SPEECH COMPRESSION WITH PRESERVATION OF SPEAKER IDENTITY

John Leis[†], *Mark Phythian*[†] *and Sridha Sridharan*[‡]

[†]Faculty of Engineering
University of Southern Queensland
Toowoomba, Queensland, AUSTRALIA
leis@usq.edu.au

[‡]Speech Research Laboratory
Signal Processing Research Centre
Queensland University of Technology
Brisbane, Queensland, AUSTRALIA

ABSTRACT

Although much effort has been directed recently towards speech compression at rates below 4 kb/s, the primary metric for comparison has, understandably, been the amount of spectral distortion in the decompressed speech. However, an aspect which is becoming important in some applications is the ability to identify the original speaker from the coded speech algorithmically. We investigate here the effect of speech compression using multistage vector quantization of the short-term (formant) filter parameters on text-independent speaker identification. It is demonstrated that in cases where the speech is stored in a compressed database for retrieval, the speaker model should be constructed from the raw speech before spectral compression. Additionally, Gaussian models of sufficiently high order are able to reduce the negative effects of spectral vector quantization upon speaker identification accuracy.

1. PROBLEM FORMULATION

When attempting to identify speakers from their voice, spectral features (linear transformations or derived from predictor coefficients) have been found to be more effective than prosodic features (pitch, stress and articulation rate) [5]. In considering the evaluation of the effect of spectrum compression on speaker identification, four possible scenarios arise as shown in Table 1. These are :-

(i) The "benchmark" for all cases, using raw speech in the identification process. No compression is performed on either the incoming or reference speech data.

(ii) The speech database is compressed (for example, on CD-ROM) and the incoming speech is available in uncompressed form. This situation arises in forensic speech processing where the database of suspects has been archived and a new suspect is to be compared.

(iii) The incoming speech is compressed, but the reference is not. This problem may arise in telecommunications applications. Note that in this case the speaker identification parameters *may* be precomputed and stored (depending on the identification algorithm), allowing the speech database to be compressed without substantially compromising the speaker identification accuracy.

(iv) Both the database and the incoming speech are compressed.

We present results for each of these cases in Section 5. Although the effect of both population size and non-ideal recording conditions has been reported in the literature [2], the availibility of the speech in digital form enables the means of identification to be based on the encoded voice model, rather than on an analog reconstruction of the speech.

2. SPECTRUM REPRESENTATION

The short-term speech predictor is used for the purposes of both coding and identification. This predictor models the spectral envelope of the speech. The short-term analysis filter is represented as

Table 1: Compression and Speaker Identification.

Condition	*Speech Database*	*Incoming Speech*
(i)	16-bit PCM	16-bit PCM
(ii)	Spectral VQ	PCM
(iii)	PCM	Spectral VQ
(iv)	Spectral VQ	Spectral VQ

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_m z^{-m} \qquad (1)$$

where the m coefficients a_i must be coded and transmitted for the coding operation. It should be pointed out that the coding problem requires minimization of the predictor size m, whereas the speaker identification problem is not normally constrained in the number of parameters, and the identification accuracy increases with the model order.

There is a considerable body of theoretical and experimental results to indicate that better performance in compression is obtained with a transformation of the predictor $A(z)$ into the Line Spectrum Frequency (LSF) representation [4]. Given the Linear Predictive Coding (LPC) model with coefficients a_i, the LSF representation is found by decomposing $A(z)$ into two polynomials $P(z)$ and $Q(z)$, as follows:

$$\left.\begin{matrix} P(z) \\ Q(z) \end{matrix}\right\} = A(z) \pm z^{-(m+1)} A(z^{-1}) \qquad (2)$$

The resulting LSF's are interleaved on the unit circle, with the roots of $P(z)$ corresponding to the odd-numbered indices and the roots of $Q(z)$ corresponding to the even-numbered indices. The quantization properties of the LSF's have been well documented in recent literature [3] [4].

3. VECTOR QUANTIZATION

The coding method examined in this work involves Vector Quantization (VQ) of the LSF's. This method produces very large compression of the short-term spectral information, at the expense of a far more complex vector coding operation and increased distortion. The coding of the LSF's is examined in more detail in [3]. The vector coding of the LSF's reduces, in the simplest case, to determining the optimal index assignment k at time t subject to a distortion criteria:

$$\hat{\mathbf{x}}_t(k) = \mathsf{argmin}\,\{\mathcal{D}(\mathbf{x}_t, \mathbf{y}_i)\} \quad \forall\ \mathbf{y}_i \in \mathbb{C} \qquad (3)$$

where $\mathcal{D}(\cdot)$ represents the distortion criteria, $\mathbf{x}_t$ is the vector to be encoded at time t, $\mathbf{y}_i$ is the i^{th} candidate vector and $\mathbb{C}$ represents the vector codebook. The codebook design must be sufficiently robust against all possible permutations of the input vector to ensure adequate coverage of the vector space. Because of the computation and storage requirements necessary for acceptable distortion, a full-search VQ codebook cannot be used. Some method which reduces the computational complexity and storage requirements is normally employed. This comes at the expense of an increased rate and/or distortion [4]. The VQ method employed in this research is the multistage VQ [3]. Thus the single index k in (3) is replaced by a set of indices, one per sub-codebook.

4. SPEAKER IDENTIFICATION

Speaker identification involves the identification of a speaker from the voice alone, using a distance metric. Text-independent identification (the focus of this paper) is more difficult than text-dependent speaker identification, but has potentially far greater application. Several measures of distance have been proposed in the literature. In this study, we have utilized the Gaussian speaker model, in which a statistical model is constructed for each speaker in the population. This method has been shown to produce near 100% identification accuracy for speech recorded under ideal conditions [2]. The effect of telephone conditions (band-limiting, microphone nonlinearity and channel distortions) has been reported elsewhere for very large populations, and was found to be the major determinant of accuracy in speaker identification [2]. It is noted that [2] utilized the cepstral coefficients for the identification algorithm – however since the cepstral coefficients are non-invertible they are unsuitable for speech coding. Thus, we utilize the LSF represenation for our identification experiments. A benchmark (unquantized model, unquantized input speech) is therefore presented in the Results section of this paper for comparison.

4.1. Gaussian Mixture Model

The Gaussian Mixture Model creates a M^{th}-order, D-variate Gaussian model for each reference speaker [7]

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}\ i = 1, \ldots, M \qquad (4)$$

where $\boldsymbol{x}$ is a D-dimensional random vector, $b_i(\boldsymbol{x}), i = 1, \ldots, M$ are the component densities and w_i are the

mixture weights. The resulting probability density is given by

$$p(\boldsymbol{x} \mid \lambda) = \sum_{i=1}^{M} w_i b_i(\boldsymbol{x}) \quad (5)$$

Each component density is of the form

$$b_i(\boldsymbol{x}) = K \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \quad (6)$$

with

$$K = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \quad (7)$$

The mean vector of the set is $\boldsymbol{\mu}_i$ and the covariance matrix (assumed diagonal here) is $\boldsymbol{\Sigma}_i$. The set of weights satisfy $\sum_{i=1}^{M} w_i = 1$.

For T training vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, the Expectation Maximization (EM) algorithm [2] is used to iteratively estimate the model parameters. The GMM total likelihood for a vector set $\boldsymbol{X}$ is given by

$$p(\boldsymbol{X} \mid \lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid \lambda) \quad (8)$$

Each iteration of the EM algorithm updates the model weights, the model means, and the model variances.

5. RESULTS

Results were obtained using Region 2 of the TIMIT speech corpus [1] and the multistage VQ (MSVQ) compression algorithm. The original clean speech, sampled at 16kHz, was decimated to 8kHz in order to simulate telephone bandwidth conditions. In accordance with standard coding practice, the 10^{th} order LPC coefficients were derived from Hamming-weighted frames of 160 samples (20 milliseconds duration). The frame rate is thus 50 frames per second, with zero overlap. No pre-emphasis was applied (as is common in coding applications) so that the results more properly reflect the effect of the quantization process alone. The multistage VQ codebook was trained using 32768 speech frames from the "train" section of Region 2. The performance of the spectral quantizer was verified using speech outside that used for training. The identification was then carried out using 50 speakers from the "train" section, but with utterances *outside* the original set used to train the quantizer. For each test, the speech used to train the model is referred to as the "reference".

The 8^{th} order Gaussian model yielded an identification accuracy of 76%, which was substantially degraded when either the incoming speech and/or the speech used to build the model were vector quantized. The 16^{th} order Gaussian model exhibits performance comparable to that reported elsewhere for bandwidth-limited speech [2]. Comparing the first column of Tables 2 and 3 (which correspond to the "telephone identification" scenario), it is seen that the identification accuracy reduces somewhat after spectral vector quantization when a low-order Gaussian model is utilized. When a higher-order Gaussian model is employed, the accuracy does not appear to suffer a comparable reduction in performance. This variation is illustrated in Figure 1. For an 8^{th} order model, the reduction in identification accuracy is from 76% to 72%, whilst for a 32^{nd} order model, the accuracy is reduced from 100% to 98%. The *increase* in accuracy for a 16^{th} order model is thought to be due to a statistical anomaly due to the size of the candidate speaker population.

Table 2: Identification accuracy (percent) using an 8-mixture Gaussian metric.

Unknown Speaker	*Reference Speaker*	
	PCM	Quantized
PCM	76	70
Quantized	72	66

Table 3: Identification accuracy (percent) using a 16-mixture Gaussian metric.

Unknown Speaker	*Reference Speaker*	
	PCM	Quantized
PCM	92	88
Quantized	94	86

6. CONCLUSIONS

We have studied the application of speaker identification/verification methods to compressed speech. It was expected that the process of compression would lead to reduced performance of the identification algorithm. We have demonstrated that this is indeed the case if the model order is not chosen appropriately. The model order used in the Gaussian modelling process exhibits a strong influence on the identification accuracy, especially for spectrally compressed speech.

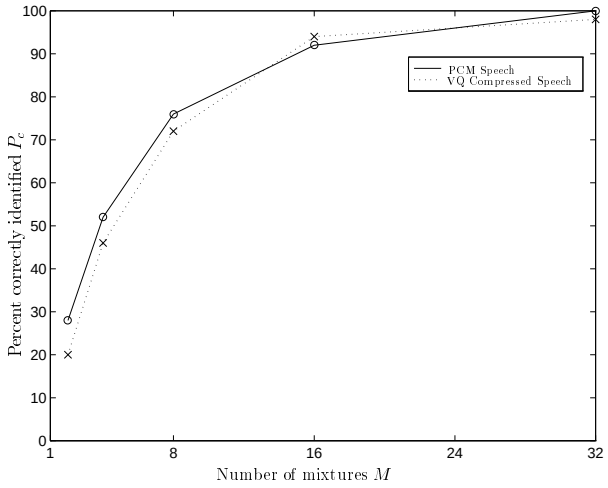

Figure 1: Compressed speaker identification using MSVQ compression and Gaussian model distance metric.

In applications where the reference speaker set is to be stored in a compressed form, considerable advantages become evident if the model is "pre-built" from the raw speech and stored alongside the compressed speech. For each speaker, $2DM + M$ parameters must be stored, as indicated in Table 4. For $D = 10^{\text{th}}$ order LSF quantization and $M = 32$ mixtures we have 672 parameters. Assuming a four byte floating point format, this is approximately 2.6 Kbytes which must be pre-computed and stored per speaker. Our results indicate that this relatively small overhead is justified if the original speech must be stored in addition to the identification model.

Table 4: Per-speaker parameters required for Gaussian model (Dimension D parameter vectors, M mixtures)

Parameter name	*Symbol*	*Size*
Mixture weights	$\boldsymbol{w}$	$M \times 1$
Means	$\boldsymbol{\mu}$	$D \times M$
Covariances	$\boldsymbol{\Sigma}$	$D \times M$

7. REFERENCES

[1] Linguistic Data Corporation, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology, 1990.

[2] D. A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech", *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[3] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[4] J. S. Collura, "Vector Quantization of Linear Predictor Coefficients", in *Modern Methods of Speech Processing*, chapter 2. Kluwer Academic Publishers, 1995.

[5] Y-H. Kao, L. Netsch, and P. K. Rajasekaran, "Speaker Recognition over Telephone Channels", in *Modern Methods of Speech Processing*, chapter 13. Kluwer Academic Publishers, 1995.

[6] H. Gish and M. Schmidt, "Text-Independent Speaker Identification", *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–31, Oct. 1994.

[7] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.