GENDER ADAPTED SPEECH CODING

David F. Marston

Ensigma Ltd.
Turing House, Station Rd.
Chepstow, UK
david@ensigma.com

ABSTRACT

Speech coders that are optimized to the characteristics of a particular set of speakers will outperform a speech coder which caters for all speakers; providing that the speaker using it is one of that particular set. This paper describes how speech coders that are optimized to either male or female speech can be an improvement over unoptimised coders. These improvements are bit-rate reduction, speech quality and robustness. A reliable gender identifier is described, which would be practical for the most demanding applications, achieving 95% accuracy after 1 second of speech. The improvements in terms of gender specific speech coding are shown in LSF quantisation with bit-saving, and pitch detection with both bit-saving and robustness.

1. INTRODUCTION

The current speech coding standards have all been designed to work effectively for as wide a range of speakers as possible. Naturally some coders perform better for some types of speakers than others, for example it is known that Mandarin speech [3] performs well in LSF quantisation tests. If a speech coder is optimized towards a particular set of speakers, then its performance should be greater than that of a coder optimized for all speakers; providing it is only used with that set of speakers. One problem with speaker specific coding is that the type of speaker must be identified in some way. In typical mobile applications only a few seconds of speech may be available to make such an identification. The only practical selection that would be possible is the binary decision of gender (more specifically high and low pitch speakers). Gender identification can be reliably carried out after a short burst of voiced speech using just a pitch estimator [2]. The robustness and accuracy of the pitch detector will have some bearing on the performance of the identifier.

Pitch and LSF quantisation have been chosen to demonstrate the advantages of gender specific coding. The range of pitch values is smaller for a particular gender than it is for both genders, this be utilized to provide bit-rate reduction, or more refined estimation (sub-sample estimates for example). The pitch detector will be more robust if the range of pitch values is constrained. Speech coders which rely on a reliable pitch estimate for analysis will gain by this extra robustness, so overall speech quality can improve. LSF quantisation was chosen for gender specific coding for two reasons: (1) LSF quantisation exists in most modern speech coders [1], so any improvements here will be relevant to many applications; (2) LSFs represent the temporal aspect of speech, which has a certain amount of speaker dependency, so the distribution of LSFs will differ between male and female speakers.

2. SPEECH DATABASE

The speech database which was used in all of the experiments was the Subscriber database [4]. The gender of the speaker for each speech file is known. Table 1 outlines the use of the speech database. The training and testing sets are independent. The speech files vary in length, but there is an average of 30 seconds of speech per speaker. There are the same number of male and female speakers.

Speech Set	Number of speakers	Total speech/mins
Pitch training	200	100
Pitch testing	200	100
LSF training	400	200
LSF testing	200	100

Table 1: Use of speech database

3. GENDER IDENTIFIER

3.1. Pitch Detector

Pitch detection for gender identification does not have to be particularly precise, robustness being a more valuable asset. The pitch detector used in these experiments is a Cepstral based detector. This is a reasonably reliable method, and not too computationally intensive. The position of the peak in the cepstrum determines the pitch period (within possible pitch value range), and the size of the peak determines the confidence of this estimate. Only pitch estimates which are above a certain confidence threshold are considered. High confidence scores only occur in voiced speech, and are virtually free of pitch doubling and halving. The detection is carried out on the speech signal, not a LP residual as this proves to be more reliable. Figure 1 shows the distribution of pitch estinates in voiced speech for male and female (adult) speakers. The speech is sampled at 8kHz, and the pitch periods are in samples. The pitch training set of speech was used.

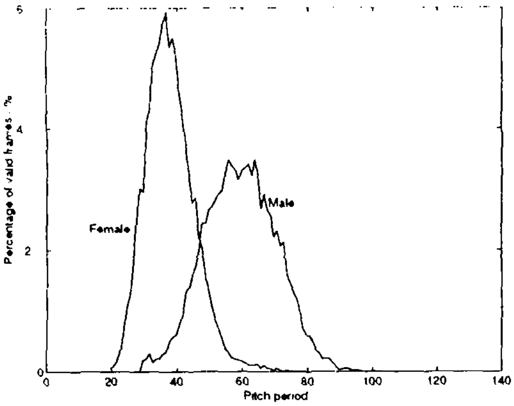

Figure 1: Distribution of pitch estimates for male and female speakers.

3.2. Identification from Pitch Statistics

Averaging confident pitch estimates over a number of frames is enough for adequate gender identification in the described applications. By taking the average pitch over 10 confident frames (20ms long frames) and using a simple threshold decision around 95% accuracy can be achieved. Figure 2 shows the distribution of average pitch scores for 200 different speakers over 10 voiced frames of speech.

By taking a threshold of 41 samples (i.e. male speech averages over 41 samples, otherwise female), the following accuracy results were achieved:-

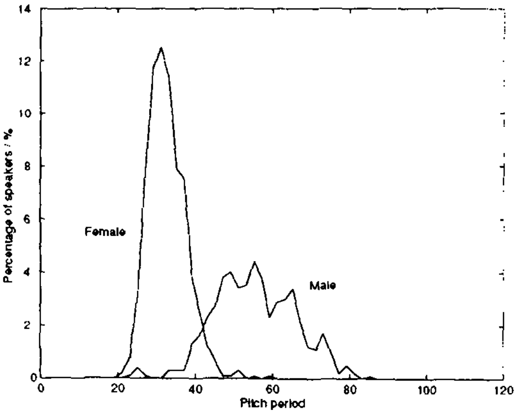

Figure 2: Distribution of mean pitch estimates for male and female speakers.

Gender of test set	Accuracy
Male	94.25%
Female	95.37%

This accuracy has proved adequate for the purposes of pitch detection and LSF quantisation improvements.

4. IMPROVED LSF QUANTISATION

LSF quantisers trained on a particular gender should perform better with that gender than LSFs trained with both genders. Three sets of LSFs were trained using male, female, and both sets of speakers. The test speech was independent of the training speech.

4.1. Vector Quantisation Techniques

Two methods of LSF vector quantisation were tried, a straightforward 10-dimension VQ and a 4-3-3 split VQ (similar to [3]). Both were trained using the LBG algorithm with a Euclidean distance metric. For the 10-dimension VQ, codebooks from 5 bits to 12 bits were generated. For the split VQ each codebook from 2 bits to 7 bits were generated, producing total codebook sizes ranging from 6 bits to 21 bits.

4.2. Results

The performance of the VQ is tested by measuring its Log Spectral Distortion (LSD, equation 1) for each of the codebooks.

$$LSD = \sqrt{\frac{1}{\pi}\int_0^{\pi}\left[10\log_{10}S(\omega) - 10\log_{10}S'(\omega)\right]} \quad (1)$$

where $S(\omega)$ is the unquantised LPC power spectrum, and $S'(\omega)$ is the quantised LPC power spectrum.

Figures 3 and 4 compare the performance of codebooks trained on a specific gender to one trained on both genders. The graphs show how LSD varies with codebook size. It is clear that there is a significant improvement in using a gender specific codebook. For a given LSD level around 2 bits can be saved for male speakers, and 2.5 bits for female speakers.

Figures 5 and 6 show the performance of the split VQ. It can be seen it is not as good as the 10-dimension VQ. Female speech only showing 1.5 bits improvement, and male speech barely makes any difference.

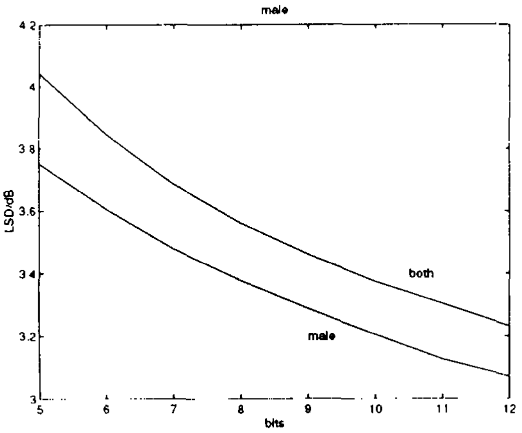

Figure 3: Comparing a male specific VQ with a general one.

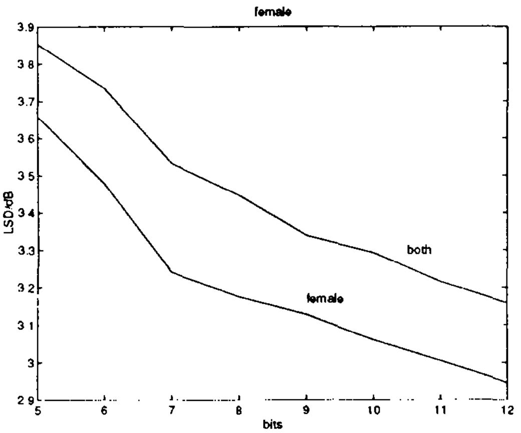

Figure 4: Comparing a female specific VQ with a general one.

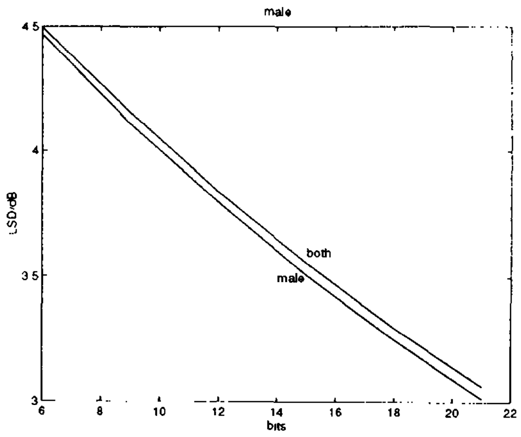

Figure 5: Comparing a male specific split VQ with a general one.

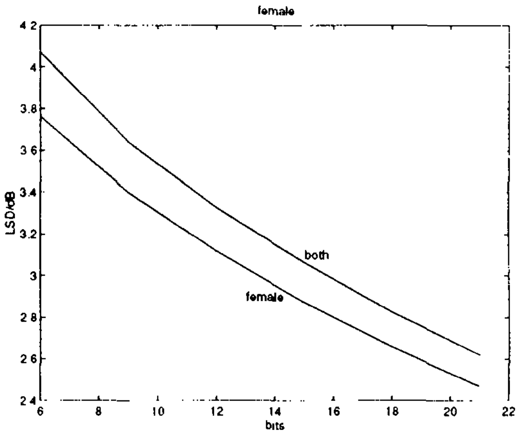

Figure 6: Comparing a female specific split VQ with a general one.

5. ROBUST PITCH DETECTION

The robustness of pitch detection can be improved if the range and possible pitch values is constrained. By using the gender of the speaker it is possible to set a more constrained range of pitch values. From the graph in Figure 1 it can be shown that male pitch periods occur mainly between 36 and 95 samples, and female pitch between 20 and 58 samples. Constraining pitch searches will reduce the chance of pitch halving and doubling, and may speed up the pitch detection process, as the search is smaller.

One bit saving can be achieved with the scheme shown in table 2.

Female speakers can also benefit from sub-sample resolution. Figure 7 shows how constraining male pitch

Gender	Pitch Range	Resolution	Bits
Male	36-99	1 sample	6
Female	20-35.5 36-67	0.5 sample 1 sample	6
Both	20-120	1 sample	7

Table 2: Pitch bit allocation scheme

detection improves robustness.

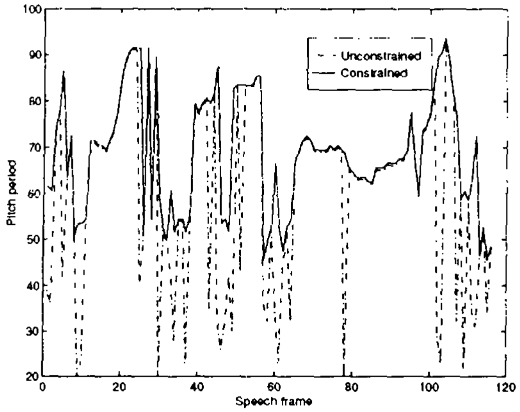

Figure 7: Constrained and unconstrained pitch tracks for some male speech.

6. APPLICATIONS

Figure 8 shows how the gender identifier would be added as a front end to a speech coder. The gender identifier outputting a flag to the coder's pitch detector and LSF quantiser. The coder will also transmit this flag to the decoder in some way. It would not be particularly efficient to transmit this bit every frame, so a reserved codebook entry in one of the parameters which could be transmitted in a silence frame would be a more efficient solution. The set-up shown assumes the gender of the speaker will not change throughout the use of the coder; this may not be be case in some applications. Either a user-controlled reset could be implemented or an automatic system, where after a pre-defined amount of non-speech activity occurs a reset can be done. Other speech coder specific parameters may also benefit from gender specific quantisation, such as the harmonic amplitudes in MBE coders. Overall bit allocation may be altered on a gender dependent basis, for example one bit less for female pitch than male pitch.

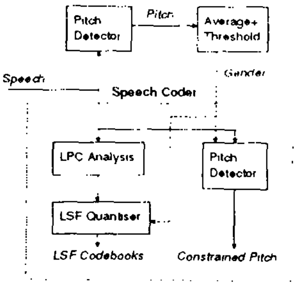

Figure 8: Combining gender identifier with a speech coder.

7. CONCLUSIONS

The results in this paper indicate that by the application of gender specific quantization of LSF and pitch parameters, improvements in coding efficiency can by gained. The gains in pitch detection robustness can have also improve analysis in other parts of the speech coder. The gender identifier can be made into a simple front-end to most speech coders with very little processing overhead. Reliable identification can be made after just a second of speech, so it is practical for mobile applications.

8. ACKNOWLEDGEMENTS

This work has been supported under the DTI's LINK directive. The project partners are BT, Liverpool University and Ensigma Ltd.

9. REFERENCES

[1] K.K.Paliwal, B.S.Atal, Efficient Vector Quantization of LPC Parameters at 24 bits/frame. ICASSP 1991, pp.661-664.

[2] E.S.Parris, M.J.Carey. Language Independent Gender Identification. ICASSP, 1996., pp.685-688.

[3] J.J.Parry, I.S.Burnett, J.F.Chicharo. A Cross-Language Performance Study of Vector Quantisation. IEEE Workshop on Speech Coding, 1997.

[4] A.D.Simons, K. Edwards. Subscriber: A Phonetically Annotated Telephony Database, Proc. IOA (Speech and Hearing), Vol 14, part 6, Nov 1992.