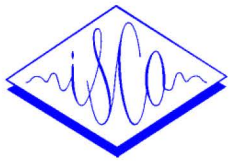

First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

INSTANTANEOUS AND TRANSITIONAL PERCEPTUALLY-BASED FEATURES IN SPEAKER IDENTIFICATION

L. Xu and J.S. Mason

Department of Electrical and Electronic Engineering
University College, SWANSEA, UK.

ABSTRACT

A recent comparison of features and distance measures [3] shows the perceptually based linear prediction, PLP, together with the appropriate distance measure to be consistently better than other widely used standard combinations. This paper investigates the PLP-derived cepstra representing the instantaneous spectral information and the time slope of the cepstra representing the transitional spectral information in automatic speaker identification (ASI). The root-power-sum (RPS) distance and the inverse variance (INV) weighted distance are discussed. The experiments relate to a vector quantization (VQ) based digit-independent ASI. The study shows the first 8 coefficients of the PLP features are the most important in distinguishing inter-speaker differences. The RPS distance and the INV weighted distance perform similarly well, and significantly better than the unweighted cepstral distance. Also, The overall advantage of PLP features over the LPC features are demonstrated.

INTRODUCTION

Feature extraction is considered one of the most important steps in successfully achieving speaker recognition. Earlier work which investigated and compared a number of features can be found in [1]. The coefficients from linear predictive coding (LPC), especially LPC-derived cepstral coefficients, have been accepted as highly efficient parameters for representing short-time spectral estimates of speech, containing both speaker-specific and linguistic information, see [1][2].

LPC-derived cepstral coefficients and a time-function of these coefficients, representing the instantaneous and transitional spectral information respectively, was first proposed by Furui [2] for using in speaker recognition. A further study of these two spectral information components is given in [4]; they find using a linear combination of the two can achieve a better speaker recognition performance than using either of them separately. [4] also shows that cepstral coefficients, with inverse variance weighting, gives better recognition performance than the unweighted form in their vector quantization (VQ) approach to automatic speaker identification (ASI).

Recently, a perceptually-based linear prediction feature, PLP, proposed originally by [5] for speech recognition, was investigated for use in ASI by Xu *et al.* [3]. They assess standard LPC and PLP with appropriate distance measures. Using single-digit test utterances PLP cepstra with the root-power-sum (RPS) distance, (or the inverse variance (INV) distance, the two are found to give similar results) gives significant improvements over standard LPC-based front-ends. Meanwhile, the experimental results in [3] emphasise the importance of using the RPS or INV weighting on PLP cepstral coefficients to harness not only the inter-speaker variation in the lower orders but also in the higher orders. The weighted cepstral distances for PLP perform significantly better than the unweighted cepstral distance in ASI, as they do also in speech recognition [6][13][10].

In this paper, we extend the work of [3] and investigate the effects of PLP derived transitional and instantaneous spectral features in ASI performance. The weighting influence for PLP cepstral distance is discussed. The purpose of this paper is to further examine the advantage of PLP-based features over standard LPC-based features in speaker recognition.

BACKGROUND

PLP

PLP analysis [5] comprises two basic phases:

1. auditory spectrum calculation, here interpreted mathematically as:

$$Q_k = \{E(\omega_k) \int_0^{\pi} C_k(\omega) P(\omega) d\omega\}^{1/r}, \qquad (1)$$

where $C_k()$ is the critical band weighting coefficient, $P()$ is the speech power spectrum, $E()$ is an equal-loudness weighting, r is a parameter chosen by using Steven's power law [7] for intensity-to-loudness conversion, and k = 1,2,..,17, the upper limit chosen to cover frequency range of $0 \leq f \leq 5kHz$.

2. all-pole approximation of the auditory spectrum using the LP approach.

Distance Measures

Cepstral spectral distance measures can be described as,

$$d = \sum_{i=1}^{n} (w_i * (c_i - c_i'))^2, \qquad (2)$$

where c_i and c_i' are ith cepstral coefficients, w_i is a weighting coefficient. $w_i = 1$ gives the well known Euclidean or unweighted cepstral distance, d_{CEP}. Statistical characteristics of the LPC

and PLP cepstral coefficients [8] [3] are such that the variances of the higher order coefficients are much smaller than the variances of the lower order ones. Soong [4] gives a set of histograms of the intra- and inter-speaker distance components $(c_i - c_i')^2$ in d_{CEP}. It is suggested that higher order cepstral coefficients are as important as the lower order ones in their ability to separate one speaker from another. In order to equalize the contributions from the individual components, non-unity weighting coefficients w_i are used in the distance measure (2). One of the most popular weightings is the inverse variance (INV) and the weighted distance measure, d_{INV}, is given as (2) when w_i^2 is the inverse variance of the ith coefficients. The corresponding $w_i, (1 \leq i \leq 14)$ for LPC and PLP cepstral coefficients are shown in Fig.1. Both [8][3] show the values are relatively consistent across speakers.

The spectral slope measure was proposed by Klatt [9] and successfully used in speech recognition by Nocerino, *et al.* [11]. The application of the spectral slope distortion concept to all-pole model spectra of PLP and standard LPC is discussed in [12]. This measurement, for PLP and LPC spectra, can be efficiently approximated by the root-power-sum (RPS) distance measure, d_{RPS}, which is considered as a special case of (2) when $w_i = i$ [8].

The RPS distance for PLP spectra is examined in [12][10][13]. They show the RPS distance measure is consistently better than the unweighted cepstral distance measure. Furthermore, [3] examines both RPS and INV weighting in ASI, and find the two weightings give very similar results.

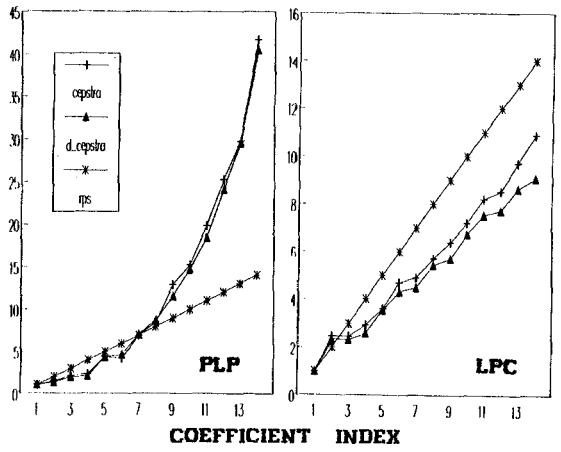

Fig.1 Inverse standard deviations of PLP and LPC derived cepstral, Δcepstral coefficients.

Transitional Features and Regression Coefficients

The transitional spectral information is approximated by the linear regression coefficients, namely the first-order orthogonal polynomial coefficients denoted by:

$$\Delta c_i(t) = \frac{\sum_{k=-K}^{K} k c_i(t+k)}{\sum_{k=-K}^{K} k^2}, \qquad (3)$$

which represents the slope of the time-function of each coefficient in the vector [2]. In (3), $c_i(k)$ is the ith cepstral coefficient, and $2K+1$ is the transitional feature window length. These features are referred to as Δcepstra.

EXPERIMENTS

The ASI System and Database

A VQ-based speaker identification system is employed in our study, similar to the one proposed by Soong [4]. Each speaker-dependent model is represented by two separate digit-independent codebooks which describe the instantaneous and transitional spectral characteristics of the speakers.

The isolated word digit database is collected from 10 speaker (5 male and 5 female), similar aged, English speaking university students. Each speaker recorded 100 digits, 10 versions of each. The recordings for each speaker were spaced over a period of weeks, and sampled at 10 kHz.

Arrangements for training and testing are the same as in [3]: for each speaker 5 versions are used for training and the other 5 versions are used for testing; each test utterance is one digit long. Training and testing data are interchanged, so that the experimental results given here are averages of 1000 digit test utterances.

PLP vs LPC based ASI

As highlighted in [3], PLP-RPS or PLP-INV yields a better ASI performance than any of the examined LPC-based feature-distance measures combinations. In this section we extend the examination of PLP features for ASI to include the:

- effects of window length for PLP Δcepstra,
- influence of PLP Δcepstra on performance.

Fig.2 compares recognition error percentages of LPC and PLP based transitional features. Window lengths of 3, 5, 7, 9, 11, 13, 15 and 17 are examined. The curve for LPC Δcepstra agrees with that of Soong [4], in that the 11 window length is the best for LPC Δcepstra. This is also the best for PLP Δcepstra. It can also be seen in Fig.2 that PLP Δcepstra give much better performance than LPC Δcepstra.

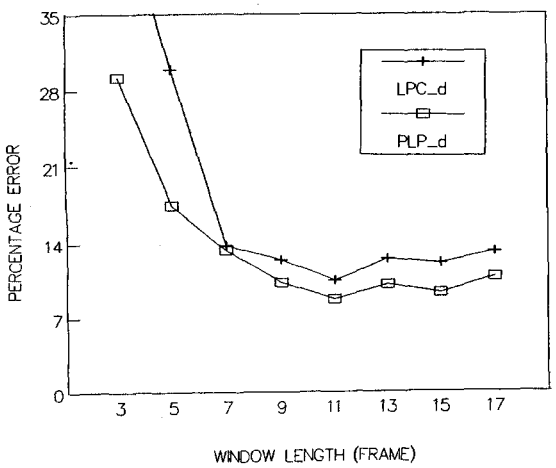

Fig.2 ASI error rate versus window length for PLP and LPC Δcepstra.

The lowest PLP Δcepstra error is 8.7 % compared with 4.7 % (Fig.3) using the direct cepstra of PLP. Here the codebook size is 32 and model order is 14.

The effects of PLP and LPC derived cepstra, Δcepstra and the combination of these two spectral representations are shown in Fig.3, for model order 14 and codebook size 32. Clearly, the best recognition performance is achieved by jointly using PLP derived cepstra and Δcepstra; PLP features are significantly better than LPC features in all cases.

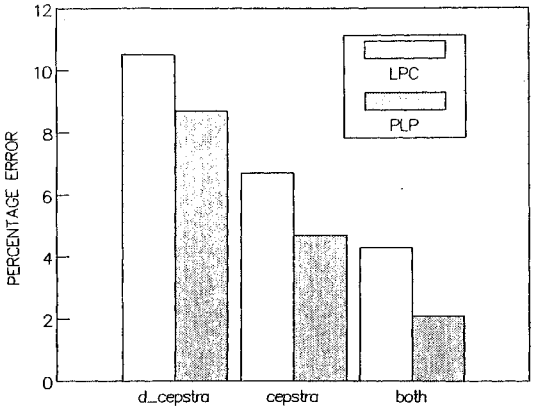

Fig.3 ASI error rate for PLP and LPC cepstra, Δcepstra and combination.

PLP-derived Coefficients and Weighting

The effects of weighted and unweighted distance measures for PLP derived cepstra, Δcepstra and their combination are given in Fig.6. The unweighted distance measure performs much worse than the weighted distance measure in all cases. This indicates clearly that higher order coefficients of PLP do carry useful information for ASI as predicted by Xu *et al* [3].

In order to study the effectiveness of each coefficient, a distance ratio R_i, which has been used by Furui [2], is introduced and defined as,

$$R_i = \frac{E(d_i^{jk})}{E(d_i^{jj})}. \qquad (4)$$

This represents the inter- and intra-speaker distance ratio of the ith coefficients. $d_i^{jk} = (w_i * (c_i^j - c_i'^k))^2$, is the distance of ith coefficients between a speaker j's feature vectors pooled in speaker k's codebook. $E()$ estimates the mean of such distances over all speakers and their feature vectors. The greater the value of R_i statistically the more ability the ith coefficient has to separate one speaker from the others.

Values of R_i for the first 14 terms of PLP-CEP, PLP-RPS and PLP-INV are given in Fig.4. It shows all of these coefficients are larger than unity and therefore potentially useful for ASI; values for cepstra are larger than values for Δcepstra. Importantly, looking at the cepstra, it is seen from Fig.4 that, in any case of weighting or no weighting, values for the first 8 coefficients are considerably larger than those for the the higher order ones.

The measure,

$$D_i = (E(d_i^{jk}) + E(d_i^{jj}))/2, \qquad (5)$$

where d_i^{jk} is as defined for Equation (4), indicates the contribution that the ith term, $(w_i * (c_i - c_i'))^2$ has on to the computation of (2).

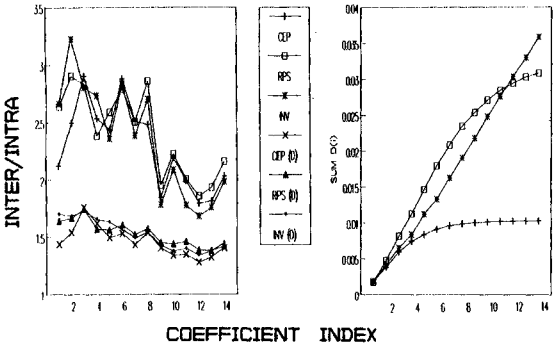

Fig.4 Value of inter- and intra-speaker distance ratio versus coefficient index.

Fig.5 Distance magnitude versus number of distance terms for d_{CEP}, d_{RPS} and d_{INV}.

The value, $\sum_{i=1}^{p} D_i$, integrates the measure from 1 up to p coefficients, showing the value of (2) when just p terms of distance are used. The values for PLP cepstral coefficients with unity, RPS and INV weighting are plotted in Fig.5. Note the CEP curve is magnified (factor of 5) for plotting purposes. It is seen that the values for CEP increase slowly after $p = 4$ and the curve become flat when $p > 7$. Thus the contributions to d_{CEP} from higher orders (> 5) are negligible. The linear increasing values for INV show that the equal contribution from each term towards d_{INV} is as expected. The curve for RPS in Fig.5 rises linearly and slightly faster than the one for INV up to $p = 8$ and then tends to flatten after $p = 14$. It indicates that d_{RPS} equally emphasizes the effectiveness of the first 8 coefficients and the coefficients above that become less important.

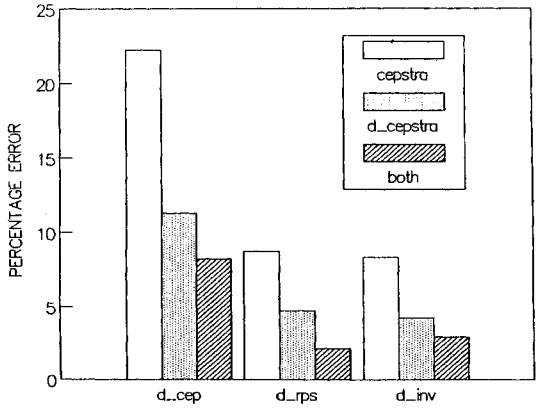

Fig.6 ASI error rate versus distance measure for PLP features.

The ASI performance using PLP cepstral representation and distance measures, d_{CEP}, d_{RPS} and d_{INV}, are reported in [3]. The d_{CEP}, d_{RPS} and d_{INV} performance with PLP Δcepstra, cepstra and the combination of both are shown in Fig.6. Codebook size 32 and model order 14 are used. Again, similar and good performance comes from d_{INV} and d_{RPS}, much better than that for the unweighted distance, d_{CEP}.

Male-Female Error Analysis

A comparison of cross-sex speaker confusion between PLP and LPC based ASI is given in Fig.7. The portion of error caused by the confusions of male-to-male, female-to-female, male-to-female, and female-to-male are indicated. Here d_{RPS} and 8th order PLP and LPC are used. When LPC cepstra or the combination of cepstra and Δcepstra are used, about 25 % of total error is caused by cross-sex speaker confusions. This figure is to be compared with about 4 % when PLP features are used.

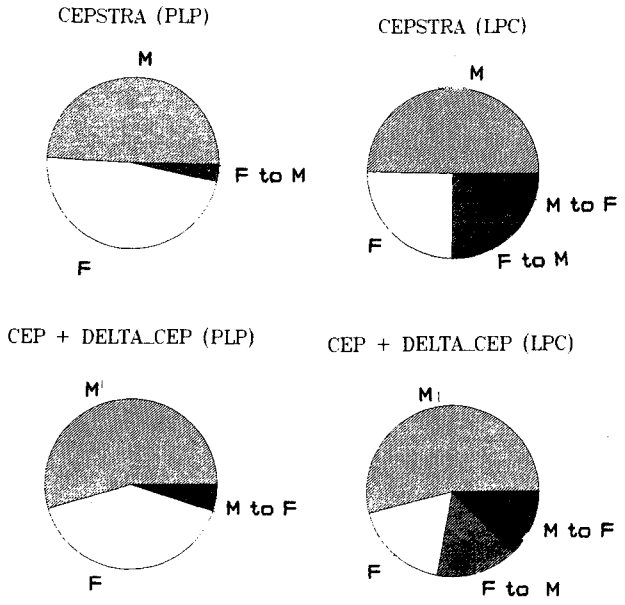

Fig.7 Percentage of error from male and female speakers.

CONCLUSIONS

In this paper we investigated the PLP features and distance measures for ASI. It is found that PLP-derived cepstra give better performance than the Δcepstra; the best results are achieved by jointly using both. The information conveyed in the first 8 coefficients of PLP cepstra is very important in the discriminating process, and the RPS weighting is seem to balance its contributions of each. The INV weighting gives very similar results to the RPS, with slightly higher bias placed on its higher order coefficients. These two weighting are significantly better than with no weighting. The superiority of PLP features to LPC features is evident whatever cepstra, Δcepstra, or combinations are used. Also, the PLP features can better distinguish between male and female speakers.

REFERENCES

[1] Atal, B, *Automatic recognition of speakers from their voices,* Proc. IEEE, vol.64, pp.460-475, 1976.

[2] Furui, S., *Cepstral Analysis Technique for Automatic Speaker Verification*, IEEE Tran. ASSP, vol.29, pp.254-272, 1981.

[3] Xu,L., Oglesby,J. and Mason,J., *The Optimization of Perceptually-based Features for Speaker Identification*, Proc. ICASSP, pp.520-523, 1989.

[4] Soong,F.K. and Rosenberg,A.E., *On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition*, Proc. ASSP, pp.877-879, 1986.

[5] Hermansky, H., Hanson, A. B and Wakita, H., *Perceptually Based Linear Predictive Analysis of Speech*, Proc. ICASSP, pp. 509 - 603, 1985.

[6] Gu,Y. and Mason,J.S., *A Comparsion Between Vocal Tract and Auditory Feature Analysis in ASR System*, Proc. ECST, pp.132-135, 1987.

[7] Stevens,S.S., *The Psychological Review*, Psychological Review, vol.64, pp.153-181, 1957.

[8] Tohkura,Y., *A Weighted Cepstral Distance Measure for Speech Recognition*, Proc. ICASSP, pp.761-765, 1986.

[9] Klatt,D., *Predication of Perceived Phonetic Distance From Critical Band Spectra: A First Step*, Proc. ICASSP, pp.1278-1281, 1982.

[10] Applebaum,T., Hanson,A. and Wakita,H., *Weighted Cepstral Distance Measures in Vector Quantization Based Speech Recognition*, Proc. ICASSP, pp.1155-1158, 1987.

[11] Nocerino,N.S., Soong,F.K., Rabiner,L.R. and Klatt,D., *Comparative Study of Several Distortion Measures for Speech Recognition*, Proc. ICASSP, pp.25-28, 1985.

[12] Hanson,B.A., and Wakita,H., *Spectral Slope Based Distortion Measures for All-pole Models of Speech*, Proc. ICASSP, pp.757-760, 1986.

[13] Hermansky,H., *An Efficient Speaker-independent Automatic Speech Recognition by Simulation of Some Properties of Human Auditory Perception*, Proc. ICASSP, pp.1159-1162, 1987.