IDIAP RESEARCH REPORT

LP-TRAP: Linear predictive temporal patterns

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

IDIAP–RR 04-59

November 2004

published in
International Conference on Spoken Language Processing ICSLP-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 04-59

LP-TRAP: Linear predictive temporal patterns

Marios Athineos Hynek Hermansky Daniel P.W. Ellis

November 2004

published in

International Conference on Spoken Language Processing ICSLP-04, Jeju, Korea, Oct 2004

Abstract. Autoregressive modeling is applied for approximating the temporal evolution of spectral density in critical-band-sized sub-bands of a segment of speech signal. The generalized autocorrelation linear predictive technique allows for a compromise between fitting the peaks and the troughs of the Hilbert envelope of the signal in the sub-band. The cosine transform coefficients of the approximated sub-band envelopes, computed recursively from the all-pole polynomials, are used as inputs to a TRAP-based speech recognition system and are shown to improve recognition accuracy.

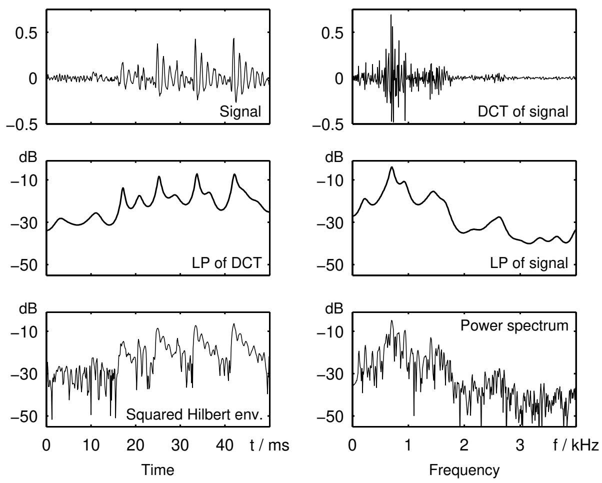

Figure 1: *The two dual forms of linear prediction.* On the left column (time) we plot 50 ms of a speech signal, the FDLP all-pole fit and corresponding squared Hilbert envelope. On the right column (frequency) we display the DCT of the same signal, the conventional time-domain LP all-pole fit and the corresponding power spectrum. Both models use 28 poles.

1 Introduction

The speech signal is not stationary but carries information it its dynamics. To enable the use of processing techniques that assume signal stationarity, short segments of the signal (10-30 ms) are used to derive short-term features for pattern classification in automatic speech recognition (ASR). The signal dynamics are then represented by a sequence of the short-term feature vectors with each vector representing a sample from the actual underlying dynamic process, in a manner similar to the way motion in movies is represented by a sequence of static shots. The issues of windowing, time-frequency resolution compromises, proper sampling of the short term representation, emulating the unequal frequency resolution of hearing, etc., are typically addressed in an ad hoc manner.

To parameterize short-term spectral envelopes, a rich inventory of techniques has evolved. Among them, the linear predictive (LP) regression model offers a convenient way of approximating the underlying short-term power spectrum of speech in terms of its dominant peaks. There are a number of alternative ways to describe the autoregressive model. In particular, it can be computed directly from the power spectrum of the signal [1] and modifications of the power spectrum prior to LP modeling can be used to advantage (e.g. [2]).

Some recent work has looked into better ways to exploit local speech dynamics in speech recognizers. It has been shown a number of times (see e.g. [3]) that the important linguistic information lies in the 1-16 Hz modulation frequency range. In order to use information in the modulation spectrum at those frequencies, one has to look at signals over relatively long time scales. Therefore, in the TRAP-TANDEM technique [4, 5] temporal trajectories of spectral densities in individual critical bands over windows as long as 1 sec are used as features for pattern classification. However, the temporal dynamics are still described by a sequence of short-term features; it would be interesting and elegant to model these trajectories more directly, and frequency-domain linear prediction (FDLP) [6, 7] is a technique that allows for that. The technique was originally applied to very short segments of speech to emulate some effects of temporal masking in hearing [6], and later used for extracting temporal features from larger segments of the speech signal for ASR [7].

The current paper presents a further evolution of the FDLP technique and its application in modeling long Hilbert envelopes of a signal in critical bands for TRAP-TANDEM based ASR. Since we use linear prediction polynomials in order to parameterize each TRAP, we call this model linear predictive temporal patterns or *LP-TRAP*.

In the next section we motivative our model and present the building blocks for this novel parameterization. In section 3 we describe the TRAP-TANDEM setup and evaluate it using features from the new model. Lastly in section 4 we discuss the results and present the conclusions.

2 Parameterizing the temporal envelopes

Almost all current ASR systems represent temporal information by a sequence of feature vectors from short-time Fourier analysis. To emulate the non-equal spectral resolution of human hearing, the frequency resolution of the Fourier spectra is typically modified by Mel or Bark frequency energy grouping to a small number of sub-bands. The temporal resolution of such a representation is the same at all frequencies and is given by the applied analysis window (typically around 25 ms) which acts as a lowpass filter on the temporal trajectories.

An alternative way of deriving the short-term speech representation (applied e.g. in the original Spectrograph) could be using the rectified output from a bank of band-pass filters. Spectral resolution could be then controlled by band-pass filter design, and the temporal resolution could be different at different frequencies depending on the lengths of impulse responses of the individual filters.

2.1 Frequency-domain linear prediction (FDLP)

There is a third, perhaps less obvious way of deriving the short-term spectral representation. Just as a squared Hilbert envelope (squared magnitude of the analytic signal) represents instantaneous energy in a signal, the squared Hilbert envelopes of the sub-band signals are a measure of the instantaneous energy in the corresponding sub-bands. To get the Hilbert envelope would normally involve the use of either the Hilbert operator in the time domain (whose infinite impulse response presents some practical issues) or the double use of the Fourier transform with modifications to the intermediate spectrum [8].

An interesting and practical alternative is to get the all-pole approximation of the Hilbert envelope by computing a linear predictor on 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). In the same way TDLP fits an all-pole model to the power spectrum of a signal, FDLP fits an all-pole model to the squared Hilbert envelope. Since the cosine transform represents the Fourier transform of the even-symmetrized time signal, the "spectrum" of the resulting predictor gives an approximation to the Hilbert envelope of the signal (in the same way as the spectrum of the predictor derived in the time domain is an approximation of the power spectrum of the signal). To get an all-pole approximation of the Hilbert envelope for a specific sub-band, the prediction needs to be derived only from the appropriate part of the cosine-transformed signal.

The parametric all-pole description of the temporal trajectory offers control over the degree of smoothing of the Hilbert envelope. Moreover, the fit can be controlled by applying transform techniques introduced in [2]. The easily-computable "cepstrum" of the time-domain all-pole model represents in this case the spectrum of the logarithmically-compressed temporal envelope and is related to the cosine transform of the original TRAP which has been found useful in ASR [9].

The duality between the power spectrum and the squared Hilbert envelope is essential to the understanding of FDLP. Figure 1 illustrates these two dual forms of linear prediction. On the upper left pane we display 50 ms of speech that we want to model using the two dual forms of linear prediction. Conventional linear prediction (TDLP in our terminology) approximates the power spectrum of the signal, as shown in the middle panel on the right (frequency) side of the figure, which is the TDLP of the top-left (time) signal. The full Fourier power spectrum to which this is an approximation is plotted directly below, in the bottom-right pane.

FDLP on the other hand operates on the DCT of the signal (top right pane) and results in an LP model describing the *temporal* envelope, shown in the middle left (time) panel. Directly below it is

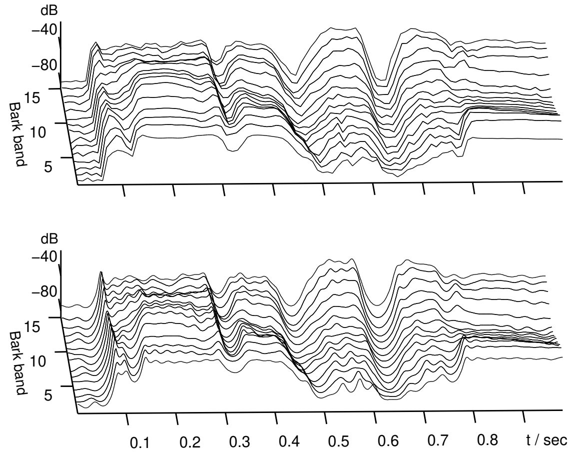

Figure 2: *Auditory spectrogram versus all-pole trajectories.* The first pane displays the short-time auditory spectrogram whereas the second pane shows the FDLP-approximated Hilbert envelopes using 80 poles per band.

plotted the corresponding squared Hilbert envelope that is being estimated. Each column provides three alternative representations of each domain. Whereas TDLP exploits the spectral structure of the signal to construct an efficient predictor of the temporal signal, FDLP exploits the temporal structure of the signal to predict spectral values.

The concept of FDLP was to our knowledge first introduced by Herre [6] as a method for efficient coding of transients in transform coders. Kumaresan has independently discovered and extensively worked on FDLP, a method which he calls linear prediction in the spectral domain or LPSD [10].

2.2 Linear predictive temporal patterns (LP-TRAP)

In this paper we extend the FDLP model to speech segments up to 1 sec long. We seek here to summarize the temporal dynamics rather than capture every single nuance of the temporal envelope. Taking the DCT of a 1 sec speech segment at 8 kHz sampling rate generates 8000 frequency domain samples. Instead of fitting one predictor on the whole frequency series as we do in figure 1, we first apply 15 Bark-spaced overlapping Gaussian windows. We then apply FDLP separately on each of the 15 bands. Each predictor then approximates the squared Hilbert envelope of the corresponding sub-band. This is the "sub-band FDLP" introduced in [7] but here we extend the time window to even longer speech segments and use overlapping windows.

We compute the auditory spectrogram over the 1 sec windows by stacking the individual temporal trajectories (rather than by stacking the individual frequency vectors as done in the conventional short-term spectral analysis). This is demonstrated in figure 2. The top panel shows the auditory spectrogram obtained by short-term Fourier transform analysis and Bark scale energy binning to 15 critical bands. In the second panel we fit fifteen 80-pole FDLP models, one for each Bark band, and display the 15 estimates of the squared Hilbert envelopes.

2.3 Spectral transform linear prediction (STLP)

Spectral transform linear prediction was introduced as method to adjust the relative fit of the conventional (TDLP) predictor to the peaks and dips of the speech spectrum [2]. By raising the power spectrum to an arbitrary power, the *compression factor*, one can adjust the peak-hugging property of linear prediction. STLP is an integral part of the well-known perceptual linear prediction (PLP)

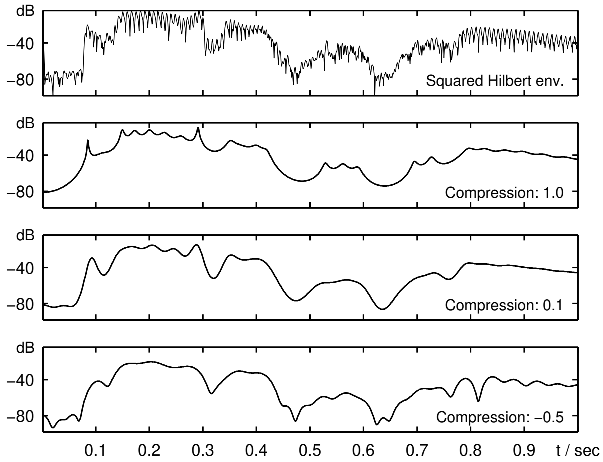

Figure 3: *The effect of the compression factor.* The top pane shows the squared Hilbert envelope of the sixth Bark band of figure 2. In the second pane FDLP using no compression fits the peaks. Using moderate compression in the third pane FDLP achieves a better fit to dips in the envelope. The fourth model fits a compressed version of the inverse spectrum, thereby fitting the dips in preference to the peaks. The number of poles is 50 in all three cases.

technique; the cube-root compression of the power spectrum in PLP prior to prediction is an instance of STLP with compression factor 1/3.

We borrow this idea and we apply it on the Hilbert envelopes of the Bark sub-bands instead of the power spectra. In some sense this method could now be dubbed temporal transform linear prediction (TTLP). In figure 3 we demonstrate the effect of the compression factor. We take the sixth Bark band from figure 2 and this time we keep the number of poles fixed to 50. On the top pane we display the logarithm of corresponding squared Hilbert envelope. On the second pane we plot the FDLP sub-band envelope with compression factor 1.0 which amounts to no compression. The all-pole model fits the peaks much better than the dips. A moderate compression of 0.1 still gives a better model of the peaks but of a greatly compressed Hilbert envelope. This time the dips are much better modeled. Lastly for compression of -0.5 the all-pole model fits a compressed version of the *inverse* Hilbert envelope. The dips are now accurately modeled. Spectral expansion using compression factors greater than 1 is also possible but it may result in ill-conditioned solutions due to the extreme sharpness of the peaks; we do not consider spectral expansions here.

2.4 Feature extraction

In [7] we identified two broad approaches to extracting features from the FDLP polynomials. The first derives features directly from the poles of the polynomial, since the angle of the pole corresponds to very accurate timing information and the magnitude is a measure of the energy of the signal. In [7] we showed benefits from using pole sharpness as a measure of the local dynamics of the temporal envelope, especially in the recognition of stops.

The second approach seeks to derive features directly from the temporal envelopes. But instead of sampling the DFTs of the envelopes and taking the DCT, we use the cepstral recursion as an elegant and computationally efficient way to convert our all-pole models of the temporal trajectories into modulation spectra. The recursion allows for the calculation of an arbitrary number of cepstral coefficients.

In a conventional TRAP-TANDEM setup each sub-band is fed to a TRAP which describes the phoneme in the center of the pattern [4], as estimated by a Multi-Layer Perceptron (MLP) trained on labeled data. The TRAP outputs from all sub-bands are combined together with a "merger" MLP, generating further phoneme detection estimates to be fed to a GMM-HMM sequence recognizer, operating at a 10 ms frame rate [11, 9]. This means that in the LP-TRAP we need to calculate FDLP polynomials from 1 sec DCTs every 10 ms. Our current feature extraction still operates in real-time but we believe that some computational short-cuts might exist for the calculation of the FDLP envelopes in a more efficient manner.

3 Evaluation

The TRAP-TANDEM approach combines the extraction of phoneme information from long temporal windows in narrow frequency regions [4, 5] with a learned discriminative feature transformation feeding into a conventional GMM-HMM recognizer [11]. The TRAP-TANDEM recognizer used in this work consisted of sub-band TRAP MLPs trained on OGI Stories, followed by a TANDEM MLP and HTK-based GMM-HMM recognizer trained on OGI Numbers95. Testing was performed on the test part of OGI Numbers95.

Our baseline system uses a standard TRAP front end. Temporal trajectories of 1 sec duration are derived from short-time Fourier transform analysis and Bark binning to 15 bands. This is displayed on the top pane of figure 2. The temporal trajectories are decorrelated via a DCT (along time) and truncated to 50 points before being fed as the input to the per-band TRAP MLPs, thence to the merger MLP, thence to the GMM-HMM recognizer. The word error rate for this baseline system is 5.9%.

In our experiments we substituted the standard analysis frontend with FDLP to create LP-TRAPs. We experimented with parameter sets to obtain autoregressive envelopes that accurately approximated the auditory spectrum. While keeping the number of Bark bands fixed to 15 for the 8 kHz sampling rate of our database we evaluated the effect of different model orders and different compression factors.

To remove the effect of different-sized input layers on the LP-TRAP MLPs we truncated temporal DCT representation of the input trajectories to 50 coefficients, independent of the number of poles or compression factor. Note that because the LP representation of the temporal envelope is not intrinsically bandlimited, there is no limit on the order of the cepstra that can be derived from the LP representation. Lastly we excluded C_0 (the energy term) from each DCT; pilot experiments that included C_0 gave us worse performance.

3.1 LP-TRAP results

First we fix the number of poles (at 50) and sweep the compression factor. Recognition results are presented in the left pane of figure 4 and in table 1. Negative compression factors, corresponding to models that concentrate on the dips in the Hilbert envelopes, gave worse performance. A small positive compression factor of 0.1, corresponding to highly compressed envelope peaks, resulted in the best performance.

Cmpr	-1.0	-0.5	-0.1	0.1	0.5	1.0
WER (%)	8.0	6.4	5.8	5.3	5.6	5.8

Table 1: *LP-TRAPs word error rate when varying the envelope compression factor. Each temporal envelope is fit with 50 poles.*

Next we fix the compression factor to 0.1 and vary the number of poles. The results are on the right pane of figure 4 and in table 2. Extreme smoothing of the envelopes resulting from very low model orders hurts performance. However, using between 50 and 80 poles per band (for a 'pole rate' of around 0.1 pole/ms) gives good performance. We conclude that these models are sufficient to capture the necessary temporal dynamics for ASR.

# Poles	15	20	30	40	50	60	80
WER (%)	7.1	6.0	5.6	5.7	5.3	5.4	5.3

Table 2: *Error rates as a function of temporal model complexity as determined by the number of poles per sub-band. Compression is fixed at 0.1.*

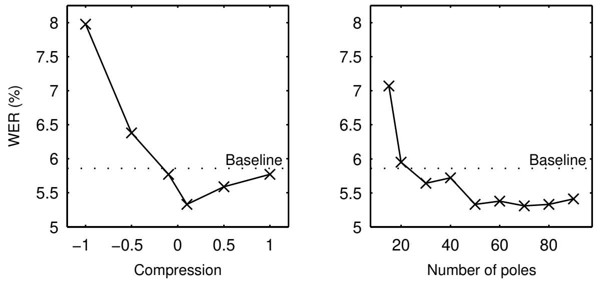

Figure 4: *LP-TRAP results.* Word error rate as a function of compression factor and number of poles. The compression variants all use 50 poles and the pole rate variants use a compression factor of 0.1. The dashed line represents the baseline.

With this combination of compression and pole rate, the LP-TRAP features consistently performed in the vicinity of 5.3% WER which represents a 10% relative improvement over the 5.9% baseline. Increasing the number of poles up to 200, close to the limit imposed by the length of the narrowest Bark band, reduces the accuracy to 5.7%. Subsequent experiments with 0.5 sec LP-TRAP showed similar performance as the 1 sec LP-TRAP discussed here.

We have also performed initial experiments with alternative features including direct use of the FDLP coefficients as well as alternative parameterizations such as the line spectral pairs (LSPs). So far we have not found any of these to offer improvements in recognition accuracy.

4 Discussion and conclusions

All-pole approximations of the Hilbert envelopes in critical-band sub-bands are an elegant and interesting alternative to the *ad hoc* weighted averaging of the short-term Fourier spectrum used in conventional ASR. Issues of the short-term windowing are avoided and numerous new possibilities appear, particularly given the rich literature of techniques and variants associated with linear prediction. So far, we have explored only a small fraction of these directions.

In the current work, this technique is used to derive features for the TRAP-TANDEM system, where, after some optimization, it yields about 10% relative improvement in error rate on our standard OGI Numbers task. We are confident that future investigations will reveal many more interesting and valuable applications in speech processing and recognition.

5 Acknowledgments

We were supported by DARPA under the EARS Novel Approaches grant no. MDA972-02-1-0024. This work took place while the first author was a visiting researcher at IDIAP as part of the collaboration within the EARS-NA team. We would like to thank Frantisek Grézl and Petr Fousek for their help with the TRAP-TANDEM baseline recognizer.

References

[1] J. Makhoul, "Spectral linear prediction: Properties and applications," in *Trans. ASSP*, vol. 23:3, Jun 1975, pp. 283–296.

[2] 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.

[3] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, 1994.

[4] H. Hermansky and S. Sharma, "Traps - classifiers of temporal patterns," in *Proc. ICSLP*, Sydney, Australia, 1998.

[5] H. Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech," in *Proc. IEEE ASRU*, St. Thomas, USVI, 2003.

[6] J. Herre and J. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," in *Proc. 101st AES Conv.*, Nov 1996.

[7] M. Athineos and D. Ellis, "Frequency-domain linear prediction for temporal features," in *Proc. IEEE ASRU Workshop*, S. Thomas, US Virgin Islands, Dec 2003.

[8] J. L. Marple, "Computing the discrete-time 'analytic' signal via fft," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 47, Sep 1999.

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

[10] R. Kumaresan and A. Rao, "Model based approach to envelope and positive instantaneous frequency estimation of signal with speech applications," *The Journal of the Acoustical Society of America*, vol. 105, 1999.

[11] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, Istanbul, 2000.