THE CONSEQUENCES OF LINGUISTIC PERCEPTION ON LOW-RATE SPEECH CODING

John J. Parry, Ian S. Burnett and Joe F. Chicharo

Department of Electrical and Computer Engineering
University of Wollongong, NSW, Australia
http://www.whisper.elec.uow.edu.au
jack@snrc.uow.edu.au

ABSTRACT

This paper considers the issue of the effect of languages and linguistic perception on low rate speech coding. Current algorithms exploit the redundancies of speech but these redundancies are not common across all languages. Similarly speech coder evaluation techniques do not take into account the nuances of linguistic perception across languages. This paper illustrates some of the linguistic sensitivities experienced by low-rate coders and explores approaches to low-rate coder design. This is achieved through an evaluation of cross-language spectral distortion measures which account for specific linguistic peculiarities influencing linguistic perception.

1. INTRODUCTION

Speech coders are designed to minimize the bandwidth required for speech communication, the latest algorithms [1,2] allowing speech transmission at 2.4kb/s. Speech is highly redundant. But as compression ratios increase the exploited redundancies are not universal across all languages [3]. Speech coder evaluations [4] have shown coder performance biases towards certain languages. In each case the evaluation results favour the language group of the country developing the coder [2,5]. Initial investigations [3] looked into the origins of these linguistic biases and identified that one source of linguistic bias is spectral quantisation. Line Spectral Frequencies (LSFs) are used in most low-rate coders to represent spectral behaviour. In this work, differences were seen in the spanning nature of LSFs across language groups. The acoustic nature of languages is quite complex and employs many distinct auditory processes in different ways. Kawahara [6] has shown that humans adopt modes of perception optimized for the acoustic nature of a language. Current low-rate coders [1,2] have bit allocation schemes that give primary importance to spectral structure but little emphasis to the other auditory factors. This choice, while previously appearing circumspect, is brought into question by research results [6] indicating the perceptual relevance of other factors in some languages. This paper illustrates some of the linguistic sensitivities of low-rate speech coders and then presents the results of a cross-language study on the effects of language-specific spectral quantisation.

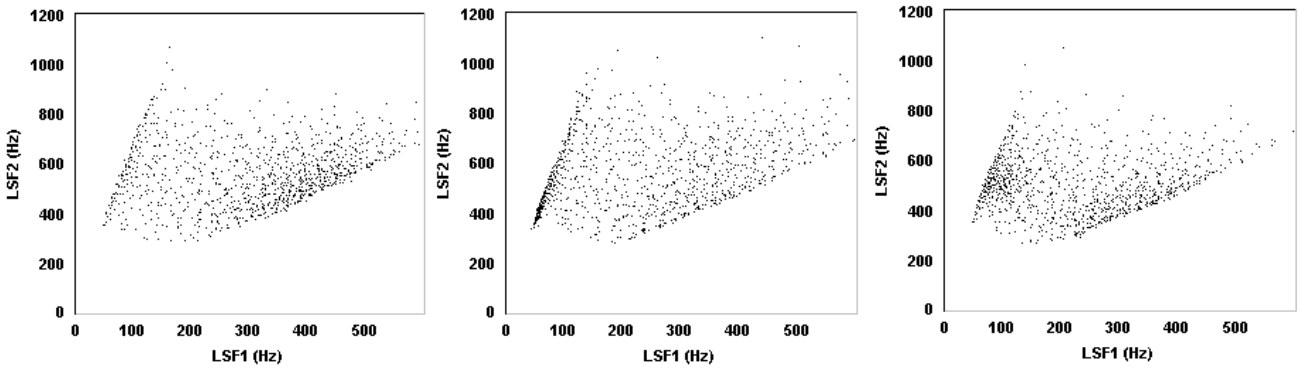
2. LINGUISTIC SENSITIVITIES OF LOW RATE SPEECH CODERS

The development of low-rate coders is being carried out in many countries of the world. It is apparent however that in the effort to lower bit rates, linguistic robustness may have been overlooked. Many advances in bit rate reduction can be attributed to linguistic redundancies [5,7] and consequently may prove very limited in global applications such as future cellular systems. Previous work [3] has illustrated the linguistic sensitivities of some low-rate speech coders [1,2]. This work shows that the paradigms used in LSF quantisation are not linguistically robust.

2.1 LSF Spectral Quantisation

In modeling the formant structure of speech, most current speech coders use Linear Prediction filters. In low-rate coders these filters are described in terms of LSFs, a more robust representation of Linear Prediction Coefficients. Major bit rate reductions achieved in modern low rate speech coding are attributed in a large part to the use of quantised LSFs. Using LSFs for quantisation permits a 10 coefficient LPC representation using as few as 24 bits/frame [11]. LSF quantisation is a process whereby a finite set of LSF vectors are compared to the input LSF vector and the closest codebook vector to the input vector is used to represent that input. The codebook is generated through a process where quantised LSFs are extracted from a sequence of training speech.

The nature of the training speech thus governs the nature of the LSF quantisation. Figure 1 shows scatter plots of the first two LSF quantisation vectors used in three example codebooks using a 30 bit split VQ algorithm (where 10 bits are used to quantise the first 3 LSFs). The codebooks in Figure 1 were trained using 60 minutes of uninterrupted male and female speech in English, Spanish and Mandarin respectively. Some languages contain distinctive speech sounds with formant structures not found in other languages. As such, the process of quantisation using codebooks trained on different languages will result in the modification of original formant structures. Thus the true sound of the original speech will be modified. The degree of cross-language spectral modification is therefore put into question.



(a) English Codebook (b) Mandarin Codebook (c) Spanish Codebook

Figure 1 Split VQ Distributions for LSF1 vs. LSF2

3. LINGUISTIC PERCEPTION AND LOW-RATE CODER DESIGN

A fundamental problem with the development of speech processing systems is the treatment of human speech perception. The concept of "natural speech" is relative to the linguistic background of a speaker, and languages employ different aspects of speech to convey meaning. The relative importance of acoustic information thus differs in each language. With maturity, humans develop the ability to selectively mask speech so as to place emphasis on perceptually relevant information. Knowledge of perceptual weighting is thus, of paramount importance in the processing of speech in different languages. For example, in the perception of Japanese words, speech energy plays a primary role [6] whereas it is considered to have little perceptual importance in other languages. Subjective tests have shown [6] that certain spectrally distinct phones are perceptually distinguishable for some language groups but not others. Thus, the observations of the acoustic characteristics of languages show quantitative differences, and, in order to design coders and correctly evaluate their performance, perceptual performance needs to be considered. Many quantitative approaches exist for the evaluation of low-rate coders and some consider the perceptual boundaries of the listener. However, none take into consideration the varying perceptual characteristics of other language groups. Current low-rate speech coders represent speech in terms of factors such as spectral structure (LP coefficients), pitch, energy, gain and voicing information. In most cases the main emphasis (bit allocation) is placed upon the spectral structure. However this may not be desirable for all languages. For example, work on low-rate coders using tonal languages [5] has focused upon pitch estimation whilst using only half the resources of standard coders for spectral structure. This is due to the more important role played by pitch in intelligibility of tonal languages. Perceptually weighted spectral distance measures in the training of LSF codebooks have been found to improve the "clarity" of speech in English. "Clarity" is however a perceptual factor and thus the degree of weighting should be dependent on the perceived importance of "clarity" to a language. All languages have specific perceptual weightings. Thus a low-rate coder, when perceptually weighted towards a language, will provide a more "natural" sounding speech to a native listener.

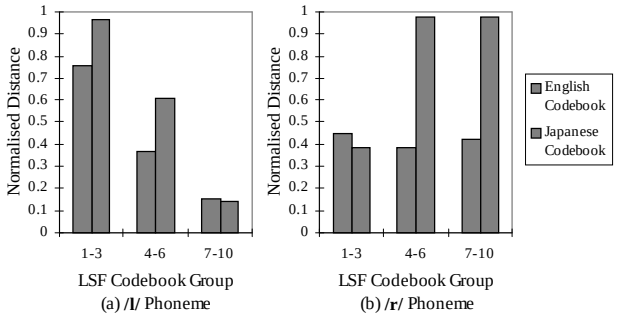


Figure 2 Codebook Distortions of Perceptual Contrasts /r/ and /l/

3.1 Distance Measures and Perception

Measuring cross-language codebook distortion provides us with a clear indication of the linguistic sensitivities of spectral quantisation. Studies of perceptual learning [8] suggest that the development of the ability to perceptually contrast speech sounds is determined by the acoustic characteristics of the stimulus language. The **/ r /-/ l /** perceptual contrast of Japanese speakers is well documented [8]. The nature of this contrast can be illustrated using codebook distortion measures (Figure 2). It can be seen that distances are similar for both English and Japanese codebooks for the phoneme **/l/**, but considerable differences exist in the relative distances of the phoneme **/r/**. This difference indicates that the English phoneme **/r/** is not well represented in the Japanese codebook reflecting its absence in the Japanese language [9].

4. RESULTS

While the above example is particularly clear, the principle can also be extended across languages. It is then clear that significant spectral quantisation performance variation will be evident across languages. To illustrate this, exhaustive tests were carried out on a substantial amount of language data. In this evaluation, a 10 language database, consisting of continuous telephone quality speech in English, Farsi, French, German, Hindi, Japanese, Mandarin, Spanish, Tamil and Vietnamese was used. The codebook of each language was used to quantise

LSFs generated from samples of the 10 language set (each sample being 10,000 frames long consisting of male and female speech). The quantised LSFs were compared with the original LSFs and mean squared distance measures were calculated. Distance measures were taken for low, middle and high order LSFs (corresponding to the 3,3,4 split VQ codebook scheme). Fifteen separate trials were carried out on each language codebook, with 90% confidence levels. The results are shown in Figures 3-5.

4.1 Lower Order Spectrum Trials

An English Codebook displays minimal cross-language distortion in the low order spectrum (Figure 3(a)), but it was found that the Mandarin codebook had the lowest overall distortion (Figure 3(c)). Mandarin had notably low distortion levels across codebooks as well. Thus a Mandarin trained codebook is an attractive model for cross-language codebook design in the lower order spectrum. Japanese and Vietnamese (Figure 3(b)) exhibited the highest codebook distortion for the lower order spectrum.

4.2 Middle Order Spectrum Trials

The nature of the middle order spectrum was not as varied as the higher and lower spectrum. Again the English codebook has low distortion in the middle order spectrum (Figure 4(a)). The Hindi codebook gave the lowest distortion levels (Figure 4(c)) and the Tamil codebook (Figure 4(b)) had marginally the highest distortion levels.

4.3 High Order Spectrum Trials

In high frequency spectrum trials on cross-language codebook distortion, English exhibited a significantly high distortion in all cases except when using the English codebook. Cross-language distortions varied from 18% to 33% higher than that of the English codebook (Figure 5(a)). This indicates that the high order spectral features of English are distinct from those of the other languages tested. Previous work on Split VQ [10,11] has reported that bit reductions can be made, by placing quantisation constraints on the higher frequency spectrum (of English), without affecting perceptual performance. The reductions achieved in [10,11] may be attributed to this characteristic.

Excluding English, all the codebooks displayed the same distortion levels (within confidence limits). The Farsi codebook was marginally the lowest (Figure 5(c)) and the Japanese codebook (Figure 5(b)) marginally the highest. In all cases observed, the minimum overall distance was seen when languages were coded using their respective codebooks, in particular English had the lowest overall levels.

5. CONCLUSIONS

This paper has considered the effect of languages and linguistic perception on low rate speech coding. Pertinent linguistic sensitivities of low-rate coders were discussed and, through example, the importance of linguistic perception was explored. The investigations presented in this paper have shown how the spectral characteristics of languages influence vector quantisation in codebook training. It was found that overall, codebooks trained on English displayed the minimum overall distortion of the languages tested. It is interesting to note that in most speech coding research English has been used, somewhat blindly, as a training language, with good results. Results obtained show that while English is an acceptable codebook training language, other languages contain spectral structures that can further reduce spectral distortion. It should be noted that while quite clear, these results have been derived from a single database. Additional trials must be carried out using different language databases before any solid conclusions can be made.

Future work in codebook design should therefore be directed to this end (for example, combining the codebooks of Mandarin, Hindi and English for the low, middle and high order spectrum codebooks, respectively, provides an immediate improvement). The outcome of this work will be a linguistically robust perceptual weighting scheme. Other avenues of future work include investigation into pitch tracking algorithms, which have also been identified as a source of linguistic sensitivity [4]. The overall work will lead to linguistically optimized low-rate coders which take into account the relative importance of all the facets of languages.

6. ACKNOWLEDGEMENTS

We wish to thank The Oregon Graduate Institute for the use of their Multi-Language Telephone Speech corpus. This work was supported in part by The Australian Research Council (ARC) under their small grants program. Mr Parry is funded under an Australian Government postgraduate research scholarship.

7. REFERENCES

[1] I.S. Burnett, G. J. Bradley, “New techniques for Multi-Prototype Waveform Coding at 2.84kb/s”, *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Detroit, pp. 261-264, 1995.

[2] J.P Campbell et al, “The Proposed Federal Standard 1016 4,800 bps Voice Coder : CELP” *Speech Technology*, pp 58 - pp 64, April/May 1990.

[3] I.S. Burnett, J.J.Parry, "On The Effects of Language and Accents on Low Rate Speech Coding", *Proc. Int. Conf. Spoken Lang. Process.*, Philadelphia, pp.291-294, 1996.

[4] J.J.Parry, “Accent Classification for Speech Coding”, Honors thesis, The University of Wollongong, 1995.

[5] T. Wang et al. “A High Quality MBE-LPC-FE Speech Coder at 2.4kbps and 1.2kbps”, *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Atlanta, pp. 208-211, 1996.

[6] H. Kawahara, “Interactions between Speech Production and Perception under Auditory Feedback Perturbations on Fundamental Frequencies.” *J. Acoust. Soc. Jpn.*, Vol.15, No.3, pp.201-202, 1994.

[7] S.H.Chen, Y.R. Wang, “Vector Quantisation of Pitch Information in Mandarin Speech”, *IEEE Transaction on Communications*, Vol. 38, No. 9, September 1990.

[8] J. S. Logan et al. “Perceptual Learning of Nonnative Speech Contrasts” In J. C. Goodman, H. C. Nusbaum, “The Development of Speech Perception”, Chap. 4, pp.121-166. *The MIT Press*, London, England. 1994.

[9] M. Ruhlen, “A Guide to The Languages of The World”, *Stanford University Press*,1976

[10] W.B. Kleijn, “On Memoryless quantisation in Speech Coding”, *IEEE Signal Processing Letters*, Vol. 3, No. 8, pp.228-230, August 1996.

[11] K.K. Paliwal, B.S. Atal, “Efficient Vector Quantisation of LPC Parameters at 24 Bits/Frame”, *IEEE Trans. of Speech and Audio Proc.*, Vol. 1, No. 1,pp 3-14 January 1993.

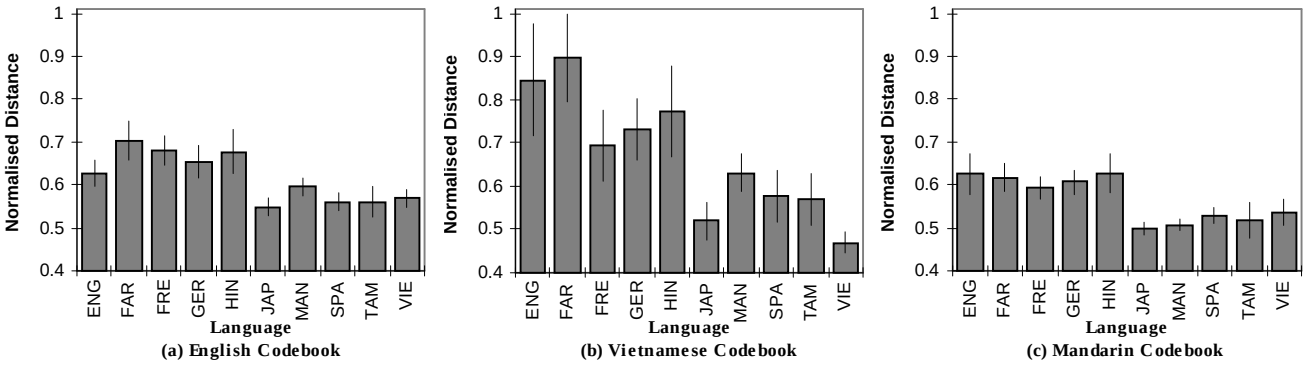


Figure 3 Codebook Distortion of Line Spectral Frequencies 1-3

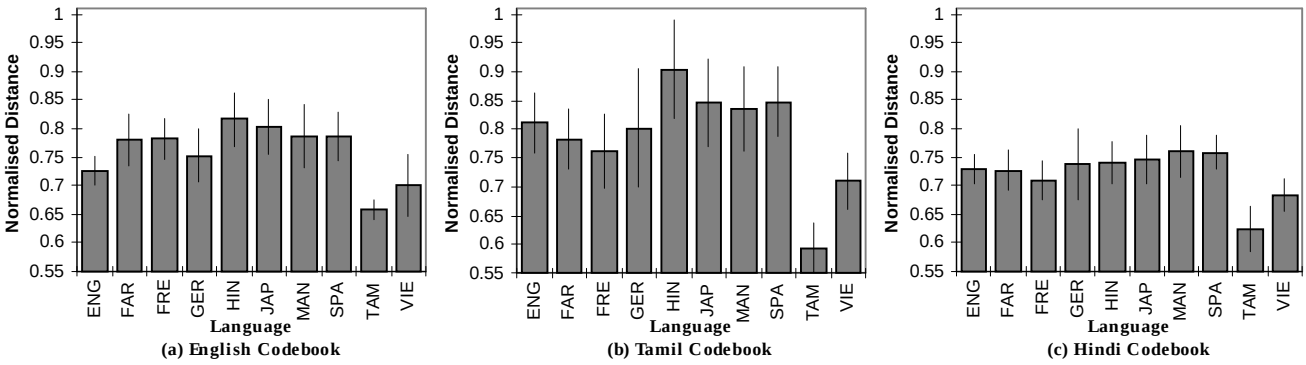


Figure 4 Codebook Distortion of Line Spectral Frequencies 4-6

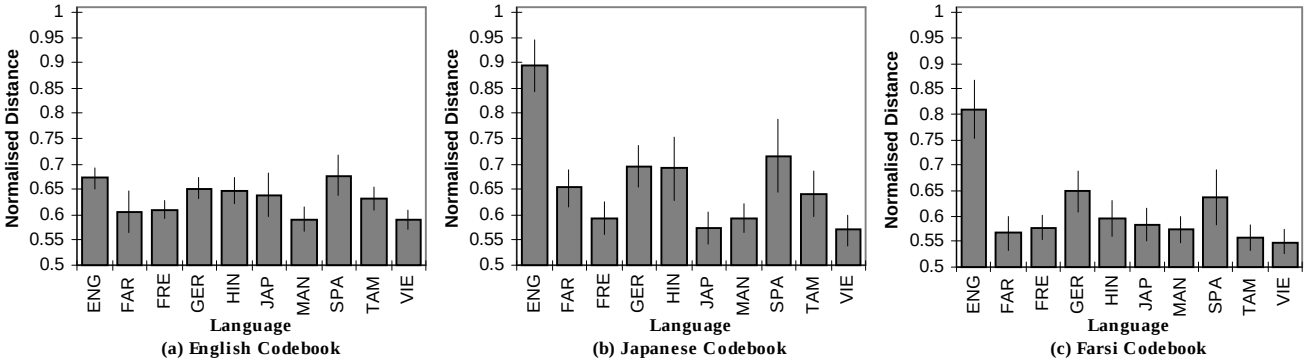


Figure 5 Codebook Distortion of Line Spectral Frequencies 7-10