Perceptually based processing in automatic speech recognition

PERCEPTUALLY BASED PROCESSING IN AUTOMATIC SPEECH RECOGNITION

Hynek Hermansky, Kazuhiro Tsuga, Shozo Makino, and Hisashi Wakita

Speech Technology Laboratory,
3888 State Street,
Santa Barbara, California 93105, U.S.A.

Abstract

The perceptually based linear predictive (PLP) speech analysis method is applied to isolated word automatic speech recognition (ASR). Low dimensionality of the PLP analysis vector, which is otherwise identical in form to the standard linear predictive (LP) analysis vector, allows for computational and storage savings in ASR. We show that in speaker-dependent recognition of the alpha-numeric vocabulary, the PLP method in VQ-based ASR yields similar recognition scores as does the standard ASR system. The main focus of the paper is on cross-speaker ASR. We demonstrate in experiments with vowel centroids of two male and one female speakers that PLP speech representation is more consistent with the underlying phonetic information than the standard LP method. Conclusions from the experiments are confirmed by superior performance of the PLP method in cross-speaker isolated word recognition.

1. Introduction

Most speech analysis methods used in ASR originated in low-bit rate speech transmission. Such methods usually model process of speech production and their most useful property in ASR is the information rate reduction of the speech signal. However, speech is perceived by human auditory system and therefore modeling of process of speech perception seems to be a more reasonable approach to ASR speech analysis. Another reason for use of the perceptually based analysis in ASR is pragmatic. Modeling of many properties of the human auditory system makes sense from the engineering point of view.

Perceptually based speech analysis methods can be divided into two categories. One approach attempts to model the physiology of the human auditory organs as measured in the periphery of the auditory systems of mammals. Some very promising cues have been obtained by this approach (see e.g. [6, 14, 17, 18]). However, the complexity of the following processing of those cues so far discourages practical use of the physiologically based analysis methods in ASR. The second approach to perceptually based speech analysis is psychophysical (see. e.g. [11, 20, 5, 16, 3, 1, 13, 9]). It treats the human auditory system as a whole, attempting to model reported response of the human being to the acoustic stimuli. Some of the well established psychophysical properties of the human auditory perception, as is e.g. the nonlinearity of spectral resolution, wide bands of spectral energy integration, or the nonlinear compression of the spectral acoustic energy are useful engineering concepts in speech analysis.

Needless to say that most of the efforts to integrate perceptually based analysis into a practical ASR system met only with limited success. As pointed out by Blomberg et al. [3] the failure to demonstrate advantages from use of the perceptually based analysis in ASR might be partly because the rest of the ASR system is not modified for the perceptually based analysis. Furthermore, the improvement is being sought in the areas in which the standard analysis technique performs the best, instead in the areas in which the standard technique unquestionably fails.

In our paper we will discuss applications of previously proposed perceptually based LP (PLP) method [9] in multi-speaker ASR. The PLP method belongs to the second, psychophysical, category of perceptually based analysis methods and produces results in the form of a low-dimensional all-pole model. That allows for utilizing many processing techniques, developed for LP analysis. We demonstrate that PLP analysis significantly improves recognition accuracy in multi-speaker recognition.

2. PLP Speech Analysis Method

The PLP speech analysis method models the speech auditory spectrum by the spectrum of low-order all-pole function. The auditory spectrum is obtained by critical-band spectral analysis which integrates the speech energy spectral density over 18 bands in the 0 - 5 kHz frequency range, followed by equal-loudness pre-emphasis which emphasizes the middle and the upper part of the speech spectrum, and by intensity-to-loudness cubic compression, which reduces dynamics of the speech spectrum. Eighteen samples of the auditory spectrum, obtained in this way, are transformed through the inverse discrete Fourier transform into the autocorrelation domain. Five coefficients of the 5th order all-pole model are computed from the Yule-Walker relations. The block diagram of the PLP speech analysis method is shown in Fig. 1. Further details of the method can be found in [9].

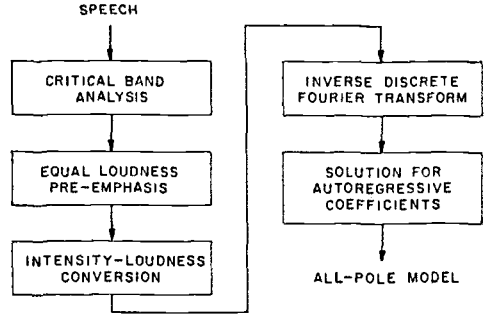

Fig. 1 Block diagram of PLP speech analysis method

ICASSP 86, TOKYO CH2243-4/86/0000-1971 $1.00 © 1986 IEEE

Spectrum of the resulting all-pole model is linear in the physiological tonality-loudness domain (as compared to the linearity in the frequency-power spectral density domain of the standard LP method) and has at most two spectral peaks. The PLP analysis vector is about half size of the analysis vector from the typical LP method. It models, consistently with human auditory perception, more detail in the lower part of the speech spectrum and has rather broad spectral peaks.

On the other hand the low dimensionality of the PLP analysis vector raises the question of how adequate is the phonetic information extracted by PLP analysis. Fig. 2 shows the PLP peak trajectories of the analysis of an utterance, spoken by male and female speakers. The peak trajectories represent simplified speech spectrograms with ordinate on the Bark scale. Different speech segments are easily recognizable in those spectrograms. Uniformity of male and female spectrograms is rather good. This indicates that the PLP analysis might provide efficient means for speech analysis in ASR. The rest of our paper is devoted to this topic.

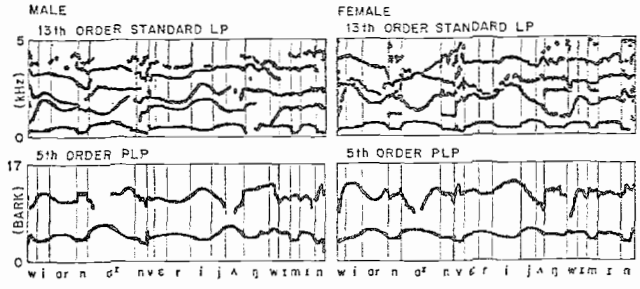

Fig. 2 Spectral peak trajectories from LP and PLP analysis for male and female speakers.

3. Distance Metrics for the PLP-Based ASR System.

Parameter extraction in ASR is followed by a component in which the analysis vectors are compared with some precomputed standard. The distance metric applied in this comparison stage has a significant impact on the ASR system performance. The distance metric must respect character of the speech representation obtained in the speech analysis stage.

When the recently proposed root-power sums (RPS) distance measure [14] is used with PLP analysis, the recognition result is within limits of statistical variance for both the PLP-based and the LP-based ASR systems. On the other hand, when the standard LP cepstral distance measure [7] is applied to the vector comparison in PLP-based ASR, the recognition result is significantly inferior to the result of the standard LP-based ASR system [8]. Our experiments confirmed those findings.

Characteristics of the above mentioned distance metrics are shown in Fig. 4. The figure shows measured distances when one spectral peak of the compared all-pole model changes its position with respect to the stationary reference model. The peak bandwidths investigated varied within limits of typical PLP model bandwidths (4.5 Bark-9.5 Bark). Cepstral and RPS representations were truncated to 14 coefficients in both examples.

The RPS distance measure is more sensitive to the bandwidths of the spectral peaks. Consequently, it puts more emphasis on movements of sharp peaks than on movements of broad peaks and is more sensitive to small changes in sharp peak position than is the LP cepstral distance measure. Compared to the LP method spectral peaks, the bandwidth of the PLP spectral peaks is relatively stable and carries important phonetic information about the spread of the original spectral cluster [9]. Distance metrics which are sensitive to the bandwidth value can utilize this information. The larger sensitivity to the movements of the sharp spectral peaks is a perceptually consistent feature. We adopt the RPS distance metrics for all experiments described in this paper.

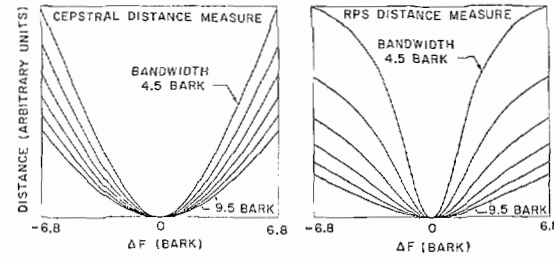

Fig. 3 Sensitivity of spectral distance metrics to movement of spectral peak in all-pole model

4. Speaker-Dependent Isolated Word Recognition

We applied the PLP analysis technique with the RPS distance metric in the double-SPLIT VQ coding ASR system [18] and carried out an initial series of experiments with data from one male speaker. Five repetitions of the alphanumeric data were used. Each repetition served as training data for recognition of the remaining four repetitions. Fifth order PLP analysis using 20 msec Hamming window in 10 msec analysis steps and 14 RPS coefficients was used in all experiments. Results of the experiments, displayed in Fig. 5, show that the total VQ distortion was lowest (which implies that the recognition accuracy was close to optimal) when the clustering threshold was set so that about 10% of the training vector space were not used in the codebook construction. Since this phenomenon was observed for all codebook sizes, this indicates that it is desirable to exclude some atypical analysis vectors from the codebook construction. However, more systematic study with different databases is still needed to accept this hypothesis.

Recognition accuracy codebooks of 64 or more codes is practically the same as the accuracy without VQ coding. Better than 90% recognition accuracy on the highly confusable alphanumeric data indicate similar performance of the PLP method with the standard LP method in this recognition task.

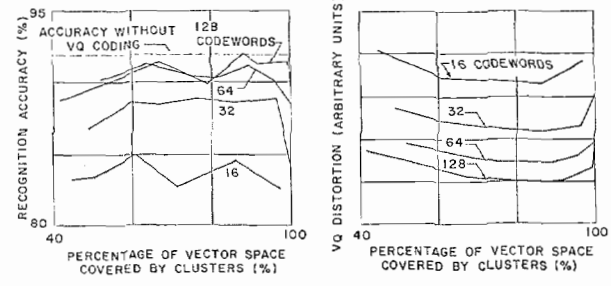

Fig. 4 Dependency of recognition accuracy and VQ introduced distortion on percentage of training vector space involved in clustering

5. Cross-Speaker Isolated Word ASR.

It is known that across speakers, the formant representation of speech estimated by LP analysis is not very consistent with the underlying phonetic information. Consequently LP-based recognizers perform best in the speaker dependent mode. Results reported in [10] indicate that the PLP speech representation is more isomorphic across speakers than the LP representation. In order to gain further insight into the speaker normalizing properties of PLP analysis, we carried out series of cross-speaker experiments.

Speech of two male and one female speakers, each pronouncing five times the complete typewriter keyboard vocabulary (104 words in each repetition) was manually labeled so that the typical frame for each vowel in the data were known. Centroids for all 12 vowels were obtained by averaging the autocorrelation coefficients of typical frames for a given vowel. The all-pole models were computed for each centroid by both the 14th order LP and the 5th order PLP methods. In this way, 12 vowel-like LP and PLP speech representatives of each of three speakers, i.e. 36 LP vectors and 36 PLP vectors, were available for further experiment.

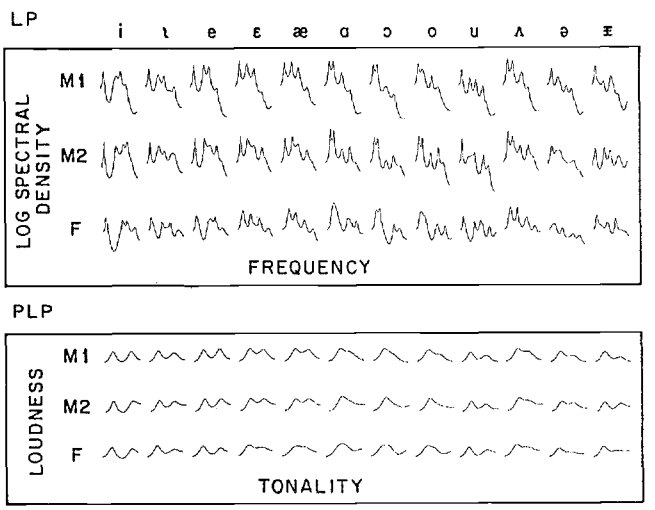

Fig. 5 Spectra of LP and PLP models of vowel-like centroids for two male and one female speakers

Spectra of the LP vectors, shown in the upper part of Fig. 5, exhibit the typical vowel formant structure, the spectra of the PLP vectors, shown in the lower part of Fig. 5, have at most two spectral peaks, consistently with the effective second formant theory [4]. Visually, there are indications that PLP analysis compensates for some discrepancies of the standard formant representation. In order to quantify our observation we designed a following experiment:

RPS distances between vectors from the LP and PLP methods were computed and arranged in matrix forms. Six matrices for each of the analysis methods were obtained from the data. Three of them contain distances between vectors from analysis of vowel centroids of one speaker and are symmetrical and have the zero diagonal. Another three matrices contain distances between vectors of different speakers. In the case of ideal speaker-independent and perceptually consistent speech representation, all matrices would be the same and have zero diagonals. The real situation is that no speech representation is entirely speaker independent. Consequently, all distances in the cross-speaker matrices are nonzero. Matrices are also asymmetric, reflecting the perceptual inconsistency of the speech representation across the speakers. We have chosen two criteria for evaluation of cross-speaker matrices.

1) Distances between centroids with identical phonetic values but from different speakers do not necessarily have to yield the smallest distances. However, distances along diagonals of cross-speaker matrices should be in average smaller then off-diagonal distances, since they represent distances between centroids of the same vowel, uttered by a different speaker. The criterion for evaluation of both analysis methods is the ratio of the mean distance from off-diagonal entries to the mean distance from diagonal entries.

2) All matrices should be similar, since they would ideally be the same if the speech representation was speaker-independent. Cross-speaker matrices should be similar to one-speaker matrices, since it is known that both the LP method and the PLP method work well in the single-speaker ASR. The criterion for evaluation of similarity is the value of the correlation coefficient between matrices.

Some typical matrices obtained by the LP and PLP methods are shown in Fig. 6. The distance between abscissa and ordinate elements is indicated by the size of the asterisk at the coordinates of the compared elements. The higher similarity of the PLP matrices is obvious even from the visual inspection of the figure.

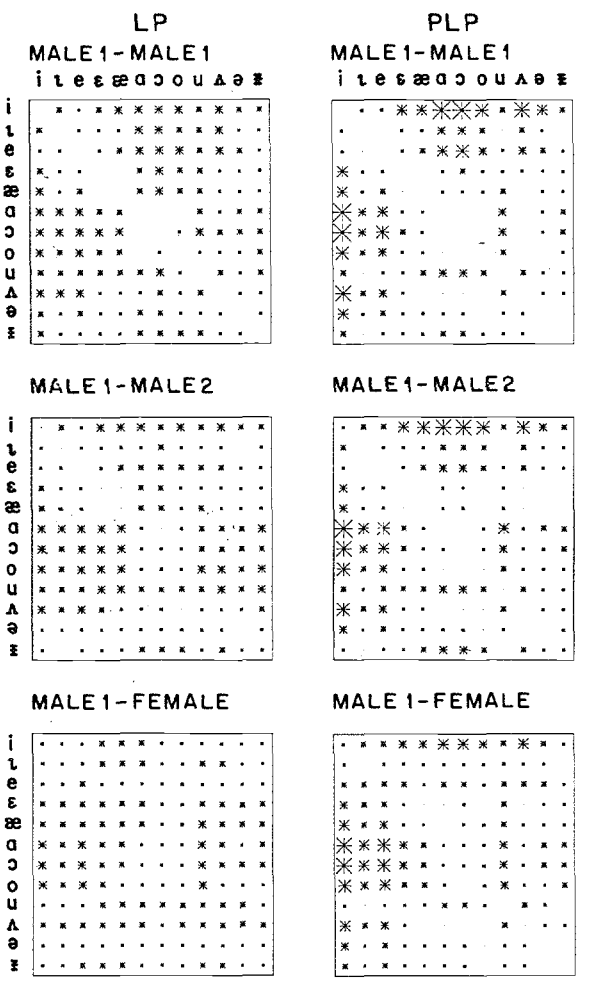

Fig. 6 Single speaker and cross-speaker spectral distance matrices

Table I shows ratios of the off-diagonal average distance to the on-diagonal average distance. According to this criterion, the PLP method has higher discriminating power than the LP method. Table II shows correlation coefficients between all matrices. Better similarity of the PLP matrices was confirmed in this test.

TABLE I: SEPARATION COEFFICIENTS S OF CROSS-SPEAKER SPECTRAL DISTANCE MATRICES

	m1-m2	m1-f	m2-f
LP	2.72	1.09	1.12
PLP	5.94	2.25	2.45

$$S = \frac{1}{(N-1)} \times \left(\sum_{i \neq j} a_{ij} \; / \; \sum_{i=j} a_{ij} \right)$$

TABLE II: CROSS-CORRELATION r_{A-B} BETWEEN SPECTRAL DISTANCE MATRICES

m2		f		m1 – m2		m1 – f		m2 – f		Type of
LP	PLP	LP	PLP	LP	PLP	LP	PLP	LP	PLP	Matrix
.83	.94	.81	.92	.75	.95	.29	.82	.33	.81	male1(m1)
		.78	.91	.80	.95	.26	.78	.37	.81	male2(m2)
				.72	.91	.40	.81	.32	.64	female (f)
						.40	.83	.55	.79	m1 – m2
								.83	.95	m1 – f

$$r_{A-B} = \frac{\sum_{j=1}^{N} \sum_{i=1}^{N} (a_{ij} - \bar{a})(b_{ij} - \bar{b})}{\sqrt{\sum_{j=1}^{N} \sum_{i=1}^{N} (a_{ij} - \bar{a})^2 \times \sum_{j=1}^{N} \sum_{i=1}^{N} (b_{ij} - \bar{b})^2}}$$

The power of the PLP analysis in the cross-speaker speech representation can be practically utilized in multi-speaker ASR. We carried out a series of cross-speaker recognition experiments. The standard configuration of the recognizer with LP analysis on the front end has been compared with the PLP front end recognizer. The analysis vectors were compared using the RPS distance measure. The LP cepstral distance measure was also applied in some initial comparative experiments but was found inferior to the RPS distance measure.

Five repetitions of the 36 word alpha-numeric database for each of two male and one female speakers were tested in the cross-speaker mode. Each repetition of one speaker served as the reference which was compared to all five repetitions of another speaker. The word end points were determined manually. The 14th order LP and 5th order PLP analyses were applied; 20 msec Hamming window and 10 msec analysis step were used in both analysis methods. RPS representations of the LP and PLP models were truncated at 14 coefficients. Template-matching with the standard fixed end-point dynamic time warping was used. A single template was used for each reference word. The word template with minimal accumulated distance determined the recognized word. No rejection threshold was applied.

Results are summarized in TableIII. The agreement of the cross-speaker isolated word recognition results with tendencies observed in the vowel-like centroid comparison is remarkable. The perceptually based PLP method is superior on the 99% level of statistical confidence to the standard LP method in all comparisons. It is encouraging to see a rather dramatic improvement in cross-sex recognition without any speaker normalization except for the partial overall spectral slope compensation by the RPS method [12] and the anticipated normalization by the auditory-like processing [2] of the PLP analysis.

TABLE III: RECOGNITION ACCURACIES IN CROSS SPEAKER RECOGNITION

Recognition Accuracy for Top 3 Candidates [%] (Accuracy for Top 1 Candidate in Brackets)						
Analysis Method	Speaker Combination					
	m1-m2	m2-m1	m1-f	f-m1	m2-f	f-m2
LP	83 (59)	83 (60)	21 (9)	25 (11)	15 (6)	15 (3)
PLP	88 (70)	90 (69)	71 (40)	58 (30)	63 (37)	63 (35)

6. Conclusion

Perceptually based PLP speech analysis method has been used in the isolated word speech recognition. PLP analysis yields analysis vectors of the same form as the standard LP analysis does but the size of the vectors is about half of the size of standard LP analysis vectors. This allows for computational and storage savings in ASR. When applied in a VQ coding speaker-dependent template-matching ASR system, it yields similar accuracies as the standard LP analysis. However, a significant improvement in the recognition accuracy has been observed when PLP analysis was substituted for the standard LP analysis in cross-speaker ASR. The improvement was most significant in male-female cross-speaker recognition.

References

[1] Anderson J. et al: Proc. ICASSP-84, 1364-1367, 1984
[2] Bladon et al.: Language and Comm. 4, 59-69, 1984
[3] Blomberg et al.: Proc. ICASSP-85, 17.9.1-17.9.4, 1984
[4] Carlson et al.: STL-QPRS 2-3, 19-35, 1970
[5] Davis et al.: IEEE-ASSP 28, 357-379, 1980
[6] Delgutte et al.: JASA 75, 866-886, 1984
[7] Gray A.H. et al.: IEEE-ASSP 24, 380-391, 1976
[8] Hanson B.A.: Personal communications, 1985
[9] Hermansky et al.: Proc. ICASSP-85, 509-512, 1985
[10] Hermansky et al.: Speech Comm. 4, 181-187, 1985,
[11] Itahashi et al.: Proc. ICASSP-76, 310-313
[12] Klatt: Proc. ICASSP-82, 1278-1281, 1982
[13] Koljonen et al.: Proc. ICASSP-84, 1.9.1-1.9.4, 1984
[14] Lyon: Proceedings ICASSP-85, 981-984, 1985
[15] Paliwal: Speech Comm. 1, 151-154, 1982
[16] Schroeder et al.: "Frontiers of Speech Comm. Res.", Lindblom ed., 217-229, Acad. Press, 1978
[17] Seneff: Proceedings ICASSP-85, 36.2.1-36.2.4, 1985
[18] Shamma: JASA 78, 1612-1632, 1985
[19] Sugamura et al.: J.Ac.Soc.Japan (E)5, 243-252, 1984
[20] Zwicker et al.: Jasa 65, 487-497, 1979

Acknowledgements

The result of the speaker-dependent ASR without VQ coding was obtained from Brian Hanson. Debbie Szeczei provided programming support. Ted Applebaum also contributed to the software development.