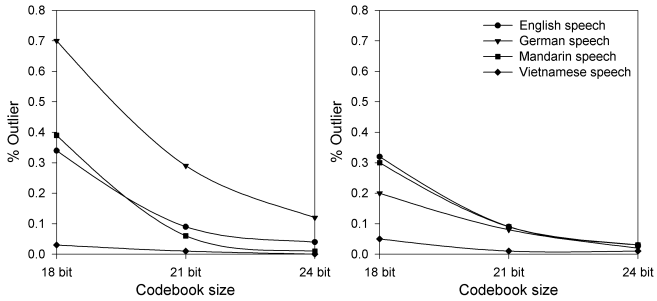

Figure 3: Type 2 outliers present when quantising 4 OGI test languages (60,000 vectors for each language) using split VQ LSF codebooks trained on (a) Vietnamese and (b) Mandarin

From these results it is evident that even between two languages picked from the large number of World languages substantial differences in spectral content exist. The exploitation of such differences across all languages requires a common approach to selecting language, and phonetically, sensitive codebooks. Here we suggest the construction of codebooks from smaller trained phonetic unit codebooks. This allows control the phonetic content and the flexibility to cater for the known phonetic variations across languages.

4. PHONETIC MAKE-UP OF LSF SPACE

LSFs, unlike LPCs, provide us with a perceptually meaningful representation of a section of speech. The frequency values of 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 [14]. This allows an isolated investigation into the phonetic components of language in the context of speech coding parameters. 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 [15] (Figure 4 graphically illustrates this behaviour).

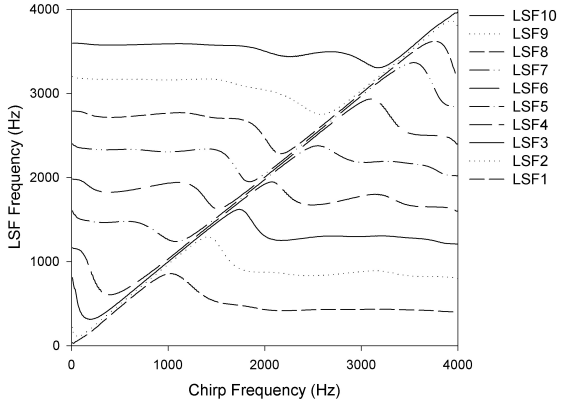

Figure 4: The regions of activity of LSFs. By converting a 0 to 4kHz simulated formant frequency sweep into LSFs, distinct regions inside which LSFs contribute to the representation of formants can be seen.

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 (in the LSF domain this corresponds to the juncture of subsequent LSFs). Figure 5 was created using a Bi-Variate LSF histogram of separate samples of phonemes extracted from TIMIT-TRAIN. It shows the primary regions of LSF activity of a number of phonemes (concentrating on LSFs 1 and 2). It is apparent that voiced phonemes occur inside tight zones, while unvoiced phonemes occur in relatively larger zones. The regions of voiced activity shown in Figure 5 are clearly oriented along a diagonal region. There is a progression of vowel zones along the diagonal of Figure 5, from closed vowels at the bottom to open vowels at the top. This explains the high density of LSF vectors along the diagonal of Figure 1 as Mandarin has a high voiced speech content. 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 5 that the regions of unvoiced activity are considerably wider spread. While there is considerable overlap in the distribution of LSF vectors among unvoiced phonemes it can be seen that they occur in distinct, identifiable regions.

5. EXPERIMENTS

To further investigate the nature of the phonetic structures described in Section 4 a series of codebooks were generated in a process whereby individual phonemes from the TIMIT-TRAIN database were segmented out and clustered together for training. The unaltered TIMIT-TRAIN was also used to train a set of standard MSE split VQ codebooks. Further, to verify the role of phonetic segmentation, TIMIT-TRAIN was also segmented and re-clustered together in an arbitrary manner so as to train a set of "control" codebooks.

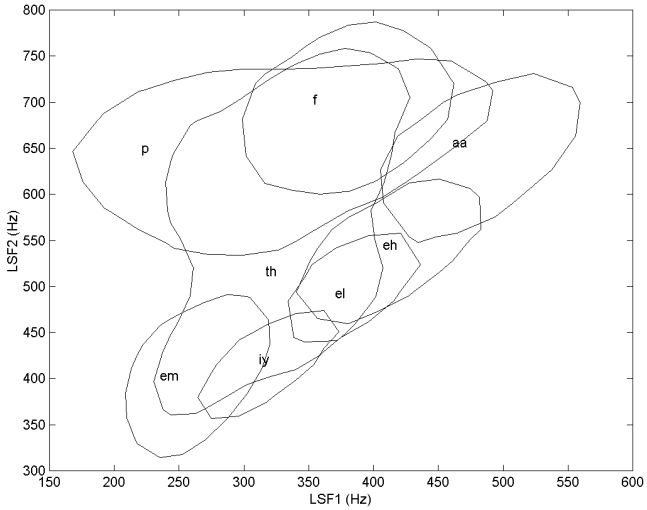

Figure 5: The phonetic composition of LSF(1-2) space.

5.1 Speech Database

The speech data used in this investigation were extracted from the TIMIT database [16]. The data used in training comprised of speech from 460 speakers (female and male). The speech was re-sampled to 8kHz and then segmented into phonetic units using the TIMIT phonetic labeling information. (the chosen labeling scheme used by TIMIT uses 51 distinct phonetic units).

5.2 Quantisation Performance

Objective Tests

Codebook quantisation distortion between the original LPC spectra and the phonetically quantised LPC spectra was measured using the standard spectral distortion (SD) measure (in dB). The entire set of TIMIT-TEST sentences was used in objective testing across a range of codebook sizes. Figure 6 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 was selected from TIMIT-TEST. 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 7. For equivalent quality the MSE and phonetic codebooks would score equal rankings in the subjective pair-wise testing. Alternatively listeners were allowed to express no preference between the two. 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 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.

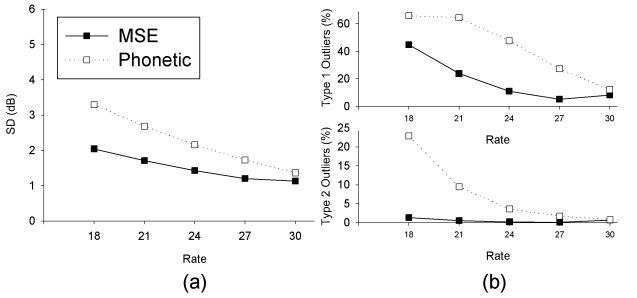

Figure 6: SD Objective performance tests; MSE cw. phonetic segmentation (a) Mean SD (b) SD Outliers.

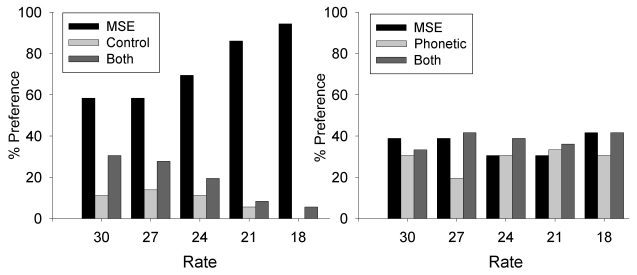

Figure 7: Pair-wise subjective performance tests comparing standard MSE codebooks with phonetically segmented codebooks

6. CONCLUDING REMARKS

This paper has looked at the statistical characteristics of language behaviour in LSF space and demonstrated that there are notable differences in the distribution and concentration of LSFs between languages. The consequences, in a multi-lingual environment, of training LSF codebooks using speech of a single language were discussed. This is particularly reflected in type 2 outlier performance results. The concept of designing codebooks using individual phonetic elements was introduced. An investigation into the use of phonetic structure in the design of LSF codebooks was then used to illustrate, through subjective and objective tests that significant redundancies are present in the standard MSE style approach to quantiser design. The reported subjective test results indicate favourable as compared to MSE codebooks. However, it must be emphasied that these results use codebooks constructed of an un-optimised mix of phonetic segments. The use of phonetic density information to improve the performance of phonetically constituted codebooks is currently under investigation.

7. ACKNOWLEDGEMENTS

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

8. 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”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. Y.K. Muthusamy et al.,"The OGI Multi-language Telephone Speech Corpus" Proceedings of the Int. Conf. on Spoken Lang. Process., October 1992.
10. TITR Whisper Labs, University of Wollongong web page: http://www.whisper.elec.uow.edu.au/Xlanguage/icassp99/
11. W.P. LeBlanc et al. “Joint design of multi-stage VQ codebooks for LSP quantization with applications to 4 kbit/s speech coding”, in: B.S. Atal et al., Speech and Audio Coding for Wireless and Network Applications (Kluwer Academic Publishers), 1993
12. R. Montagna, “Selection phase of GSM half rate channel” in Proc. IEEE Speech coding workshop, pp.95-96, 1993.
13. C.R. South et al. “Subjective performance assessment of CCITT’s 16 kbit/s speech coding algorithm” Speech Communication, vol.12,pp.113-133,1993
14. F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals”, J.Ac.Soc.Am., Vol.57 ,1975.
15. G.S. Kang and L.J. Fransen, “Low-bit rate speech coders based on line spectral frequencies (LSFs)”, Naval Research Laboratories Report 8857
16. J.S. Garofolo et al. "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM"