

Average Instantaneous Frequency (AIF) and Average Log-Envelopes (ALE) for ASR with the Aurora 2 Database

Yadong Wang, Jesse Hansen, Gopi Krishna Allu, Ramdas Kumaresan

Department of Electrical and Computer Engineering
University of Rhode Island, RI, USA

ydwang, hansenj, gopi, kumar@ele.uri.edu

Abstract

We have developed a novel approach to speech feature extraction based on a modulation model of a band-pass signal. Speech is processed by a bank of band-pass filters. At the output of the band-pass filters the signal is subjected to a log-derivative operation which naturally decomposes the band-pass signal into analytic (called $\hat{\alpha}(t) + j\hat{\alpha}$) and anti-analytic (called $\hat{\beta}(t) - j\hat{\beta}$) components. The average instantaneous frequency (AIF) and average log-envelope (ALE) are then extracted as coarse features at the output of each filter. Further, refined features may also be extracted from the analytic and anti-analytic components (but not done in this paper). We then evaluated the Aurora 2 task where noise corruption is synthetic. For clean training, (compared to the mel-cepstrum front end, with 3 mixture HMM back-end,) our AIF/ALE front end achieves an average improvement of 13.97% with set A and 17.92% improvement with set B and -31.72% (negative) 'improvement' with set C. The overall improvement in accuracy rates for clean training is 7.97%. Although the improvements are modest, the novelty of the front-end and its potential for future enhancements are our strengths.

1. Introduction

Virtually every speech-recognition system that engineers have built uses framewise feature vectors (obtained from a sequence of ≈ 20 ms frames of a speech signal). The feature vectors are derived from short-term spectral envelopes computed by linear prediction (LP) analysis or by using a bank of bandpass filters (BPFs). When speech is degraded by noise, interference, and channel effects (such as telephone speech, reverberation etc..) perturbations at one frequency affect the entire feature vector rendering the extracted features vulnerable. This type of framewise spectral feature extraction is at odds with how the auditory system processes and recognizes speech.

Since Helmholtz the perceptual basis of speech has been thought to be associated with the distribution of energy of the speech signal across frequency. This traditional view of speech perception agrees well with the prevalent but over simplified opinion that the inner-ear is a spectrum analyzer. Further, human speech production system is thought to be analogous to a slowly time-varying filter excited by the vocal chord vibrations or by turbulent noise. The above views on speech production, reception and perception have motivated the development of traditional spectrum analysis based front-ends for ASR. However, of late, this entire view of spectrum-oriented speech representation has come under critical scrutiny [1, 2, 3]. Large amounts of auditory physiological and psychophysical evidence, as well as experience with today's ASR systems point to the fact that, more than the spectrum, the time evolution of the spectrum which man-

ifests itself in slowly varying modulation properties of a speech signal are the primary carriers of information. This view has been advanced by many but most eloquently by Greenberg in [4]. We translated this vision into a new computational feature extraction approach/algorithm which eventually resulted in improved ASR systems and also help advance our understanding of the functioning of the auditory system.

The complex modulation representation we investigated has been inspired both by the mathematics of signal processing and by the physiology of the auditory system. Our method focuses on characterizing the slow temporal (envelope and phase) modulations in a speech signal. This was achieved by using a non-linear and hierarchical signal processing method which we call as the Log-Derivative Filter-bank Tree (LDFT). We use this feature extraction method to improve the performance of speech recognizers in the presence of noise and interference.

1.1. The Temporal Dynamics/ Modulation Information is Crucial:

Faithful preservation of the spectrum of a speech signal is not absolutely necessary for intelligibility. This fact has been demonstrated time and again. Alterations of the speech spectra by heavy filtering, suppressing large portions of the speech spectrum does not seem to affect speech intelligibility significantly. Instead what is important is the evolution of the spectrum or portions of the spectrum with time. This information manifests itself in the form of temporal envelope and phase modulations. There is significant evidence that extracting features related to temporal modulations is accomplished by the auditory system and that ASR methods also improve their performance when such features are incorporated in the feature vector.

1.2. Psychophysics of Speech Supports Importance of Temporal Information

A key result from the study of speech intelligibility in humans is the importance of slow changes in the speech spectrum. These changes are confined to a bandwidth of about 16 Hz or so. This is not surprising considering that the articulators in the human vocal tract can not move at a higher rate. Greenberg et al. [3] have demonstrated the inherent robustness of low frequency modulations to noise by proposing a modulation spectrogram which displays magnitude of the energy in the lowest modulation band (1 to 8 Hz) as a function of frequency. The modulation spectrograms are remarkably stable under noise and reverberation suggesting that low frequency modulations are sufficient to encode linguistically relevant information. We agree with their arguments but believe that better methods can be devised to extract this information.

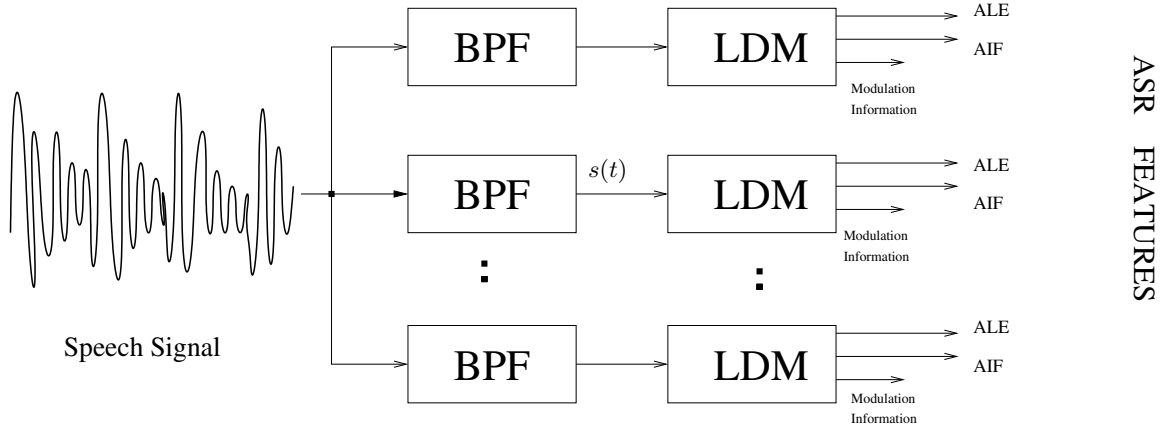


Figure 1: **Overview:** Speech signal is separated by a bandpass filter bank. Each output from the BPFs is decomposed by LDM (Log Derivative Module) (Details in Fig. 4.) into three parts: $\log A_c$, Ω_c and modulation information which include both α and β . ALE (Average or low-pass filtered Log-Envelopes, $\log A_c$) and AIF (Average Instantaneous Frequency, Ω_c) are the features we used in the HTK training and recognition experiments. In this paper we have not used the modulation details in α and β .

1.3. Dynamic Temporal Features Tend to Improve ASR

The most common feature extraction methods for ASR are the so called Mel cepstrum and perceptual linear prediction (PLP) introduced decades ago. They emulate some of the very basic textbook properties of human auditory perception. Hermansky discusses some drawbacks of these features. Furui introduced the important step of incorporating temporal changes in the cepstral features into the feature vector. These are velocity and acceleration measures of the spectra, usually called delta and double-delta features. This step demonstrably improved the performance of ASR. RASTA (Relative Spectral) method generalizes Furui's idea. A number of other hyphenated methods such as RASTA-PLP, RASTA-PLP-LDA-NN have proliferated. But the key point is that these methods further emphasize the importance of temporal evolution of the spectrum. Further Arai et al. demonstrate à la Drullman et al. that if we retain the low frequency information in the cepstral coefficients then speech intelligibility is not terribly degraded. Many phonetic distinctions depend on temporal envelopes produced by amplitude modulations in the 0-30 Hz range. Experiments with reduced spectrum speech sounds (1-6 frequency channels) have demonstrated that a high degree of speech intelligibility can be achieved on the basis of relatively limited spectral information coupled with low frequency amplitude modulation information.

2. Envelope and Phase Modulations of a Bandpass Signal

Before we extract modulation information from a speech signal, the logical first step is to characterize and understand the phase and envelope modulations of a bandpass signal. With this goal in mind we have developed models for a bandpass signal in references [5, 6]. We summarize these results below.

Consider a real-valued band-pass signal $x(t)$. Most general form of $x(t)$ will have both envelope and phase variations. Thus a model for $x(t)$ is

$$x(t) = A_c a(t) \cos(\Omega_c t + \phi(t)). \quad (1)$$

A_c is the amplitude scale factor. $a(t)$ represents the normalized envelope variations. Ω_c is the carrier frequency. $\phi(t)$ represents the phase variations. In order to visualize the signal $x(t)$ one

might imagine that $x(t)$ is obtained by filtering a speech signal with a relatively broad bandpass filter centered around one of the formant frequencies. Typically, as the formant meanders in the time-frequency plane, the quantities A_c and Ω_c (roughly the formant frequency) vary slowly, whereas the phase $\phi(t)$ and envelope $a(t)$ vary more rapidly.

It is often convenient to work with the complex version (called the analytic signal of $x(t)$). Let us denote the analytic signal corresponding to $x(t)$ by $s(t)$, i.e.

$$s(t) = x(t) + j\hat{x}(t) = A_c a(t)e^{j(\Omega_c t + \phi(t))}. \quad (2)$$

The $\hat{\cdot}$ denotes the Hilbert Transform operation. Often, using such a model one computes the envelope as $|s(t)|$ and the instantaneous frequency (IF) as $\frac{1}{2\pi} \frac{d}{dt} \angle s(t)$. Although $s(t)$ itself is bandlimited, the corresponding envelope and IF are typically band-unlimited functions. The early phase vocoder due to Flanagan used filtered versions of the envelopes and IFs of several band-pass filtered outputs to encode speech signals. Unfortunately, the model of a bandpass signal in Eq.(2) does not lead to any insight into the relationship between phase and envelope functions. To further understand the anatomy of the modulations in a bandpass signal we invoke models borrowed from standard linear-time-invariant systems theory.

In traditional engineering literature, a linear time-invariant (LTI) discrete-time system is characterized by a system function $H(z)$. $H(z)$ is a polynomial or a ratio of polynomials in z . z is the complex-frequency variable. The frequency response of the system is the function $H(z)$ evaluated around the unit circle $|z| = 1$. It is a 2π -periodic complex-valued function, and specifies the magnitude and phase responses of the system. Since our aim is to characterize the phase and envelope functions of a signal $s(t)$ in Eq.(2) (over a time window of T seconds), we might borrow some of the terminology from LTI systems theory. Thus we define a new complex-time plane, and call it the ζ -plane, and define a complex-valued function $s(\zeta)$ in that plane. *The complex-time ζ -plane is the dual of complex-frequency z -plane*, and is suitable for modeling complex-valued periodic signals. In this case a signal function $s(\zeta)$ is defined as a ratio of polynomials in ζ . We obtain a corresponding periodic signal $s(t)$ by evaluating $s(\zeta)$ around the unit circle $|\zeta| = 1$ i.e. $\zeta = e^{-j\Omega t}$, where $\Omega = 2\pi/T$ is the fundamental frequency.

Hence, the unit circle in the ζ -plane corresponds to the time interval 0 to T seconds, just as the unit circle in the z -plane corresponds to an angular frequency of 0 to 2π radians. The location of poles and zeros of $s(\zeta)$ in the ζ -plane, decide the shape of phase and envelope functions of $s(t)$ over the interval of 0 to T seconds. We normally use signal models that have only zeros, since inclusion of poles in the model implies a spectrum with infinite support, just like its dual, the IIR filters. Using this perfect dualism between complex-time representation of the signal and an LTI system's frequency response, we can then talk of minimum phase (MinP), maximum phase (MaxP) and all-phase (the equivalent of all-pass filters) (AllP) signals. If a signal $s_{MinP}(t)$ is MinP, as in systems theory, the phase of the signal is the Hilbert transform of its log-envelope. That is the signal has the form

$$s_{MinP}(t) = e^{\alpha(t) + j\hat{\alpha}(t)}. \quad (3)$$

Similarly for a maximum phase signal with spectrum confined to the positive side of the frequency axis ($\Omega \geq 0$)

$$s_{MaxP}(t) = e^{j\Omega_c t} e^{\beta(t) - j\hat{\beta}(t)}. \quad (4)$$

More generally, an arbitrary bandpass analytic signal as in Eq.(2) then can be modeled as a product of MinP and MaxP signals. That is

$$s(t) = A_c e^{\alpha(t) + \beta(t) + j(\Omega_c t + \hat{\alpha}(t) - \hat{\beta}(t))}. \quad (5)$$

Thus, we model analytic bandpass signals over a time window of T seconds as polynomial functions (since we allow it to have only zeros) in the complex-time variable ζ . This is in contrast to the standard speech production models where a signal is modeled as the output of a linear system with an all-pole or pole/zero transfer function in the complex-frequency or z domain. Our model is equally applicable for (bandpass filtered) signals which are deterministic or stochastic. For deterministic signals like voiced signals, a natural choice for T would be the pitch period. For stochastic signals (unvoiced) we implicitly work with the periodic extensions of the T -second (may be appropriately weighted) segment to apply the model. T may be different for different frequency bands.

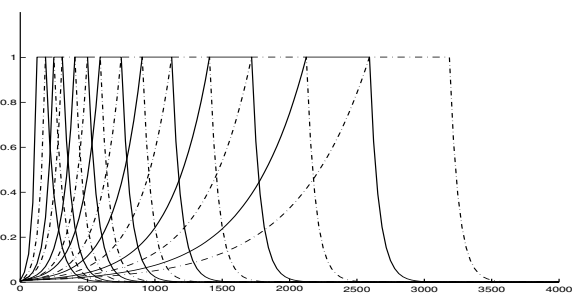


Figure 2: Frequency responses of 14 bandpass filters. A special feature of our 'trapezoid' shaped filterbank is the significant overlap of the frequency responses at the low frequency end.

3. Extraction of Temporal Modulation Features for Robust ASR

Speech is processed first by a bank of 14 band-pass filters. The filter bank we used (in Fig. 2) is dramatically different from

the MFCC filter bank and PLP filter bank. A number of other cochlear filterbank models can be used in place of our filterbank. A special feature of our filterbank is the significant overlapping of the frequency responses at the low frequency end, which is crucial to noise robustness.

In the experiments described here, our front-end extracted the following features: 1) Average instantaneous frequency (AIF), 2) decorrelated average log envelope (ALE). Taking the natural logarithm of $s(t)$ in Eq.(5) we get

$$\log(s(t)) = \log A_c + j\Omega_c t + \alpha(t) + j\hat{\alpha}(t) + \beta(t) - j\hat{\beta}(t). \quad (6)$$

Further taking the time derivative of $\log(s(t))$ we have

$$\frac{d}{dt} \log(s(t)) = j\Omega_c + \dot{\alpha}(t) + j\dot{\hat{\alpha}}(t) + \dot{\beta}(t) - j\dot{\hat{\beta}}(t). \quad (7)$$

Thus the log and derivative operations on the bandpass signal produces three components. First is the slowly varying magnitude modulation $\log A_c$ which is obtained by low pass filtering $\log|s(t)|$. Second is the slowly varying component Ω_c that corresponds to the frequency of the most dominant component in $s(t)$. The third is the modulation information $\dot{\alpha}(t) + j\dot{\hat{\alpha}}(t) + \dot{\beta}(t) - j\dot{\hat{\beta}}(t)$. $\dot{\alpha}(t) + j\dot{\hat{\alpha}}$ is an analytic signal which has a spectrum on the positive frequency side. $\dot{\beta}(t) - j\dot{\hat{\beta}}$ is an anti-analytic signal which has a spectrum on the negative frequency side. The analytic and anti-analytic components of the signal $\frac{d}{dt} \log(s(t))$ occupy non-overlapping frequency bands and can be separated by filtering.

Average filtered Log-Envelopes (ALE) ($\log A_c$) and Average Instantaneous Frequency (AIF) Ω_c 's are extracted as features for ASR. The ALEs contain the slowly varying magnitude modulation information in the speech signal, while AIFs contain the slowly varying frequency modulation information. The log-derivative operation is a form of automatic gain control because it removes the scale factor A_c from $s(t)$ thereby normalizing the modulation components. The ALEs are de-correlated further by DCT before the standard reference HTK training.

The details of the Log-Derivative Module (LDM) are shown in Figure 4. A complex-valued band-pass signal $s(t)$ is presented to the LDM. It computes the log-envelope of $s(t)$ as $\log(|s(t)|)$ and low-pass filters it to obtain $\log A_c$. It also computes in effect the complex logarithm of $s(t)$ and takes its derivative, separates Ω_c and outputs the modulation component which consists of both the analytic and anti-analytic components to the next level. Well known methods can be used to perform the operations specified in LDM without resorting to phase unwrapping.

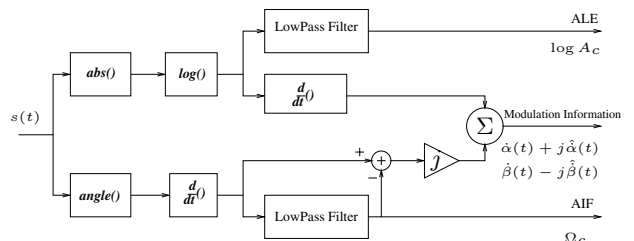


Figure 4: Log-Derivative Module performs $\frac{d}{dt} \log(s(t))$. It computes the log-envelope of $s(t)$ as $\log(|s(t)|)$ and low-pass filters it to obtain $\log A_c$. It also computes, the angle of $s(t)$ and takes its derivative then low-pass filters it, to get Ω_c .

Features: Average Instantaneous Frequency (AIF) and Average Log-Envelopes (ALE)														
Clean Training - Results														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.83	98.67	98.42	98.64	98.64	98.83	98.67	98.42	98.64	98.64	98.68	98.19	98.44	98.60
20 dB	96.75	92.65	96.60	94.42	95.11	89.38	95.71	90.78	95.65	92.88	86.95	87.03	86.99	92.59
15 dB	92.88	83.22	93.11	88.28	89.37	78.72	89.90	84.43	91.95	86.25	74.09	76.57	75.33	85.32
10 dB	82.71	63.27	78.77	73.40	74.54	61.77	76.36	67.58	79.36	71.27	54.93	58.56	56.75	69.67
5 dB	63.34	36.67	45.90	49.77	48.92	39.79	52.51	41.57	50.54	46.10	35.28	40.51	37.90	45.59
0 dB	40.13	15.66	17.30	29.99	25.77	17.96	28.48	18.82	22.34	21.90	18.21	21.89	20.05	23.08
-5dB	17.38	8.13	9.78	14.44	12.43	7.80	14.00	9.99	9.94	10.43	10.01	12.03	11.02	11.35
Average	75.16	58.29	66.34	67.17	66.74	57.52	68.59	60.64	67.97	63.68	53.89	56.91	55.40	63.25
Clean Training - Relative Performance														
	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	-9.35%	-33.00%	-51.92%	-70.00%	-41.07%	-9.35%	-33.00%	-51.92%	-70.00%	-41.07%	-53.49%	-75.73%	-64.61%	-45.78%
20 dB	-10.17%	25.38%	-31.27%	-54.57%	-17.66%	-6.09%	-0.70%	1.50%	17.61%	3.08%	-99.54%	-166.32%	-132.93%	-32.42%
15 dB	-9.37%	36.05%	30.82%	-47.24%	2.57%	10.44%	12.55%	32.27%	50.76%	26.51%	-95.84%	-111.27%	-103.56%	9.08%
10 dB	18.75%	27.37%	35.65%	-9.29%	18.12%	15.48%	28.12%	29.74%	48.02%	30.34%	-72.68%	-62.06%	-67.37%	5.91%
5 dB	23.37%	13.47%	17.92%	8.95%	15.93%	12.73%	22.84%	16.13%	31.38%	20.77%	-32.81%	-17.13%	-24.97%	9.69%
0 dB	19.08%	7.03%	3.32%	14.57%	11.00%	7.86%	12.95%	5.15%	12.18%	9.54%	-9.67%	-1.32%	-5.50%	7.12%
-5dB	6.98%	6.66%	0.43%	5.35%	4.86%	4.49%	3.95%	1.92%	1.63%	3.00%	-2.05%	0.99%	-0.53%	3.04%
Average	18.60%	16.78%	14.55%	5.14%	13.97%	10.40%	18.38%	15.80%	27.81%	17.92%	-36.27%	-27.17%	-31.72%	7.97%

Figure 3: Clean training results on Aurora 2 database using ALE/AIF frontend. Our results indicate a substantial improvement for certain tasks, especially for SNRs of 0dB to 15 dB. Average recognition rates were improved for every task in sets A and B.

4. Performance Using ALE & AIF:

Experiments with the Aurora 2 database were conducted to determine the level of robustness for mismatched conditions, i.e. when the models were trained on clean speech and tested on noisy utterances. Our results (compared to the mel-cepstrum front end, with 3 mixture HMM back-end, in Fig. 3) indicate a substantial improvement for certain tasks, especially for SNRs of 0dB to 15 dB. Average recognition rates were improved for every task in sets A and B. Accuracy rates for set C were a bit disappointing, substantially under-performing the standard set by the reference front-end. This is a result that we are still investigating. Accuracy rates improved by an average of 13.97% for set A, and by 17.92% for set B, and by -31.72% for set C. The overall accuracy rates for clean training improvement is 7.97%.

5. Discussion

In this paper we have not used the analytic and anti-analytic components in Eq.7. Note that Ω_c s often ‘capture’ the strongest component’s frequency nicely. Even if it does not, one need not worry since the α and β carry the difference frequency information anyway. $\hat{\alpha}(t) + j\hat{\alpha}$ is an analytic signal (spectrum is non-zero for positive frequencies) and $\hat{\beta}(t) - j\hat{\beta}$ is also an analytic signal but its spectrum is non-zero only for negative frequencies. We call $\hat{\beta}(t) - j\hat{\beta}$ as the anti-analytic component. Thus the analytic and anti-analytic components of the signal $\frac{d}{dt} \log(s(t))$ occupy non-overlapping frequency bands and can be separated by filtering. Once these two components are separated, and the slowly varying Ω_c removed, then the analytic/anti-analytic components may be treated just like the original bandpass analytic signal and subjected to filter/log-derivative operations once more. This results in a tree-like decomposition which we call the Log-Derivative Filterbank Tree (LDFT). We can repeat our log derivative processing at the second level and use the captured difference frequencies as additional features for speech recognition. This technique leads to added redundancy and a

recombination of valuable phonetic information across frequencies. Initial experiments indicate that these second level features can improve recognition performance on both clean and noisy speech. These issues are being currently investigated.

6. Acknowledgements

This research was supported by a grant from the National Science Foundation under grant number EIA-0130793, and CCR-0105499. The authors would like to thank J. Swaminathan for his help with the reference file. The authors also thank Dr. Cariani of Eaton Peabody Laboratory, and Dr. de Boer of U. of Amsterdam for many insightful discussions.

7. References

- [1] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, Apr. 1990.
- [2] B. E. D. Kingsbury, *Perceptually Inspired Signal-Processing Strategies for Robust Speech Recognition in Reverberant Environments*. PhD thesis, University of California at Berkeley, 1998.
- [3] S. Greenberg and B. Kingsbury, “The modulation spectrogram: In pursuit of an invariant representation of speech,,” in *ICASSP’97*, pp. 1647–1650, 1997.
- [4] S. Greenberg, “Understanding speech understanding: Towards a unified theory of speech perception,,” in *ESCA workshop on Aud. Basis of Speech Percept*, pp. 1–8, 1996.
- [5] R. Kumaresan and A. Rao, “Model-based approach to envelope and positive-instantaneous frequency of signals and application to speech,,” *Journal of the Acoustical Society of America*, vol. 105 (3), pp. 1912–1924, (March) 1999.
- [6] R. Kumaresan and Y. Wang, “On representing signals using only timing information,,” *Journal of the Acoustical Society of America*, vol. 110, pp. 2421–2439, Nov 2001.