

On the Effects of Accent and Language on Low Rate Speech Coders

I. S. Burnett and J.J. Parry

Department of Electrical and Computer Engineering,
University of Wollongong, NSW, Australia

ABSTRACT

Telecommunications networks are exposed to a plethora of accents and languages. Fundamental to current and future systems are low rate speech coders. This paper examines the problems associated with speech coding of different languages and accents. Our investigations show that most low-rate (8kb/s and below) speech coders show bias towards non-accented English.

When the coders are used for heavily accented English or other languages, significant performance degradation is noted. This paper examines the reasons for such variations and some approaches for improving coder performance.

1. INTRODUCTION

Previous work in the fields of language and accent classification [1] has focused upon making probabilistic decisions as to the origins of the speaker.

In this work, however, the origins are of secondary importance to the effects of accent and language on the performance of speech coders. A corpus of accented English speech was used to investigate the linguistic sensitivities of low bit rate speech coders. Listening tests confirm the sensitivity of various coders (e.g. Federal Standard 1016, Multi-Prototype Waveform coders) to accented speech.

Speech coders are designed to minimize the bandwidth required for speech communication, the latest algorithms [2][3] allowing speech transmission at 2.4kb/s.

However, as the compression ratios increase it is apparent that the exploited redundancies are not universal across all accents (and languages).

This paper considers the reasons for, and solutions to, such problems. Initial investigations have concentrated on the quantisation of Pitch and formant structure in low rate coders. Pitch detection and tracking algorithms have been identified to be linguistically sensitive showing low tolerance to irregular pitch velocities.

Most coders quantise the formant structure using Linear Prediction, with the coefficients quantised as Line Spectral pair Frequencies (LSFs). Observations of the behavior of LSF distributions of accented speech have revealed distinct differences between accents from broad linguistic-acoustic

groups. The latter has significant impact on the nature of trained LSF codebooks.

Quantisers trained on limited linguistic-group databases show significant performance gaps when applied to speakers from linguistic-acoustic groups outside of the training set. LSF techniques can thus be used to classify speakers into linguistic-acoustic groups; these are sufficient for speech coding purposes.

It has been found that LSFs are an efficient means to make such a classification. The information gained can be used either to improve the training algorithms for LSF quantisers or, alternatively, utilized in robust coders which adapt to the linguistic-acoustic grouping of the speaker.

2. PROBLEMS WITH ACCENTED SPEECH AND LANGUAGES

The English language has had a dominating influence in the advance of telecommunications. With many of the major developments coming from primarily English speaking areas there is the risk that these advances may not be linguistically robust.

Accented English is of particular interest to telecommunications as English is widely accepted as the language of business and research. This means that a very high proportion of English spoken in the world, and thus telecommunication traffic, is of an accented nature. Particularly in Australia, where this research is being carried out, high migrant populations cause accented speech to represent an important percentage of telecommunication voice traffic.

Problems may occur when accented speech contains sounds, inherent to the speaker's mother language, that do not exist in English. As speech coding is achieved through the modelling these sounds it is possible that speech quality of non-English languages and their associated accents may suffer in terms of quality and intelligibility.

3. LOW RATE SPEECH CODING

The object of speech coding is to represent speech in a form that uses as few resources as possible for transmission and/or storage. Speech can be described in terms of its prosody (long term changes in pitch) and its formant structure (short term changes in pitch). In speech coding, prosody is modelled using long term prediction techniques. Formant structure models generally use linear prediction techniques.

3.1. The Linguistic Sensitivities of Low Rate Prosody Modelling.

All the low rate speech coders examined in this investigation use a form of pitch tracking to model the prosody of speech. Pitch tracking algorithms contribute to the minimisation of low rate speech coder bandwidth use[2].

Pitch Quantisation These algorithms use the correlation between pitch periods, based upon quantised thresholds, to ascertain smooth pitch contours. These thresholds, designed to detect spurious pitch periods, may not be suitable in some of the algorithm used.

Initial investigations [5] have shown some indication that these algorithms have led to an increase in pitch doubling and tripling with some forms of speech due to misjudgment of certain pitch periods and of voiced/unvoiced boundaries. Speech displaying high pitch velocities (as is frequently the case in some accented English speech) have induced such behaviour. In extreme cases some coders have even failed to regain and maintain pitch tracking [5].

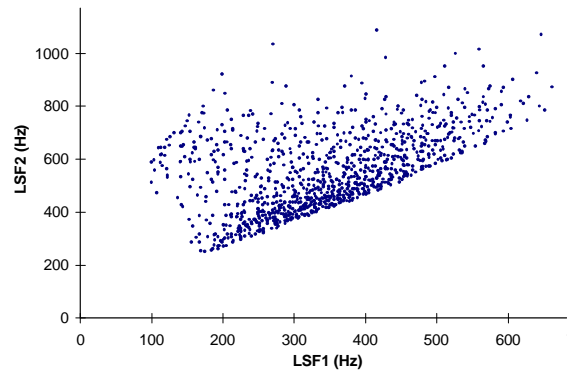
3.2. The Linguistic Sensitivities of Low Rate Formant Modelling.

In modelling the formant structure of speech, speech coders generally use Linear Prediction filters. The nature of these filters is described in terms of a set of coefficients known as Linear Prediction Coefficients (LPCs). Modern low rate speech coders represent the LPCs through the use of LSFs, a more robust representation of LPCs[3]. Most other representations of LPCs are highly sensitive to small errors due to quantisation or channel errors[4]. 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 [4].

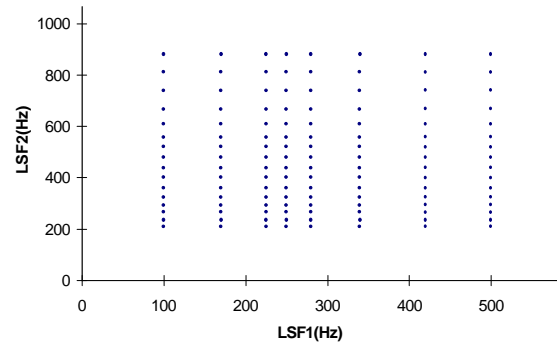
LSF Quantisation. LSF quantisation is a process whereby a finite set of LSF vectors are compared to the input LSF vector and the closest codebook vector to the input vector is used to represent that input. The codebook is generated through a process where quantised LSFs are extracted from a sequence of training speech. The nature of the training speech thus governs the nature of the LSF quantisation [5].

Figure 1 shows scatter plots of the first two LSF quantisation vectors used in two different codebooks. Figure 1(a) show the first codebook of a 30 bit split VQ algorithm [2] where 10 bits are used to quantise the first 3 LSFs. It was derived through training with the TIMIT database using the LBG algorithm. Figure 1(b) shows the quantisable LSF1-LSF2 patterns from the scalar quantiser used in the Federal Standard 1016 4.8kbs CELP based speech coder [6].

Certain speech sounds in other languages have different formant structures to that of English. As such, the process of quantisation may result in the modification of their original formant structures. Thus coding may result in the modification of the true sound of the original speech.



(a) Multi-Prototype Waveform



(b) Federal Standard 1016

Figure 1: LSF1-LSF2 Quantiser Distributions

4. RESULTS

Speech Corpus. In this initial investigation speakers representing the following language groups were examined;

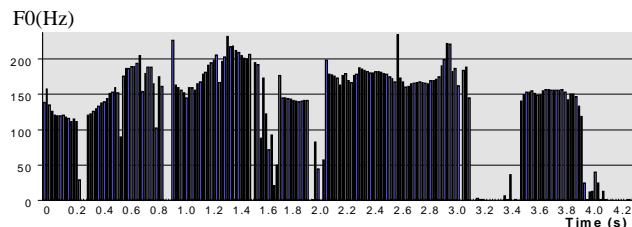
1. Indo-European
2. Sino-Tibetain
3. Dravidian

A number of males and females from each language group were used. These speakers had varying levels of experience with the English language. The data consisted of a set of ten phonetically dense English phrases recorded as telephone quality speech. This in line with the conditions expected by speech coders. This corpus is not statistically adequate for drawing any solid conclusions but it is sufficient to illustrate the nature and repercussions of low rate coder linguistic sensitivities.

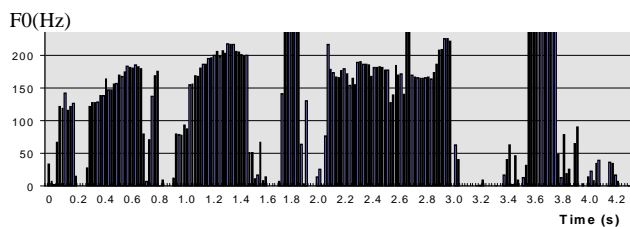
4.1. Pitch Contours

The pitch contour in Figure 2 (a) shows an example of accented speech of a Dravidian group speaker. The pitch quantiser, being attuned to the relatively flat contours of English, has simplified to a certain degree any fast undulations at the beginning or end of voiced sections (Figure 2(b)). Where doubling and tripling has occurred the pitch tracking algorithm has failed to regain

normal levels. These effects were confirmed using time domain analysis and listening tests [5].



(a) Accented Pitch Contour before coding



(b) Accented Pitch Contour after coding

Figure 2: Pitch doubling and tripling effects using MPW coding

4.2. LSF Distributions

The first three formants give an indication of a speaker's vowel behaviour. In accented speech, vowels are more often distorted than consonants. The behaviour of the first three LSFs corresponds to the nature of this vowel distortion. Figure 3 shows LSF1-LSF2 distributions for a representative set of speakers from each language group. The left hand column shows the original distributions. The right hand column shows the same distributions after split VQ coding.

It was observed that with milder accents, the distributions were not as clustered and suffered little distortion when coded, as was the case with the neutral accent (Figure 3(a) - 3(b)). Figure 3(c) to Figure 3(g) shows some examples of heavy accents. Figure 3(e), a Sino-Tibetan speaker (Mandarin), has only five short vowels which do not exist in the English phonetic set. It can be seen that a considerable amount of information spanned by the quantiser is not required. Moreover, as the signal is simple, it shows that the distortion of coding has moved this simple cluster to a position clearly out of the original range.

Figure 3 shows that coding has resulted in the displacement of any clustering patterns. The manner by which this has occurred was dependent upon the algorithms used for vector selection in quantisation. This clustering behaviour was only observed in quite heavily accented speech. This may reflect upon the limited nature of the speaker's sound set. The limitations imposed have led in some cases to a characteristic distribution. As such more heavily accented speech may prove to be an interesting base upon which further codebook training be carried out. A clear problem exists when the original LSF distribution lies outside the range of the quantiser. Some algorithms[7] use inappropriate quantising paradigms when vectors are out of range. Initial

tests[5] show that some language groups (such as Dravidian) induce certain coder malfunctions. To the best of our knowledge, no investigations have been carried out upon the LSF distributions of pure languages as yet but training LSF codebooks based upon such groups may prove to be fruitful.

5. CONCLUSIONS

The observations presented in this paper are only the results of an initial investigation. While in no way conclusive they do illustrate the following points;

1. The LSF clustering distributions are markedly different across the language groups represented.
2. The process of LSF quantisation clearly modifies the distributions of accented speech.
3. Significant differences exist in pitch contours between accents and languages.

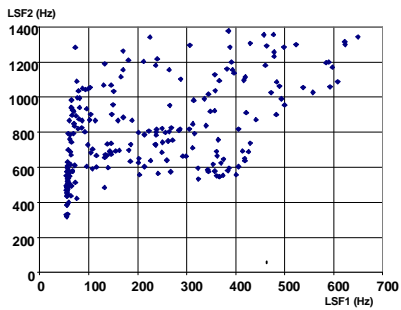
With a greater understanding and knowledge of the LSF behaviour, across pertinent linguistic groups, new linguistically tuned codebooks could be trained and integrated to make higher performance adaptive speech coders. LSFs may prove to be a very attractive approach to accent classification. An LSF based recognition system may be an appropriate means of achieving this due to the existence of distinct linguistic LSF patterns and the inherent efficiency of employing LSFs.

6. ACKNOWLEDGEMENTS

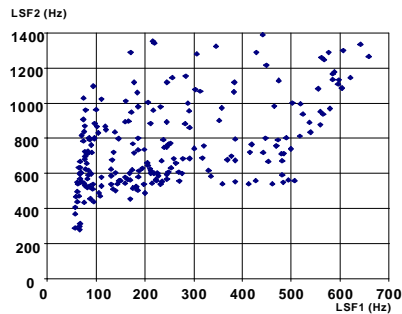
This work was supported in part by The Australian Research Council (ARC) under their small grants program.

7. REFERENCES

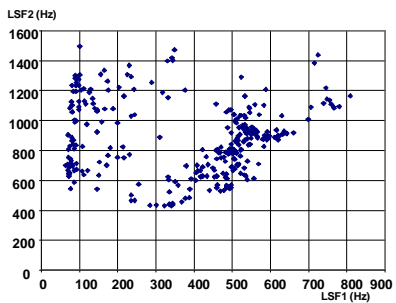
1. J.H.L. Hansen, L.M. Arslan, "Foreign Accent Classification Using Source Generator Based Prosodic Features", *Proc. Int. Conf. Acoust. Speech Sign. Process., Detroit, pp. 836-839, 1995.*
2. I.S. Burnett, G.J. Bradley, "New techniques for Multi-Prototype waveform coding at 2.84 kb/s", *Proc. Int. Conf. Acoust. Speech Sign. Process., Detroit, pp. 261-264, 1995*
3. W.B. Kleijn, J. Haagen, "Transformation and decomposition of speech signals for coding", *IEEE Sig. Proc. Letters, vol. 1, no. 9, pp. 136-138, 1994.*
4. I.S. Burnett, "Hybrid Technology for Speech Coding", *PhD. thesis; Chapter 3, University of Bath, 1992*
5. J.J. Parry, "Accent Classification for Speech Coding", *Honours thesis, The University of Wollongong, 1995*
6. 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.*



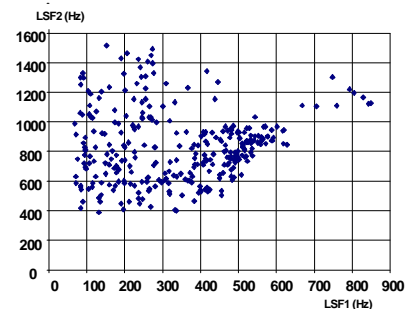
(a) Neutral English Accent



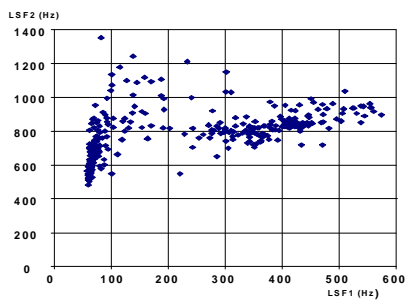
(b) Coded Neutral English Accent



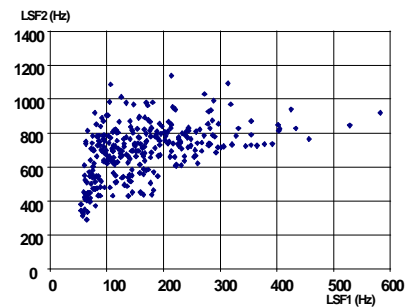
(c) Indo-European Accent



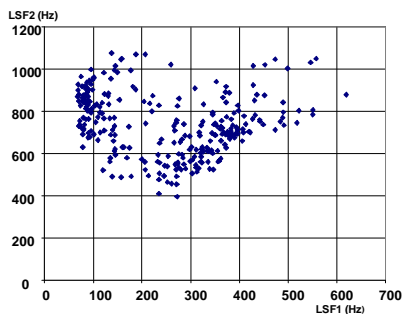
(d) Coded Indo-European Accent



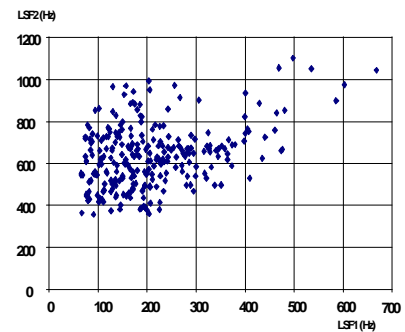
(e) Sino-Tibetan Accent



(f) Coded Sino-Tibetan Accent



(g) Dravidian Accent



(h) Coded Dravidian Accent

Figure 3 LSF1 vs. LSF2 Distributions for the representative accent linguistic groups examined (using split VQ coding).