ELSEVIER

Speech Communication 25 (1998) 117–132

SPEECH COMMUNICATION

Robust speech recognition using the modulation spectrogram

Brian E.D. Kingsbury [a,b,*], Nelson Morgan [a,b], Steven Greenberg [a,b]

[a] *International Computer Science Institute, Suite 600, 1947 Center Street, Berkeley, CA 94704-1105, USA*
[b] *University of California at Berkeley, Berkeley, CA 94720, USA*

Received 1 September 1997; received in revised form 1 January 1998; accepted 1 March 1998

Abstract

The performance of present-day automatic speech recognition (ASR) systems is seriously compromised by levels of acoustic interference (such as additive noise and room reverberation) representative of real-world speaking conditions. Studies on the perception of speech by human listeners suggest that recognizer robustness might be improved by focusing on temporal structure in the speech signal that appears as low-frequency (below 16 Hz) amplitude modulations in subband channels following critical-band frequency analysis. A speech representation that emphasizes this temporal structure, the "modulation spectrogram", has been developed. Visual displays of speech produced with the modulation spectrogram are relatively stable in the presence of high levels of background noise and reverberation. Using the modulation spectrogram as a front end for ASR provides a significant improvement in performance on highly reverberant speech. When the modulation spectrogram is used in combination with log-RASTA-PLP (log RelAtive SpecTrAl Perceptual Linear Predictive analysis) performance over a range of noisy and reverberant conditions is significantly improved, suggesting that the use of multiple representations is another promising method for improving the robustness of ASR systems. © 1998 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Die Performanz heutiger Spracherkennungssysteme wird stark von akustischen Interferenzen (z.B. additivem Rauschen und Hall) beeinträchtigt, die typisch für reelle Sprechbedingungen sind. Untersuchungen zur menschlichen Sprachwahrnehmung zeigen, daß die Robustheit von Spracherkennern möglicherweise durch Konzentration auf die zeitliche Struktur des Sprachsignals verbessert werden könnte, die als tieffrequente (unter 16 Hz) Amplitudenmodulation in den Frequenzkanälen der kritischen Bandanalyse auftritt. Es wurde eine Sprachsignalrepräsentation, das sogenannte Modulationsspektrogramm (modulation spectrogram), entwickelt, die diese zeitliche Struktur betont. Visualisierungen von Modulationsspektrogrammen zeigen eine relativ große Stabilität auch bei hochgradig verrauschter Sprache und bei starkem Hall. Die Verwendung des Modulationsspektrogramms als Vorverarbeitungsmethode in einem automatischen Spracherkenner liefert eine signifikante Verbesserung bei der Erkennung verhallter Sprache. Eine Kombination des Modulationsspektrogramms mit log-RASTA-PLP (log RelAtive SpecTrAl Perceptual Linear Predictive analysis) erzielt eine signifikante Verbesserung der Performanz bei einer Reihe von verschiedenen Rausch- und Hallbedingungen. Dies deutet darauf hin, daß eine Kombination verschiedener Signalrepräsentationen eine vielversprechende Methode zur Verbesserung der Robustheit automatischer Spracherkennungssysteme ist. © 1998 Elsevier Science B.V. All rights reserved.

[*] Corresponding author. Tel.: +1 510 642 4274 ext 143; fax: +1 510 643 7684; e-mail: bedk@icsi.berkeley.edu.

PII: S 0 1 6 7 - 6 3 9 3 (9 8) 0 0 0 3 2 - 6

Résumé

La performance des systèmes actuels de reconnaissance de la parole automatique est considérablement compromise par des niveaux d'interférence acoustique (telle que du bruit additif et de la réverbération) qui sont représentatifs de conditions réelles. Des études sur la perception de la parole par des êtres humains et une analyse des bandes fréquencielles critiques suggèrent que la robustesse des systèmes de reconnaissance pourrait être améliorée en se focalisant sur la structure temporelle du signal qui apparaît comme des modulations d'amplitude de basse fréquence (moins de 16 Hz) dans les sous-bandes. Une représentation de la parole soulignant cette structure temporelle, appelé "spectrogramme de modulation" (modulation spectrogram), a été développée. Des visualisations de la parole utilisant le spectrogramme de modulation sont relativement stables, malgré des niveaux élevés de bruit de fond et de réverbération. L'utilisation du spectrogramme de modulation apporte une amélioration de performance importante en présence de beaucoup de réverbération. La combinaison du spectrogramme de modulation avec le codage log-RASTA-PLP (log RelAtive SpecTrAl Perceptual Linear Predictive analysis) permet d'obtenir des améliorations significatives pour de nombreuses conditions de bruit et de réverbération. Ceci suggère que l'utilisation de plusieurs représentations est une méthode prometteuse pour améliorer la robustesse d'un système de reconnaissance de la parole automatique.

Keywords: Robust speech recognition; Reverberation

1. Introduction

Automatic speech recognition (ASR), which was once little more than a laboratory curiosity, is now being deployed in a wide range of tasks, including telephony and desktop dictation. As ASR technology emerges from the laboratory, however, fundamental shortcomings in the most widely used recognition algorithms and speech representations are becoming apparent. While humans are capable of reliably understanding speech across a broad range of environmental conditions and speaking styles, automatic recognition systems are much more fragile. Recognition methods developed on corpora of carefully-enunciated speech collected under laboratory conditions yield impressive results on "clean" speech, but their results are much less encouraging on material that contains realistic levels of acoustic interference such as background noise, spectral shaping, and room reverberation. Although research on channel robustness in ASR has produced techniques that greatly mitigate the deleterious effects of acoustic distortions on recognizer accuracy, these techniques generally fail when different, unanticipated distortions are encountered.

While it is likely that ASR systems will always work best on data similar to that on which they have been trained, considerable progress can still be made towards reducing the sensitivity of ASR systems to mismatches between training data and speech data received during actual operation. We believe that a key to progress in this area is the development of speech representations and recognition algorithms that better use information based on specific temporal properties of speech. These representations and recognition methods should be more robust because the temporal characteristics of speech appear to be less variable than static characteristics in the presence of acoustic interference. In this paper, we focus on the development of a speech representation, the modulation spectrogram, [1] that emphasizes the temporal structure in speech. We begin with a review of current ASR technology, and next describe characteristics of human speech perception that point to the importance of temporal information. Then, we discuss the implementation of a temporally-oriented speech representation and its use as

[1] It should be noted that, while the representation described here and the speech representation described in Kollmeier and Koch (1994), share the name "modulation spectrogram", they are quite distinct in the details of their signal processing and overall motivations.

a stable visual display of speech. Finally, we describe ASR experiments comparing its utility to that of the RASTA-PLP representations for recognizing clean, reverberant, and noisy speech.

2. Speech recognition by machines

Although the details of their implementations differ, most ASR systems employ the same basic approach to recognize speech. First, an incoming signal is analyzed to produce feature vectors that describe its short-time spectral envelope. Usually, the spectral analysis is based on a segment of roughly 20 ms, and the feature vectors are produced at a rate of around one every 10 ms. The feature vectors are then classified, with the output of the classifier being a set of distances between the feature vector and a set of phone classes. This classification is often made on the basis of features from a single frame, possibly augmented with differential features derived from three to five frames. Finally, a search based on dynamic programming or best-first search is used to find the most likely sequence of words, given the sequence of phone distances, a set of hidden Markov models characterizing the words in the recognizer's vocabulary, and a grammar that describes the structure of the word sequences that the recognizer is expected to process.

This framework is powerful and flexible, enabling the development of a broad array of recognizers, from digit recognizers that work over telephone lines to desktop dictation systems capable of recognizing tens of thousands of different words. However, recognizers based on this framework, including our own, still fall well short of human performance in natural conditions, including moderate to high levels of background noise, moderate or greater levels of room reverberation, and spontaneous speech. While this gap between human and machine performance has many causes, an especially important one is that the automatic systems do not take full advantage of the relatively invariant temporal encoding of phonetic information in speech because they focus almost exclusively on short segments of the acoustic signal.

3. Speech recognition by humans

A central result from the study of human speech perception is the importance of slow changes in the speech spectrum for speech intelligibility. These changes appear as low-frequency amplitude modulations with rates of below 16 Hz in subband signals following spectral analysis. The first evidence for this perspective emerged from the development of the channel vocoder in the early 1930s (Dudley, 1939), when Homer Dudley and his colleagues at Bell Labs found that they could synthesize high quality speech from spectral shape estimates that were lowpass filtered at 25 Hz. Since then, studies on speech intelligibility in noisy and reverberant rooms (Houtgast and Steeneken, 1973, 1985) and over communication channels that may impose nonlinear distortions (Steeneken and Houtgast, 1980) have demonstrated the link between intelligibility and the fidelity with which these slow modulations are transmitted. Direct perceptual experiments have shown that modulations at rates above 16 Hz are not required, and that significant intelligibility remains even if modulations at rates of 6 Hz and below are the only ones preserved (Drullman et al., 1994).

A second key to human speech recognition is the integration of phonetic information over relatively long intervals of time. In recognizing speech with inserted gaps of silence, listeners appear to be capable of associating sounds where onsets are separated by as much as 200 ms (Huggins, 1975). When speech is temporally decorrelated by analyzing it into critical bands, randomly time-shifting the bands with respect to one another, and resynthesizing the speech from the shifted bands, it is found that listeners are remarkably tolerant to the distortion. While the quality of the speech is clearly affected, the intelligibility of the speech is 90% for shifts of 100 ms and nearly 70% for shifts of 160 ms (Arai and Greenberg, 1998; Greenberg and Arai, 1998).

This focus on long time segments in human speech recognition may explain much of its robustness to acoustic interference. Basing speech recognition on modulations – changes in the signal – reduces sensitivity to relatively stationary forms of interference such as spectral coloration and some forms of noise. The use of relatively slow modulations and the integration of phonetic information over hundreds of milliseconds reduces the sensitivity of listeners to transient interference. It has also been shown that the slower modulations in speech are the least influenced by reverberation (Houtgast and Steeneken, 1973).

4. Incorporating temporal information into automatic speech recognition

It seems likely, then, that the robustness of ASR systems could be enhanced by using longer-time information, both at the level of the front-end speech representation, and at the level of phonetic classification. Indeed, many improvements in recognizer robustness have already been attained by using temporal information. Biphones, triphones, and other context-dependent phonetic units provide one means for modeling coarticulation and integrating information over longer periods of time. Many robust front-end processing methods such as delta features (Furui, 1986), RASTA and other modulation filtering methods (Langhans and Strube, 1982; Hirsch, 1988; Hermansky and Morgan, 1994), cepstral-time matrices (Milner and Vaseghi, 1995), and articulatory front ends (Ramsay and Deng, 1995) provide an enhanced representation of the dynamics of the speech signal.

As direct precursors of the modulation spectrogram, the log-RASTA-PLP and J-RASTA-PLP front ends are of particular interest. The two RASTA-PLP algorithms are temporal processing extensions to perceptual linear prediction (PLP) (Hermansky et al., 1985; Hermansky, 1990), an algorithm for producing a spectral shape estimate that reflects properties of the human auditory system. In PLP, incoming speech is segmented into overlapping, fixed-length frames that are typically around 20 ms in duration with about a 10 ms overlap. A windowed FFT is performed on each frame, and the square of the magnitude of each component is computed to produce a power spectrum. Each power spectrum is convolved with a set of overlapping, trapezoidal filters that are equally spaced on the Bark scale, to approximate critical-band frequency resolution. The critical-band power spectra are mapped to an approximation of perceptual loudness via a static equal-loudness weighting and cube-root compression. The perceptually-warped power spectra are then approximated by an autoregressive all-pole model, and the resulting linear prediction coefficients are converted into cepstral coefficients. The first eight to twelve cepstral coefficients plus the energy term comprise the PLP output.

The general RASTA-PLP algorithm (Hermansky and Morgan, 1994) is an extension to PLP that attempts to model the sensitivity of human speech perception to preceding context (Summerfield et al., 1984) and apparent insensitivity to absolute spectral shape (Lea and Summerfield, 1994). This is accomplished by interposing a compressive nonlinearity, a bandpass filter, and an expansive nonlinearity between the output of the critical-band filtering and the input to the loudness approximation in PLP. The bandpass filter is a differentiator followed by a leaky integrator, and passes modulations between 1 and 12 Hz. log-RASTA-PLP, which increases the robustness of ASR systems to spectral shaping of speech, uses a logarithm for the compressive nonlinearity and an exponential for the expansive nonlinearity. J-RASTA-PLP (Morgan and Hermansky, 1992), which increases the robustness of ASR systems to joint spectral shaping and additive noise, uses the function $y = \ln(1 + Jx)$ for the compressive nonlinearity and $x = e^y/J$ for the expansive nonlinearity. The compressive nonlinearity is approximately linear for small values of Jx and approximately logarithmic for large values of Jx. During recognition, the J parameter is varied in inverse proportion to an estimate of the noise power in the incoming speech signal, so that channels with low power relative to the estimated noise are processed to suppress the noise, while channels with high power relative to the estimated noise are processed to suppress spectral shaping.

Our research group is currently pursuing two approaches to making better use of temporal information in ASR systems. First, we are investigating the development of new front-end representations for speech recognition that attempt to capture and represent the temporal structure of speech using signal processing methods suggested by studies of human speech perception. This work is being done in a flexible framework that is a generalization and extension of the RASTA-PLP algorithms. This framework uses an FIR filterbank instead of the short-time Fourier transform for frequency analysis, performs automatic gain control and modulation filtering separately, and uses different modulation filters than RASTA-PLP. We have used both visualization and ASR experiments to develop these representations (Greenberg and Kingsbury, 1997; Kingsbury and Morgan, 1997; Kingsbury et al., 1997). In parallel work described elsewhere (Wu et al., 1998; Greenberg, 1997), the use of syllables as a fundamental unit of speech recognition is being explored. The use of multiple streams of information, either from different front-end representations or from recognizers using different units of recognition, is central to both lines of inquiry.

5. Visualizing speech with the modulation spectrogram

We began our current study of speech representations that emphasize temporal structure with the development of a visual speech representation insensitive to room reverberation and noise. The result of this development, which we call the modulation spectrogram, displays the distribution of slow modulations in the speech signal across time and frequency. Although it is not a detailed model of auditory processing, a number of processing steps are incorporated into the modulation spectrogram that capture key aspects of the auditory cortical representation of speech (cf. Schreiner and Urbas, 1988). Namely, it displays amplitude modulations at rates of 0–8 Hz, with a peak sensitivity at 4 Hz, in roughly critical-band-wide channels, and includes automatic gain control and peak enhancement mechanisms.

Fig. 1 illustrates our implementation of the modulation spectrogram. A spectral analysis into roughly critical-band-wide channels is performed on an incoming speech signal, sampled at 8 kHz, using an eighteen-channel FIR filterbank. The filters have a roughly trapezoidal magnitude response, with minimal overlap between adjacent channels. The filter bandwidths are set according to a slightly modified version of Greenwood's cochlear place-to-frequency mapping function (Greenwood, 1961). In each channel, an ampli-

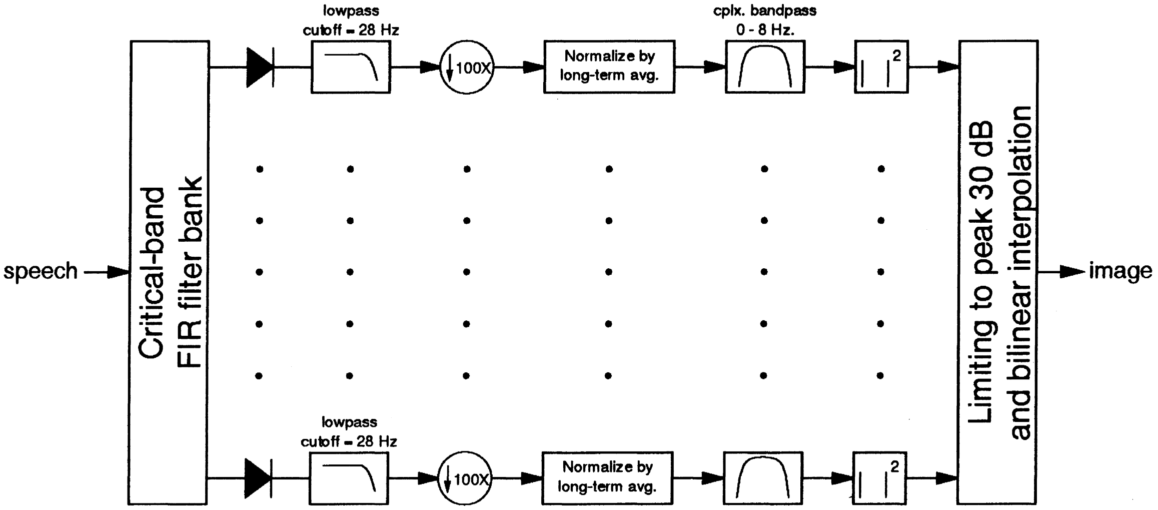

Fig. 1. Diagram of the signal processing used to produce modulation spectrographic displays of speech. The processing attempts to capture, in a simple manner, key aspects of the auditory cortical representation of speech. These aspects are critical-band frequency resolution, automatic gain control (adaptation), sensitivity to low-frequency amplitude modulations and enhancement of spectro-temporal peaks.

tude-envelope signal is computed by half-wave rectification and lowpass filtering with a cutoff frequency of 28 Hz. Each amplitude envelope signal is then downsampled by a factor of 100 to reduce subsequent processing time and to match with the typical data rate into the phonetic classification stage of an ASR system. Each envelope signal is normalized by computing the average level over an entire utterance and dividing by this magnitude. The slow modulations in each normalized envelope signal are then analyzed by filtering the signal through a complex bandpass filter and taking the log of the magnitude of the output. The filter response is a Hamming window modulated by a 4 Hz complex exponential, so the result of the filtering and magnitude calculation is equivalent to computing the FFT over the signal, windowed with a 250-ms Hamming window, and computing the log magnitude of the 4-Hz component. The filter has a passband of 1.25–6.72 Hz, and passes significant modulation energy in the 0–8 Hz range. The log magnitudes are then plotted in spectrographic format, with a normalization and thresholding applied such that the peak level over all channels is set to 0 dB, and levels more than 30 dB below this peak are all set to −30 dB. Bilinear smoothing is used to produce the final image.

Displays of clean and corrupted versions of the same utterance appear more alike in the modulation spectrographic format than in the traditional spectrographic format, and the syllabic structure of the speech is emphasized in the modulation spectrographic display. Fig. 2 illustrates these properties of the modulation spectrogram. It shows modulation spectrograms and wideband spectrograms of clean and noisy versions of the telephone-bandwidth utterance “seventy-three” produced by a male speaker. The noisy version of the utterance was produced by mixing the clean utterance with pink noise at an overall signal-to-noise ratio of 0 dB. The wideband spectrograms were produced by filtering the input signal, sampled at 8 kHz, with a pre-emphasis filter with a transfer function $H(z) = 1 - 0.94z^{-1}$, then computing 256-point FFTs over an 8 ms Hamming window, zero-padded to a length of 32 ms, once every 2 ms. To facilitate comparison with the modulation spectrogram, the peak value in each wideband spectrogram was normalized to 0 dB, and levels below −30 dB were set to −30 dB. Note that Fig. 2 is much clearer in color. A color version of this figure is available on the World-Wide Web at `http://www.icsi.berkeley.edu/real/specom_fig2_color.gif`.

The wideband spectrogram of the clean speech portrays a significant amount of spectro-temporal detail. Sharp onsets and pitch pulses are clearly visible, as are formant trajectories. In the wideband spectrogram of the noisy speech, however, only a general indication of the energy distribution of the speech, including formant trajectories, is evident. The modulation spectrogram of the clean speech is much less detailed than the wideband spectrogram. It provides a rough picture of the energy distribution over time and frequency, but there is no representation of harmonic structure, and pitch pulses and onsets are smeared out by the modulation filtering. The gross distribution of energy over time and frequency, however, is the information that is best preserved in the presence of acoustic interference. As a result, the modulation spectrogram of speech is more stable than the wideband spectrogram.

The vertical black bars in the spectrograms in Fig. 2 indicate the onsets of the syllables [s eh], [v ih n dcl], [d iy], and [th r iy] that make up the utterance. Most of the energy that appears in the modulation spectrographic display falls between these onsets, corresponding to syllabic nuclei.

This enhancement of the segmental structure of speech and representational stability are produced by several processing steps. The filtering that emphasizes modulations at rates of 0 Hz to 8 Hz, with peak sensitivity at 4 Hz, acts as a matched filter for signals with temporal structure characteristic of speech. Modulations at rates around 4 Hz correspond to syllables (Houtgast et al., 1980; Greenberg et al., 1996). The critical-band-like frequency resolution expands the representation of the low-frequency, high energy portion of the speech signal, and the thresholding emphasizes spectro-temporal peaks in the signal that rise above the noise floor.

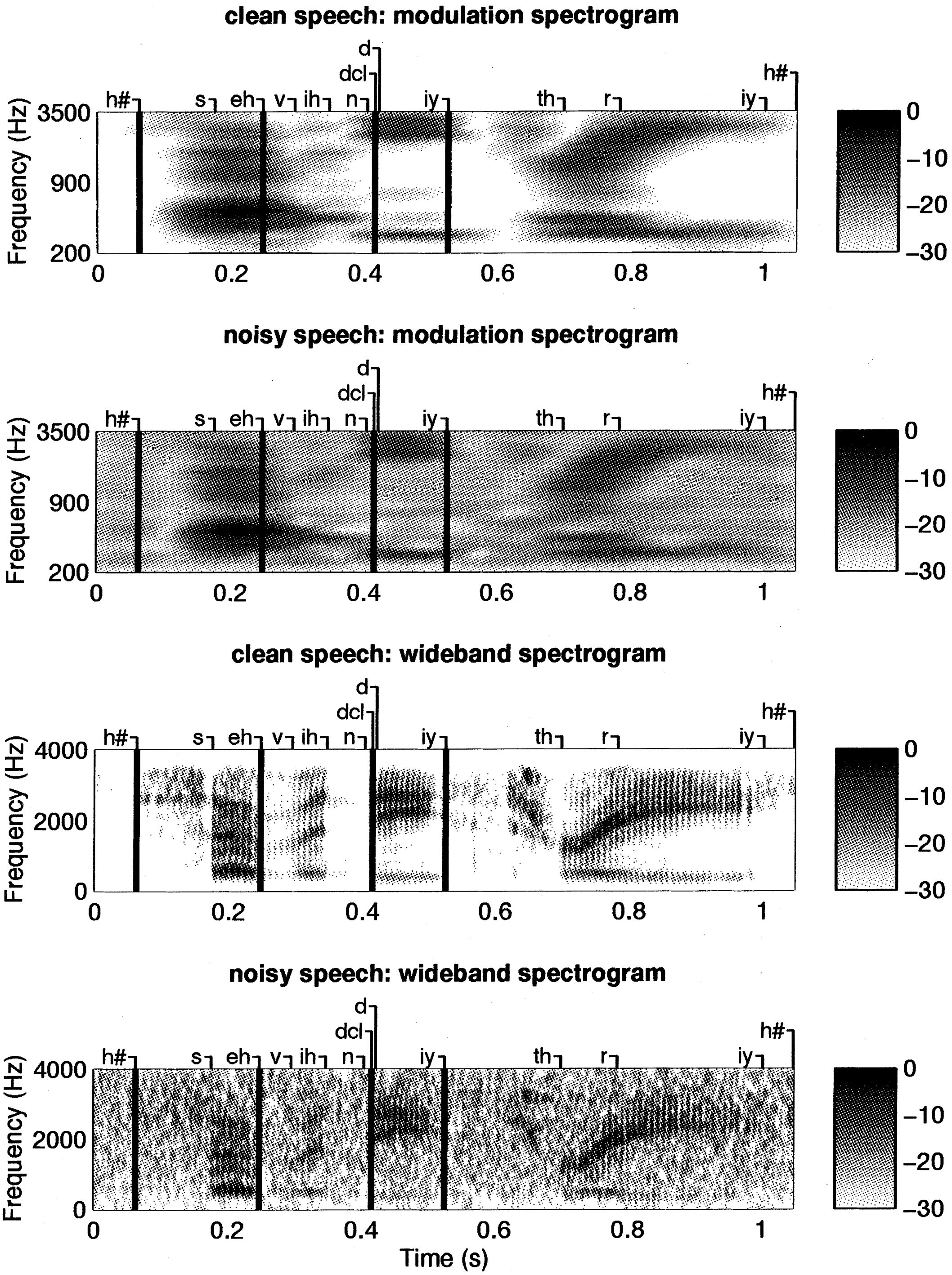

Fig. 2. Modulation spectrograms and wideband spectrograms of clean and noisy (0 dB SNR pink noise) versions of the telephone-bandwidth utterance "seventy-three". The black vertical bars mark syllable onsets. A time-aligned phonetic labeling using ARPAbet symbols is provided at the top of each spectrogram. The symbol "dcl" marks a d-closure. This figure is much clearer in color. A color version of it is available on the World-Wide Web at `http://www.icsi.berkeley.edu/real/specom_fig2_color.gif`.

6. Automatic speech recognition with the modulation spectrogram

The stability of the modulation spectrogram is useful not only for producing relatively invariant displays of speech, but also for more robust ASR. Table 1 shows the results of a set of experiments comparing the performance of recognizers that use PLP, log-RASTA-PLP, J-RASTA-PLP, or the modulation spectrogram, either singly or in combination, on six different test sets: a clean set, a reverberant set produced by convolving all the clean test set utterances with a real room impulse response having a reverberation time of 0.5 s and a direct-to-reverberant energy ratio of 0 dB, and four noisy test sets generated by adding pink noise from the NOISEX database to the clean utterances at SNRs of 30, 20, 10 and 0 dB. Because the recognizers are trained only on clean speech, these results illustrate the relative invariance of the different front ends in the presence of acoustic interference. Note that all recognition results are reported in terms of word error rate, which is the sum of the number of word substitutions, deletions and insertions divided by the total number of words in the test set. Substitutions, deletions and insertions are found by aligning the reference and recognized word strings using a dynamic programming routine.

The experiments summarized in Table 1 were performed using a connected-word, hybrid hidden Markov model/multilayer perceptron (HMM/MLP) recognizer in which phone probabilities are estimated from acoustic features using an MLP and speech decoding is performed with a Viterbi search (Bourlard and Morgan, 1994). A variant of the modulation spectrographic features, described in more detail below, was compared against PLP, log-RASTA-PLP, and J-RASTA-PLP features supplemented with delta features. The PLP, log-RASTA-PLP and J-RASTA-PLP front ends used a 25 ms window for spectral analysis, and produced a vector of nine cepstral and nine delta-cepstral coefficients (including the energy term) at a rate of one vector every 10 ms. The delta-cepstral features were calculated from the cepstral features by regression over a nine-frame window centered on the current frame. The modulation-spectrographic front end produced an output vector of thirty features, produced by applying two different modulation filters to 15 channels, at a rate of one every 10 ms. The MLP phonetic classifier for the PLP, log-RASTA-PLP, and J-RASTA-PLP features had 162 input units, corresponding to the current frame, the four previous frames and the four next frames, 488 hidden units, and 56 output units, while the MLP phonetic classifier for the modulation spectrogram features had 270 input units, also corresponding to the current frame, the four previous frames and the four next frames, 328 hidden units, and 56 output units. A multiple-pronunciation lexicon with simple duration modeling and a backoff bigram grammar was used for recognition. To ensure a good match between the

Table 1
Percent word error rates for clean, moderately reverberant (T60 = 0.5 s) and noisy speech using PLP, log-RASTA-PLP, J-RASTA-PLP and the modulation spectrogram. Statistically significant differences range from 0.9 on the clean test to 1.7 on the 0 dB SNR test. The criterion for statistical significance is $p < 0.5$ using a one-tailed significance test based on a normal approximation to a binomial distribution

Experiment	Features	Clean	Reverb	30 dB SNR	20 dB SNR	10 dB SNR	0 dB SNR
Baseline	PLP	6.4	37.6	28.3	43.5	60.7	78.8
	Log-RASTA	6.4	26.0	11.4	16.3	27.8	51.6
	J-RASTA	6.6	27.9	15.6	23.5	35.7	54.4
	Mod. spec.	8.5	27.3	14.6	22.9	38.7	61.5
Combined probabilities	PLP & Log-RASTA	5.7	26.9	15.9	26.6	43.7	67.3
	PLP & Mod. spec.	6.1	29.1	20.5	36.1	53.5	71.3
	Log-RASTA & mod. spec.	5.5	20.1	10.4	14.7	23.2	44.7
Double num. of MLP parameters	Log-RASTA	5.9	26.1	10.8	16.4	29.7	54.7
	Mod. spec.	8.2	27.9	14.4	22.1	39.8	65.3

MLP phone probability estimator and the lexicon, an embedded Viterbi training procedure was used in which a trained recognizer was used to label the training set via a forced-alignment procedure and a new recognizer was trained on the new labels.

The modulation-spectrographic features used for these recognition experiments differ from those used to produce speech displays, such as those in Fig. 2. Some changes do not significantly alter the representation. For example, the features used for the recognition experiments employed a filterbank with quarter-octave frequency resolution and modulation filters based on a Kaiser window rather than a Hamming window. Other changes are more significant. The recognition features used no thresholding of modulation energy, while the features used for producing displays had a threshold of 30 dB below a global peak level. Also, the recognition features used cube-root compression of the output instead of log compression and used two modulation filters corresponding to the real and imaginary components of the complex modulation filter used to produce the visual displays. These changes were inspired by a series of ASR experiments described in Section 7.

The experiments were performed on a subset of the Numbers corpus, a collection of speech data distributed by the Center for Speech and Language Understanding at the Oregon Graduate Institute. The Numbers corpus consists of spontaneous, continuous utterances collected over the telephone and digitized at an 8-kHz sampling rate with a 16-bit A/D converter. The subset used for these experiments does not contain any words cut off by the automatic segmentation used in the data collection, nor does it contain any out-of-vocabulary words. All of the material was phonetically transcribed by experienced transcribers. The vocabulary comprises thirty words and two filled pauses, and is restricted to numbers, including confusable sets such as “seven,” “seventeen” and “seventy.” A sample utterance from the corpus is “thirty-one thirty-five.” For the recognition experiments, a training set containing 3590 utterances (a total of 13873 words) and a disjoint test set of 1206 utterances (a total of 4673 words) was used. Utterances in the subset range in length from one to ten words, with a mean of 3.9 words. The distribution of utterance durations is roughly Gaussian, with a mean of 1.7 s and a standard deviation of 0.7 s.

Two sets of experiments were performed. In the baseline experiments, a recognizer was trained on clean speech and its performance on the six different test sets measured. In a second set of experiments, phone probability estimates from pairs of MLPs trained on PLP, log-RASTA-PLP, and the modulation spectrogram were combined and then used for recognition. The MLPs used as phonetic classifiers estimate posterior probabilities, that is, the probability of a phone given the acoustic data. These posterior probabilities are converted to scaled phone likelihoods by dividing by the phone priors, estimated from a labeling of the MLP training set. The combination of phone probabilities is accomplished by multiplying together scaled phone likelihoods. This method works best when the MLPs tend to make independent errors and produce relatively flat output distributions for difficult-to-classify inputs. Because recognizers using these combined probabilities effectively have twice as many parameters in their phonetic classifiers, the results of these tests are compared with recognizers that use a single feature set and have twice as many weights in their MLP phonetic classifiers to provide a fair reference.

This combination of evidence occurs on a frame-by-frame basis. Other combination methods have also been used successfully in combining the RASTA-PLP and modulation spectrogram features, including combination on a whole-utterance basis via rescoring of *N*-best lists (Wu et al., 1998) and combination on a syllable-by-syllable basis using a two-level dynamic programming search (Wu, 1998). All three methods achieve similar performance on clean speech. On reverberant speech the frame-by-frame and whole utterance methods achieve comparable levels of performance, and the syllable-by-syllable method performs significantly better, with a word error rate of 17.6% on the test set used in the experiments in Table 1.

In the baseline experiments the log-RASTA-PLP features give the best performance across all conditions, including reverberation. The J-RASTA-PLP and modulation spectrographic features give roughly equal performance on the reverberant, 30-dB SNR, and 20-dB SNR cases, with J-RASTA-PLP outperforming the modulation spectrographic features on the clean, 10-dB SNR, and 0-dB SNR cases. Except on clean speech, where it matches the performance of log-RASTA-PLP, the simple PLP front end exhibits the worst performance. In this experiment, the front ends that incorporate some form of temporal processing emphasizing slow modulations, log-RASTA-PLP, J-RASTA-PLP, and the modulation spectrogram, are more robust to acoustic interference than PLP, the one front end that performs no such modulation filtering. We are not certain why log-RASTA-PLP outperforms the modulation spectrogram and J-RASTA-PLP matches it on these tasks, when earlier experiments have shown that J-RASTA-PLP has better performance than log-RASTA-PLP on noisy speech and that modulation spectrogram features are better than log-RASTA-PLP and J-RASTA-PLP on highly reverberant speech. In any case, these results should be interpreted with care because the more sophisticated, multiple-pronunciation lexicon with duration modeling used in the most recent tasks was trained on a recognition that used log-RASTA-PLP features with deltas, while the earlier experiments used a lexicon that was not specialized for any of the feature sets.

When phone probability estimates from pairs of MLPs are combined and used for recognition, the combination of the modulation spectrogram and log-RASTA-PLP is the only one to provide a significant improvement over the baseline log-RASTA recognition. This improvement is achieved on the reverberant, 20-dB SNR, 10-dB SNR, and 0-dB SNR tests. Performance on the clean and 30-dB SNR tests is also improved, but the difference is not statistically significant. The performance of the recognizers using log-RASTA-PLP or modulation spectrogram features and twice as many MLP weights are either the same as or slightly worse than their respective baseline recognizers.

The PLP, log-RASTA-PLP, and modulation spectrogram features have significantly different temporal properties. PLP does no temporal processing, while log-RASTA-PLP strongly enhances the onsets of sounds (cf. Fig. 8 in Hermansky and Morgan, 1994) and the modulation spectrogram enhances syllabic nuclei, as shown in Fig. 2. This may explain the success of the log-RASTA-PLP and modulation spectrogram combination. Both feature sets have enhanced robustness due to their focus on slow modulations in speech, but they tend to emphasize distinct temporal portions of the signal, and are thus somewhat independent.

7. Optimizing the modulation spectrogram for automatic speech recognition

The modulation spectrogram was originally developed as a *visual* representation of speech. In general, representations of speech intended for visual display do not give the best performance when they are used as representations for ASR. Therefore, before performing the experiments described in Section 6 the modulation spectrographic representation was optimized for use in ASR. A series of recognition experiments were performed in which different variants on the modulation spectrogram were tested on clean and reverberant speech. The best variant representation was then used for the recognition experiments described above.

To accomplish a faster turnaround, these preliminary recognition experiments used a smaller subset of the Numbers corpus, with a training set of 875 utterances (a total of 3315 words) and a test set of 657 utterances (a total of 2426 words). This subset used a slightly larger vocabulary of 36 words and no filled pauses. The vocabulary of this subset included all the words in the subset described in Section 6, plus the words, “a”, “and”, “dash”, “double”, “hyphen” and “thousand”. Two versions of the test set were used: a clean set and a reverberant set that was generated by digitally convolving all of the utterances in the clean set with an impulse response designed to match the gross acoustic characteristics of a highly reverberant hallway approximately 6.1 m long, 2.4 m high and 1.7 m wide with concrete walls, floor and ceiling. Reverberation times in different frequency bands were estimated from a simultaneous recording of speech produced in the hallway using

Table 2
Estimated subband reverberation times for a highly reverberant hallway

Freq. band (Hz)	Reverb. time (s)
0–250	3.1
250–500	2.6
500–1000	2.2
1000–2000	1.6
2000–4000	1.4

a head-mounted, close-talking microphone and an omnidirectional microphone located on the floor, roughly 2.5 m from the speaker. Table 2 summarizes the estimated reverberation times. A sample of Gaussian white noise was analyzed into identical subbands, and each noise band was modulated with a decaying exponential envelope matched to the estimated reverberation times. The modulated noise bands were added together to produce the reverberant tail of the impulse response, while the early reflections were estimated using a time-domain image expansion simulation. The ratio of direct to reverberant energy was set manually to match the recording from the omnidirectional microphone. The resulting impulse response has a gross reverberation time of about 2 s and a direct-to-reverberant energy ratio of −16 dB.

These experiments also used a simpler recognition system than the earlier experiments. The PLP, log-RASTA-PLP, and J-RASTA-PLP representations were not supplemented with delta features and the energy term (zeroth-order cepstral coefficient) was not included in the feature vector. Also, all four front ends produced one feature vector every 12.5 ms instead of one vector every 10 ms. To compensate for the lack of delta features, a longer acoustic context window of fifteen frames was used as input to the MLP phonetic classifiers. The MLPs were roughly 15% smaller, with the PLP, log-RASTA-PLP, and J-RASTA-PLP classifiers containing 120 input units, corresponding to the current frame, the previous seven frames and the next seven frames, 512 hidden units, and 56 output units, and the modulation spectrogram classifier containing 225 input units, corresponding to the current frame, the previous seven frames and the next seven frames, 320 hidden units, and 56 output units. For recognition, a single-pronunciation lexicon of word models with no duration modeling and a class bigram grammar were used. Embedded Viterbi training was used to ensure a good match between the MLP phone probability estimator and the lexicon.

7.1. Initial recognition experiments

In the first set of experiments, summarized in Table 3, three different types of experiment were carried out. In the baseline experiment, a recognizer for PLP, log-RASTA-PLP, J-RASTA-PLP, and the version of the modulation spectrogram used to produce the visual displays, was trained on the clean training set then tested on the clean and highly reverberant test sets. In the second type of experiment, phone probability estimates from

Table 3
Percent word error rates for clean and highly reverberant (specified in Table 2) speech using the PLP, log-RASTA-PLP, J-RASTA-PLP and modulation spectrogram features. Statistically significant differences range from 1.6 for the clean test with combined probabilities to 2.2 for the reverberant test with the baseline recognizers. The criterion for statistical significance is $p < 0.5$ using a one-tailed significance test based on a normal approximation to a binomial distribution

Experiment	Features	Clean error	Reverb error
Baseline	PLP	15.8	70.1
	Log-RASTA	14.5	72.7
	J-RASTA	15.1	77.3
	Mod. spec.	30.1	65.2
Combined probabilities	PLP & Log-RASTA	11.9	68.7
	PLP & Mod. spec.	13.6	64.1
Train on reverb	PLP	72.5	48.5
	Mod. spec.	45.4	43.5

MLPs trained in the baseline experiment on different feature sets were combined and used for recognition. In the third type of experiment, recognizers using the modulation spectrogram and PLP features were trained on a reverberant version of the training set, then tested on the clean and reverberant test sets.

On the baseline clean tests, there is no significant difference in performance between the PLP, log-RASTA-PLP and J-RASTA-PLP recognizers, while the modulation spectrogram recognizer performs significantly worse. On the reverberant test, however, there are significant differences between all scores, with the modulation spectrogram recognizer performing best. When phone probabilities from the PLP and log-RASTA-PLP MLPs are combined, the best recognition performance on the clean test set is attained, and performance on the reverberant test set is significantly improved. Combining probabilities from the PLP and modulation spectrogram MLPs gives the second-best performance on the clean test set and the best performance on the reverberant test set. The improvement in performance observed when the probabilities are combined is not the result of an increased number of parameters in the phonetic classifier. Using only one feature set and doubling the number of weights in the MLP does not improve recognizer performance. The difficulty of the reverberant test set is illustrated by the results of the third experiment in which reverberant speech was used for training. Even when the training and testing conditions for the recognizer are matched, the word error rates from the test set are still in the 44–50% range.

We also performed a human listening experiment in order to obtain a measure of human recognition performance under comparable conditions. For this experiment, three subjects who were native speakers of American English, had no known hearing impairments, and were experienced at phonetic transcription of speech were asked to perform word-level transcription of the reverberant test set. The subjects were given a list of the words present in the test set, to provide them with the same knowledge available to the automatic recognition systems. The order of presentation of the utterances was randomized to prevent the listeners from learning speaker characteristics. The subjects were allowed to listen to each utterance as many times as they wished, and were provided with an initial training on ten utterances from the reverberant training set to familiarize them with the task. During transcription of the test set, the subjects had no feedback on their transcription accuracy. The utterances were produced by the 16-bit D/A converter in a SPARC-5 workstation at a sampling rate of 8 kHz, and were presented through headphones at a comfortable listening level in a quiet office. The listeners' transcriptions were scored by the program used for scoring machine recognition results.

The human listeners had considerably less difficulty on the reverberant test set than the automatic recognizers did. The human listeners had an average word error rate of 6.1%, roughly ten times less error than the best automatic system. Although transcription accuracy on the clean test set was not measured, we note that the word error rate for humans transcribing utterances from the TI DIGITS corpus was 0.105% (Leonard, 1984).

7.2. *Finding the important processing steps*

The results of these initial ASR experiments were encouraging, with the modulation spectrogram features giving the best performance on highly reverberant speech and when combined with PLP giving good performance on clean speech. However, it would be preferable for the modulation spectrogram features to give good performance on clean speech without combining them with other feature sets. We therefore performed a second series of experiments aimed at understanding which steps in the modulation spectrogram processing were important for robustness in reverberation and for improving the performance on clean speech.

First, different variants on the modulation spectrogram in which some processing steps were omitted were used as ASR front ends, and the performance of recognizers using these variant representations was measured on clean and reverberant speech. The steps that were optionally omitted in the variants were

Table 4
Percent word error rates for clean and highly reverberant speech using variants of the modulation spectrogram features. The presence of an 'X' in a cell indicates that the corresponding processing step was included in the feature calculation. Statistically significant differences range from 1.6 for the best clean tests to 2.2 for the reverberant tests. The criterion for statistical significance is $p < 0.5$ using a one-tailed significance test based on a normal approximation to a binomial distribution

Env. norm	Cplx filt	Glob peak	Out. thrsh	Clean error	Reverb error
X	X	X	X	30.1	65.2
X		X	X	30.6	67.8
X	X	X		17.5	66.1
X		X		13.6	69.9
	X			18.3	68.8
				16.1	73.5

1. the normalization of the envelope signals by their long-term averages,
2. the complex filtering that measures modulation levels in the 0–8 Hz range,
3. the normalization of the global peak to 0 dB, and
4. the thresholding of levels more than 30 dB below the global peak.

When all of these steps are omitted the resulting front end produces the log of the squared subband envelope signals – a simple filterbank-based power spectrum. Table 4 summarizes the results of these experiments.

The most significant result is that thresholding, which is vital for the production of stable visual displays of speech in low signal-to-noise ratio or highly reverberant conditions, has deleterious effects on the performance of automatic recognition systems. Eliminating the thresholding reduces the error rate on the clean test set by about half, while only slightly increasing the error rate on the reverberant test set. Because the thresholding is based on a global peak level, it is likely that it impairs representation of low-energy segments of speech. It is also clear that the complex filtering operation is vital for good recognition performance on the reverberant test set. Comparing the recognition scores of all pairs of recognizers that differ only by the presence or absence of the complex modulation filter, the recognizers with the filter perform significantly better on the reverberant test set than the recognizers without the filter.

Second, the role of the complex modulation filter was examined. Although the complex filter has the advantage of producing a strictly positive output that may be converted to a decibel scale, it requires twice as much computation as a real filter of the same length. Also, the temporal resolution of the complex filter is lower than that of either its real or imaginary part, as illustrated in Fig. 3. We therefore investigated the use of the real part of the filter, the imaginary part of the filter, or both parts without the magnitude calculation. Because the individual parts of the filter could produce negative outputs, the log compression was replaced with a cube-root compression. When the outputs of both the real and imaginary filters were used, the front end produced an output vector with thirty spectral features, so an MLP phonetic probability estimator with 450 input units, 176

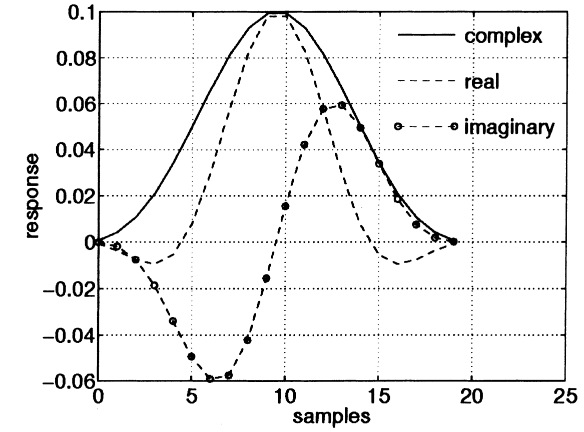

Fig. 3. The magnitude of the impulse response of the complex modulation filter is compared with the impulse responses of its real and imaginary components. Note that the temporal response of the complex filter is broader than that of either component.

Table 5
Percent word error rates for clean and highly reverberant speech using the modulation spectrogram features computed with different compression methods and modulation filters. Statistically significant differences range from 1.8 for the best clean tests to 2.2 for the reverberant tests. The criterion for statistical significance is $p < 0.5$ using a one-tailed significance test based on a normal approximation to a binomial distribution

Filter	Compress.	Clean error	Reverb error
Complex	Log	17.8	63.8
Complex	Cube root	17.8	67.2
Real	Cube root	16.5	68.3
Imaginary	Cube root	17.3	64.3
Real and imag.	Cube root	14.7	63.5

hidden units, and 56 output units, was used. This MLP had roughly the same number of weights as the MLPs used for the other recognizers. Table 5 summarizes the results of these experiments.

When the complex modulation filter was used with cube-root compression, the performance on the clean test set was unchanged, but the performance on the reverberant test set was significantly degraded. This may be a result of the small degree of compression at high signal levels afforded by the cube-root function relative to log-based compression. When cube-root compression is used there is no significant difference in performance between the complex filter and its real part on either the clean or reverberant test set. Using the imaginary part of the filter also has no significant effect on performance for the clean test set, but it does provide a significant improvement on the reverberant test set. This is probably due to the enhancement of changes in the amplitude envelope produced by the imaginary part of the filter. Finally, using the outputs of the real and imaginary parts of the filter together provides the best performance on the clean test set. Performance on the reverberant test set is equivalent to that of the complex filter with log compression. Using the outputs of both filters is similar to using both spectral and delta-spectral features. The real part of the filter is a smoothing filter, while the imaginary part is a differentiator.

8. Conclusions

Focusing on the temporal structure in speech is a promising direction for work on robust speech representations. Even a very simple set of processing steps that capture basic aspects of the auditory cortical representation of speech – critical-band frequency analysis, automatic gain control, and sensitivity to slow modulations – is sufficient to produce relatively robust visual displays of speech and to significantly improve the performance of an ASR system on highly reverberant speech. Although the modulation spectrogram is not as robust as log-RASTA-PLP in the tests on moderate reverberation or additive noise, the combination of the two representations, both emphasizing slow modulations in somewhat different ways, yields a significant improvement in performance over log-RASTA-PLP on its own.

Research towards improving the modulation spectrogram is continuing. The current emphasis is on the development of an entirely on-line version of the algorithm, the reduction of the spectral resolution of the final representation, and the integration of the front end with a syllable-based ASR system. An on-line gain control mechanism would enhance the representation of onsets, and thus may further improve robustness of the representation, as suggested by the results of combining log-RASTA-PLP and the off-line spectrogram features. Preliminary results from experiments with an on-line modulation spectrogram support this hypothesis. On its own, an on-line modulation spectrogram front end achieves a word error rate of 19.1% on the reverberant test set used in the experiments summarized in Table 1, and in framewise combination with log-RASTA-PLP features a word error rate of 17.6% is achieved. In both cases performance on the clean set is comparable to the results reported here for the off-line algorithm.

Acknowledgements

We are grateful to Jim West and Gary Elko, from Bell Labs, and Carlos Avendaño, now at the University of California, Davis, for collecting a set of room impulse responses and making them available to us, to Hynek Hermansky of the Oregon Graduate Institute for many valuable conversations on the robust representation of speech, to Katrin Kirchhoff for translating the abstract of this paper into German and to Holger Schwenk for translating the abstract of this paper into French. Support for this project came from a European Community Basic Research grant (Project Sprach) and from the International Computer Science Institute.

References

Arai, T., Greenberg, S., 1998. Speech intelligibility in the presence of cross-channel spectral asynchrony. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing.

Bourlard, H., Morgan, N., 1994. Connectionist Speech Recognition: A Hybrid Approach. Kluwer Academic Publishers, Dordrecht, pp. 155–183.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95 (2), 1053–1064.

Dudley, H., 1939. Remaking speech. J. Acoust. Soc. Amer. 11 (2), 169–177.

Furui, S., 1986. Speaker-independent isolated word recognition based on emphasized spectral dynamics. Proceedings of the 1986 IEEE-IECEJ-ASJ International Conference on Acoustics, Speech and Signal Processing, pp. 1991–1994.

Greenberg, S., 1997. On the origins of speech intelligibility in the real world. Proceedings of the ESCA–NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 23–32.

Greenberg, S., Arai. T., 1998. Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. Proceedings of the Joint Meeting of the Acoustical Society of America and the International Congress on Acoustics.

Greenberg, S., Hollenback, J., Ellis, D., 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. Proceedings of the Fourth International Conference on Spoken Language Processing, pp. S24–27.

Greenberg, S., Kingsbury, B.E.D., 1997. The modulation spectrogram: In pursuit of an invariant representation of speech. Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1647–1650.

Greenwood, D.D., 1961. Critical bandwidth and the frequency coordinates of the basilar membrane. J. Acoust. Soc. Amer. 33, 1344–1356.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. on Speech and Audio Processing 2 (4), 578–589.

Hermansky, H., Hanson, B.A., Wakita, H., 1985. Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain. Speech Communication 4, 181–187.

Hirsch, H.G., 1988. Automatic speech recognition in rooms. In: Lacoume, J.L., Chehikian, A., Martin, N., Malbos, J. (Eds.), Signal Processing IV: Theories and Applications, Proceedings of the EUSIPCO-88, Fourth European Signal Processing Conference, Elsevier, Amsterdam, pp. 1177–1180.

Houtgast, T., Steeneken, H.J.M., 1973. The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acustica 28, 66–73.

Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility. J. Acoust. Soc. Amer. 77 (3), 1069–1077.

Houtgast, T., Steeneken, H.J.M., Plomp, R., 1980. Predicting speech intelligibility in rooms from the modulation transfer function I. General room acoustics. Acustica 46 (1), 60–72.

Huggins, A.W.F., 1975. Temporally segmented speech. Perception and Psychophysics 18 (2), 49–157.

Kingsbury, B.E.D., Morgan, N., 1997. Recognizing reverberant speech with RASTA-PLP. Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1259–1262.

Kingsbury, B.E.D., Morgan, N., Greenberg, S., 1997. Improving ASR performance for reverberant speech. Proceedings of the ESCA–NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 87–90.

Kollmeier, B., Koch, R., 1994. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. J. Acoust. Soc. Amer. 95 (3), 1593–1602.

Langhans, T., Strube, H.W., 1982. Speech enhancement by nonlinear multiband envelope filtering. Proceedings of the 1982 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 156–159.

Lea, A.P., Summerfield, Q., 1994. Minimal spectral contrast of formant peaks for vowel recognition as a function of spectral slope. Perception and Psychophysics 56 (4), 379–391.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. Proceedings of the 1984 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 42.11.1–42.11-4.

Milner, B.P., Vaseghi, S.V., 1995. An analysis of cepstral-time matrices for noise and channel robust speech recognition. EUROSPEECH 95, Proceedings of the Fourth European Conference on Speech Communication and Technology, pp. 519–522.

Morgan, N., Hermansky, H., 1992. RASTA extensions: Robustness to additive and convolutional noise. ESCA Workshop on Speech Processing in Adverse Conditions, pp. 115–118.

Ramsay, G., Deng, L., 1995. Maximum-likelihood estimation for articulatory speech recognition using a stochastic target model. EUROSPEECH 95, Proceedings of the Fourth European Conference on Speech Communication and Technology, pp. 1401–1404.

Schreiner, C.E., Urbas, J.V., 1988. Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. Hearing Research 32 (1), 49–63.

Steeneken, H.J.M., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Amer. 67 (1), 318–326.

Summerfield, Q., Haggard, M., Foster, J., Gray, S., 1984. Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. Perception and Psychophysics 35 (3), 203–213.

Wu, S.-L., 1998. Incorporating information from syllable-length time scales into automatic speech recognition. Ph. D Thesis, University of California, Berkeley.

Wu, S.-L., Kingsbury, B.E.D., Morgan, N., Greenberg, S., 1998. Incorporating information from syllable-length time scales into automatic speech recognition. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing.