TEXT INDEPENDENT SPEAKER IDENTIFICATION USING LSP CODEBOOK SPEAKER MODELS AND LINEAR DISCRIMINANT FUNCTIONS

R. D. Zilca and Y. Bistritz

Department of Electrical Engineering
Tel Aviv University
Tel Aviv 69978, ISRAEL
bistritz@eng.tau.ac.il

ABSTRACT

The popularity of Line Spectra Pairs (LSP) in speech processing has been supported recently by theoretical studies of their statistical properties. It has been shown that LSP frequencies are uncorrelated, and have a diagonal sensitivity matrix with respect to spectral distortion. Therefore, LSP is suitable for Vector Quantization (VQ) schemes with simple weighted Euclidean distance measures. This was further supported by an analysis of the LSP probability density function, shown to be appropriate for distance-based recognition frameworks. This paper reports our study on developing improved methods for using LSP in VQ based speaker recognition. We used Linear Discriminant Analysis (LDA) to explore the speaker-discrimination statistics of LSP, and to transform them into a speaker-discriminative space. Further enhancements include the use of special VQ distance measures such as F-ratio weighting and Inverse Harmonic Measure (IHM). Performance evaluation experiments were conducted on very short speech sessions, using a database of 32 male speakers, taken from the TIMIT and NTIMIT. Identification results of 100% for clean speech and 70% for telephone speech were achieved.

1. INTRODUCTION

Line Spectra Pairs were proposed originally for speech coding and have been studied extensively in this context. Few studies have examined LSP for speaker recognition. In the first relevant research the authors are aware of [1], LSP and some LSP derived features were shown to provide excellent speaker recognition rates for a text restricted experiment (digits). More recent work [2] used LSP for text independent speaker recognition, modeling each speaker by a single multivariate Gaussian and deploying a divergence-based distance measure between speaker models. The results of these works are supported by some theoretic analyses performed on the statistical properties of LSP. It was shown that LSP are appropriate for recognition tasks (especially with centroid-distance rules) due to the convex nature of their distribution [3]. In addition, it was shown that they are suitable for diagonal weighting based frameworks [4]. In particular, the Inverse Harmonic Measure (IHM) diagonal weighting [5] was shown to approximate the log spectral distortion [6]. The findings of [6] also support the use of diagonal weighting, since the Mahalanobis distance degenerates into a simple inverse variance weighting for uncorrelated variables.

The desirable statistical properties of LSP and the previous good results of LSP-based speaker recognition systems motivated us to study the use of LSP as feature vectors in computationally less demanding settings of such systems. VQ classifiers may perform worse than other more elaborate classifiers [2]. However, our study reveals that the statistical properties of LSP may compensate on the limitations of the VQ method, at the expense of relatively little additional computation. We also find that exploring the speaker-discrimination statistics of LSP prior to the construction of the recognition system may improve performance. This insight is missing in former statistical analyses.

The VQ speaker recognition schemes with the enhancements that were proposed achieved in our experiments up to 10% of improvement compared to a simple Euclidean LSP baseline system. The resulting identification rate reached 100% for clean speech and 70% for telephone speech. These results are comparable to the performance reported for more elaborated speaker recognition methods, such as Gaussian Mixture Models (GMM) used in similar experimental settings [7, 8].

This paper is organized as follows: first the principles of VQ based speaker recognition are briefly reviewed. Then the proposed enhancements are described and theoretically justified. This is followed by a report of the linear discriminant analysis performed for LSP and the resulting pre-processing transformation. Finally, the experimental setting and identification results are summarized and discussed.

2. VECTOR QUANTIZATION BASED SPEAKER RECOGNITION

Speaker recognition systems may be classified by two basic characteristics: text dependency, and closed set vs. open set. The open set scenario is termed *speaker verification*. In this case the speaker makes a claim of his/her identity and the system is required to confirm or reject this claim. In the closed set scenario, termed *speaker identification,* the speaker states no claim regarding his/her identity and the system determines the identity from a predetermined set of reference speakers.

In VQ based speaker identification tasks, the features extracted from the test utterance (in this case – LSP) are compared to all of the speakers' codebooks, and the best matching codebook is selected. A block diagram describing a VQ based speaker identification system is shown in Figure 1. The VQ scheme may also be adapted for speaker verification, where an identity claim is stated, and the test utterance is compared to the claimed speaker codebook to check whether a required similarity threshold is exceeded.

VQ based speaker recognition was introduced in the mid-eighties ([9], [10], [11]) and was studied mainly in text restricted experiments, using cepstral features. The main advantage of VQ as a classification scheme is its computational simplicity. Training the speaker models may

Proc of the 6th European Conference on Speech Communication and Technology, Eurospeech'99, pp. 799-802, Sept. 5-10, 1999, Budapest, Hungary.

be performed using the K-means algorithm [12], and recognition is simply performed by choosing the codebook whose average distance from each incoming feature vector is minimal.

Following our goal to improve speaker recognition performance while maintaining the computational simplicity of VQ schemes, we concentrated on the next two topics:

1. Trying more specialized distance measures than a simple Euclidean. However, in order to keep simplicity we restricted our search to simple diagonal weightings that are attentive to the properties of the LSP features vectors.
2. Pre-processing of a feature vector is advantageous for any type of classifier and may enhance performance. We focused on LDA as a means for analyzing the speaker-discrimination statistics of LSP and obtaining an optimal transform of LSP into a speaker discriminative space.

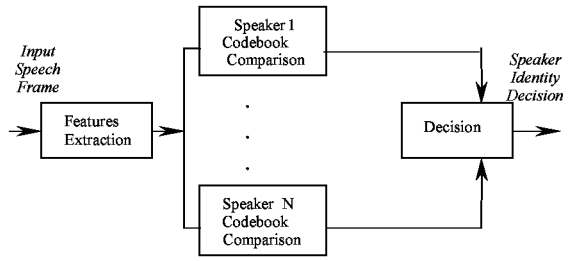


Figure 1. VQ based Speaker Identification System

3. ENHANCED VECTOR QUANTIZATION BASED SPEAKER IDENTIFICATION USING LSP

Improved Distance Measures

We used two special weighted Euclidean distance measures; F-ratio and IHM weighting. The use of a non-Euclidean distance measure affects both the construction of speaker codebooks during training and the calculation of the accumulated distance measure during recognition [12].

The use of F-ratio weights stems from statistical theory. Each weight is calculated by a Multivariate Analysis Of Variance (MANOVA) and equals the variance of speaker means divided by the average intra-speaker variance for the respective feature. The F-ratio weighting may be viewed as an improvement of the inverse-variance weighting, formerly used for VQ speaker recognition [11], yet it also may be viewed as a simplification of the more general LDA transform when inter-feature second order interactions are taken into consideration, as will be shown later.

The IHM weighting was introduced in [5] for more efficient LSP quantization. For a n dimensional LSP vector, $(f_1, f_2, \ldots, f_n)$, the IHM weights are defined as

$$w_i = \frac{1}{f_i - f_{i-1}} + \frac{1}{f_{i+1} - f_i} \tag{1}$$

where w_i serves as the weight for the squared error in the i^{th} LSP frequency f_i, for $i = 1, \ldots, n$, $f_0 = 0$ and $f_{n+1} = \frac{f_{sampling}}{2}$. The logic of this weighting scheme is that two adjacent LSP frequencies that are close together relate to a peak near that frequency in the speech spectral envelope. Gardner and Rao performed a comparison of various LSP weightings for vector quantization coding [4] and found that IHM weighting performs better than all other proposed weights, except to an optimal, but complicated, sensitivity weighting. They have also extended the concept of spectral sensitivity weighting [13] into a (generally non-diagonal) sensitivity matrix weighting. The sensitivity matrix of LSP parameters was shown to be diagonal, meaning that the log spectral distortion may be approximated by diagonal weighting of the LSP parameters. [6] provided further support for this finding, by showing that a first order approximation to the log spectral distortion results in an inverse-variance diagonal weighting of LSP. They have also shown that IHM weights approximate these variances because they approximate the formant bandwidths.

Preprocessing using Linear Discriminant Analysis (LDA)

Scatter Matrices

In order to formulate a criterion for class separability, we consider the *within-class* and *between-class* scatter matrices for an M-class pattern recognition problem.

The within-class scatter matrix, S_w, is defined as the next $n \times n$ matrix:

$$S_w = \sum_{i=1}^{M} P(\omega_i) E\{(\mathbf{x} - M_i)(\mathbf{x} - M_i)^t \mid \omega_i\} = \sum_{i=1}^{M} P(\omega_i)\Sigma_i \tag{2}$$

The between-class $n \times n$ scatter matrix, S_b, is defined by:

$$S_b = \sum_{i=1}^{M} P(\omega_i)(M_i - M_0)(M_i - M_0)^t \tag{3}$$

where $P(\omega_i)$ are the a-priori probabilities for each of the M classes. The M_i's are the class expected vectors, defined by

$$M_i = E(\mathbf{x} \mid \omega_i) \tag{4}$$

and M_0 is the expected vector of the mixtures:

$$M_0 = E\{\mathbf{x}\} = \sum_{i=1}^{M} P(\omega_i) M_i \tag{5}$$

S_b indicates the deviation between the expected vectors for each pair of classes, while S_w shows the scatter of samples around the expected vector of their own class. The scatter matrices are designed to be invariant under coordinate shifts.

Separability Criteria

Once the scatter matrices are defined, we can use them to derive a single scalar measure that indicates class separability. Various methods may then be used to find a transformation to be performed on the feature vectors so that this value would be maximized. A well-known class

separability measure is the trace of the discrimination matrix $S_w^{-1}S_b$ viz,.

$$J(n) = tr(S_w^{-1}S_b) \tag{6}$$

The measure $J(n)$ is invariant under any nonsingular linear transformation of the feature vectors. Yet, suppose we are interested in selecting a reduced number of m $(m < n)$ features, by applying a $m \times n$ transformation matrix A to the original $n-$dimensional vector. We would then like to choose the matrix A such that $J(m)$ of the transformed $m-$dimensional space is maximized. It can be shown that this is achieved by selecting the first m eigenvectors of the discrimination matrix $S_w^{-1}S_b$, whose eigenvectors $(\Phi_i, i = 1, 2, ..., n)$ are ordered by dominance of their eigenvalues

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n \tag{7}$$

i.e. A is

$$A = [\Phi_1, \Phi_2, \cdots, \Phi_m]^t \tag{8}$$

The class separability measure that will be obtained after the transformation is

$$J(m) = \sum_{i=1}^{m} \lambda_i \tag{9}$$

Following the above explanation, LDA might be expected to be useful only for dimensionality reduction, since the class separability measure $J(n)$ is invariant to linear transformation. However, $J(n)$ is not an ultimate clas separability measure, since it does not take int consideration the type of classifier that is used and assume a unimodal type of within-class distributions. Therefor LDA may also enhance class separability by transformin features *without* dimensionality reduction.
The F-ratio weights discussed earlier, are defined by:

$$w_j = \frac{\frac{1}{M}\sum_{i=1}^{M}[(M_i)_j - (M_0)_j]^2}{\frac{1}{M}\sum_{i=1}^{M}(\Sigma_i)_{jj}} \tag{10}$$

where subscripts on vectors and matrices indicate th addressed entries. Each weight equals the variance of th class means divided by the variance of the class. Equation (10) may also be expressed in terms of the scatter matrices. Taking only the diagonals of S_w and S_b and ignoring the off-diagonal elements will result in a diagonal discrimination matrix with the F-ratio coefficients on the diagonal.

4. PERFORMANCE EVALUATION EXPERIMENTS

Experiment Settings

The sampled speech from the TIMIT and NTIMIT databases was downsampled to 8Khz and segmented into 25 millisecond frames with 50% overlap. Each frame was multiplied by a Hamming window and a 10^{th} order LPC analysis was performed to extract the LSP frequencies. Training and testing used only voiced speech frames.
The speech database included 32 different male speakers, four speakers from each of the eight dialect regions. We chose to construct a single gender database, since it has been clearly shown that cross gender recognition errors are very rare in speaker recognition tasks. The training session for each speaker was composed of seven sentences (two "sa" sentences, three "si" and two "sx" sentences). The total duration of speech training intervals varies among speakers, and is 20 seconds on average (12 seconds of voiced speech). Testing was performed on the remaining three "sx" sentences, 2 to 3 seconds long each (about 1.5 seconds of voiced speech).

Speaker Discrimination Statistics

The scatter matrices and the discrimination matrix for the LSP frequencies of the 32 speakers were calculated and are shown in Figure 2, for clean speech and telephone speech. As seen, S is approximately diagonal, expressing the fact that the LSP frequencies tend to be uncorrelated. However, the between-class scattering matrix, S_b, is far from diagonal. Consequently $S_w^{-1}S_b$ is even more remote from diagonality. For this reason, the diagonalization of the discrimination matrix by LDA produces effective separation features. The above observation also anticipates that transforming LSP via LDA, before classification, will perform better than weighting the LSP distance measure by F-ratios.

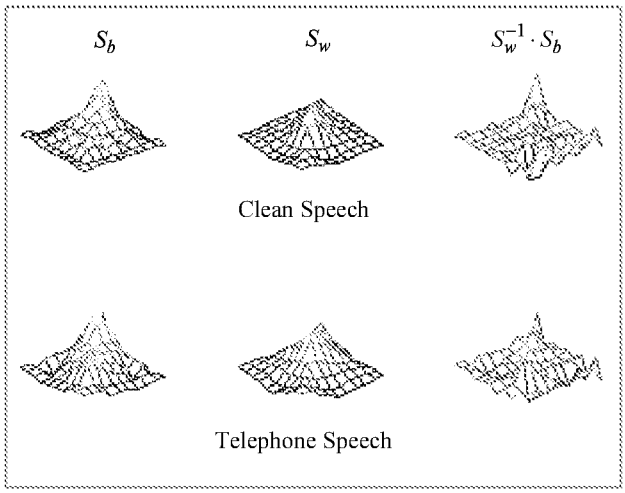


Figure 2. Scatter and Discrimination Matrices of LSP

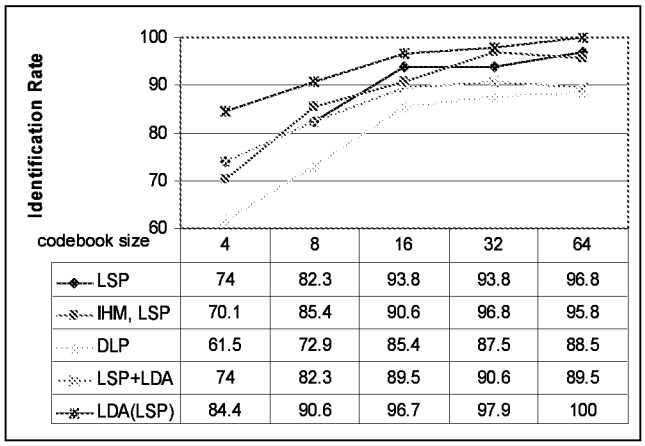
Identification Performance

The identification rate obtained for clean speech and telephone speech is shown in Figures 3 and 4, respectively. The results for the F-ratio weighted distance are indicated by "LSP+LDA" and the LDA transformed LSP are indicated by

"LDA(LSP)". The DLP (Difference weighted Line spectra Pairs) are obtained by multiplying each of the LSP elements with its IHM weight. This is performed at the feature-level, as opposed to distance based IHM weighting which is performed on the classifier level.

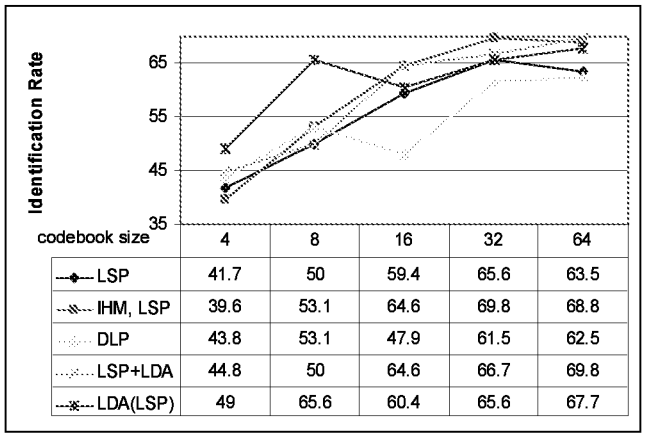
It is seen that transforming the LSP using LDA prior to classification improves identification rate significantly, while F-ratio weighting provides slight improvement for telephone speech but the performance degrades for clean speech. This finding is consistent with the speaker discrimination statistics.

Applying IHM at the feature level was found to degrade performance, while the use of IHM weighted distance maintained the baseline performance for clean speech and improved performance considerably for telephone speech, exceeding the performance obtained by the LDA transform for large codebooks.



codebook size	4	8	16	32	64
LSP	74	82.3	93.8	93.8	96.8
IHM, LSP	70.1	85.4	90.6	96.8	95.8
DLP	61.5	72.9	85.4	87.5	88.5
LSP+LDA	74	82.3	89.5	90.6	89.5
LDA(LSP)	84.4	90.6	96.7	97.9	100

Figure 3. Identification Performance – Clean Speech



codebook size	4	8	16	32	64
LSP	41.7	50	59.4	65.6	63.5
IHM, LSP	39.6	53.1	64.6	69.8	68.8
DLP	43.8	53.1	47.9	61.5	62.5
LSP+LDA	44.8	50	64.6	66.7	69.8
LDA(LSP)	49	65.6	60.4	65.6	67.7

Figure 4. Identification Performance – Telephone Speech

5. CONCLUSIONS

We have shown that VQ based speaker identification using LSP achieves fair recognition rates for short training and testing sessions, with a very low computational cost. On a database of 32 male speakers, with an average of 20 seconds training data and with 2.5 seconds testing, we achieved 100% identification for clean speech, and 70% recognition for telephone speech. These results compare well with other published results that use statistical parametric models like Gaussian Mixture Model (GMM) at similar size single gender populations, and short training and testing sessions [7,8].

The two most successful manners of using LSP in VQ based speaker identification were weighting the LSP distance by IHM and using LDA transformed LSP as feature vectors.

The IHM weighting of LSP is successful in particular in improving the identification rates for telephone speech.

The LDA transformation of the LSP features achieves good identification rates because the speaker-discrimination matrix of LSP is far from diagonal. Our experimental results also indicate that the identification rates of LDA transformed LSP degrade gracefully with decreasing of the codebook size. These observations admit flexibile trade off in meeting constraints on training duration and computational complexity in the design of VQ based speaker recognizers.

REFERNCES

[1] C.S. Liu, M.T. Lin, W.J. Wang and H. C. Wang, "Study of Line Spectrum Pair Frequencies for Speaker Recognition," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.* 1990.

[2] J.P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings IEEE*, Vol. 85, No. 9, pp. 1437-1462, Sept. 1997.

[3] J. Y. Tourneret, "Statistical Properties of Line Spectrum Pairs," *Signal Processing*, 1998.

[4] W. R. Gardner and Bhaskar D. Rao, "Theoretical Analysis of the High Rate Vector Quantization of LPC Parameters," *IEEE Trans. Speech. Audio. Proc*, Vol. 3, No. 5, Sept 1995.

[5] R. Laroia, N. Phamdo, and N. Farvaradin, "Robust and Efficient quantization of Speech LSP Parameters using Structured Vector Quantizers," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc*, pp. 641-644, 1991.

[6] J. S. Erkelens and P. M. T. Broesen, "On the Statistical Properties of Line Spectrum Pairs," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.* 1995.

[7] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models," *IEEE Trans. Speech. Audio. Proc*, Vol. 3, No. 1, Jan. 1995.

[8] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, pp. 91-108, 1995.

[9] F.K. Soong, A.E, Rosenberg, L.R. Rabiner, and B.-H Juang, "A Vector Quantization Approach to Speaker Recognition," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, Tampa, FL, pp. 387-390, 1985.

[10] F.K. Soong, A.E, Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," *IEEE Trans. Speech and Audio Proc.*, Vol. SAP-36, No. 6, pp. 871-879, June 1988.

[11] A. E. Rodsenberg and F .K. Song, "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, 1986.

[12] A. Gersho, R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[13] F. K. Soong and B.H. Juang, "Optimal Quantization of LSP Parameters," in *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc*, 1988.