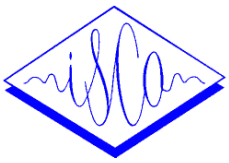


2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

Optimization of Perceptually-based Spectral Transforms in Speaker Identification

L. Xu[†] *and J.S. Mason*[‡]

[†] Dept of Computer Science, University of Manchester, Manchester. UK.
[‡] Dept of Electrical and Electronic Engineering,University of Wales, Swansea, UK.

ABSTRACT

It is recently reported in [1][2] that perceptually-based linear prediction, PLP, features achieve significantly better speaker recognition results than when using standard LPC features. The superiority of the PLP model is attributed to a series of perceptually-based spectral transforms, applied prior to deriving feature sequences from the standard linear prediction process. This paper investigates further the use of PLP features in speaker identification, focusing on the contributions of each of the perceptual factors. PLP, as proposed originally by Hermansky [3] was optimised for *speech* recognition. This paper demonstrates that, not surprisingly, different optimum conditions apply for *speaker* recognition. In particular we show the distinct benefit of increasing the number of critical bands (from the original 17 up to 64). The increased spectral detail is clearly important in this task, and ASI experiments based on 1000 single-digit tests in digit-independent codebook scheme gives a 2.7% error rate for the modified PLP method, compared with 4.7% and 6.5% when using the original PLP and standard LPC models respectively. Furthermore, it is found that all the perceptual weightings considered in the PLP model to some extent enhance the performance, and in agreement with Gu's findings in *speech* recognition [4], ASI performance is shown to be relatively insensitive to the precise masking pattern.

1. Introduction

An important issue in any *speech* or *speaker* recognition system is the front-end feature extraction. Even though the two applications require almost diametrically opposite information, to a large extent the same front-ends have been employed by most researchers. Recent years has seen the widespread use of perceptually weighted cepstral representations, the most popular of which is the the *mel*, often coupled with an inverse variance weighting.

The work here examines variants of a different feature, namely Hermansky's perceptually-based linear prediction (PLP), [3]. The motivation comes from the recent success in the application of this feature to speaker identification, [1][2]. The goal in this case is to somehow extract the speaker-specific information, ideally across a wide range of text. Our earlier work highlights the importance of high-order analysis, the best results coming from PLP-14 and PLP-16. This is in clear contrast to the low order optimum of PLP-5 for *cross-speaker speech* recognition reported in [5], suggesting that the fine spectral detail is more important in *speaker* recognition.

In this paper we extend the hypothesis on finer spectral detail, and also examine the contributions of each of the perceptual parameters in the computational model of PLP.

2. PLP Spectral Transforms

2.1 The Perceptual Model

The perceptually-based spectral transforms are designed on the basis of empirical studies of the responses of the human ear. In the computational model, the goal is to convert the short-time spectra from a conventional linear form to a so-called perceptual domain, with the resultant spectra referred to as 'auditory spectra'.

A number of the facts gleaned from auditory psychophysical experiments are employed in the perceptually-based transforms of PLP, including concepts of: perceived pitch, critical band masking, equal-loudness and intensity-loudness. The computational model of the perceptually-based transforms is illustrated in Figure 1.

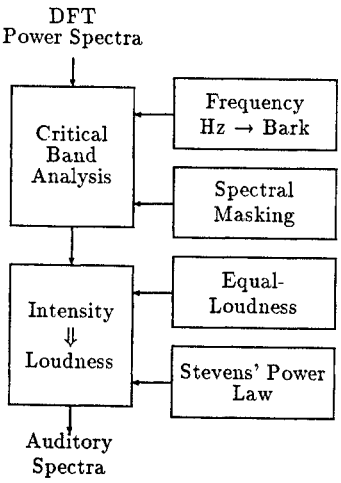


Figure 1: Perceptually-based transforms associated with PLP

The model takes account the ear's nonlinear transformations in *frequency* and *amplitude*. The first is encompassed in the critical band analysis, which has two components: warping of the frequency scale (Hz to $Bark$ scale [1]), and critical band masking

[1] The *mel* perceptual scale is based on pitch doubling perception,, whereas the *Bark* is based on tone discrimination; both have a logarithmic relation to Hz.

which accounts for the masking effects of in-band tones. The conversion of power spectra to loudness gives a representation of the loudness sensation of the auditory system, and is accomplished in two steps. First, a perceived equal-loudness versus frequency function is applied to the critical band output. This is followed by a frequency-*independent* perceived loudness scaling using cubic-root power function relating to physical intensity [6]. A full account of PLP is given in a recent paper by Hermansky [11]

2.2 Spectral Resolution

The auditory spectra are computed from a set of auditory bandpass filters, the number and shape (including desired masking pattern) of which determine the spectral resolution of the model. In the original form of PLP, 17 auditory filters cover the frequency range 0 - 5 kHz. Based on this number of filters example auditory spectra are shown in Figure 2-a.

These spectra are derived from first linearly interpolating the output from filterbank. In the first case (dashed line) this gives the spectrum directly, without further computation. In the second case (solid line) the autocorrelation function is derived from the auditory spectrum and then a standard 14th order all-pole model is derived. This in turn is transformed into a cepstral representation which leads to the characteristically smoother spectra.

We address the question of spectral resolution by considering the number of auditory filters. In related work on perceptually-based analysis, Bladon and Lindblom [7] employ a large number of successive auditory filters. Figure 2-b displays the equivalent auditory spectrum derived from 128 auditory filters, again with a 14th order LP-derived cepstrum (solid line). Comparing the two Figures, 2-a and 2-b, it is clear that both curves in Figure 2-b are smoother and depict more spectral detail.

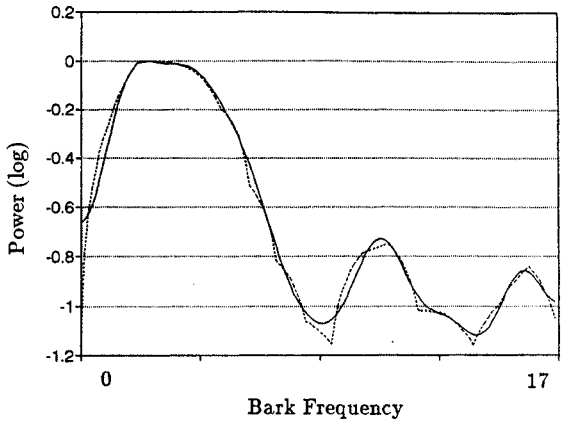
Masking, on the other hand, is an operation which results in *degrading* the spectral resolution. The extent of degradation is affected by the masking pattern used. The masking patterns in the original PLP are designed to have 'flat-top' bandwidths (rather than 3dB bandwidths) of one *Bark*. Here, we consider three masking patterns, shown in Figure 3, as proposed by Schroeder [8], Davis and Mermelstein [9], and Hermansky *et.al* [3]. The smoothed, asymmetric masking curve from Schroeder approximates the experimental data of critical band masking with a one *Bark* 3dB bandwidth. For the symmetrical triangular pattern, we set the bandwidth to be one *Bark*, and the flat-topped asymmetric curve, having a more than one *Bark* bandwidth, is as used in the original PLP analysis [3].

These are three very different masking patterns. Related experiments concerning these shapes and the number of auditory filters are described below.

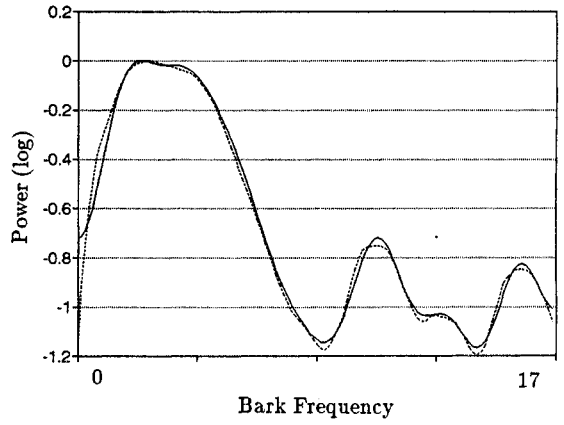
3. A Speaker Identification System

A vector quantization (VQ) codebook approach , as proposed by Soong [10], is chosen for digit-independent speaker identification experiments.

The database contains 10 speakers (5 males and 5 females) and 1,000 isolated digit utterances, 10 versions per person, sampled at 10 kHz, and collected over a period of weeks. Training of models uses either the first or second five versions of the data, in time recording sequence; the testing uses the other five versions.



(a)



(b)

Figure 2: Auditory spectra computed from (a) 17 and (b) 128 filters. The dashed lines show the direct filterbank auditory spectra (after linear interpolation), and the solid lines indicate an approximation by 14th order LP-derived cepstra.

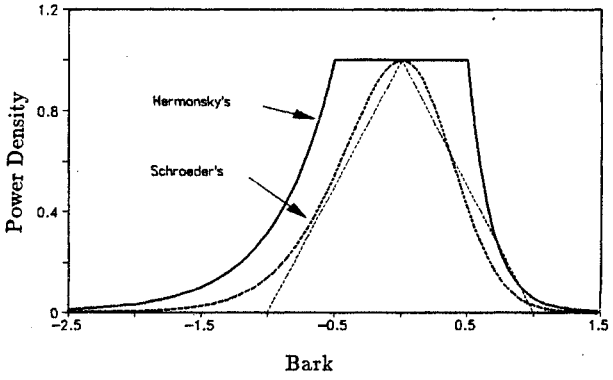


Figure 3: Three different masking patterns: the smooth pattern given by Schreoder [8], the triangle pattern used by Davis [9], and the 'flat-top' in Hermansky's PLP [3].

Such combinations are known to give higher error rates than time mixing the modelling and testing versions, but are more realistic.

The pre-processing of speech signal includes a 25.6 ms Hamming window, but excludes pre-emphasis. Speech frames have a 50% overlap. The PLP analysis uses 14th order all-pole approximation, shown to give good results in [1]. The resultant feature vectors are represented in cepstral coefficient form, and the root-power-sum (RPS) distance measure, approximating spectral slope differences, is employed for the similarity measurement. The RPS leads to a spectral slope interpretation, and has also been shown by Xu *et al* to give very similar results to the inverse variance cepstral weighting. Hunt very recently postulated [12] that the good performance of the RPS might just be the fortuitous similarity between the index weighting of RPS and the inverse variance.

4. Experiments and Results

Throughout the paper, experimental results are quoted as an average of 1,000 single-digit testings.

4.1 Effect of Auditory Filter Number

Here, the effect of filter number on recognition performance is examined. Figure 4 shows the speaker identification error rate versus the filter number, i.e. 8, 16, 32 64, and 128 filters, when using a VQ codebook of size 32. It is seen that the error rate decreases sharply when increasing the number of filters from 8 to 16, with smaller improvements upto 64 filters. The overall improvement is from 11.7% to 2.7%. For comparison, the equivalent result when using the original formulation of 17 filters is 4.7(±1.10)%, and is shown by the horizontal line.

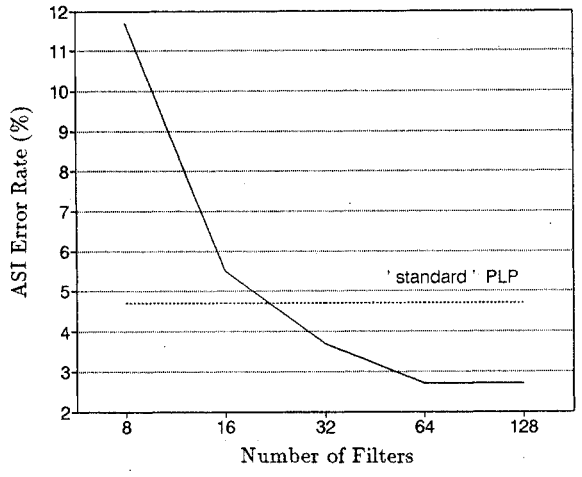


Figure 4: The speaker identification error rate vs. the number of filters. The horizontal line indicates the performance given by original PLP using 17 filters.

4.2 Effect of Masking Patterns

Table 1 gives identification error rates associated with different critical band patterns. It should be noted that 128 filters are used to compute the auditory spectra. However, 3 different masking patterns are employed, as described in Section 3. The term 'non-masking' here means that the operation of the critical band masking is not used in the transform.

Masking pattern	Error rate (90% c.i.)	VQ level
Non-masking	3.8% (± 0.99%)	32
Hermansky's pattern	2.7% (± 0.84%)	
Schroeder's pattern	3.0% (± 0.88%)	
Triangle pattern	2.9% (± 0.87%)	
Non-masking	2.8% (± 0.85%)	64
Hermansky's pattern	1.8% (± 0.69%)	
Schroeder's pattern	2.0% (± 0.72%)	
Triangle pattern	2.3% (± 0.78%)	

Table 1: Speaker identification error rates for various critical band masking conditions.

The results show that the worst performance comes from the case of non-masking, with error rates of 3.8% and 2.8% for codebook sizes 32 and 64 respectively; the best scores are achieved using critical band masking proposed by Hermansky, with error rates are 2.7% and 1.8% for codebook sizes 32 and 64 respectively. It is worth noting that this is the widest of the masking patterns considered.

4.3 Effect of Perceptually-based Operations

The 4 separate factors included in PLP, as shown in Figure 1, are: (1) Hz to Bark frequency warping, (2) critical band masking, (3) equal-loudness conversion, and (4) Steven's power law. Here we examine the contribution of each of these to the identification performance, and again the number of the auditory filters used here is 128.

Table 2 summarises the results. Benchmarks for all-pole modelling without any perceptual weightings are 7.0% and 6.5% for codebooks of 32 and 64 respectively. The features here are the same as those from standard LPC-derived cepstra.

With linear-to-Bark frequency warping the identification error rates are reduced to 5.0% and 3.5%. It can be seen from the Table that performance is generally improved by adding more perceptually-based operations. The exception is the combination of 'frequency warping + critical band masking'. However, results in Table 1 indicate that, with the two remaining factors (equal-loudness and power law), the masking is beneficial.

Critical band		Intnsty⇒Loud		Error rate	
Freq warp	Masking	Eq-loud	Pwr-law	VQ 32	VQ 64
				7.0%	6.5%
√				5.0%	3.5%
√	√			9.6%	7.4%
√	√	√		5.0%	3.2%
√	√		√	3.2%	2.5%
√	√	√	√	2.7%	1.8%

Table 2: Speaker identification error rates for various combinations of perceptual factors.

5. Discussion and Conclusion

PLP analysis is accomplished in two phases: computing an auditory spectrum and the all-pole modelling of that auditory spectrum. The effect of a higher resolution, from using a higher number of bandpass filters to compute the auditory model, has been examined. The resultant smoother, more accurate spectra (Figure 2) encompass detail which proves to be important in speaker recognition. However, it is interesting to note that the smearing process inherent in the critical band operation, is also useful in *speaker* recognition. The identification results become worse when this component is removed, and the broadest of the smearing functions examined gives the best results.

Auditory spectrum from 128 filters, i.e. many more than 17 filters in original PLP, provides a much higher resolution, and the resultant finer detail, together with the relatively high analysis order of PLP-14 which is known to be beneficial in speaker recognition [1], is demonstrated here to give good results.

This finding provides an interesting contrast to Hermansky's hypothesis [11] in the context of *speech* recognition. He argues that phonetic information of speech can well be determined by the rough contour of auditory spectra generated by 17 auditory filters; and the 5th order PLP spectra, having two peaks, are consistent with some vowel perception theory described in [13]. On the other hand, our study on *speaker* recognition suggests the importance of more accurate representations of the auditory spectrum.

We also show that the full perceptually-based model, i.e. all four factors, gives the best results. Interstingly Gu [4] finds a similar result in his study of PLP applied to to speech recognition.

References

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

[2] Xu, L. and Mason, J. S., *Instantaneous and Transitional Perceptually-based Features in Speaker Identification*, Proc. Eurospeech, pp. 271-274, September 1989.

[3] Hermansky, H., Tsuga, K., Makino, S. and Wakita, H., *Perceptually Based Processing in Automatic Speech Recognition*, Proc. ICASSP, pp. 1971-1974, April 1985.

[4] Gu, Y. and Mason, J.S., *Perceptual-based features in ASR*, IEE Colloquium on Computer Speech Processing, London, 1988.

[5] Hermansky, H. and Junqua, J.C., *Optimization of Perceptually-Based Processing ASR Front-End*, Proc. ICASSP, pp.219-222, April 1988.

[6] Stevens. S. S., *Measurement of Loudness*, JASA, Vol.27, pp. 815-829, September 1955.

[7] Bladon, R. A. W. and Lindblom B., *Modeling the Judgement of Vowel Quality Differences*, JASA, Vol.69, pp. 1414-1422, May 1981.

[8] Schroeder, M. R., Atal, B. S. and Hall, J. L., *Objective Measure of Certain Speech Signal Degradation Based on Masking Properties of Human Auditory Perception*, in Frontiers of Speech Communication Research, edited by B. Lindblom and S. Öhman, Academic Press, New York, 1979.

[9] Davis, S. B. and Mermelstein, P., *Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Trans. ASSP, Vol. ASSP-28, pp. 357-366, August 1980.

[10] Soong, F. K., Rosenberg, A. E., Rabiner, L. R. and Juang, B. H., *A Vector Quantization Approach to Speaker Recognition*, Proc. ICASSP, pp. 387-390, April 1985.

[11] Hermansky, H., *Perceptual Linear Preductive (PLP) Analysis of Speech*, JASA, Vol. 87, 1990.

[12] Hunt, M. J., Richardson, S. M., Bateman, D. C., Piau, A., *An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination*, Proc. ICASSP, pp.881-884, May 1991.

[13] Fant, G. and Risberg, A., *Auditory Matching of Vowel with Two Formant Synthetic Sounds*, STL-QPRS 4, KTH, Stockholm, pp7-11.