Proceedings of the Institute of Acoustics

A COMPARISON OF PLP-RPS AND PLP-IMELDA ACOUSTIC REPRESENTATIONS IN AUTOMATIC SPEECH RECOGNITION

Stephen M. Richardson and *Melvyn J. Hunt*

Marconi Speech & Information Systems
Airspeed Road, The Airport
Portsmouth, P03 5RE, U.K.

SUMMARY

This paper explores two acoustic representations, IMELDA and PLP-RPS, both of which had given good results in speech recognition tests. IMELDA is examined in the context of some related representations developed at NTT, Lincoln Laboratory and IBM. Experimental results suggest that the effectiveness of PLP-RPS stems not from its modelling of perceptual properties but from its approximation to a desirable statistical property attained exactly by IMELDA. A combined PLP-IMELDA representation is found to be more effective than PLP-RPS, but not clearly better than an IMELDA representation derived directly from a filter-bank. Some preliminary experiments with PLP incorporating dynamic information are described.

1. Introduction

Two acoustic representations, PLP-RPS [1] and IMELDA[2] have recently been reported to give particularly good results in speech recognition experiments. This paper explores the properties of the two representations, compares their performance directly, and tests their potential for combination.

2. PLP

PLP — Perceptual Linear Prediction — is a representation of the short-term speech power spectrum that attempts to incorporate a set of perceptual features. The speech signal is first analysed by a critical-band filter-bank. The channel energies emerging from the filter-bank are then scaled to reproduce the varying sensitivity of the ear at different frequencies. The cube roots of the resulting intensities are then taken to make them closer to perceptual loudness values.

In Linear Predictive Coding (LPC) the autocorrelation properties of a waveform are used to carry out an analysis that amounts to fitting an all-pole model to the power spectrum. In PLP, an all-pole model is fitted to the perceptually processed power spectrum just described. This is achieved by exploiting the Wiener-Khinchine relationship, which states that the inverse Fourier transform of the power spectrum of a signal is its autocorrelation function. The autocorrelation function corresponding to a perceptually transformed spectrum is therefore obtained simply by taking its inverse Fourier transform.

The all-pole spectral fit produced by LPC, although corresponding to a least-squares fit in the time domain, does not correspond to what would be obtained using a conventional least-squares error criterion on the spectrum. Rather, it corresponds to a criterion in which errors in intense parts of the spectrum are given more weight in the minimisation process. Since this can result in details in low-intensity parts of the spectrum being ignored, it resembles simultaneous masking in the ear and it might be expected to reduce sensitivity to low-level interfering noise in the recognition process. Taking the cube root of the spectrum before applying the fit reduces this effect. It has been suggested [3] that this may have the desirable result of reducing sensitivity to harmonics of the fundamental frequency in voiced speech sounds.

Since an n'th-order analysis can model a spectrum with at most n peaks, PLP can be seen as a means of smoothing the log power spectrum. The maximum order of the PLP analysis is set by the number of channels in the filter-bank.

Note that if the all-pole spectral fitting is not applied, but the results of the earlier stages are simply expressed on a log scale, then the cube rooting simply divides all the log energy values by three and the

frequency sensitivity scaling simply adds constants to the log channel energies. As neither of these effects will influence decisions in a recogniser, in the absence of spectral modelling they are irrelevant.

The result of a PLP analysis is generally expressed as cepstrum coefficients. These are equivalent to the coefficients of the cosine transform of the log of the fitted all-pole spectrum, but they can be computed directly from the PLP predictor coefficients. The number of coefficients used is normally set equal to the PLP analysis order.

In recognition tests elsewhere[4] particularly good results have been reported when the sets of weighted PLP cepstrum coefficients are compared using a Euclidean distance measure. The weighting scheme, known as *root power sums* (RPS), consists simply of multiplying each cepstrum coefficient by its index number: C_3 by 3, C_4 by 4, *etc.*.

3. IMELDA

Linear Discriminant Analysis (LDA) is a technique used in statistical pattern classification where an unknown sample is to be assigned to one of a set of discrete classes according to the values of a set of numerical parameters describing the sample. The parameters are assumed to have identical multivariate Gaussian distributions about the class centroids. The distribution is completely specified by a within-class covariance matrix, **W**. Under these assumptions, the probability of a sample's belonging to a given class can be computed. The simplest way of doing this is to perform a linear transformation on the parameters corresponding to the rotation and scaling needed to turn the within-class distributions into uniform distributions, *i.e.* such that the class-conditional probability falls off with distance from the centroid at the same rate in all directions and **W** becomes an identity matrix. The log probability of belonging to a class is then linearly related to the Euclidean squared distance from the centroid. Distances in this transformed space are called *Mahalanobis* distances.

LDA further assumes that the class centroids are themselves distributed according to a different multivariate Gaussian distribution, represented by a between-class covariance matrix, **B**. Excluding the unlikely case of **B** being a simple multiple of **W**, some directions in the parameter space will be more effective at discriminating between classes than others. In the transformed space in which **W** is an identity matrix, the set of orthogonal axes that run from the directions that give most discrimination to those that give least consists of the eigenvectors of the transformed **B** matrix. Computation can be reduced by dropping the resulting least effective transformed parameters; and, somewhat surprisingly, this dimensional reduction generally improves discrimination.

As early as 1979 it had been argued[5] that the distance measures in speech recognition should be based on within-class variances in speech sounds, and that even though it might not be possible to define the "classes" of speech sounds — or even decide whether discrete classes existed — it was still possible to estimate a lumped within-class covariance matrix, **W**, of the kind used in linear discriminant analysis. The matrix was to be derived by dynamic programming non-linear time alignment of individual examples of words to their corresponding averaged templates or word models. The vectors describing the distribution about the class centroids were taken to be the differences between the set of spectral parameters in a frame of a particular example and the aligned template frame lumped over all frames in the example, all examples of the word, and all words in the vocabulary. The template frames thus played the rôle of class centroids. Although each template frame could not be considered to be a separate class, an estimate of the between-class covariance matrix could be made by computing the covariance over the parameters in the frames of all the templates in the vocabulary. A set of acoustic parameters describing the spectrum could thus be transformed to a set of parameters designed to optimise discrimination between template frames when used with unweighted Euclidean distances.

The technique just described is particularly attractive when different kinds of acoustic information — for example, static and dynamic spectral information — are to be used in combination. In such cases LDA can provide an optimal combination, where otherwise empirical methods must be used[6]. Indeed, the technique was first used to derive an effective representation for speech recognition from the two separate spectral

representations generated by an auditory model[7].

The LDA method was next used with an FTT-derived mel-scale filter-bank [2]. Since the method can integrate disparate types of information, and since it uses the perceptually motivated *mel* frequency scale, the resulting representations were called *IMELDA* — Integrated Mel-scale representation with LDA. A version using only static spectral information was called IMELDA-1, and a version incorporating dynamic spectral information IMELDA-2.

Although both speaker-dependent and speaker-independent experiments were carried out, the IMELDA transform was generally computed using templates averaged over all speakers in the set used. Computing an IMELDA transform separately for each speaker showed only a slight advantage in speaker-dependent tests, and it would not in any case be practicable in many applications.

To give the acoustic representation produced by the transform some immunity to signal degradations, the following procedure was adopted. Three copies were made of the speech used to derive the transform. The first copy was left undegraded and was used to produce the templates and the B matrix. The second copy had white noise added to give a 15dB SNR, and the third copy was passed through a pre-emphasis filter to produce a 6dB per octave spectral tilt. All three copies were then time aligned to the templates to compute a single W matrix. Because of fears that corresponding speech sounds might not be reliably aligned with the degraded copies, the alignment path determined from the undegraded copy was used throughout. Experiments showed that this process greatly improved recognition performance with degraded speech while only slightly reducing performance with undegraded speech. Moreover, subsequent experiments showed that the method conferred resistance to other degradations — for example those encountered on the telephone and in a helicopter — though admittedly not quite so great as is obtained by using the specific degradations to be encountered in recognition in the derivation of the IMELDA transform.

4. Relationship of IMELDA to Other Acoustic Representations

In speech recognition systems incorporating dynamic programming time alignment or Hidden Markov Models (HMM's) with Viterbi decoding, the spectral distances used are implicitly assumed to correspond to class-conditional log probabilities. If Euclidean distances are used, the acoustic parameters should have class-conditional covariance properties corresponding to an identity matrix. We have seen in the previous section that IMELDA provides this property under the (probably unreasonable) assumption that the acoustic parameters have the same class-conditional distributions in all speech sounds.

On the other hand, the log energies generated by a mel-scale filter-bank do not have the desired properties because adjacent channels are highly correlated. The class-conditional correlation can be greatly reduced by applying a cosine transform to give a mel-cepstrum representation. However, the class-conditional variances of mel-cepstrum coefficients are not uniform but rather decrease with increasing quefrency. Various semi-empirical schemes have been proposed for scaling up the variances of the higher quefrency coefficients. In particular, Tohkura[8] at NTT proposed scaling to make the *total* variance of cepstrum coefficients uniform. While this procedure has been claimed to provide Mahalanobis distances, such distances, as we have seen, require the class-conditional variances to be uniform. Data presented elsewhere [2][9] show that the Tohkura weighting scheme gives too much weight to higher quefrency coefficients relative to what would be needed for Mahalanobis distances.

Continuous-parameter HMM's [10] offer the possibility of converting the class-conditional covariance matrix to an identity matrix separately for each state in each model. This avoids the oversimplification in IMELDA that all states (or template frames) have the same covariance matrix. However, such an approach is rarely used in practice because of the high computational cost in the recognition process and the enormous amount of training material needed for each word in the vocabulary. Instead, the individual state covariance matrices are often assumed to be diagonal, so that the state-specific statistics are specified entirely by variances. Furthermore, workers at Lincoln Laboratory [11] have reported better results when these variances are pooled

over all states to produce what they call *grand variances*.

Grand variances thus share IMELDA's assumption of identical state-specific probability distributions and add the further assumption that the pooled covariance matrices are diagonal. Pooling variances rather than covariances is well motivated only for acoustic parameters that have little correlation between them. Also, since the method simply results in a scaling of the input acoustic parameters, computational limits on the number of parameters that can be used in the recognition process also represent limits on the number of acoustic parameters that can be introduced. It may have been a combination of these two factors that led the Lincoln Lab workers to confine their experiments to static spectral features.

Workers at IBM [12] used a linear transformation on the output of a perceptually based filter-bank that is extremely close to IMELDA. They took as their representation the top few eigenvectors of $\mathbf{W}^{-1}\mathbf{T}$, where $\mathbf{T}$ is the total covariance matrix. LDA gives the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$, but since $\mathbf{T} = \mathbf{B} + \mathbf{W}$, the two sets of eigenvectors are identical, even though the eigenvalues are of course different. Dynamic features were incorporated in the representation by concatenating log power spectra from adjacent 10ms frames. This is equivalent to differencing adjacent frames over time, since the two are related by a linear transformation. However, dynamic information is known to be more effective when derived from longer temporal separations, as it has been in the IMELDA work. This and the absence of the application of signal degradations from the IBM work constitute the two major differences from the IMELDA work.

The IBM work supports our experience that dropping the bottom few discriminant vectors helps recognition performance. However, the improvement is modest in scale compared with that obtained from the first stage of LDA, namely the transformation of W into an identity matrix.

5. Implementation, Verification and Optimisation of PLP

All experiments reported in this paper were carried out on a nine-speaker digit database with a quasi-isolated-word recognition system, both described elsewhere [7]. To assess robustness under degradations, recognition performance was measured with speech to which steady white noise had been added to give a 15dB SNR and with speech to which a 6 db/octave tilt had been applied as well as with undegraded speech. The averaged templates were, however, always derived from undegraded examples.

The first set of experiments sought to determine the optimal analysis order with a representation consisting of static-spectrum RPS-weighted PLP coefficients. As shown in Figure 1, the overall optimum recognition performance was obtained for an order around fifteen. This is inconsistent with results published by Hermansky [4] indicating an optimal order in speaker-dependent experiments of eight and in cross-speaker experiments of five, though it is consistent with more recent work by Hanson and Applebaum[13]. The discrepancy may be partly explained by our use of averaged templates, which, particularly in our speaker-independent tests where the averaging was over eight speakers, causes a smoothing of the spectrum. This may remove any advantage that low-order PLP provides in smoothing the spectrum. Also, the advantage we find for higher-order analyses is most marked for degraded speech, while Hermansky's tests were exclusively with undegraded speech.

Whatever the reason for our finding that higher-order analyses are preferable, it has an impact on the interpretation of the apparent effectiveness of PLP. It has been suggested[14] that humans may analyse speech into two major resonances labelled $F1$ and $F2'$, with $F2'$, the effective second formant, corresponding to the resonance of the front cavity of the vocal tract. This front cavity is sometimes associated with the second formant and sometimes with the third; and speakers, the argument goes, may be able to adjust it so as to produce the same $F2'$ for the same speech sound despite differences in the size of their vocal tracts. Fifth-order PLP has been shown [15] to produce something close to an $F1$–$F2'$ analysis. This seemed to provide a particularly elegant explanation of the effectiveness of fifth-order PLP in the cross-speaker experiments. Our results and those of Hanson and Applebaum favouring much higher order analyses where a full set of formants can be observed are inconsistent with this explanation.

PLP-RPS AND PLP-IMELDA ACOUSTIC REPRESENTATIONS

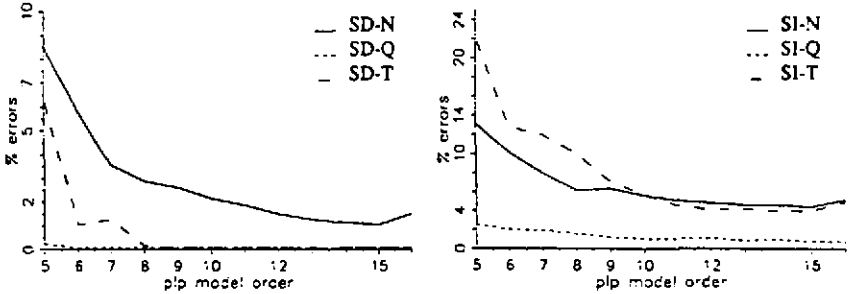

Figure 1 Percentage error rates for speaker-dependent (SD) and speaker-independent (SI) quasi-isolated word recognition experiments using the PLP-RPS acoustic representation. Undegraded (Q), noise degraded (N) and with spectral tilt (T).

We next sought to verify the advantage of RPS weighting; and here are results are entirely consistent with Hermansky's. Table 1 shows that while unweighted PLP is no better than an unweighted mel-cepstrum representation, PLP with RPS weighting is much better. It is also much better than an equivalently weighted mel-cepstrum representation. We will return to the question of why RPS weighting is so effective with PLP after first discussing results with IMELDA-1 and a combination of PLP and IMELDA.

Representation	# Coeffs.	Speaker Dependent			Speaker Independent		
		Q	N	T	Q	N	T
plp-cepstrum	12	0.59	29.82	69.29	5.71	47.48	75.89
quef. wtd. plp-cep	12	0.07	1.48	0.07	1.11	4.82	4.08
mel-cepstrum	12	0.37	29.82	68.47	3.71	43.92	78.49
quef. wtd. mel-cep.	12	0.00	26.34	23.74	3.86	39.69	66.17
IMELDA-1	12	0.00	2.00	0.00	1.11	3.93	2.00
PLP-IMELDA	12	0.07	1.26	0.07	0.89	4.30	2.08

Table 1 Comparison of error rates for weighted and unweighted mel-cepstrum and PLP acoustic representations, and an IMELDA-1 representation derived from the output of the mel-scale filterbank.

6. Tests with IMELDA and PLP-IMELDA

The derivation of an IMELDA representation from the log channel energies of a mel-scale filter-bank described in Section 3 can equally well be applied to a set of PLP cepstrum coefficients. We call such a representation PLP-IMELDA, and Figure 2 compares the performance obtained with this representation and the PLP-RPS representation. The differences are small, but PLP-IMELDA has proved consistently better than PLP-RPS. However, it is not clear that the PLP-IMELDA representation is any better than IMELDA-1 derived directly from the mel-scale filter-bank (Table 1). The speaker-dependent PLP-IMELDA results in noise may show the hoped-for advantage over the direct IMELDA, but it is not maintained in the speaker-independent results in noise.

7. Interpretation of PLP-RPS Results

It is striking in Table 1 that while unweighted PLP cepstrum coefficients are no more effective than unweighted mel-cepstrum coefficients, RPS-weighted PLP cepstrum coefficients are much more effective than equivalently weighted mel-cepstrum coefficients. Indeed, they are almost as effective as the PLP-IMELDA representation. We need to ask why RPS weighting seems so well suited to the PLP cepstrum.

PLP-RPS AND PLP-IMELDA ACOUSTIC REPRESENTATIONS

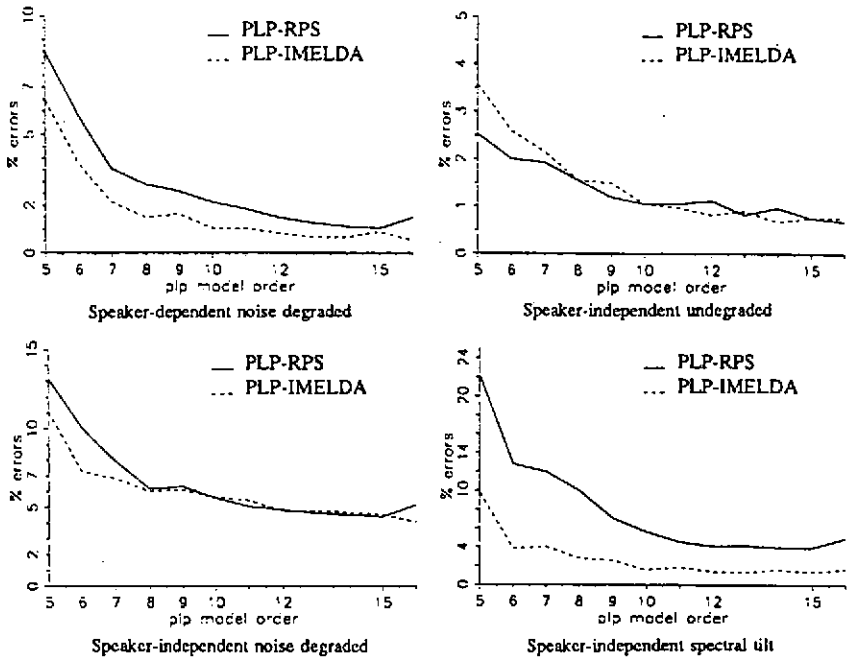

Figure 2 Percentage error rates for speaker-dependent † and speaker-independent quai-isolated word recognition experiments using the PLP-RPS and PLP-IMELDA acoustic representations.

† Speaker-dependent experiments with undegraded and spectral tilt speech data have almost zero error rate.

Weighting according to quefrency — RPS weighting — is known to give too much weight to the higher quefrency coefficients[8][9]. That is, the class-conditional variances scaled up by this weighting scheme are too large. For this reason, weighting schemes of this kind limit the weighting beyond the first few coefficients. The higher coefficients are sensitive primarily to spectral fine structure. Since, as we have seen, PLP has the effect of smoothing the spectrum, we might expect that the variance of the higher quefrency cepstrum coefficients would be reduced. Measurements on our database confirm this expectation. Indeed, as Figure 3 shows, the class-conditional variances of RPS-weighted cepstrum coefficients to be almost constant for 12'th-order PLP. By contrast, equivalently weighted mel-cepstrum coefficients show

steeply rising variances, while IMELDA coefficients are the most accurately constant, as one would expect since it is part of their definition. Moreover, IMELDA coefficients are accurately uncorrelated.

It appears that the effectiveness of a PLP-RPS representation stems from the fortuitous combination of the low-pass liftering imposed by PLP and the high-pass liftering of RPS weighting which balance each other to give almost constant class-conditional variances.

8. Tests with Dynamic Spectral Information

Section 3 pointed out that dynamic spectral information can be incorporated in an IMELDA representation. Recognition tests with dynamic PLP parameters have also been reported[16][13].

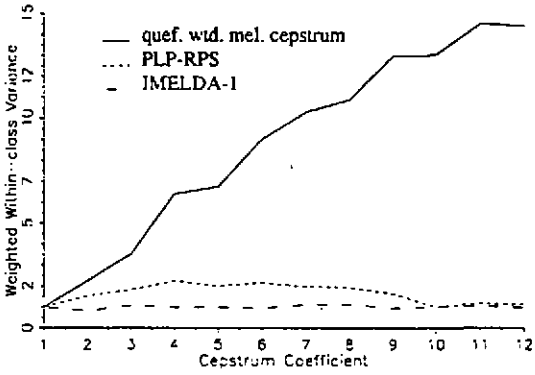

Figure 3 Within-class variances of the first twelve cepstrum coefficients derived from the mel-scale filterbank and a 12^{th} order PLP analysis, weighted by their squared quefrency index, and within-class variances of the first twelve IMELDA coefficients derived from the mel-scale filterbank.

Given the good performance of PLP-IMELDA, we therefore investigated a PLP-IMELDA representation including dynamic coefficients. The dynamic coefficients were obtained from linear regression over seven consecutive 6.4 ms frames.

Figure 4 shows the performance of the PLP-IMELDA-2 representation to be consistently better than the PLP-IMELDA-1 representation. However, it appears to offer no clear advantage over a "conventional" IMELDA-2, though the error rates in both cases are so low that it it is questionable whether the performance differences are real. We are about to begin tests with all the representations on a larger, more difficult database, which should allow definite conclusions to be drawn.

Hanson and Applebaum[13] have recommended low-order PLP for dynamic coefficients. Our preliminary results with mixed-order static and dynamic representations appear consistent with these recommendations.

9. Conclusions

1. An explanation has been offered for the effectiveness of the PLP-RPS representation. It appears not to be related to $F1$–$F2'$ modelling or other perceptual phenomena but to an accidental statistical property.
2. This statistical property is optimised in IMELDA, and PLP-IMELDA gives better performance than PLP-RPS when only static spectral parameters are used, though not clearly better than a direct IMELDA-1, which is computationally simpler.
3. It is not yet clear if PLP offers any advantages when incorporating dynamic information. Results to settle this issue will be announced shortly.

Acknowledgments

The work described here was partly in fulfilment of a master's degree requirement for SMR and partly a contribution to the IED ISTD Project in collaboration with HUSAT and CSTR. WE are grateful to the National Research Council of Canada for providing access to databases and software, to Alain Piau for help with some of the experiments, and to Dr. Hynek Hermansky for useful discussions.

PLP-RPS AND PLP-IMELDA ACOUSTIC REPRESENTATIONS

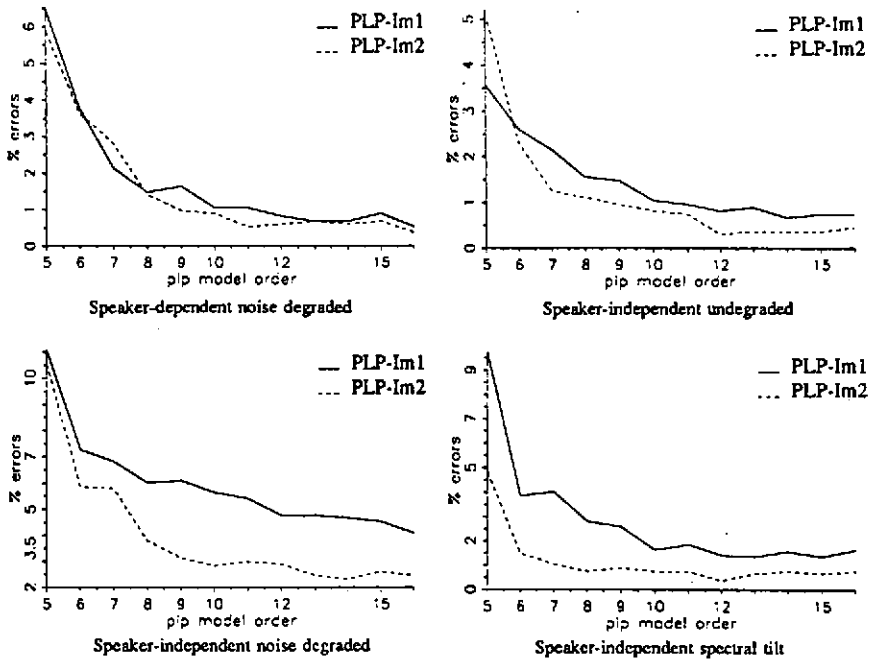

Figure 4 Percentage error rates for speaker-dependent and speaker-independent quasi-isolated word recognition experiments using PLP-IMELDA-1 and PLP-IMELDA-2 acoustic representations.

References

1. Hynek Hermansky, Brian A. Hanson and Hisashi Wakita, "Perceptually Based Linear Predictive Analysis of Speech," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-85*, pp. 509-512, Tampa, Florida, March 1985.
2. Melvyn J. Hunt, Claude Lefèbvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-89*, vol. S1, pp. 262-265, Glasgow, Scotland, May 1989.
3. Hynek Hermansky, Hiroya Fujisaki and Yasuo Sato, "Analysis and Synthesis of Speech Based on Spectral Transform Linear Predictive Method," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-83*, pp. 777-780, Boston, Massachusetts, April 1983.
4. Hynek Hermansky, "An Efficient Speaker-Independent Automatic Speech Recognition by Simulation of Some Properties of Human Auditory Perception," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-87*, pp. 1159-1162, Dallas, Texas, April 1987.

5. Melvyn J. Hunt, *A Statistical Approach To Metrics For Word And Syllable Recognition*, Nov 1979. Presented at the 98th Meeting of the Acoustical Society of America, Salt Lake City, Utah

6. Kiyoaki Aikawa, Sadaoki Furui, "Spectral Movement Function and Its Application to Speech Recognition," *PRoc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-88*, pp. 223-226, New York City, USA, April 1988.

7. Melvyn J. Hunt, Claude Lefèbvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-88*, pp. 215-218, New York City, USA, April 1988.

8. Yoh'ichi Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-86*, pp. 761-764, Tokyo, Japan, April 1986.

9. Melvyn J. Hunt, Claude Lefèbvre, *Distance Measures for Speech Recognition*, National Aeronautical Establishment, Canada, March 1989. NAE-AN-57, NRR No.30144

10. S. E. Levinson, L. R. Rabiner and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell System Tech. J.* , vol. 62, pp. 1035-1074, 1983.

11. R. P. Lippmann, E. A. Martin, "Discriminant Clustering Using an HMM Isolated-Word Recognizer," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP-88*, pp. 48-51, New York, April 1988.

12. L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Speech Recognition with Continuous Parameter Hidden Markov Models," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, ICASSP-88*, pp. 40-43, New York, April 1988.

13. Brian A. Hanson, Ted H. Applebaum, "Robust Speaker-Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments With Lombard and Noisy Speech," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICCASP-90*, pp. 857-860, Albuquerque, New Mexico, April 1990.

14. G. M. Kuhn, "On The Front Cavity Resonance And Its Possible Role In Speech Perception," *JASA*, vol. 58, No.2, pp. 428-433, 1975.

15. Hynek Hermansky, David J. Broad, "The Effective Second Formant F2' and the Vocal Tract Front-Cavity," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing*, pp. 480-483, Glasgow, Scotland, May 1989.

16. Hynek Hermansky, Jean-Claude Junqua, "Optimization of Perceptually-Based ASR Front-End," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-88*, pp. 219-222, New York City, USA, April 1988.