FREQUENCY-DOMAIN LINEAR PREDICTION FOR TEMPORAL FEATURES

Marios Athineos and Daniel P.W. Ellis

LabROSA, Dept. of Electrical Engineering, Columbia University, New York, NY 10027
{marios,dpwe}@ee.columbia.edu

ABSTRACT

Current speech recognition systems uniformly employ short-time spectral analysis, usually over windows of 10-30 ms, as the basis for their acoustic representations. Any detail below this timescale is lost, and even temporal structure above this level is usually only weakly represented in the form of deltas etc. We address this limitation by proposing a novel representation of the temporal envelope in different frequency bands by exploring the dual of conventional linear prediction (LPC) when applied in the transform domain. With this technique of frequency-domain linear prediction (FDLP), the 'poles' of the model describe temporal, rather than spectral, peaks. By using analysis windows on the order of hundreds of milliseconds, the procedure automatically decides how to distribute poles to best model the temporal structure within the window. While this approach offers many possibilities for novel speech features, we experiment with one particular form, an index describing the 'sharpness' of individual poles within a window, and show a large relative word error rate improvement from 4.97% to 3.81% in a recognizer trained on general conversational telephone speech and tested on a small-vocabulary spontaneous numbers task. We analyze this improvement in terms of the confusion matrices and suggest how the newly-modeled fine temporal structure may be helping.

1. INTRODUCTION

Contemporary analysis techniques for acoustic modeling in automatic speech recognition are spectrum-based. Spectral structures such as formants convey important linguistic information. Nevertheless this is only a partial representation of speech signals.

We believe that temporal structure in sub-10 ms transient segments contains important cues for both the perception of natural sounds [1] as well the understanding of stop bursts in speech. At the other extreme, the gross temporal distribution of acoustic energy in windows of up to 1 sec has proven to be a successful domain for the recognition of complete phonemes and the description of their dynamics [2].

Speech features based on short 10-30 ms time frames frequently incorporate temporal information in some form e.g. through delta features. Hermansky and Morgan showed that relative spectral (RASTA) post-processing of critical-band energy time trajectories using IIR bandpass filters increased robustness and removed inter-frame spectral analysis artifacts [3].

We consider systems that start by breaking the signal into time frames are treating time as a secondary dimension. We propose to alleviate this shortcoming with a new signal representation that treats time as the primary processing dimension, without the distortion of an intermediate frame rate.

In the next section we present a generic model that fulfills our goals by capturing temporal information adaptively. In section 3 we show how we can extract features from our model that can be used successfully in ASR. Those features are evaluated in section 4 by using a standard HTK testbed. In section 5 we discuss our findings and present our conclusions.

2. TIME-ADAPTIVE MODEL

The goal of our model is a parametric description of the temporal dynamics of speech. We want adaptively to capture fine temporal nuances with millisecond accuracy while at the same time summarize the signal's gross temporal evolution in timescales of 500 ms or more.

The part of the model responsible for the time-adaptive behavior is frequency-domain linear prediction (FDLP). The discrete cosine transform (DCT) provides a frequency-domain representation that is real-valued on which we apply linear prediction. By duality with the way we model spectral envelopes using linear prediction in time, we can model temporal envelopes using linear prediction in frequency.

This model has several advantages. Fine time-adaptive accuracy can be used to pin-point important moments in time such as those associated with transient events like stop bursts. At the same time, the long-timescale summarization power of the temporal envelopes gives us the ability to train recognizers on complete linguistic units lasting longer than 10 ms and possibly even learning acoustically-feasible

Submitted to *ASRU-2003, Nov 30–Dec 04, 2003, St. Thomas USVI*

phoneme sequences – a step that has been traditionally left for the domain of sequential state models.

Conceptually, our method is comprised of two parts. The DCT is applied on long-time frames, then linear prediction is carried out on the output of the DCT. A discussion of some key properties of the DCT will provide insight and build an intuition for the method.

2.1. Discrete Cosine Transform (DCT)

The DCT enjoys wide use in the speech community in applications such as transform coding and, most prominently, as an approximation to Karhunen-Loeve Transform (KLT) decorrelation. As part of the cepstral transformation, the DCT appears as a post-processing step in virtually all feature extractors for ASR. Formally, the forward DCT of an N point real sequence $x[n]$ can be defined as

$$X_{DCT}[k] = a[k] \sum_{n=0}^{N-1} x[n] \cos\left(\frac{(2n+1)\,\pi k}{2N}\right) \quad k = 0, 1, \ldots, N-1 \tag{1}$$

where:

$$a[k] = \begin{cases} 1 & k = 0 \\ \sqrt{2} & k = 1, 2, \ldots, N-1 \end{cases} \tag{2}$$

A less well-known property of the DCT is its ability to approximate the envelope of the discrete Fourier transform (DFT). Denoting as $X_{DFT}[k]$ the DFT of a length $2N$ zero-padded version of $x[n]$, it has been shown [4] that the envelope of the DCT is bounded by the envelope of the zero-padded DFT and in fact they are exactly related by

$$X_{DCT}[k] = a[k]\,|X_{DFT}[k]| \cos\left(\theta[k] - \frac{\pi k}{2N}\right) \quad k = 0, 1, \ldots, N-1 \tag{3}$$

where $|X_{DFT}[k]|$ and $\theta[k]$ are the magnitude and phase of the zero-padded DFT respectively.

The above mentioned property helps us understand figure 1. On the left we see the spectrogram of a 2 sec speech sample and on the right we see the spectrogram of a DCT transform of the whole sample (treating the DCT output sequence as a sequence in time). One can notice that the DCT spectrogram looks like a mirror image of the regular spectrogram over the axis $time = frequency$. It is important to realize that the two figures are not exact mirrors, due to the cosine modulating term in equation 3. (To listen to these DCT waveforms, please visit `http://www.ee.columbia.edu/~marios/projects/dct_listening/`.) Having introduced DCT and its relevant properties we are now ready to proceed with the discussion of FDLP.

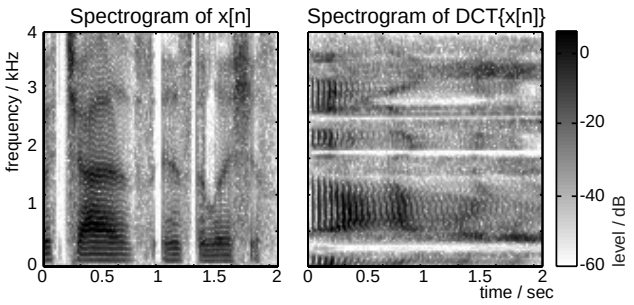

Fig. 1. Time-frequency analysis of a 2 sec speech sample. On the left the regular spectrogram and on the right the spectrogram of the 2 sec DCT.

2.2. Frequency-domain linear prediction (FDLP)

As mentioned above, FDLP is the part of the model that provides the time adaptive behavior we want to achieve. FDLP is the frequency-domain dual of the well-known time-domain linear prediction (TDLP). In the same way that TDLP estimates the power spectrum, FDLP estimates the temporal envelope of the signal, specifically the square of its Hilbert envelope,

$$e(t) = \mathcal{F}^{-1}\left\{\int \tilde{X}(\zeta) \cdot \tilde{X}(\zeta - f) d\zeta\right\} \tag{4}$$

i.e. the inverse Fourier transform of the autocorrelation of the single sided (positive frequency) spectrum $\tilde{X}(f)$. We use the autocorrelation of the spectral coefficients to predict the temporal envelope of the signal.

Although FDLP is applied in the spectral domain it still has the same peak-hugging properties as TDLP, only the peaks are in the temporal, not spectral, envelope. FDLP model order selection is guided by the temporal structure of speech the same way TDLP model order is dictated by formant structure. The difference here is that we can choose an arbitrarily long temporal window on which to apply FDLP so we need to define a 'pole rate', i.e. the number of model poles used per time unit.

To illustrate, figure 2 shows a 256 ms long speech segment at 8 kHz sampling rate. After taking the 2048 point DCT of the whole sample we fit a single FDLP polynomial and extract the temporal envelope of the segment. Notice the tradeoffs involved in model order selection: In the case of 10 poles the envelope is too smooth and gives only a loose approximation. In the case of 40 poles, the envelope is starting to fit the pitch pulses—something we wish to avoid (for English-language ASR applications anyway). The case of 20 poles strikes a good balance, capturing both the gross variation as well as the stop burst transients in the beginning of the sample. This combination defines a pole rate of 20/256 ms $\approx$ 0.1 pole/ms. Note, however, that the poles are

distributed *adaptively* within the 256 ms window. This flexibility in deploying its modeling power is a key strength of the model.

A second example is displayed on figure 3. In this case we use the same 256 ms long sample but this time we apply FDLP on 4 logarithmically-split octave bands, namely 0-0.5, 0.5-1, 1-2 and 2-4 kHz. We use the same pole rate of 20 poles per 256 ms for each band as in our previous example. Notice that the high frequency band is resolving the transient while the low frequency band is capturing the gross spectral variation. We call this method "subband FDLP".

FDLP is in fact a repurposing of a system first introduced by Herre and Johnston [5]. Dubbed temporal noise shaping (TNS), its principal application was in the elimination of pre-echo artifacts associated with transients in perceptual audio coders. Using D*PCM coding of the frequency-domain coefficients, the authors showed that coding noise could be shaped to lie under the temporal envelope of the transient.

In their original paper, Herre and Johnston presented a case where FDLP was applied on four separate spectral bands modeling four independent temporal envelopes, essentially what we call subband FDLP. But because the time windows they considered were very short (as used in MPEG2-AAC) and because they were mainly concerned with time-localized transients which have a nearly flat spectrum, they concluded that the temporal envelopes in different bands were highly correlated and virtually identical. The power of our method comes from extending subband FDLP to consider long time windows. By transforming longer, 256 ms blocks of signal (extensible to seconds or more), we capture enough variation to manifest itself as significantly different temporal envelopes between bands.

To tie this back to the property of the DCT that we presented in equation 3 and in figure 1, consider being handed the signal that led to the right-hand spectrogram of figure 1, without knowing that it was the result of a DCT. If we were to model this one-dimensional sequence using our standard frame-based LPC techniques, we would be finding approximations to the vertical (spectral) structure in vertical slices (short-time windows) of the spectrogram. But because of the spectrogram-domain reflection effected by the DCT, the envelopes we recover in fact describe the *temporal* structure of "short-frequency" regions of the spectrum of the original (pre-DCT) signal, i.e. subbands.

This method gives us a new parameter space from which we can extract novel features for use in ASR. We will now present our initial feature extraction approach.

3. FEATURE EXTRACTION

There are many ways in which the temporal envelope information modeled in FDLP could be converted into features for use in current speech recognizers; the approach described below can only be regarded as a first foray. One can identify two families of parameters that can be extracted. Firstly, we can use the temporal envelopes directly: The envelopes in figure 2 are sampled DFTs of the impulse responses (IR) of the all-pole filters that have been fit to the frequency domain. The basic LP representation may be suitable for direct transformation into temporal-based features such as modulation spectra, and relationships such as the direct transformation from prediction coefficients to cepstra [6] can give us decorrelated features describing the temporal behavior in different subbands.

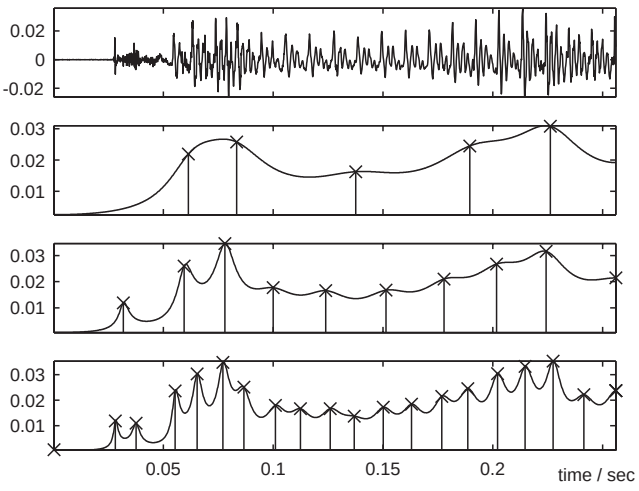

Fig. 2. Bracketing the pole rate. Original 256 ms speech segment and temporal envelopes fit using 10, 20 and 40 poles respectively. Notice that the first envelope captures the gross variation but doesn't give us any information about the transient burst. The the third envelope uses so many poles that it resolves individual pitch pulses. The middle panel using 20 poles for 256 ms captures the transient and the gross envelope while avoiding the pitch pulses.

The second approach seeks to derive features from each individual pole in the model i.e. the roots of the predictor polynomial. The angle of the pole on the z-plane corresponds to very accurate timing information, and the magnitude can provide knowledge about the energy of the signal, keeping in mind that this is a smoothed approximation to the 'true' Hilbert envelope. The sharpness of the pole (i.e. how closely it approaches the unit circle) relates to the dynamics of the envelope: a sharper pole indicates more rapid variation of the envelope at that time.

In our initial experiments, our goal was to extract features at a 10 ms frame-rate so that we could concatenate them with standard PLP features as enhancements to a baseline recognition system. In that sense we have not fully exploited the richness of the FDLP representation, but merely

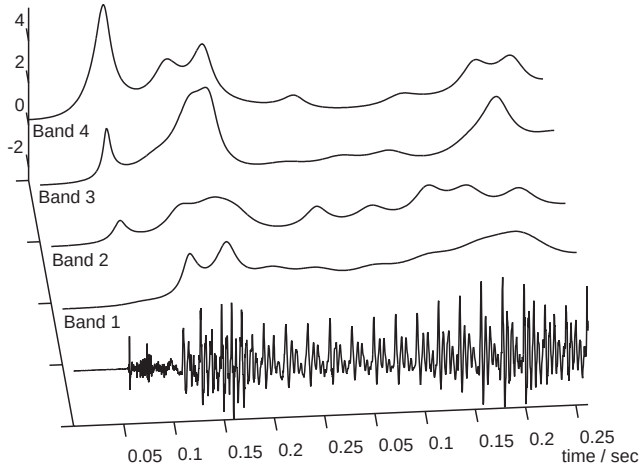

Fig. 3. Subband FDLP. The same speech segment as in figure 2 using 20 poles but this time with FDLP performed separately on 4 logarithmic bands. The split is 0-0.5, 0.5-1, 1-2 and 2-4 kHz. Notice that the highest frequency band (band 4) clearly captures the transient while band 1, reflecting the lowest frequencies, captures a smoother energy variation.

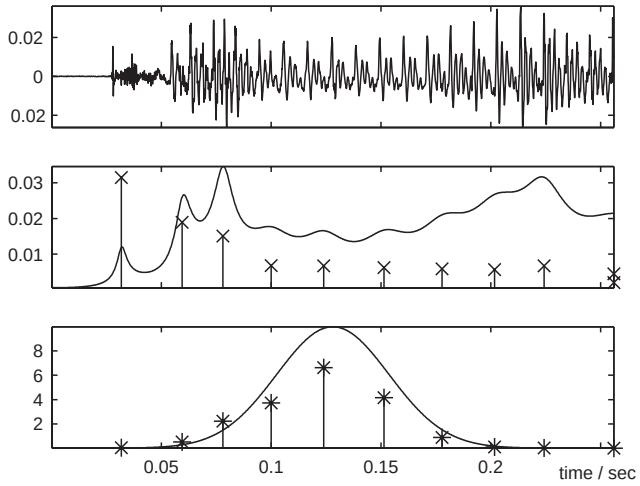

Fig. 4. Feature extraction process. The top pane shows the original 256 ms segment of speech. The middle pane shows the envelope modeled by FDLP, and the sharpness indices for the 10 positive-time poles. The bottom pane shows the effect of the time-localizing Gaussian window; features are calculated as the maximum of these values per band per frame.

tested its usefulness in a conventional recognition system.

Adopting the second, pole-based approach, we examine an index of sharpness of the FDLP poles $\{p_i\}$ defined by

$$\rho_i = \frac{1}{1 - |p_i|} \tag{5}$$

Thus, as pole magnitudes grow from zero to approach the unit circle, ρ_i grows from 1 to an unbounded large positive value.

For each analysis frame in time we take the full DCT and perform FDLP on 4 log bands using 20 poles per band. The choice of a 256 ms analysis window (2048 samples at 8 kHz) is, without loss of generality, dictated by computational considerations. Subbands are formed by breaking up the DCT into subranges that are exact powers of two, e.g. 256, 256, 512 and 1024 points for a 4-way split. After modeling with 20 poles per band per frame we calculate our sharpness index of equation 5. We then scale the ρ_is using a Gaussian window to achieve a finer time resolution than the 256 ms window, as illustrated in figure 4, and keep the maximum value in each band in each frame. The purpose of the window is to localize the sharpness values in the vicinity of the center of the frame. Figure 5 visually compares these pole sharpness features with direct measures of the subband energy. After examining the distributions of the sharpness parameters, we added a logarithmic transform to make the distributions closer to Gaussian, and thus a better match to our statistical models.

4. EVALUATION

We used a conventional HTK recognizer related to the systems described in [7] and [8]. We trained GMM-HMM models on a mixture of conversational and read speech, using a combination of Switchboard, Callhome, and Macrophone databases. To explore generalizability and to simplify the testing procedure, we tested on OGI-Numbers95, a 35-word task consisting of spontaneous numbers extracted from prompted telephone interactions.

Table 1 shows our recognition Word Error Rate (WER) results. The first line, “PLP12”, is our baseline system employing 12th order PLP features (plus deltas and double deltas). Subsequent systems augment these features with FDLP sharpness features in various guises. “FDLP-4log” adds four elements to each feature vector, derived from 4, logarithmically-spaced octave subbands (0-500 Hz, 500 Hz-1 kHz, 1-2 kHz, and 2-4 kHz). We found that performing a final DCT decorrelation on each frame of FDLP features improved recognition, shown in the “FDLP-Xlog+dct” lines. We tried using between 2 and 5 octave bands (where 2 ‘octaves’ is simply 0-2 kHz and 2-4 kHz, and 5 bands goes down to 0-250 Hz) to find the best compromise between signal detail and model accuracy (since narrow frequency bands contain fewer frequency samples with which to estimate the LP parameters). We also tried dividing the frequency axis on a Bark scale; this allowed us to use more bands (since Bark bands do not get narrow so quickly in the

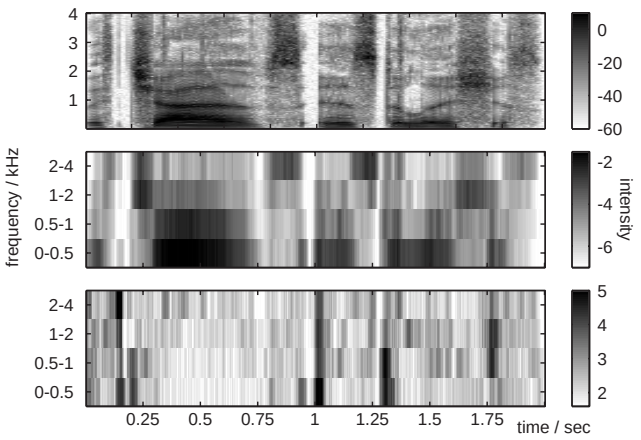

Fig. 5. Feature examples. The top pane shows a conventional spectrogram of the speech sample. The middle pane shows the per-frame maximum of the temporal envelopes extracted in each band by FDLP, using 256 ms frames stepped by 10 ms. Note the similarity to the energy in the spectrogram. The bottom pane plots the sharpness index features, calculated as described in the text. Notice that they pick the the important moments in time, like transient stop bursts, while almost completely ignoring high-energy but stable parts of the signal such as vowels.

low frequencies) but it was not beneficial.

We report results for two systems. Our initial experiments were done with a system, "raw 20k", trained on approximately 16 hours of speech (20,000 utterances), and were tested using the raw Numbers95 utterances. Training features were mean- and variance-normalized within all the utterances identified as belonging to a single speaker, but the test set normalization was performed only within utterance.

For our later experiments we sought to improve the baseline, so the training set was increased to 64 hours (85,000 utterances, "pad 85k"). We found that padding each end of our test utterances with 100 ms of artificial 'background noise' silence was beneficial (common practice for the Numbers95 corpus). We also normalized across all test set utterances marked as coming from the same speaker. These changes together effected a relative WER improvement of almost 45% in our PLP baseline, from 4.97% to 2.75%.

4.1. FDLP results

For the "raw 20k" system, we see that any kind of FDLP-derived information improved word error rate, with the greatest improvement coming from augmenting the PLP features with decorrelated 4 octave-subband FDLP sharpness features ("FDLP-4log+dct"). The WER change from 4.97% to 3.81% represents a 23.3% relative improvement. The advantage of using DCT decorrelation was quite clear, and the

Features	raw 20k	pad 85k
PLP12	4.97%	2.75%
FDLP-4log	4.08%	2.90%
FDLP-2log+dct		2.82%
FDLP-3log+dct		2.61%
FDLP-4log+dct	3.81%	2.63%
FDLP-5log+dct		2.69%
FDLP-8bark+dct	4.38%	

Table 1. Recognition WER results. "PLP12" is the 12th order PLP cepstrum baseline, whose features are augmented with FDLP features from between 2 and 8 subbands in the other lines. "log" indicates log-spaced (octave) subbands, which "bark" is for Bark-spaced subbands. "+dct" indicates DCT decorrelation applied to the small FDLP vector. The "raw 20k" column gives results from the smaller, initial system, and "pad 85k" results come from the improved training and testing sets.

lack of any advantage from using more, Bark-spaced bands ("FDLP-8bark+dct") was also clearly shown.

With the larger, better-performing "pad 85k" system, the improvements due to FDLP are smaller, with the best improvement of 2.75% baseline WER to 2.61% for 3 subband decorrelated features ("FDLP-3log+dct") constituting a 5% relative improvement. We are, however, getting to the limits of this test set at this level of performance: With only 4757 words in the test set, a simple binomial significance test requires an absolute word error rate difference of about 0.8% for significance at the 5% level. Thus, although the improvement from FDLP in the "raw 20k" system is statistically significant, none of the "pad 85k" results are significantly different from each other by this measure. In future experiments we will be using a larger test set, as well as a more difficult task, to provide greater insight into performance differences.

Some error analysis of these results is revealing, however. Figure 6 compares the word-level confusion matrices for the baseline "raw 20k" PLP system, and for the best-performing "FDLP-4log+dct" system. Looking at the absolute differences in error counts (middle pane), we see the greatest differences in for the words "four" (fewer confusions with "forty"), "eight" and "six" (fewer deletions), and "five" (fewer confusions with "nine"). We note that most of these main differences involve stops (/t/ in "eight" and "forty", and /k/ is "six"); this is consistent with our initial motivation for the FDLP sharpness features, of capturing information about short-duration transients in the speech signal.

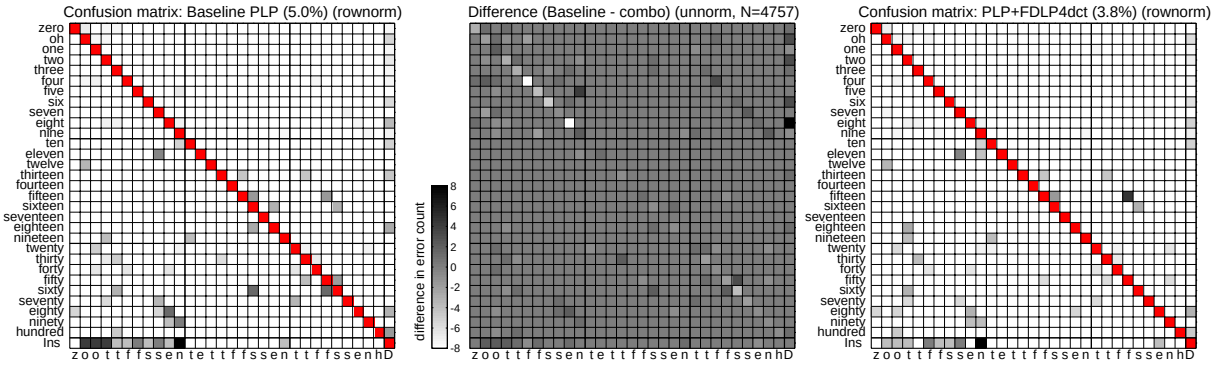

Fig. 6. Word-level confusion matrices for the "raw 20k" system. Left: for the baseline system (each row is normalized by the prior of that word). Right: for the best FDLP-augmented system. Middle: difference between the raw confusion counts in the two matrices. Improvements appear as lighter cells in the leading diagonal (more correct recognitions) and darker off-diagonal cells (reduced confusions).

5. DISCUSSION AND CONCLUSIONS

The benefits of FDLP lie in its ability to describe temporal structure without frame-rate quantization, and its rich and flexible representation of temporal structure in the form of poles; we have but scratched the surface of the transformations possible in this domain.

One possibility is to use FDLP as a direct temporal envelope estimate. By converting LPC coefficients directly to cepstral-like temporal envelope coefficients, FDLP could be married to novel modeling schemes such as TRAPS [2], a direction we are investigating.

There is a vast background on linear prediction methods; investigating how these work when applied to FDLP should lead to many further novel representations. For example the peak-hugging property of LPC can be adjusted, or even inverted to become a valley-hugging property, using the spectral transformations introduced in [9].

In conclusion, we have presented a new way to estimate temporal envelopes that exploits the dual of linear predictive coding applied in the frequency domain. This flexible, adaptive representation of the temporal structure can be analyzed across the full-band or for arbitrarily-spaced subbands, and presents many possibilities for novel speech recognition features. Our first experiments with extracting 'sharpness' parameters for individual poles gave promising and significant improvements on a relatively large speech task, and we are continuing to experiment with different feature derivations and more demanding recognition tasks.

6. ACKNOWLEDGMENTS

This project was supported by DARPA under the EARS-NA program. Thanks to Ozgur Cetin, Qifeng Zhu, and Barry Chen for help with the baseline recognizer.

7. REFERENCES

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

[2] H. Hermansky and S.Sharma, "Temporal patterns (TRAPs) in ASR of noisy speech," in *Proc. ICASSP*, Mar 1999, vol. 1, pp. 289–292.

[3] H. Hermansky and N. Morgan, "RASTA processing of speech," in *Trans. Speech and Audio Processing*, Oct 1994, vol. 2:4, pp. 578–589.

[4] J. Tribolet and R. Crochiere, "Frequency domain coding of speech," in *Trans. ASSP*, Oct 1979, vol. 27, pp. 512–530.

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

[6] L. Rabiner and R. Schafer, *Digital processsing of speech signals*, Prentice Hall, 1978.

[7] Ozgur Cetin and Mari Ostendorf, "Cross-stream observation dependencies for multi-stream speech recognition," in *Eurospeech*, Geneva, 2003.

[8] P. Somervuo, B. Chen, and Q. Zhu, "Feature transformations and combinations for improving ASR performance," in *Eurospeech*, Geneva, 2003.

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