Presented at ICASSP-97, Munich, vol. 3 pp. 1647-1650.

THE MODULATION SPECTROGRAM: IN PURSUIT OF AN INVARIANT REPRESENTATION OF SPEECH

*Steven Greenberg** and *Brian E. D. Kingsbury*†

*†International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704, USA
*Department of Linguistics
†Department of Electrical Engineering and Computer Sciences
University of California at Berkeley, Berkeley, CA 94704, USA
{steveng,bedk}@icsi.berkeley.edu

ABSTRACT

Understanding the human ability to reliably process and decode speech across a wide range of acoustic conditions and speaker characteristics is a fundamental challenge for current theories of speech perception. Conventional speech representations such as the sound spectrogram emphasize many spectro-temporal details that are not directly germane to the linguistic information encoded in the speech signal and which consequently do not display the perceptual stability characteristic of human listeners. We propose a new representational format, the modulation spectrogram, that discards much of the spectro-temporal detail in the speech signal and instead focuses on the underlying, stable structure incorporated in the low-frequency portion of the modulation spectrum distributed across critical-band-like channels. We describe the representation and illustrate its stability with color-mapped displays and with results from automatic speech recognition experiments.

1. INTRODUCTION

Human listeners are able to reliably decode phonetic information carried by the speech signal across a wide range of acoustic conditions and speaker characteristics. This perceptual stability is not captured by traditional representations of speech which tend to emphasize the minute spectro-temporal details of the speech signal. Speaker variability and distortions such as spectral shaping, background noise, and reverberation that typically exert little or no influence on the intelligibility of speech drastically alter such conventional speech representations as the sound spectrogram. This disparity between perceptual stability and representational lability constitutes a fundamental challenge for models of speech perception and recognition. A speech representation insensitive to speaker variability and acoustic distortion would be a powerful tool for the study of human speech perception and for research in speech coding and automatic speech recognition.

A key step for representing speech in a stable fashion is to focus on the elements of the signal encoding phonetic information. By suppressing phonetically irrelevant elements of the signal, the variability of the representation is reduced. There is significant evidence that much of the phonetic information is encoded by slow changes in gross spectral structure that characterize the low-frequency portion of the modulation spectrum of speech. In the late 1930's the developers of the vocoder found that it was possible to synthesize intelligible, high-quality speech based on a ten-channel spectral estimate with roughly 300-Hz resolution that was low-pass filtered at 25 Hz [1]. More recently, in a study on the intelligibility of temporally-smeared speech, Drullman and colleagues have demonstrated that modulations at rates above 16 Hz are not required for speech intelligibility [2]. A representation that focuses on slow modulations in speech also has compelling parallels to the dynamics of speech production, in which the articulators move at rates of 2–12 Hz [3], and to the sensitivity of auditory cortical neurons to amplitude-modulations at rates below 20 Hz [4].

2. THE MODULATION SPECTROGRAM

We have developed a new representational format for speech, the modulation spectrogram, that displays and encodes the signal in terms of the distribution of slow modulations across time and frequency. Although not intended as an auditory model, the representation captures many important properties of the auditory cortical representation of speech. The modulation spectrogram represents modulation frequencies in the speech signal between 0 and 8 Hz, with a peak sensitivity at 4 Hz, corresponding closely to the long-term modulation spectrum of speech. The modulation spectrogram is computed in critical-band-wide channels [5] to match the frequency resolution of the auditory system, incorporates a simple automatic gain control and emphasizes spectro-temporal peaks.

Figure 1 illustrates the signal processing procedure used to produce the modulation spectrogram. Incoming speech, sampled at 8 kHz, is analyzed into approximately critical-band-wide channels via an FIR filter bank. The filters are trapezoidal in shape, and there is minimal overlap between adjacent channels. Within each channel the signal envelope is derived by half-wave rectification and low-pass filtering (the half-power cutoff frequency is 28 Hz). Each channel envelope signal is downsampled to 80 Hz and then normalized by the average envelope level in that channel measured over the entire utterance. The modulations of the normalized envelope signals are analyzed by computing the FFT over a 250-ms Hamming window every 12.5 ms in order to capture the dynamic properties of the signal. Finally, the squared magnitudes of the 4-Hz coefficients of the FFTs are plotted in spectrographic format, with log energy encoded by color. Note that the display portrays modulation energy from 0–8 Hz. The effective filter response for the 4 Hz component is down by 10 dB at 0 and 8 Hz. A threshold is used in the energy-to-color mapping: the peak 30 dB of the signal is mapped to a color axis, while levels more than 30 dB below the global peak are mapped to the color for -30 dB. Bilinear smoothing is used to produce the final image.

3. REPRESENTATIONAL STABILITY

The modulation spectrographic representation of speech is more stable than the conventional spectrographic represen-

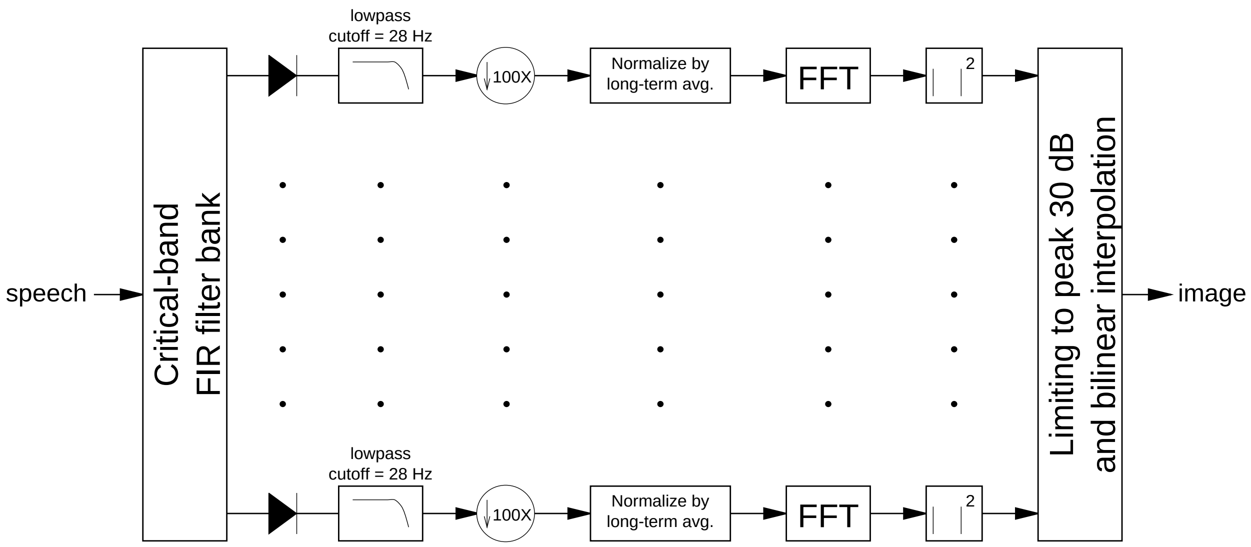


Figure 1. Diagram of the processing currently used to produce modulation spectrograms

tation in low signal-to-noise ratio (SNR) and reverberant conditions. Several processing steps contribute to this stability. The emphasis of modulations in the range of 0–8 Hz with peak sensitivity at 4 Hz acts as a matched filter that passes only signals with temporal dynamics characteristic of speech. The critical-band-like frequency resolution of the representation expands the representation of the low-frequency, high-energy portions of the speech signal, while the thresholding used in the color mapping emphasizes the spectro-temporal peaks in the speech signal that rise above the noise floor.

Figure 2 illustrates the stability of the modulation spectrographic representation of speech by comparing conventional narrow-band spectrograms[1] and modulation spectrograms for clean and noisy versions of the utterance "Tell me about the Thai barbecue." The noisy sample was produced by mixing the clean sample with pink noise at a SNR of 0 dB. Both the modulation spectrograms and narrow-band spectrograms cover approximately the same range of frequencies. However, the modulation spectrogram frequency axis is nonlinear in accordance with the human spatial frequency coordinates described in [5].

While the narrow-band spectrogram of the clean speech sample clearly portrays features of the speech signal such as onsets, formant trajectories, and harmonic structure, these features are all but lost in the narrow-band spectrogram of the noisy speech, where only a few spectro-temporal peaks stand out above the noise. In contrast to the spectrographic representation, the modulation spectrogram of the clean speech provides only a coarse picture of the energy distribution in time and frequency. The fine details, such as harmonicity, are not preserved. A comparison of the modulation spectrograms for the clean and noisy speech samples illustrates the stability of the representation. The major features of the modulation spectrogram observed for clean speech are preserved in the modulation spectrogram for speech embedded in intense levels of noise.

[1] The narrow-band spectrograms were computed by pre-emphasizing the speech, sampled at 8 kHz, with the filter $H(z) = 1 - 0.97z^{-1}$, then performing 512-point FFTs with a 64-ms Hamming window and a 16-ms window step. A lower threshold of -30 dB was applied in the energy-to-color mapping.

4. AUTOMATIC RECOGNITION BASED ON MODULATION SPECTROGRAPHIC FEATURES

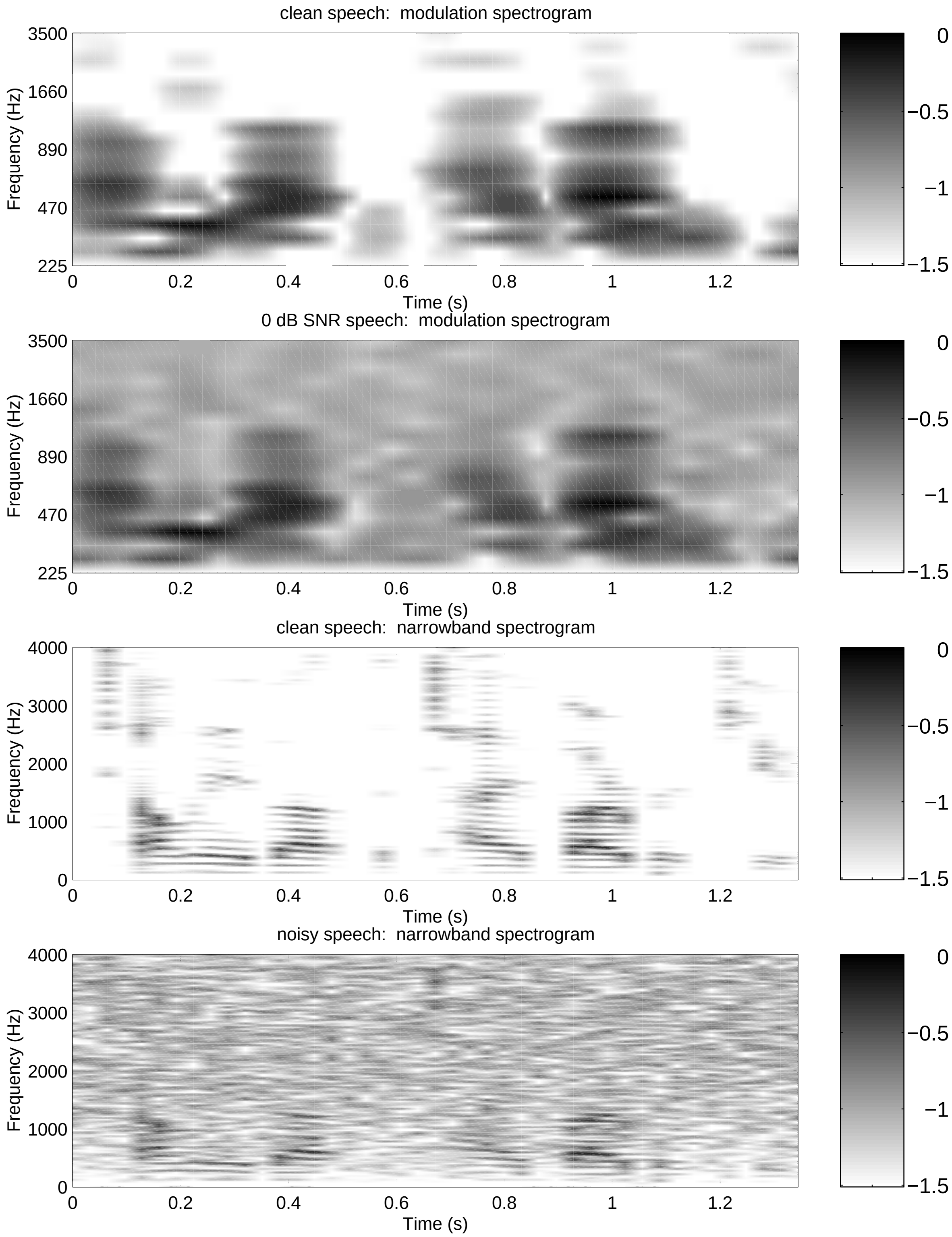
A similar representational stability is observed for reverberant speech, and has been demonstrated in tests with an automatic speech recognition system. In these tests, the performance of a hybrid hidden Markov model/multilayer perceptron (HMM/MLP) recognizer trained on clean speech is measured on clean and reverberant speech. The performance of recognizers using different front-end feature extraction methods is compared on the two test sets. Aside from the front-end processing, the recognizers are identical, using similarly-sized MLPs for phonetic probability estimation, and the same HMM word models and class bigram grammar for speech decoding. Further details on the recognition experiments are provided in [6].

Table 1 compares the performance of a recognizer using features based on the modulation spectrogram[2] with the performance of a recognizer that uses PLP features [7].

5. THE IMPORTANCE OF THE SYLLABLE IN SPEECH RECOGNITION

A central problem in speech science is the explication of the process by which the brain is able to go from sound to meaning. The traditional models posit a complex and somewhat arbitrary series of operations that advance from the acoustic signal to phonemic units, from phonemic units to words, and from words to meaning through a language's grammar. However, even a cursory examination of the statistical properties of speech indicates that the relationship between sound and symbol is anything but arbitrary. Instead, it appears that speech is organized into syllable-like units at both the acoustic and lexical levels, and that these

[2] The features are computed in quarter-octave bands, the modulation transfer function of the system is flat between 0 and 8 Hz, and no thresholding is applied to the output. The most important difference between these features is the absence of thresholding. If thresholding is used for automatic recognition, the recognition performance on clean speech degrades. However, the stability of the representational format is enhanced by some degree of thresholding.



These patterns are far more clearly delineated in the original color versions, which are available in the CD-ROM version of the proceedings and at `http://www.icsi.berkeley.edu/~bedk/ICASSP97_fig2_color.gif`

Figure 2. A comparison of the modulation spectrogram and narrow-band spectrogram for clean and noisy speech.

syllable-like units are the basis for lexical access from the acoustics of the speech signal.

It has been previously suggested that the broad peak at 4 Hz in the modulation spectrum corresponds to the average syllable rate [8]. Recently, we have found a more specific correlation between the distribution of low-frequency modulations in speech and the statistical distribution of syllable durations in spoken discourse [9]. It has also been shown that the concentrations of energy in the modulation spectrographic display correspond to syllabic nuclei. Thus, it appears that the modulation spectrogram robustly extracts information pertaining to the syllabic segmentation of speech, and that this information is of some utility in recognizing speech under adverse acoustic conditions [10].

Two common objections to a syllabic representation of English are the relatively complex and heterogeneous syllable structure of English and the large number of syllables required to cover the lexical inventory. However, these theoretical concerns are not borne out in practice. In spoken English, over 80% of the syllables are of the canonical CV, CVC, VC, and V forms, and many of the remainder reduce to these formats by processes of assimilation and reduction. In written English, only 12 syllables comprise over 25% of all syllable occurrences, and 339 syllables account for 75% of all syllable occurrences [11]. Spoken English employs a similarly reduced syllabic inventory [12, 13].

The robust encoding of syllabic structure by low-frequency modulations in speech, the sensitivity of the human auditory system to these modulations, and the statistics demonstrating that, in practice, English has a relatively simple syllabic structure and relies on a small subset of the possible syllables all support a model of real-time human speech perception in which auditory mechanisms parse the speech signal into syllable-like units and a core vocabulary of a few hundred, highly familiar syllables support efficient lexical access. This model is described in more detail in [14].

6. CONCLUSIONS

We have developed a new representational format for speech that captures many important properties of the auditory cortical representation of speech, namely selectivity for the slow modulations in the signal that encode phonetic information, critical-band frequency analysis, automatic gain control, and sensitivity to spectro-temporal peaks in the signal. These signal processing strategies produce a representation with greater stability in low SNR and reverberant conditions than conventional speech representations. The enhanced stability of the modulation spectrogram provides a potentially useful tool for research in human speech perception, speech coding, and automatic speech recognition.

	PLP		mod. spectrogram	
	clean	reverb	clean	reverb
substitutions	9.2%	33.5%	11.8%	37.7%
deletions	3.2%	33.8%	3.5%	25.9%
insertions	3.5%	2.7%	2.1%	2.6%
total	15.8%	70.1%	17.8%	66.1%

Table 1. Comparison of PLP and modulation spectrographic features for recognition of clean and reverberant speech. For the number of words in the test set (2426) the difference in performance on clean speech is not statistically significant, while the difference in performance on the reverberant speech is. Note that the performance improvement in reverberation for the modulation spectrographic features over PLP features comes almost entirely from a 23% relative reduction in the deletion rate.

REFERENCES

[1] Homer Dudley. Remaking speech. *JASA*, 11(2):169–177, October 1939.

[2] Rob Drullman, Joost M. Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *JASA*, 95(2):1053–1064, February 1994.

[3] Caroline L. Smith, Catherine P. Browman, Richard S. McGowan, and Bruce Kay. Extracting dynamic parameters from speech movement data. *JASA*, 93(3):1580–1588, March 1993.

[4] Christoph E. Schreiner and John V. Urbas. Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Research*, 21(3):227–241, 1986.

[5] Donald D. Greenwood. Critical bandwidth and the frequency coordinates of the basilar membrane. *JASA*, 33:1344–1356, 1961.

[6] Brian E. D. Kingsbury and Nelson Morgan. Recognizing reverberant speech with RASTA-PLP. In *Proc. ICASSP-97*. IEEE, 1997.

[7] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *JASA*, 87(4):1738–1752, April 1990.

[8] Tammo Houtgast and Herman J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility. *JASA*, 77(3):1069–1077, March 1985.

[9] Steven Greenberg, Joy Hollenback, and Dan Ellis. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *Proc. ICSLP-96*, 1996.

[10] Su-Lin Wu, Michael L. Shire, Steven Greenberg, and Nelson Morgan. Integrating syllable boundary information into speech recognition. In *Proc. ICASSP-97*. IEEE, 1997.

[11] Godfrey Dewey. *Relative Frequency of English Speech Sounds*, volume 4 of *Harvard Studies in Education*. Harvard University Press, Cambridge, 1923.

[12] Norman R. French, Charles W. Carter, Jr., and Walter Koenig, Jr. The words and sounds of telephone conversations. *The Bell System Technical Journal*, IX:290–325, April 1930.

[13] Steven Greenberg, Joy Hollenback, and Dan Ellis. The Switchboard transcription project. Technical report, International Computer Science Institute, 1997.

[14] Steven Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In William Ainsworth and Steven Greenberg, editors, *Proc. of the ESCA Workshop on the Auditory Basis of Speech Perception*, pages 1–8. ESCA, 1996.