IDIAP RESEARCH REPORT

# PLP$^2$

# Autoregressive modeling of auditory-like 2-D spectro-temporal patterns

Marios Athineos [a] Hynek Hermansky [b]
Daniel P.W. Ellis [a]

IDIAP–RR 04-60

November 2004

published in
Workshop on Statistical and Perceptual Audio Processing SAPA-04, Jeju, Korea, Oct 2004

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] LabROSA, Dept. of Electrical Engineering, Columbia University, New York, NY 10027, USA.
[b] IDIAP Research Institute, CH-1920 Martigny, Switzerland.

IDIAP Research Report 04-60

# PLP$^2$
# Autoregressive modeling of auditory-like 2-D spectro-temporal patterns

Marios Athineos     Hynek Hermansky     Daniel P.W. Ellis

November 2004

published in

Workshop on Statistical and Perceptual Audio Processing SAPA-04, Jeju, Korea, Oct 2004

**Abstract.** The temporal trajectories of the spectral energy in auditory critical bands over 250 ms segments are approximated by an all-pole model, the time-domain dual of conventional linear prediction. This quarter-second auditory spectro-temporal pattern is further smoothed by iterative alternation of spectral and temporal all-pole modeling. Just as Perceptual Linear Prediction (PLP) uses an autoregressive model in the frequency domain to estimate peaks in an auditory-like short-term spectral slice, PLP$^2$ uses all-pole modeling in both time and frequency domains to estimate peaks of a two-dimensional spectro-temporal pattern, motivated by considerations of the auditory system.

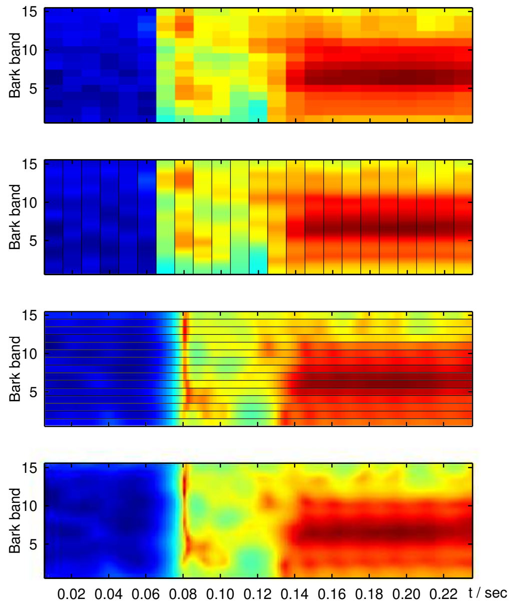


Figure 1: *Auditory STFT vs PLP vs Subband FDLP vs PLP*$^2$.

# 1 Introduction

Recent advances in understanding the physiology of the mammalian auditory cortex have revealed evidence for the existence of two-dimensional (time-frequency) cortical receptive fields – roughly equivalent, in engineering terms, to two-dimensional matched filters – that are sensitive to time- and frequency-localized stimuli. A typical receptive field can extend up to several hundred ms. [1, 2]. From this new perspective, a single slice of the short-term spectrum, as is commonly used as the basis for sound recognition systems, can hardly capture the information used by listeners; longer temporal spans of the signal seem necessary to facilitate cortical-like information extraction from acoustic signals. This strongly suggests the need for alternatives to the current short-time approach to speech and audio processing.

Speech recognition feature extraction techniques such as dynamic (delta) features, RASTA processing, or short-term cepstral mean removal, have been adopted as post-processing techniques that operate on sequences of the short-term feature vectors. Such techniques provide a "locally-global" view in which features to be used in classification are based upon a speech segment of about one syllable's length (see [3, 4] for more discussion and references).

The TRAP approach [5] is notable as an attempt to extract information from even longer segments of the acoustic signal. TRAPs are a rather extreme technique in which the trajectory of spectral energy in individual frequency bands are used for the initial classification (although arguable no more extreme than the attempts to classify sounds based on individual short-time spectra). But since the observed mammalian cortical receptive fields are two-dimensional, we infer that hearing is quite capable of integrating evidence both from larger time spans than the 10-30 ms short-term analysis window typically used, as well as from larger frequency range than a single critical band. Supporting this notion, recent work shows benefits from considering more than one temporal trajectory to obtain evidence for speech sound classification [6].

We believe that to be consistent with cortical physiology, a technique for efficient modeling of a two-dimensional auditory-like representations of acoustic signals is of interest and may prove useful in sound recognition applications. Such a technique, based on iterative sequential all-pole modeling in time and in frequency domains, is proposed and discussed in this paper.

## 2 Auditory-like spectro-temporal patterns

Replacing spectral slices of the short-term spectrum by a spectro-temporal pattern has previously been accomplished by stacking short-time feature vectors from neighboring frames to form a longer vector. This multi-frame input can be used directly for classification by multi-layer perceptrons (MLP) [7], or combined with linear discriminant analysis [8] or with MLP classifiers [9] to yield features for subsequent HMM recognition.

For the frequency axis, the underlying nonuniform critical-band representation is well justified, based on extensive physiological and perceptual results. In the time dimension, we will initially retain a linear scale (even though this may not be the only choice). In considering the appropriate temporal length of our new pattern, we note that although many earlier approaches went up to about 100 ms, the pioneering work of Fanty and Cole use the data from as much as 300 ms time spans in recognition of spoken alphabet[10], and the original TRAPs were up to 1 s long. In attempts to minimize the resulting processing delay, Sharma et al. report that reducing TRAP lengths to 400 ms results in only a minimal loss of performance [11], and on a phoneme recognition task 300 ms TRAPs were reported optimal [12].

Time spans of around 200-300 ms are well justified by many psychoacoustic phenomena that operate on such timescales (see [4] for the review), with the forward masking "critical interval" (the time-domain counterpart of the critical band in the simultaneous frequency masking) being especially relevant. Further, very recent observations from mammalian auditory cortex physiology indicate dominant temporal components around this time scale [13]. We have therefore settled on 250 ms for the current presentation.

## 3 Subband frequency-domain linear prediction (FDLP)

Just as a squared Hilbert envelope (the squared-magnitude of the analytic signal) represents the total instantaneous energy in a signal, the squared Hilbert envelopes of sub-band signals are a measure of the instantaneous energy in the corresponding sub-bands. Deriving these Hilbert envelopes would normally involve either using a Hilbert operator in the time domain (made difficult in practice because of its doubly-infinite impulse response), or the use of two Fourier transforms with modifications to the intermediate spectrum.

An interesting and practical alternative is to find an all-pole approximation of the Hilbert envelope by computing a linear predictor for the positive half of the Fourier transform of an even-symmetrized input signal – equivalent to computing the predictor from the cosine transform of the signal. Such Frequency Domain Linear Prediction (FDLP) is the frequency-domain dual of the well-known time-domain linear prediction (TDLP) [14, 15]. In the same way that TDLP fits the power spectrum of an all-pole model to the power spectrum of a signal, FDLP fits a "power spectrum" of an all-pole model (in this case in the time domain) to the squared Hilbert envelope of the input signal. To obtain such a model for a specific sub-band, one simply basis the prediction only on the corresponding range of coefficients from the original Fourier transform.

When we wish to summarize temporal dynamics, rather than capturing every nuance of the temporal envelope, the all-pole approximation to the temporal trajectory offers parametric control over the degree to which the Hilbert envelope is smoothed (e.g. the number of peaks in the smoothed envelope cannot exceed half the order of the model). Moreover, the fit can be adjusted by applying the transform techniques introduced in [16].

# 4 Auditory-like spectro-temporal patterns from subband FDLP

Having a technique for estimating temporal envelopes in individual frequency bands of the original signal permits the construction of an spectrogram-like signal representation. Just as a typical spectrogram is constructed by appending individual short-term spectral vectors alongside each other, a similar representation can be constructed by vertical stacking of the temporal vectors approximating the individual sub-band Hilbert envelopes, recalling the outputs of the separate band-pass filters used to construct the original, analog Spectrograph [17].

This is demonstrated in figure 1. The top panel shows the time-frequency pattern obtained by short-term Fourier transform analysis and Bark scale energy binning to 15 critical bands, which is the way the short-term critical-band spectrum is derived in PLP feature extraction. The second panel shows the result of PLP smoothing, with each 15-point vertical spectral slice now smooth and continuous as a result of being fit with an LP model. The third panel is based on a series 24-pole FDLP models, one for each Bark band, to give estimates of the 15 subband squared Hilbert envelopes. As with PLP, cube-root compression is applied here to the sub-band Hilbert envelope prior to computing the all-pole model of the temporal trajectory. The similarity of all these patterns is evident, but there are also some important differences: Whereas the binned, short-time spectrogram is 'blocky' in both time and frequency, the PLP model gives a smooth, continuous spectral profile at each time step. Conversely, the temporal evolution of the spectral energy in each sub-band is much smoother in the all-pole FDLP representation, constrained by the implicit properties of the temporal all-pole model.

# 5 PLP$^2$

In PLP, an auditory-like critical-band spectrum, obtained as the weighted summation of the short-term Fourier spectrum followed by cube-root amplitude compression, is approximated by an all-pole model in a manner similar to the way that conventional LP techniques approximate the linear-frequency short-term power spectrum of a signal [18]. Subband FDLP offers an alternative way to estimate the energy in each critical band as a function of time, raising the possibility of replacing the short-term critical band spectrum in PLP with this new estimate.

In doing so, a new representation of the critical-band time-frequency plane is obtained. However, comparing this new representation to the subband FDLP spectro-temporal pattern (constrained by the all-pole model along the temporal axis), the all-pole constraint is now along the spectral dimension of the pattern.

Nothing prevents us repeating the processing along the temporal dimension of the new representation to again enforce the all-pole constraints along the time axis. And the outcome of this step can be subject to another stage of all-pole modeling on the spectral axis; this alternation can be iterated until the difference between successive representations is negligible. Convergence of this process to a solution that approximates the peaks in the underlying spectro-temporal pattern has not been yet proven analytically, but our experiments so far support it.

At the end of the process, we have a two-dimensional spectro-temporal auditory-motivated pattern that is constrained by all-pole models along both the time and frequency axes. We therefore call this model *Perceptual Linear Prediction Squared (PLP$^2$)*. The "P" part (perceptual constraints) comes from the use of a critical-band frequency axis and from the use of a 250 ms critical-time-span interval; the "LP" part indicates the use of all-pole modeling, and the "squared"" part comes from the use of all-pole models along both the time and frequency axes.

# 6 Implementation Details

Taking the DCT of a 250 ms speech segment (equivalent to the Fourier transform of the related 500 ms even-symmetric signal) at a sampling rate of 8 kHz generates 2000 unique values in the frequency domain. We divide these into 15 bands with overlapping Gaussian windows whose widths and spacing

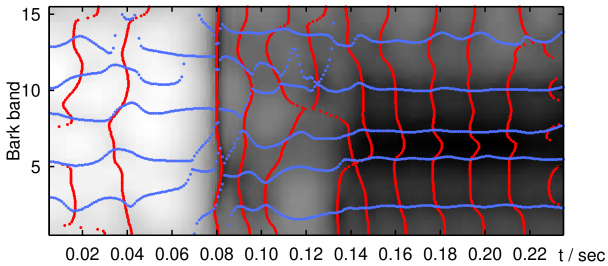


Figure 2: *$PLP^2$ pole locations. The red points (which form vertical lines) are the poles for each of the FDLP temporal envelope estimates, and blue points (creating horizontal lines) show the poles for each of the spectral estimates from the conventional PLP stage.*

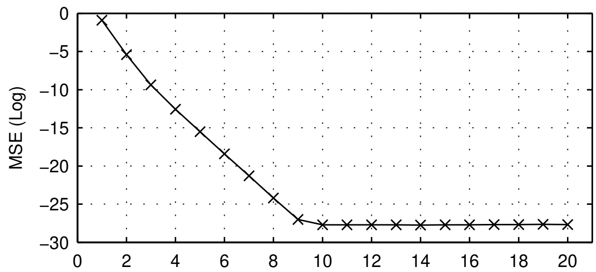


Figure 3: *Mean-squared differences between the 28,800-point log-magnitude surfaces obtained in successive iterations of the $PLP^2$ approximation.*

select frequency regions of approximately one Bark, and apply 24th order FDLP separately on each of the 15 bands such that each predictor approximates the squared Hilbert envelope of the corresponding sub-band. We compute the critical-band time-frequency pattern within the 250 ms time span by sampling each all-pole envelope at 240 points (i.e. every 1.04 ms) and stack the temporal trajectories vertically. This gives a 2-dimensional array amounting to a spectrogram, but constructed row-by-row, rather than column-by-column as in conventional short-term analysis. This time-frequency pattern is the starting point for further processing.

Next, 240 12th-order time-domain LP (TDLP) models are computed to model the spectra constituted by the 15 amplitude values in a vertical slice from the pattern at at each of the 240 temporal sample points. The spectral envelopes of these models are each sampled at 120 points (i.e. every 0.125 Bark) and stacked next to each other to form a new $240 \times 120 = 28{,}800$ point spectro-temporal pattern. Now each horizontal slice of 240 points is modeled by the same process of mapping a compressed magnitude 'spectrum' to an autocorrelation and thence to an all-pole model, to yield 120 24th-order FDLP approximations to the temporal trajectories in the new fractional-Bark subbands. Sampling these models on the same 240 point grid gives the next iteration of the 28,800 point spectro-temporal pattern. The process then repeats and has been observed to converge after a certain number of iterations, where the number of iterations required for convergence appears to depend on the models orders as well as the compression factor in the all-pole modeling process. The mean-squared difference between the logarithmic surfaces of the successive spectro-temporal patterns as a function of the iteration number is shown in figure 3, which shows stabilization after 10 iterations in this example. (Although this plot shows that the differences between successive iterations do not decline all the way to zero, we believe that the residual changes in later iterations are immaterial; inspection of the time-frequency distribution of these differences reveals no significant structure.)

The final panel of figure 1 shows the results of the new $PLP^2$ compared with conventional PLP. The increased temporal resolution in comparison with the 10 ms sampled PLP (second panel) is very clear; the second important property of the PLP$^2$ surface, which is a little harder to see in the picture, is the increased spectral resolution in comparison with the 15 frequency values at each time for the basic FDLP model (third panel).

Further insight can be obtained by plotting the pole locations on the time frequency plane. In figure 2 the pole locations are superimposed on a grayscale version of the PLP$^2$ pattern presented on the 4th pane of figure 1. Red dots show the 12 FDLP poles for each of the 120 subband envelope estimates; due to the dense frequency sampling, the poles of adjacent bands are close in value, and the dots merge into near-vertical curves in the figure. Blue dots are the 6 TDLP poles at each of the 240 temporal sample points, and merge into near-horizontal lines. (Pure-real poles are not shown, so some frames show fewer than the maximum number of possible poles.)

The blue TDLP poles successfully track the smoothed formants in the $t = 0.14$ to 0.24 s region but they fail to capture the transient at around 0.08 s. The red FDLP poles, on the other hand, with their emphasis on temporal modeling, make an accurate description of this transient. As expected, neither TDLP or FDLP models track any energy peaks in the quiet region between 0 and 0.08 s. But, while the TDLP models for these temporal slices are obliged to place their poles somewhere in this region, the FDLP models are free to shift the majority of their poles into the later portion of the time window, between 0.08 and 0.25 s, where the bulk of the energy lies.

# 7 Preliminary findings

We are currently investigating the use of these features in automatic speech recognition (ASR). In order to find a reasonable point of departure, we have attempted to create features that are very similar to the conventional PLP features used in many ASR systems, yet which still incorporate the unique features of PLP$^2$. We are also obliged to moderate the complexity of the calculations to make the feature calculation feasible for the training set sizes used in current ASR problems.

Our reduced implementation starts with a 250 ms segment of speech, then divides its DCT into 15 Bark bands. Each band is fit with a 12th order FDLP polynomial, then the resulting smoothed temporal envelope is sampled on a 10 ms grid. The central-most spectral slices are then smoothed across frequency using the conventional PLP technique, but we do not perform any further iterations; instead, the cepstra resulting from this stage are taken as replacements for the conventional PLP features as input to the recognizer.

Thus far, these features have indeed shown performance very close to standard PLP features, achieving word error that differ by less than 2% relative. (We have tested two large-vocabulary tasks; in one case PLP$^2$ was better, and in one case worse.) Although small, these differences are statistically very significant, and when we combine the results from a PLP$^2$ system with conventional system outputs using simple word-level voting, we achieve a significant improvement in overall accuracy. Full details of these experiments are currently being prepared for publication.

Calculation of these features was about 20$\times$ slower than the conventional features. This comparison, however, is somewhat unfair since we are comparing an experimental, research implementation in Matlab to long-standing, highly-optimized C-code.

# 8 Discussion and conclusions

We have introduced a new modeling scheme to describe the time and frequency structure in short segments of sound of about a quarter of a second. Based on recent physiological results and psychoacoustic evidence, we believe that a representation of about this scale is likely involved in human auditory processing. The technique of all-pole (linear predictive) modeling, applied in both the time and frequency domains, allows us to smooth this representation to adaptively preserve the most significant peaks within this window in both dimensions. The convergence of this representation after a few iterations constitutes a novel and promising canonical description of the information in each window.

In this preliminary paper we have presented the basic idea and given a simple illustration. Our current work is to exploit this new description for practical tasks such as speech recognition or other information extraction applications. Techniques for reducing the smoothed energy surface to a lower-dimensional description appropriate for statistical classifiers include conventional basis decompositions such as Principal Component Analysis or two-dimensional DCTs. A second possibility, paralleling the representations proposed in [15], is to exploit the pole locations illustrated in figure 2 as a reduced, parametric description of the energy concentrations. For instance, recording the ‘crossing points’ of the nearly-continuous time and frequency pole trajectories could provide a highly compact description of the principal energy peaks in each 250 ms spectro-temporal window.

# 9 Acknowledgments

This work was supported by DARPA under the EARS Novel Approaches grant no. MDA972-02-1-0024, by the IM2 Swiss National Center for Competence in Research managed by Swiss National Science Foundation on behalf of Swiss authorities, and the European Community AMI and M4 grants. Our thanks go to three anonymous reviewers for their constructive comments.

# References

[1] S. Shamma, H. Versnel, and N. Kowalski, “Ripple analysis in ferret primary auditory cortex: I. response characteristics of single units to sinusoidally rippled spectra,” *Aud. Neurosci.*, vol. 1, 1995.

[2] D. Klein, D. Depireux, J. Simon, and S. Shamma, “Robust spectro-temporal reverse correlation for the auditory system: Optimizing stimulus design,” *J. Comput. Neurosci*, vol. 9, 2000.

[3] H. Hermansky, “Exploring temporal domain for robustness in speech recognition,” in *Proc. of 15th International Congress on Acoustics*, vol. II, Trondheim, Norway, June 1995.

[4] ——, “Should recognizers have ears?” *Speech Communication*, vol. 25, 1998.

[5] H. Hermansky and S. Sharma, “TRAPS - classifiers of temporal patterns,” in *Proc. ICSLP*, Sydney, Australia, 1998.

[6] P. Jain and H. Hermansky, “Beyond a single critical-band in TRAP based ASR,” in *Proc. Eurospeech*, Geneva, Switzerland, Nov 2003.

[7] S. Makino, T. Kawabata, and K. Kido, “Recognition of consonant based on the perceptron model,” in *Proc. ICASSP*, Boston, MA, 1983.

[8] P. Brown, “The acoustic-modeling problem in automatic speech recognition,” Ph.D. dissertation, Computer Science Department, Carnegie Mellon University, 1987.

[9] H. Hermansky, D. Ellis, and S. Sharma, “Connectionist feature extraction for conventional hmm systems,” in *Proc. ICASSP*, Istanbul, Turkey, 2000.

[10] M. Fanty and R. Cole, “Spoken letter recognition,” in *Advances in Neural Information Processing Systems 3*. Morgan Kaufmann Publishers, Inc., 1990.

[11] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, “Feature extraction using non-linear transformation for robust speech recognition on the AURORA data-base,” in *Proc. ICASSP*, Istanbul, Turkey, 2000.

[12] P. Schwartz, P. Matejka, and J. Cernocky, “Recognition of phoneme strings using TRAP technique,” in *Proc. Eurospeech*, Geneva, Switzerland, September 2003.

[13] D. Klein, 2003, personal communication.

[14] M. Athineos and D. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *Proc. ICASSP*, vol. 5, 2003, pp. 648–651.

[15] ——, "Frequency-domain linear prediction for temporal features," in *Proc. IEEE ASRU Workshop*, S. Thomas, US Virgin Islands, Dec 2003.

[16] H. Hermansky, H. Fujisaki, and Y. Sato, "Analysis and synthesis of speech based on spectral transform linear predictive method," in *Proc. ICASSP*, vol. 8, Apr 1983, pp. 777–780.

[17] R. Koenig, H. Dunn, and L. Lacey, "The sound spectrograph," *J. Acoust. Soc. Am.*, vol. 18, pp. 19–49, 1946.

[18] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87:4, April 1990.