Comprehensive Modulation Representation for Automatic Speech Recognition

Yadong Wang,[†] *Steven Greenberg,*[‡] *Jayaganesh Swaminathan,*[*] *Ramdas Kumaresan,*[*] *David Poeppel,*[†]

[†]Cognitive Neuroscience of Language (CNL) Laboratory, University of Maryland
[‡]Department of Acoustic Technology, Technical University of Denmark
[*]Department of Electrical and Computer Engineering, University of Rhode Island
ydwang@umd.edu

Abstract

We present a new feature representation for speech recognition based on both amplitude modulation spectra (AMS) and frequency modulation spectra (FMS). A comprehensive modulation spectral (CMS) approach is defined and analyzed based on a modulation model of the band-pass signal. The speech signal is processed first by a bank of specially designed auditory band-pass filters. CMS are extracted from the output of the filters as the features for automatic speech recognition (ASR). A significant improvement is demonstrated in performance on noisy speech. On the Aurora 2 task the new features result in an improvement of 23.43% relative to traditional mel-cepstrum front-end features using a 3 GMM HMM back-end. Although the improvements are relatively modest, the novelty of the method and its potential for performance enhancement warrants serious attention for future-generation ASR applications.

1. Introduction

Ever since Helmholtz, the perceptual basis of speech has been associated with the distribution of energy of the acoustic signal across frequency. This traditional view of speech agrees well with the prevalent but oversimplified view of the ear as a spectrum analyzer. Moreover, the production of speech is thought to be analogous to a slowly time-varying filter excited by either vocal chord vibrations or by turbulent noise. Such views have influenced the design of traditional ASR front-ends.

However, this spectrum-oriented perspective has recently come under critical scrutiny. There is increasing evidence that the temporal evolution of the spectrum, as manifested in slowly varying modulation properties of the signal, is a primary carrier of information. This view has been advanced by many but most forcefully by Greenberg in [1]. We translate and upgrade this vision into a new computational feature extraction approach/algorithm that improves ASR performance and also helps advance our understanding of the auditory system.

RK's research funded by NSF under the grants CCF-0130793.
YW and DP are supported by NIH DC 05660 to DP.

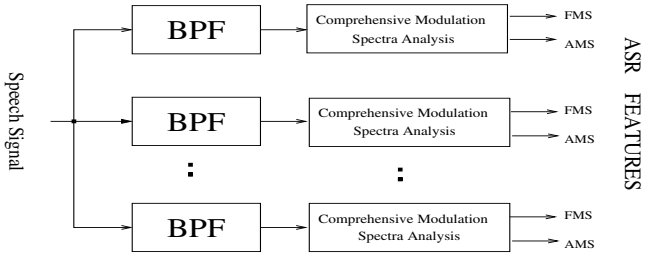

Figure 1: **Overview**: The speech signal is separated by a bandpass filter bank. Each output from the BPFs is decomposed by comprehensive modulation spectra analysis modules into AMS and FMS. (See details in Fig. 2.) AMS and FMS are the features used in the training and recognition experiments described below.

The comprehensive complex modulation representation investigated has been inspired both by the mathematics of signal processing and by the physiology of the auditory system. Our method (Figure 1) characterizes the slow temporal (both envelope and phase) modulations in the speech signal. This is achieved by using a non-linear signal-processing method referred to as the log-derivative based Comprehensive Modulation Spectra Analysis (CMSA). We use this feature extraction method to improve the performance of speech recognizers in the presence of noise and other forms of acoustic interference.

1.1. Complex modulation information is incomplete

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 psychophysical evidence that extracting features related to temporal modulations is ac-

complished by the auditory system.

Greenberg and colleagues [1] have shown that the intelligibility of spoken language depends principally on low-frequency modulation of signal energy between 3 and 10 Hz. Their recent studies [1] imply further that intelligibility is based on both phase and the amplitude components of the modulation spectrum, referred to as the complex modulation spectrum.

The modulation spectrum defined by Greenberg and Kingsbury [2] is a characterization of the way a signal's energy changes over time. It is computed by performing a spectral analysis of the signal's energy envelope and normalizing by the average energy of the signal. By adding the commensurate delta-phase (relative to the original signal) for each one-third-octave interval of the modulation spectrum, the complex modulation spectrum is obtained [1].

We will refer to this modulation spectrum as the amplitude modulation spectrum or energy envelope modulation spectrum. Because the spectral analysis has been performed only on the signal's energy envelope, only part (or half) of the relevant information is present.

1.2. Comprehensive modulation information is crucial

The lost component is the phase, or frequency-modulation part, which is also relevant for speech decoding. From our previous automatic speech recognition experiments [3] and the results of listening tests in normal-hearing and cochlea-implant subjects by Zeng [4], both the frequency modulation and amplitude modulation are important and provide complementary information.

The phase and envelope of a signal are two sides of the same coin. For very special signals (such as minimum-phase signals) the temporal envelope contains essentially all the information associated with the temporal phase. For arbitrary signals such as band-passed speech, both the temporal envelope and phase carry information. Nevertheless, temporal phase (being a wild function) has not be used effectively in ASR systems until now. We believe that the temporal evolution of the spectrum can be more directly measured by simply monitoring the phase and envelope modulations of the signal components. We argue that improved ASR performance will be achieved if low-passed amplitude modulation spectra and low-passed frequency modulation spectra are combined together in a systematic way.

2. Envelope and phase modulations of a bandpass signal

Before extracting modulation information from the speech signal, we first characterize the phase and envelope of a bandpass signal. Pertinent models have been developed for a bandpass signal [5, 6]. We summarize these results below.

The most general form of a real-valued band-pass signal $x(t)$ will have both envelope and phase variations. Thus, a model for $x(t)$ can be $x(t) = u(t)\cos\phi(t)$. 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. 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.* $s(t) = x(t) + j\hat{x}(t) = u(t)\ e^{j\hat{\phi}(t)}$. The $\hat{}$ denotes the Hilbert transform operation. By such a model, the envelope is computed 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 work of Flanagan used filtered versions of the envelopes and IFs of several band-pass filtered outputs to encode speech signals. Unfortunately, this model of a bandpass signal 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 [5].

3. Extraction of comprehensive temporal modulation features for robust ASR

In our approach the speech signal is processed first by a bank of 14 band-pass filters. The filter bank we used is different from the standard MFCC (mel-frequency cepstral coefficients) and PLP filter banks. A special feature of our filterbank is the significant overlap of the frequency responses at the low frequency end, which is crucial to noise robustness. We created a bank of complex FIR filters whose real and imaginary parts are in quadrature relation. These complex filters are directly applied to the input speech signals.

By applying the comprehensive modulation spectra analysis module (shown in Fig. 2) to the output from each of 14 filters, the front-end extracted the following features: 1)FMS as the low-pass-filtered, intensity-weighted average instantaneous frequency (IWAIF); 2)AMS as the low-pass-filtered log-envelope.

IWAIF is defined as [7]

$$\vartheta(t) \stackrel{\triangle}{=} \frac{\int_0^T \dot{\phi}(t)\ u^2(t)dt}{\int_0^T u^2(t)dt} \quad , \tag{1}$$

where $u(t)$ and $\phi(t)$ are the envelope and phase variations of the bandpass signal $s(t)$ described in the previous section. To analyze the properties of IWAIF, let us first examine the intensity-weighted instantaneous frequency (IWIF). IWIF, also called the envelope-squared weighted IF, is defined as $u^2(t)\dot{\phi}(t)$. Multiplying $\dot{\phi}(t)$ by $u^2(t)$ eliminates the denominator of $\dot{\phi}(t)$ and con-

sequently each term in IWIF is band-limited. In fact, the bandwidth of IWIF is the bandwidth of the original signal $s(t)$. Averaging IWIF over one period yields IWAIF. It is not in general the same as the carrier frequency of $s(t)$. It is equal to the carrier frequency only under special conditions. By invoking Parseval's relation, $\vartheta(t) = \int_0^\infty f|S(f)|^2 df / \int_0^\infty |S(f)^2 df$. Thus, the IWAIF of a signal is located exactly at the center of gravity of its energy density spectrum [8].

The details of CMSA are shown in Fig. 2. An analytic band-pass signal $s(t)$ (from the output of each complex filter) is presented to the CMSA. The Hilbert envelope in each band is the absolute value of the complex filter output. It computes the log-envelope of $s(t)$ as $\log(|s(t)|)$. The AMS is obtained by a low-pass filtering (with cut-off frequency at 40 Hz) and down-sampling (by 80) of the log-envelope. It also computes the angle of $s(t)$ and takes its derivative to obtain instantaneous frequency. A well-known method [9] can be used to perform the operation without resorting to phase unwrapping. The FMS is calculated by first multiplying the square of the envelope with the instantaneous frequency, then low-pass filtering (with cut-off frequency at 40 Hz) and down-sampling (by 80).

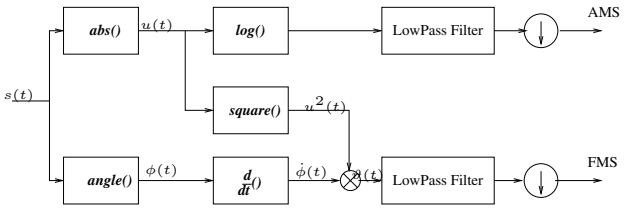

Figure 2: Comprehensive modulation spectra analysis (CMSA) module.

4. Performance using AMS and FMS

First we compare the relative performance of AMS and FMS as features separately. Initial speech recognition results with two subsets of the Aurora corpus are shown in Table 1. AMS performs best with clean speech, but not as well in noisy (SNR=20dB) conditions. FMS is relatively stable for both clean and noisy data as expected. AMS combined with FMS provides the best overall performance.

Features	Clean Data	Noisy Data
AMS	98.42%	82.02%
FMS	96.86%	95.26%
FMS & AMS	98.35%	96.93%

Table 1: Results on two subsets of Aurora data. Clean data are composed of 1000 training and 1000 testing instances extracted from the Aurora database. Noisy data are composed of 1000 training instances of clean data and 1000 testing instances of car noise with an SNR of 20 dB.

Experiments of AMS, FMS and CMS with the entire Aurora 2 database were conducted to determine the robustness for mismatched conditions (*i.e.* ,when the models are trained on clean speech and tested on noisy material). Because of space limitations, only results for CMS are shown in Figure 3. Our results, compared to the Aurora 2 standard mel-cepstrum front end, with 3 GMM HMM back-end, indicate a substantial improvement for certain tasks, especially for SNRs of 0dB to 20 dB. Average recognition rates are improved for every task in sets A and B. Accuracy rates for set C are not nearly as good, the cause of which is currently under investigation. Accuracy rates improve by an average of 27.57% for set A, and by 28.96% for set B, and by -0.48% for set C. The overall accuracy rates for clean training improvement is 23.43%.

5. Discussion

Potamianos and Maragos [10] compared short-time averages of quadratic operators, e.g., the energy spectrum, generalized first spectral moments, and short-time averages of the instantaneous frequency, to the standard front end features used in ASR. A close relationship among these feature sets was established theoretically and by experimental results. It was shown that IWAIF is an equivalent time-frequency representation, and can be expressed as the derivative of the spectral energy distribution. Front-ends using first spectral moment features, like IWAIF, were shown to perform significantly worse then standard spectral energy features [10].

The reason for this is the loss of slow amplitude modulation information. It has been shown by Greenberg [1] that the intelligibility of spoken language depends principally on low-frequency modulation of signal energy. The modulation spectrograms proposed are remarkably stable under noise and reverberation [2].

We refer to this as amplitude (or envelope) modulation spectra, because only the envelope of the signal is used, while its phase (or fine structure) is discarded directly after band-pass filtering. Analogously we can define frequency modulation spectra as

$$\Phi(f) = \frac{1}{<\vartheta(t)>} \left| \int \vartheta(t) e^{-j2\pi ft}\, dt \right|, \qquad (2)$$

where $\vartheta(t)$ is the IWAIF defined in the previous section. The energy modulation spectra are low-pass in form, with the roll-off beginning around 4 Hz. Interestingly, the frequency modulation spectra (one example taken from TIMIT) are also low-pass in form with a similar roll-off beginning around 4 Hz.

Hence, by combining these two parts of information, low-passed amplitude modulation spectra and low-passed frequency modulation spectra, together in a systematic way, it is not surprising that ASR performance improves. They are natural outputs of the product signal model [5]

Aurora 2 Clean Training - Results

	A					B					C				Percentage Improvement
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	
Clean	98.96	98.70	98.66	98.92	98.81	98.96	98.70	98.66	98.92	98.81	98.96	98.55	98.76	98.80	-24.38%
20 dB	97.85	95.50	97.73	97.19	97.07	92.63	97.25	94.09	97.38	95.34	94.84	94.44	94.64	95.89	27.19%
15 dB	96.07	87.73	95.59	93.86	93.31	84.34	94.41	89.62	94.54	90.73	88.03	88.03	88.03	91.22	38.02%
10 dB	88.92	71.70	84.94	83.96	82.38	68.71	83.49	76.17	86.24	78.65	71.97	71.49	71.73	78.76	35.59%
5 dB	73.29	43.47	53.89	62.57	58.31	44.03	59.40	52.25	62.57	54.56	50.17	49.94	50.06	55.16	26.08%
0 dB	46.79	12.82	18.82	37.30	28.93	14.74	29.66	22.99	26.78	23.54	24.50	26.39	25.45	26.08	10.97%
-5dB	19.28	-0.21	7.96	16.48	10.88	-2.52	11.55	7.84	10.34	6.80	10.81	12.61	11.71	9.41	
Average	80.58	62.24	70.19	74.98	72.00	60.89	72.84	67.02	73.50	68.56	65.90	66.06	65.98	69.42	
	36.37%	24.66%	24.35%	27.69%	27.57%	17.50%	29.43%	29.46%	40.28%	28.96%	-0.77%	-0.18%	-0.48%		23.43%

Figure 3: Clean training results on the AURORA 2 database using an AMS/FMS front-end. The bottom row shows the relative improvement across different subsets, while the last column shows the relative improvement for different SNRs. Our results indicate a substantial improvement for certain tasks, especially for SNRs of 0 dB to 20 dB. Average recognition rates are improved for every task in sets A and B. The overall improvement in accuracy rates for clean training is 23.43%.

and can be easily explained in ζ-domain formulation [6]. In addition, this combination is implemented by an elegant comprehensive modulation spectra analysis module based on a log-derivative operation.

6. Conclusions

We have demonstrated that AMS features coupled with FMS features outperforms the reference front ends in the presence of different types of additive noise, while it performs as well in noise-free conditions. Note that this is possible because:

- We do not arbitrarily discard the spectral phase information, as is done by standard MFCC method;
- We do not arbitrarily discard the temporal envelope information, as is done by pyknogram, or IWAIF [10].

The proposed features are well motivated by the auditory system and signal processing principles, and the resulting algorithm is computationally attractive. Combining the amplitude modulation spectra and frequency modulation spectra, yields a comprehensive modulation spectra. We believe that the intelligibility of spoken language depends primarily on this comprehensive modulation spectral representation and thus provides a promising approach for future-generation ASR systems.

7. References

[1] S. Greenberg and T. Arai, "What are the essential cues for understanding spoken language?" *IEICE Trans. INF SYST*, vol. E87-D, no. 5, pp. 1059–1070, 2004.

[2] S. Greenberg and B. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech." in *ICASSP'97*, 1997, pp. 1647–1650.

[3] Y. Wang, J. Hansen, G. K. Allu, and R. Kumaresan, "Average instantaneous frequencies and average log-envelopes for asr with the aurora 2 database," in *Eurospeech'03*, Geneva, Switzerland, 2003, pp. 21–25.

[4] F. G. Zeng, K. Nie, G. S. Stickney, Y. Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc Natl Acad Sci U S A*, vol. 102, no. 7, pp. 2293–8, 2005.

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

[7] J. Anantharaman, A. Krishnamurthy, and L. Feth, "Intensity-weighted average of instantaneous frequency as a model for frequency discrimination," *Journal of the Acoustical Society of America*, vol. 94, pp. 723–729, Aug. 1993.

[8] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[9] S. M. Kay, "A fast and accurate single frequency estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1987–1990, Dec. 1989.

[10] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Jun. 1999.