Sādhanā Vol. 36, Part 5, October 2011, pp. 729–744. © Indian Academy of Sciences

Speech recognition from spectral dynamics

HYNEK HERMANSKY

The Johns Hopkins University, Baltimore, Maryland, USA
e-mail: hynek@jhu.edu

Abstract. Information is carried in changes of a signal. The paper starts with revisiting Dudley's concept of the carrier nature of speech. It points to its close connection to modulation spectra of speech and argues against short-term spectral envelopes as dominant carriers of the linguistic information in speech. The history of spectral representations of speech is briefly discussed. Some of the history of gradual infusion of the modulation spectrum concept into Automatic recognition of speech (ASR) comes next, pointing to the relationship of modulation spectrum processing to well-accepted ASR techniques such as dynamic speech features or RelAtive SpecTrAl (RASTA) filtering. Next, the frequency domain perceptual linear prediction technique for deriving autoregressive models of temporal trajectories of spectral power in individual frequency bands is reviewed. Finally, posterior-based features, which allow for straightforward application of modulation frequency domain information, are described. The paper is tutorial in nature, aims at a historical global overview of attempts for using spectral dynamics in machine recognition of speech, and does not always provide enough detail of the described techniques. However, extensive references to earlier work are provided to compensate for the lack of detail in the paper.

Keywords. Carrier nature of speech; modulation spectrum; spectral dynamics of speech; coding of linguistic information in speech; machine recognition of speech; data-guided signal processing techniques.

1. Introduction

No natural system can change its state instantaneously and it is dynamics of changes that can carry an information. In the past 20 years, we have witnessed increased interest in the dynamics of temporal evolutions of the power spectrum as a carrier of information in speech. This dynamic is carried in modulation spectra of the signal. This concept has been in existence since the early days of speech signal processing and, is supported by a number of physiological and psychophysical experimental results, but was largely ignored by researchers in automatic recognition of speech (ASR). Instead, likely for historical reasons, envelopes of power spectrum were adopted as main carrier of linguistic information in ASR. However, the relationships between phonetic values of sounds and their short-term spectral envelopes are not straightforward. Consequently, this asks for complex data-intensive machine-learning techniques that are prevalent in the current state-of-the-art ASR. In spite of significant engineering advances in this direction, current ASR is still very sensitive to linear distortions, room reverberations, frequency-localized noise, or peculiarities of a particular speaker of the message – all of which are reasonably well handled by human listeners. We believe that some of these problems might be alleviated by greater emphasis on information carried in frequency-localized spectral dynamics of speech.

2. Carrier nature of speech

Over centuries of research in phonetics, there was a growing belief that the phonetic values of speech sounds are in some way related to resonance frequencies of the vocal tract in their production. Young Isaac Newton observed that when filling his tall glass with beer and a quarter-wave resonance of a column of air above the beer was increasing, he could hear a sequence of vowels going from the rounded /uh/ with its power concentrated at low frequencies to the extreme front /yi/, which has most of its power at high frequencies (Ladefoged 1967). Von Helmholtz supported Newton's observation by finding dominant resonance frequencies of his vocal tract in the production of vowels using tuning forks (von Helmholtz 1863).

However, in spite of the opinions of such highly respected scientists – as Newton and Helmhotz certainly are – the pioneering works of Homer Dudley (Dudley 1939, 1940) are very clear in his opinion about the roles of the carrier (vocal tract excitation) and the varying modulating envelope (the changing shape of the vocal tract). In his view, the vocal tract shape, slowly changing with frequency up to 10 Hz due to 'sluggishness of muscles' is reflected in the changing amounts of power in frequency bands of the signal. Excitation of the vocal tract, either by combined effects of vibrations of vocal cords and by air turbulence at vocal tract constrictions in normal speech or purely by air turbulence in whispered speech, merely makes these movements of the vocal tract audible to human hearing. Thus, Dudley is very clear about the modulation envelope being the carrier of the phonetic information. This view is evident in the vocoder design, where spectral energies in several frequency bands are low-pass filtered at 20 Hz to be transmitted to the receiving side where they modulate the carrier signal in respective frequency bands to obtain the reconstructed speech. This is even more obvious in Dudley's Voder design, where the signal amplitudes in 10 frequency sub-bands are directly controlled by the 10 fingers of a highly trained Voder operator. There is no control of resonance frequencies as in later formant syntheses. It is clearly the change of signal amplitudes in the individual bands that Dudley considers important for preserving the message in speech. Why is it his message was lost for a long time for ASR research? Some of the speculative reasons are discussed below.

3. Resonances of the vocal tract (formants of speech) and short-term spectral envelopes

The invention of the Spectrograph$^{\text{TM}}$ then emulated frequency filtering in human periphery by dividing the spectrum of a speech signal into a number sub-bands and displayed temporal trajectories of energies in these sub-bands (to help in decoding encrypted speech during World War II and to display underwater sounds originating from different ships (Schroeder 1998)) yielded speech spectrograms with clearly visible resonance frequencies of the vocal tract (formants of speech) moving in time. Relative success in visual decoding of the spectrograms (Potter *et al* 1947), with a successive flood of publications, sealed the role of the changing spectral envelope of speech as a dominant carrier of the phonetic information. It has been shown that lower formants correlate well with phonetic values of sustained sonorant sounds such as carefully produced vowels (Peterson & Barney 1952).

Digital signal processing that became dominant in the 1970s abandoned the original Spectrograph$^{\mathrm{TM}}$ technique of applying band-pass filters to the original speech signal to compute spectrograms. Instead, the digital processing revolution rediscovered the Fast Fourier Transform, which allowed for constructing spectrograms by sequencing frames of short-term spectra of speech, resulting in a two-dimensional series $S(\omega, t)$, called here as the spectrogram. Thus, the spectrogram is derived by computing the series of spectral vectors $S(\omega, t_i)$, computed from the original signal $x(t)$ within the time window Δt centred at time t_i, for each $i \in \langle 1, N\rangle$, where $N = T/\Delta t$, T being the length of the signal $x(t)$. In the digital spectrogram, short-term speech features represent samples of the signal spectral dynamics, just as the dynamics of a visual scene in a movie are emulated by sampling the scene by a sequence of still images. The minimum required sampling rate Δt of a speech spectrum has been determined by trial and error in the early days of digital speech coding (Gold 1998) to be somewhere around $\Delta t = 10$ ms and reflects the low-pass character of speech spectral envelopes resulting from inertia of dominant human vocal organs. The spectral resolution of $S(\omega, t)$ is often modified by various means to reflect spectral resolution of the human hearing periphery (Mermelstein 1976; Hermansky 1990).

The short-term spectrum frame-based approach was successfully applied in the late sixties in digital coding for speech, and it yields reasonably intelligible reconstructed speech. It was easy to adopt the frame-based techniques also in automatic recognition of speech (ASR), which started to evolve around the same time. The issues with the convolved way the information about speech sounds is coded in the short-term spectral envelope were set aside. This was not a problem in the early ASR systems, where the units of recognition were whole words. Later, large-vocabulary ASR that was based on recognizing sub-word units introduced context-dependent multi-state phoneme-like units to deal with coarticulation effects, and various compensation and adaptation techniques were applied to deal with excessive dependence of the short-term spectral envelopes on speakers and communication channels. Current ASR systems are complex examples of engineering sophistication, but the frame-based speech features derived from the short-term spectra of speech are today found in the front-ends of most state-of-the-art ASR systems.

However, even for sonorants, some well-known problems with speech spectra persist. Inertia of vocal organs produces coarticulation among neighbouring speech sounds (phones), which causes each short-term spectrum to be dependent not only on the current phone but also on the phones that surround it. Large differences exist in formant frequencies in phonetically identical sounds produced by different speakers. The ease with which the spectral envelope can be corrupted by relatively benign modifications such as linear filtering of the signal is alarming. In general, obstruents are more difficult to characterize by a single short-term spectral frame as they typically change in time rather rapidly and the only reasonable way to characterize them is by the sequence of several short-term spectral frames. However, even the sequence of short-term spectral frames fails to characterize certain obstruents such as /k/ or /h/ (Potter *et al* 1947) that are only defined in relation to the following sonorant, e.g., the /k/ is perceived whenever power of the noise burst is slightly above the major concentration of the following sonorant power. This can be at very low frequencies in the case of the syllable /k//uh/, but at much higher frequencies in /k//ih/. The /h/ has concentrations of fricative noise in the same places as the following sonorant.

In our view, the formant concept with its emphasis on the short-term spectral envelope is not wrong. After all, resonances of a vocal tract control the relative amount of power in each frequency band. Further, the values of the instantaneous power at the individual frequencies describe the short-term spectral envelope. However, just as when trying to perform long division using Roman numerals, it is the form of the representation (Marr 1982) rather than the content of the information that causes the problems. The examples above demonstrate that phonetic values of speech sounds relate to short-term speech spectra often in rather complicated ways.

3.1 *Dynamics of short-term spectra of speech*

Most natural signals such as speech change over time and the information is carried in these changes. The signal changes are reflected in the dynamics of spectral components. Yet, in current machine extraction of information from speech, speech spectral dynamics are mostly treated as a nuisance. In earlier whole-utterance template-matching systems, the spectral dynamics were arbitrarily distorted by dynamic time warping in order to compensate for variable speed of speech production. However, the utterance-level template matching at least respects the overall trends of the spectral dynamics (and uses the coarticulation patterns to its advantage). Hidden Markov Model (HMM)-based systems are even more adverse to the dynamics of the signal by approximating the dynamics by sequences of models of stationary stochastic processes. For more accurate approximations, a large number of models would be required, increasing the number of free parameters that need to be estimated from training data. To deal with the coarticulation, multi-state context-dependent speech sound models are introduced, increasing the complexity of the system. Short-term spectrum-based features in these models are complemented, almost always with advantage, by so-called 'dynamic' features (Mlouka & Lienard 1975; Furui 1981) that reflect dynamic trends of the spectral envelopes at a given instant of time computed from larger (up to about 100 ms) segments of the signal. Although in principle, the dynamic features should require different sequences of stationary stochastic models than the 'static' envelope-based features, most often the dynamic features are successfully appended to the static ones.

4. History of modulation spectrum of speech

4.1 *Defining modulation spectrum of speech*

The concept of the modulation spectrum of speech (figure 1) (Houtgast & Steeneken 1973) is consistent with Dudley's view of the carrier nature of speech. Evolution of the short-term spectrum in the spectrogram $S(\omega, t)$ at the frequency ω_0 is described by a one-dimensional time series $S(\omega_0, t)$. The discrete Fourier transform (DFT) of a logarithm of these time series within the time window ΔT centred at the time t_0 with its mean removed, i.e.

$$F(\Omega, t_0) = \Sigma_{\Delta T}(\log(S(\omega_0, t) - \log(1/\Delta T)\Sigma_{\Delta T} S(\omega_0, t))\mathrm{e}^{-j\Omega t}$$

is what we call in this article, the modulation spectrum at the time t_0. The modulation spectrum is the time series that describes the shape of the time trajectory $S(\omega_0, t)$ within the time interval ΔT. The resolution of such a modulation spectrum $1/\Delta T$ is inversely proportional to the length of the window over which the spectrum is computed. The modulation spectrum is complex, but in some applications, only the absolute values $|F(\Omega, t_0)|$ are of interest.

Since the DFT operation is linear, in many applications described in this article, the DFT step is omitted and we deal without the loss of any information only with the series $S(\omega_0, t)$ within the time window ΔT.

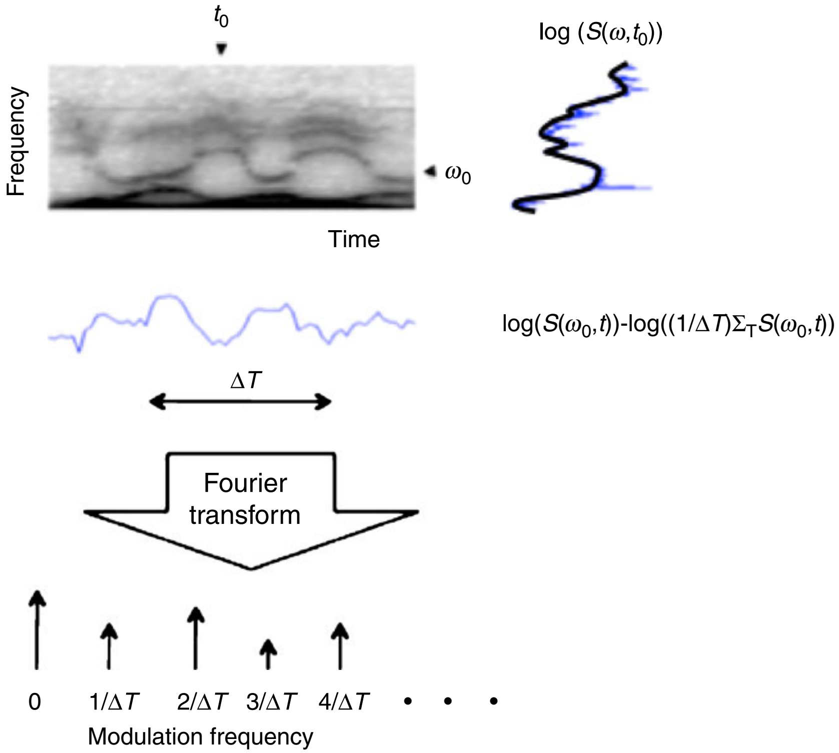

Figure 1. Principle of the modulation spectrum of speech. A conventional spectrogram consists of a sequence of short-term spectra. The short-term spectrum at the time t_0 is shown in the right part of the figure. Its spectral envelope $S(\omega, t_0)$ is indicated by the thicker line. An alternative way of looking at the spectrogram is to see it as a sequence of temporal trajectories of logarithms of spectral power $S(\omega_i, t)$. One of the trajectories at a frequency ω_0 is illustrated at the bottom of the figure. A segment of this temporal profile, centered at the time t_0, can be described by the Fourier series. When its mean is removed, the series describes just its shape. Coefficients of such Fourier series define modulation spectrum at the time t_0. Resolution of this modulation spectrum is given by the length of the segment. When the segment is extracted using the square window, 1 s of the signal is required for 1 Hz spectral resolution (as defined by the width of the main lobe of the window). Tapered windows such as the Hamming window require appropriately longer segments for the same resolution.

4.2 *Modulations and human hearing*

Since the early experiments by Riesz (1928), it is known and confirmed many times by others that human hearing is most sensitive to relatively slow modulations. Riesz's result is summarized in figure 2.

It is not surprising that most of the energy of the modulation spectrum of speech is present in the area where hearing is the most sensitive, typically peaking at around 4 Hz, reflecting the syllabic rate of speech. Expected deviations from this typical shape of the modulation spectrum resulting from noise and reverberations and measured using artificial signals (speech transmission index) have been proposed to reflect the intelligibility of speech in noisy and reverberant environments (Houtgast & Steeneken 1973). Extensions involving real speech and more involved

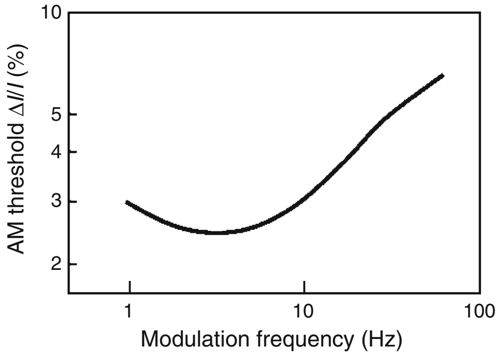

Figure 2. Results of Riesz's experiment in sensitivity of human hearing to modulations. It indicates that human hearing is most sensitive in the range of about 2–8 Hz, where only about 2.5% depth of modulation is required for the modulation to be perceived. The figure was made using Riesz's data (Riesz 1928).

spectral projections than a simple 1/3-octave integration have been proposed more recently (Kollmeier *et al* 1999; Elhilali *et al* 2003).

Attenuating components of the modulation spectrum around 4 Hz significantly lowers intelligibility of speech. This was first shown by (Drullman *et al* 1994), using a set-up that modified Hilbert envelopes of sub-band signals, and was subsequently verified by (Arai *et al* 1999), who used a residual-excited vocoder. Arai *et al* also showed that attenuation of modulation spectrum components below 1 Hz and above 16 Hz has only small effects on speech intelligibility. The results of one of their experiments are shown in figure 3. The 2-dimensional plot shows the performance surface as a function of high and low cut-offs of the modulation spectrum. The

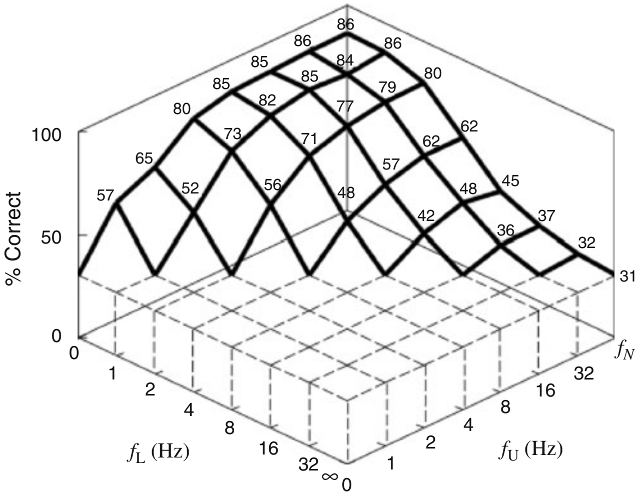

Figure 3. Recognition accuracy of phonemes in nonsense Japanese syllables as a function of frequency cutoffs of high-pass and low-pass modulation frequency filters on temporal evolutions of spectral envelopes in a residual-excited LP vocoder. The results indicate that restricting modulation frequencies in such modified speech to 1–16 Hz range has only a minimal effect on the accuracy of human recognition of phonemes in the experiment. The figure is reproduced from (Arai *et al* 1999) and used with permission.

surface remains quite flat and close to maximum as long as the modulation spectrum components between 1 and 16 Hz are preserved.

Dau and his colleagues (Dau *et al* 1997) successfully verified and promoted the earlier proposal of Houtgast (1989) on the existence of band-pass modulation frequency filters. Findings of ongoing works on the physiology of mammalian auditory cortices – see e.g., Kowalski *et al* 1996 – further support this concept.

5. RASTA processing

5.1 *How it all started*

Our interest in processing of modulation spectrum started with an anecdotal description of a simple but convincing experiment in speech perception (Cohen 1990), which goes as follows:

Extract a spectral envelope of a vowel from a spoken utterance (indicated by an arrow in the left part of figure 4) and filter the whole utterance with a filter with a frequency response that is

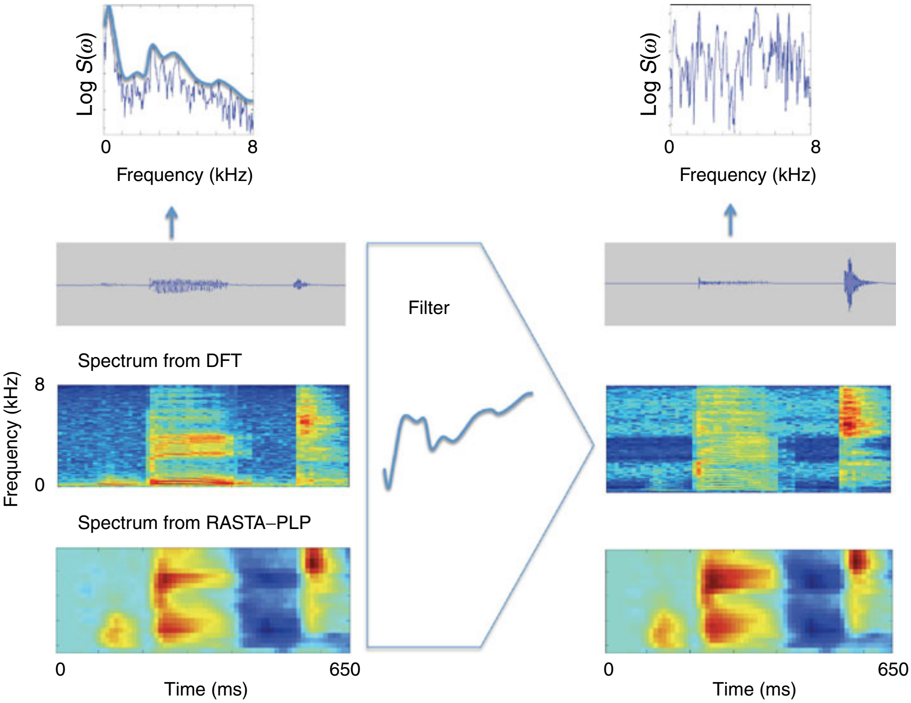

Figure 4. Left part of the figure shows the time domain signal of the utterance 'beet' (/b/ /ee/ /t/) together with its spectrogram computed by the conventional DFT analysis (left middle part of the figure) and by the RASTA–PLP technique (left bottom part of the figure). Above the speech waveform, a single spectral slice from the spectrogram, extracted at the time instant indicated by the arrow (spectrum of the vowel /ee/), is shown, together with its spectral envelope. The right part of the figure shows the speech waveform, the conventional spectrogram, the RASTA–PLP–derived spectrogram, and the spectral slice from the /ee/ vowel part after the speech waveform was filtered by the filter that has a frequency response that is the inverse of the spectral envelope of the vowel /ee/. The filtering flattens the spectral envelope of the vowel /ee/ but has only a negligible effect on the RASTA–PLP representation of speech.

the *inverse* of the extracted envelope. This makes the spectrum of the given vowel flat (shown in the right part of figure 4). In spite of that, the listeners typically report hearing an unambiguous vowel in the part of the utterance with this flattened spectrum. To emulate this human ability, we proposed an *ad hoc* but effective RASTA filtering that only passed modulation spectrum between 1 and 15 Hz to alleviate negative effects of such fixed linear distortions (Hermansky *et al* 1991; Hermansky & Morgan 1994). Figure 5 shows the frequency response of the original RASTA filter. As illustrated in the lower part of figure 4, this turned out to be very effective not only to deal with this particular effect but also to combat typical linear distortions introduced by non-flat frequency responses of communication channels. However, since the original filter is a recursive infinite impulse response filter, it introduces significant phase modifications of the modulation spectrum.

5.2 *Speech beyond 20 ms*

RASTA with its rather long (> 200 ms) time constant spurred more interest in syllable-level spectral dynamics (Hermansky 1994; Hermansky *et al* 1995). We soon realized that the spectral transforms (Fourier or cosine transform) on the temporal trajectory of the signal power that yield the modulation spectrum are a mere convenience for the subsequent processing. Thus, the term 'modulation spectrum' is actually a synonym for shapes of temporal trajectories of elements of spectral envelopes of speech, which in their turn reflect temporal movements of the vocal organs. The critical issue is the length of the signal that carries the information, which is relevant for recognizing speech sounds. Since the modulation spectrum components that are most important for perception of speech are around 4 Hz, this time interval must be at least 250 ms. This is much longer than the conventional 10–20 ms analysis windows of the short-term spectral analysis used in speech so far!

One can present many arguments for this relatively long time interval, some of which can be summarized, e.g., in Hermansky (1998c). Such a time interval comes as no surprise to any physiologist or psychophysicist, and it is surprising that it escaped the attention of most speech engineers for such a long time. It is found in many psychophysical phenomena and on higher

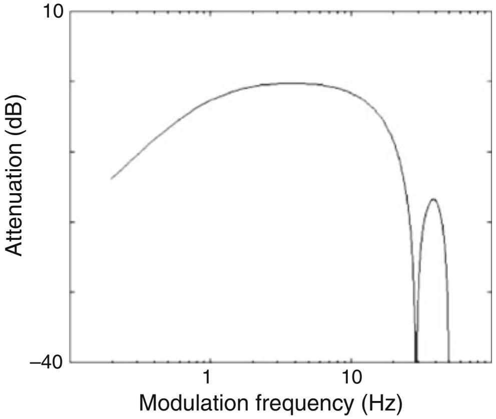

Figure 5. Logarithmic magnitude frequency response of a RASTA filter that was found optimal for recognition of telephone speech, corrupted by linear distortions. It indicates that alleviating modulation frequencies below about 1 Hz and above about 20 Hz is desirable to alleviate effects introduced by linear distortions. The figure was derived from (Hermansky & Morgan 1994).

levels of neural processing and motor control. It does not, however, imply that human perception necessarily recognizes these relatively long speech segments (syllables) (Greenberg 1999). It merely implies that, due to coarticulation, these segments carry the information about elements (speech sounds) within them (Kozhevnikov & Chistovich 1967; Hermansky 1998c).

6. Some further applications of the modulation spectrum in automatic recognition of speech

6.1 *Beyond RASTA*

A series of subsequent studies soon followed. Some of the familiar studies are mentioned here. First, Hermansky (1997, 1998a) discuss the concept of modulation spectrum in ASR. Avendano & Hermansky (1997) & Avendano (1997) discuss the application to speech enhancement. van Vuuren & Hermansky (1998) try to find the advantage of modulation spectrum for machine identification of speakers. Kajarekar *et al* (2000) attempt to find different sources of variability (information) in the modulation spectrum. Systematic experiments with filtering the modulation spectrum are performed in Kanedera *et al* (1998, 1999). These works have shown that eliminating modulation frequency components below 1 Hz can increase the performance of ASR. Kingsbury experimented with so-called MSG features (Kingsbury & Morgan 1997) that band-pass filtered the modulation spectrum into two bands. In a parallel effort with RASTA processing, Pueschel was developing his model of non-linear processing of the modulation spectrum, which later became the Oldenburg PEMO model (Dau *et al* 1996). de Veth and Boves (1997) indicated the importance of preserving the original modulation spectrum phase that is being modified by the original *ad hoc* RASTA IIR filter. To our knowledge, at least one application successfully applied RASTA processing in recognition of visual patterns (Kim *et al* 2002). van Vuuren & Hermansky (1997), Hermansky (1998c) and later Valente & Hermansky (2006) investigated a way of designing FIR RASTA filters using the linear discriminant analysis. The discriminant matrix was derived using large phoneme-labelled data from multiple speakers and conditions.

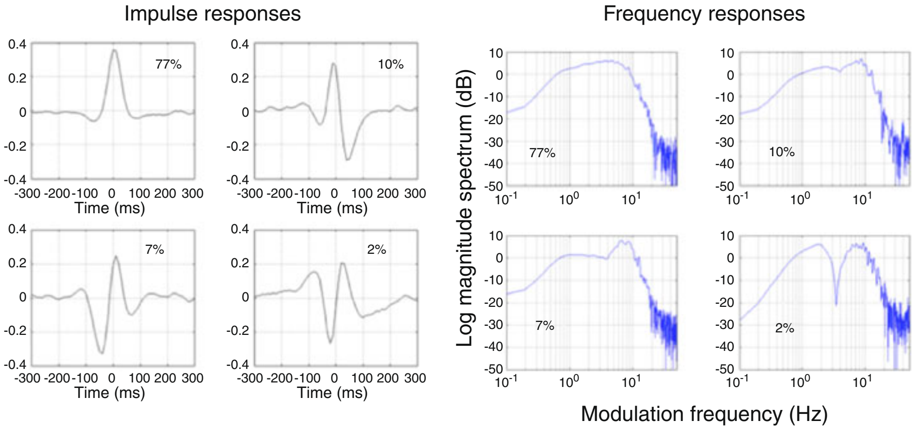

Figure 6. First four principal components of a discriminant matrix derived by linear discriminant analysis of 1 s long segments of temporal trajectories of power in critical band at 5 Barks, representing optimal FIR filters for filtering of this temporal trajectory. Results from other critical bands are very similar. All filters emphasize modulation frequency components between 1 and 10 Hz.

The first few discriminant vectors (representing impulse responses of the FIR RASTA filter-bank), together with their frequency responses, carrying most of the discriminative variability in the data, are shown in figure 6. Magnitude frequency responses of these filters are consistent with the original *ad hoc* RASTA filter; the phases are close to zero or ±π.

At about the same time, we proposed the so-called multi-stream ASR (Tibrewala & Hermansky 1997) where sub-bands on the modulation spectral domain were suggested as a way of forming the sub-streams. So, it was tentatively concluded that human (and more generally all mammalian) hearing may be not be evaluating the overall shape of the sound spectrum, but that it rather evaluates temporal profiles of signals in individual sub-bands (Hermansky 1998b, c); and one way of doing so is to evaluate the modulation spectrum within the individual sub-bands.

6.2 *TRAP and related studies*

This tentative proposal was first tested by the so-called TempoRAl Pattern (TRAP) (Hermansky & Sharma 1998), where 1001 ms long temporal trajectories of spectral power in the individual critical-band sub-bands (derived from Perceptual Linear Prediction (PLP) spectral analysis) with their means removed were first classified as belonging to phoneme categories (with a rather high error but still well above chance). The classification results from the individual sub-bands were then merged using a non-linear (NN-based) classifier, yielding results that were comparable to results from a conventional short-term spectrum-based ASR. Frequency-localized spectral power is not measured and used for the description of the spectral envelope; that is, correlations among the spectral sub-bands are not used. The power in the individual bands merely defines the local signal-to-noise ratio (SNR). The information that TRAP uses is present in the local temporal dynamics. Temporal trajectories in TRAP are (prior to any classification) often first projected on the modulation spectrum domain, either through the cosine transform (e.g. Jain 2003) or through a set of modulation spectrum band-pass filters (Hermansky & Fousek 2005).

The principle of TRAP-based processing schemes is shown in figure 7. Many variants on the original TRAP concept have been proposed and studied, and to our knowledge at least five Ph.D. theses (Sharma 1999; Jain 2003; Chen 2005; Grézl 2007; Schwarz 2008) and one habilitation thesis (Cernocky 2003) have been at least partially devoted to TRAP. The largest advantage of TRAP-based schemes is in combination with the conventional frame-based techniques where they appear to complement the information that is available in the spectral envelope. Widely used dynamic features (delta and double-delta) (Furui 1981) that are in ASR typically appended

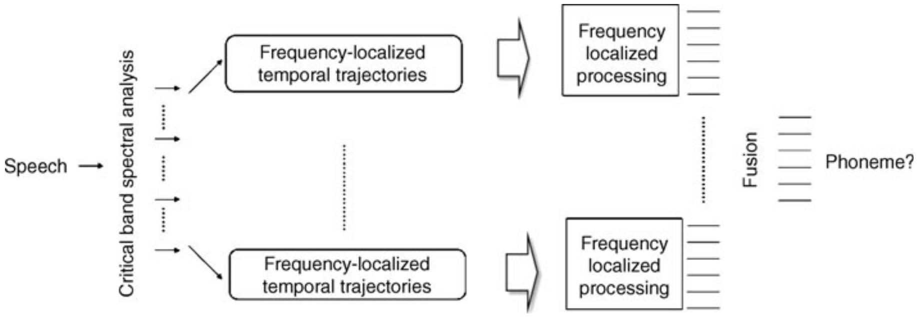
Speech
Critical band spectral analysis
Frequency-localized temporal trajectories
Frequency localized processing
Frequency-localized temporal trajectories
Frequency localized processing
Fusion
Phoneme?

Figure 7. Principle of TRAP-based feature extraction. Temporal trajectories of powers at individual frequency bands are processed to extract frequency-localized information that is relevant for classification of speech sounds. The frequency-localized information is fused to yield the final result.

to the spectral envelope-based cepstral features represent band-pass filtering by simple Finite Impulse Response (FIR) filters with pass-bands around 10 Hz.

6.3 *Modulation spectra from frequency domain perceptual linear prediction*

In most applications, the modulation spectrum is derived from temporal trajectories of spectral envelopes obtained by integrating a frame-by-frame short-term Fourier transform over critical bands of hearing. Temporal resolution of such trajectories is given by the analysis window in the short-term analysis and is typically somewhere around 10 ms. Since in modulation spectrum-based applications, we are primarily interested in temporal trajectories; hence it is tempting to abandon the short-term analysis altogether. This is possible by using the frequency domain perceptual linear prediction (FDPLP) (Athineos & Ellis 2007; Athineos *et al* 2004), where an autoregressive model is computed from a cosine transform of the signal rather than from the signal itself.

Given a real signal $s(t)$, $t = 1 \ldots N$, the real and the imaginary parts of the signal spectrum $\mathbf{DFT}[(s(t)]$ (where **DFT** stands for the discrete Fourier transform) relate through the Hilbert transform (Krammers–Kronig relation), i.e., $\mathbf{DFT}[(s(t)] = \mathrm{Re}[S(\omega)] + \mathrm{j}\mathbf{H}[\mathrm{Re}[S(\omega)]]$, where **DFT**[ ] indicates the discrete fourier transform, and **H**[ ] indicates the Hilbert transform. The power spectrum $P(\omega)$ is then given as $\{\mathbf{DFT}[(s(t)]\}^2 = \mathrm{Re}[S(\omega)]^2 + \mathbf{H}[\mathrm{Re}[S(\omega)]]^2$. The conventional autocorrelation method of the linear predictive analysis approximates the power spectrum of a signal by the autoregressive model computed by the autocorrelation method of linear predictive analysis from the signal $s(t)$ (Makhoul 1975).

Similarly, if $q(t) - s(t) + s(2N - 1 - t), t = 1, 2N - 1$ represents an even-symmetric sequence in which the first half is equivalent to $s(t)$, the cosine transform $c(\omega)$ of $s(t)$ represents the first half of the scaled inverse **DFT** of $q(t)$, i.e., $c(\omega) = (2N - 1)$ $\mathbf{DFT} - 1[q(t)]$, $\omega = 1 \ldots N$. As the $c(\omega)$ is also real, its discrete Fourier transform also obeys the Krammer–Kroning relation, i.e., $\mathrm{DFT}[c(\omega)] = q(t) + j\mathbf{H}[q(t)]$. The Hilbert envelope of the signal $s(t)$ given as **DFT** $[c(\omega)]^2 = q(t)^2 + j\mathbf{H}[q(t)]^2$ is then approximated by the autoregressive model computed by the autocorrelation method of linear predictive analysis from the $c(\omega)$.

Since the cosine transform of a time domain signal moves the signal to its frequency domain, $q(\omega)$ covers the whole frequency range of $s(t)$. To find the autoregressive model of the signal in a restricted frequency range, one can place an appropriate limited-span window on $q(\omega)$. The window span and shape determines the frequency response of the implied frequency filter. Thus, by properly windowing the cosine transform of the signal, one can directly compute autoregressive models of the Hilbert envelopes in the sub-bands over long segments of the speech signal, entirely bypassing any short-term analysis windows (Athineos *et al* 2004). The principle of the complete FDPLP computation is illustrated in figure 8.

The FDPLP model has been shown to be effective in applications that benefit from enhanced spectral dynamics such as phoneme recognition (Ganapathy *et al* 2009), recognition of large-vocabulary continuous speech (Thomas *et al* 2009), in handling linear distortions in recognition of telephone speech (Thomas *et al* 2008a), and in recognition of reverberant speech (Thomas *et al* 2008b).

6.4 *Modulation spectrum in deriving posterior-based features of speech*

Neither the modulation spectra nor the data in temporal trajectories have a normal distribution or are correlated. As such, they are not suitable for direct interface with HMM–GMM ASR

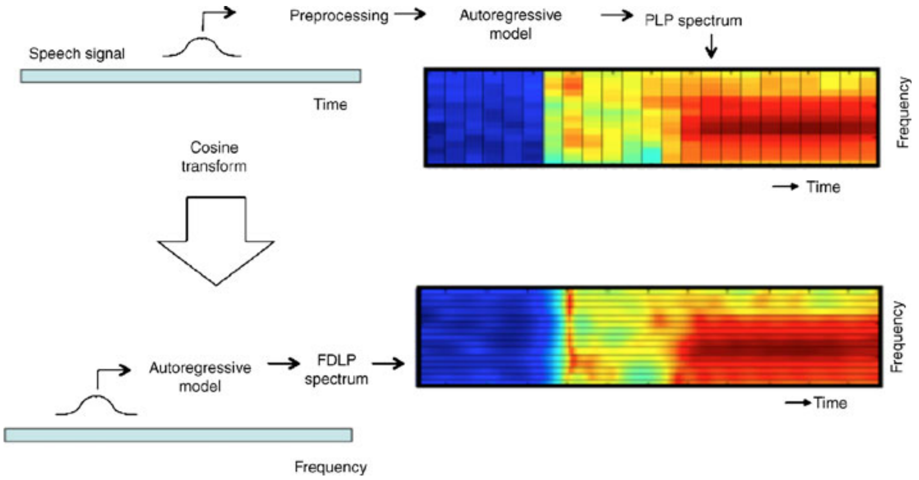

Figure 8. Frequency domain perceptual linear prediction as compared to the conventional time-domain perceptual linear prediction. The process of deriving a conventional PLP-based spectrogram is shown in the upper part of the figure. In the conventional technique, a windowed segment of the signal is used to derive the auditory-like short-term spectrum of the segment. This spectrum is approximated by an autoregressive PLP model. Stacking PLP spectra in time yields the PLP-based spectrogram shown in the upper right corner. The lower part of the figure shows the process involved in deriving the FDPLP spectrogram. The speech signal is transformed into the frequency domain by cosine transform. The window on the cosine-transformed signal determines the frequency band of the signal to be approximated by the autoregressive FDPLP model. The model approximates the temporal trajectory of power in the frequency band. Stacking the all-pole FDPLP estimates from different frequency bands yields the FDPLP spectrogram, shown in the lower right corner of the figure.

systems. We have therefore initially applied all our modulation spectrum-based techniques only in HMM/ANN hybrid recognizers, where the modulation spectrum-based features are used as an input to an artificial neural net (ANN) estimator of posterior probabilities of speech classes (Bourlard & Wellekens 1989). An important advance was the introduction of the TANDEM approach (Hermansky *et al* 2000) that applies a series of processing steps to estimates of posteriors of speech sounds from the ANN classifier, making them more suitable for the currently

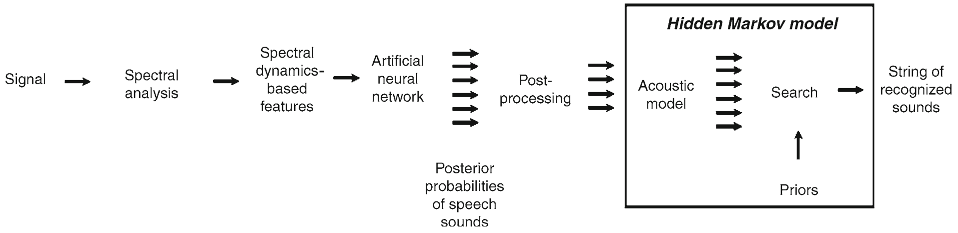
Signal
Spectral analysis
Spectral dynamics-based features
Artificial neural network
Post-processing
Posterior probabilities of speech sounds
Hidden Markov model
Acoustic model
Search
Priors
String of recognized sounds

Figure 9. Generic scheme of deriving posterior-based features in the modulation spectrum domain. Spectral analysis, either conventional or FDPLP-based, yields a signal spectrogram. Features based on spectral dynamics are derived from the spectrogram and form the input to the artificial neural network, trained on labelled data to derive posterior probabilities of speech sounds (typically phonemes). The post-processing (most often achieved by extracting values from inner layers of the trained neural net) yields posterior-based features that are suitable as an input to a Gaussian mixture-based HMM recognizer.

dominant HMM/GMM ASR technology. A generic system for computing ASR features based on the modulation spectrum is shown in figure 9. The speech signal is first converted to an auditory-like time-frequency representation, either by using conventional frame-based spectral analysis or FDPLP. Sufficiently long (typically longer than 200 ms) segments of temporal trajectories of spectral energies in the frequency sub-bands form, after some pre-processing, an input to an estimator of posterior probabilities of speech sounds that has been trained on large amounts of labelled speech data. The final features for an HMM/GMM-based state-of-the-art ASR system are derived from these posteriors by some post-processing that ensures that the features have approximately a normal distribution and are decorrelated. This post-processing may include either appropriate static non-linearities (Hermansky *et al* 2000) or the full inverse of the last layer of the ANN, in practice representing the values on the ANN hidden layer (Chen *et al* 2004; Grézl *et al* 2007). Such features based on modulation spectra are successfully used in many state-of-the-art experimental systems (Fousek *et al* 2008; Park *et al* 2009; Plahl *et al* 2009).

Using the module for converting the evidence from the signal to posterior probabilities of speech sounds (currently we use the trained ANN for this purpose) allows relatively free choice of what constitutes the 'evidence.' Currently, this evidence is typically derived by multiple projections of the time–frequency plane with varying spectral and temporal properties (e.g. Hermansky & Fousek 2005; Valente & Hermansky 2006; Thomas *et al* 2008a, b, 2009; Ganapathy *et al* 2009). Such projections are consistent with our current knowledge about properties of cortical receptive fields in mammalian brains (e.g. Kowalski *et al* 1996), and sometimes even directly derived from brain-obtained measurements (Thomas *et al* 2010). In principle, there may be large numbers of different projections, forming processing channels differently affected by different signal distortions. Exploiting this possibility for increased robustness of processing is a current research interest (Mesgarani *et al* 2011).

7. Conclusion

The dynamics of signal envelopes in frequency sub-bands are important for describing linguistic information in speech. This was the basis of the first speech coder (Dudley 1939). Unfortunately, over the years this concept was lost for ASR, which puts emphasis on instantaneous short-term spectral envelopes; spectral dynamics were treated more as a nuisance to be modified by time-aligning techniques. However, recent research unambiguously points to the importance of spectral dynamics in coding the phonetic information in speech, and the interest in spectral dynamics has started to grow again. At the time of writing of this article, posterior-based features that are derived from spectral dynamics of speech are used in most state-of-the-art experimental ASR technology. It is likely that as our appreciation of information in spectral dynamics grows, new ASR techniques will emerge. Coarticulation may be recognized as an important carrier of information in speech; recognizing speech sounds without extensive use of the top-down language constraints may become a respectable engineering endeavour; and human-like robustness of speech processing in the presence of reasonable signal degradations may become a reality.

This paper describes the work of many colleagues, most of them hopefully acknowledged by references to their earlier publications. Our own incomplete knowledge necessarily caused some fine works to be omitted, for which we apologize. Writing of the paper was partially supported by IARPA BEST and DARPA RATS grants, and by the JHU Center of Excellence in Human Language Technology.

References

Arai T, Pavel M, Hermansky H, Avendano C 1999 Syllable intelligibility for temporally filtered LPC cepstral trajectories. *J. Acoust. Soc. Am.* 105(5): 2783–2791

Athineos M, Ellis D P W 2007 Autoregressive modelling of temporal envelopes. *IEEE Trans. Signal Process.* 55(11): 5237–5245

Athineos M, Hermansky H, Ellis D P W 2004 LP-TRAPS: Linear predictive temporal patterns. *Proc. Interspeech 2004*, Jeju Island, Korea

Avendano C 1997 Temporal processing of speech in a time-feature space. Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, Portland

Avendano C, Hermansky H 1997 On the properties of temporal processing for speech in adverse environments. *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, N.Y.

Bourlard H, Wellekens C J 1989 Links between Markov models and multilayer perceptrons, in D S Touretzky (ed), *Advances in neural information processing systems I*, Morgan Kaufmann, Los Altos, CA, 502–510

Cernocky J 2003 Temporal processing for feature extraction in speech recognition. Habilitation Thesis, FIT, Brno University of Technology, Czech Republic

Chen B Y 2005 Learning discriminant narrow-band temporal patterns for automatic recognition of conversational telephone speech. Ph.D. Thesis, University of California at Berkeley

Chen B, Zhu Q, Morgan N 2004 Learning long-term temporal features in LVCSR using neural networks. *Proc. Interspeech 2004*, Jeju Island, Korea

Cohen J 1990 Personal communications at the International Computer Science Institute, Berkeley, California

Dau T, Kollmeier B, Kollrausch A 1997 Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *J. Acoust. Soc. Am.* 102(5): 2892–2905

Dau T, Pueschel D, Kohlrausch A 1996 A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.* 99(6): 3615–3622

de Veth J, Boves L 1997 Phase-corrected RASTA for automatic speech recognition over the phone. *ICASSP'97*, Munich

Drullman R, Festen J M, Plomp R 1994 Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95(5): 2670–2680

Dudley H 1939 Remaking speech. *J. Acoust. Soc. Am.* 11(2): 169–177

Dudley H 1940 The carrier nature of speech. *Bell System Tech. J.* 19: 495–513

Elhilali M, Chi T, Shamma S A 2003 A spectro-temporal modulation index (STMI) assessment of speech intelligibility. *Speech Commun.* 41(2–3): 331–348

Fousek P, Lamel L, Gauvain J 2008 Transcribing broadcast data using MLP features. *Proc. Interspeech 2008*, Brisbane

Furui S 1981 Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech Signal Process.* 29(2): 254–272

Ganapathy S, Thomas S, Hermansky H 2009 Modulation frequency features for phoneme recognition in noisy speech. *J. Acoust. Soc. Am.* 125(1): EL8–EL12

Gold B 1998 Personal communications, Berkeley, California

Greenberg S 1999 Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Commun.* 29(2–4): 159–176

Grézl F 2007 TRAP-based probabilistic features for automatic speech recognition. Ph.D. Thesis, FIT, Brno University of Technology, Czech Republic

Grézl F, Karafiat M, Kontar S, Cernocky J 2007 Probabilistic and bottle-neck features for LVCSR of meetings. *Proc. ICASSP'07*, Honolulu

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

Hermansky H 1994 Speech beyond 10 ms (temporal filtering in feature domain). *International Workshop on Human Interface Technology 1994*, Aizu, Japan

Hermansky H 1997 The modulation spectrum in automatic recognition of speech. *Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA

Hermansky H 1998a Modulation spectrum in speech processing, in Procházka A, Uhlíř J, Rayner P J W, Kingsbury N G (eds) *Signal analysis and prediction*. Boston: Birkhauser

Hermansky H 1998b Data-driven analysis of speech. *Invited Paper, Proceedings of the International Conference on Text, Speech and Dialogue*, Brno, Czech Republic

Hermansky H 1998c Should recognizers have ears? *Speech Commun.* 25(1–3): 3–27

Hermansky H, Ellis D P W, Sharma S 2000 Connectionist feature extraction for conventional HMM systems. *ICASSP'00*, Istanbul

Hermansky H, Fousek P 2005 Multi-resolution RASTA filtering for TANDEM-based ASR. *Proc. Interspeech 2005*, Lisbon, 361–364

Hermansky H, Greenberg S, Pavel M 1995 A brief (100–200 ms) history of time in feature extraction of speech. *The XV Annual Speech Research Symposium*, Baltimore, MD

Hermansky H, Morgan N 1994 RASTA processing of speech. *IEEE Trans. Speech Audio Process.* 2(4): 578–589

Hermansky H, Morgan N, Bayya A, Kohn P 1991 Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP), In *EUROSPEECH-1991*, 1367–1370.

Hermansky H, Sharma S 1998 TRAPS – Classifiers of temporal patterns. *ICSLP'98*, Sydney

Houtgast T 1989 Frequency selectivity in amplitude-modulation detection. *J. Acoust. Soc. Am.* 85(4): 1676–1680

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

Jain P 2003 Temporal patterns of frequency localized features in ASR. Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, Portland

Kajarekar S, Malayath N, Hermansky H 2000 ANOVA in modulation spectral domain. *ICASSP'00*, Istanbul

Kanedera N, Arai T, Hermansky H, Pavel M 1999 On the relative importance of various components of the modulation spectrum of speech. *Speech Commun.* 28(1): 43–55

Kanedera N, Hermansky H, Arai T 1998 Desired characteristics of modulation spectrum for robust automatic speech recognition. *ICASSP'98*, Seattle, WA, 2: 613–616

Kim J, Choi S, Park S 2002 Performance analysis of automatic lip reading based on inter-frame filtering. *Proc. 2002 Multimodal Speech Recognition Workshop*, Greensboro, NC

Kingsbury B E D, Morgan N 1997 The modulation spectrogram: In pursuit of an invariant representation of speech. *Proc. ICASSP'97*, Munich, 1259–1262

Kollmeier B, Wesselkamp M, Hansen M, Dau T 1999 Modeling speech intelligibility and quality on the basis of the "effective" signal processing in the auditory system (A). *J. Acoust. Soc. Am.* 105(2): 1305–1305

Kowalski N, Depireux D A, Shamma S A 1996 Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *J. Neurophysiol.* 76(5): 3503–3523

Kozhevnikov V A, Chistovich L A 1967 Speech: Articulation and perception. *Trans. U.S. Department of Commerce, Clearing House for Federal Scientific and Technical Information* (Washington, D.C.: Joint Publications Research Service), 250–251

Ladefoged P 1967 *Three areas of experimental phonetics* (London: Oxford University Press)

Makhoul J 1975 Spectral linear prediction: properties and applications. *IEEE Trans. Acoust. Speech Signal Process.* 23(3): 283–296

Marr D 1982 *Vision: A computational investigation into the human representation and processing of visual information* (San Francisco: W.H. Freeman and Company)

Mermelstein P 1976 Distance measures for speech recognition, psychological and instrumental, in R C H Chen (ed) *Pattern recognition and artificial intelligence*, New York: Academic Press, 374–388

Mesgarani N, Thomas S, Hermansky H 2011 Toward optimizing stream fusion in multistream recognition of speech. *J. Acoust. Soc. Am.* 130(1): EL14–EL18

Mlouka M, Lienard J S 1975 Word recognition based on either stationary items or on transitions. *Speech Commun.* 3: 257–263, Go Fant (ed.) (Stockholm: Almqvist & Wiksell Int.)

Park J, Diehl F, Gales M J F, Tomalin M, Woodland P C 2009 Training and adapting MLP features for Arabic speech recognition. *Proc. ICASSP'09*, Taipei

Peterson G E, Barney H L 1952 Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24: 175–184

Plahl C, Hoffmeister B, Heigold G, Loeoef J, Schlueter R, Ney H 2009 Development of the GALE 2008 Mandarin LVCSR System. *Proc. Interspeech 2009*, Brighton, UK, 2107–2111

Potter R K, Kopp G A, Green H C 1947 *Visible speech* (New York: D Van Nostrand)

Riesz R 1928 Differential intensity sensitivity of the ear for pure tones. *Phys. Rev.* 31(5): 867–875

Schroeder M R 1998 Personal communications, Il Ciocco NATO Advanced Study Institute

Schwarz P 2008 Phoneme recognition based on long temporal context. Ph.D. Thesis, FIT, Brno University of Technology, Czech Republic

Sharma S 1999 Multi-stream approach to robust speech recognition. Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, Portland

Thomas S, Ganapathy S, Hermansky H 2008a Hilbert envelope based spectro-temporal features for phoneme recognition in telephone speech. *Proc. Interspeech 2008*, Brisbane

Thomas S, Ganapathy S, Hermansky H 2008b Recognition of reverberant speech using frequency domain linear prediction. *IEEE Signal Process. Lett.* 15: 681–684

Thomas S, Ganapathy S, Hermansky H 2009 Tandem representations of spectral envelope and modulation frequency features for ASR. *Proc. Interspeech 2009*, Brighton, UK

Thomas S, Patil K, Ganapathy S, Mesgarani N, Hermansky H 2010 A phoneme recognition framework based on auditory spectro-temporal receptive fields. *Proc. Interspeech 2010*, Tokyo, 2458–2461

Tibrewala S, Hermansky H 1997 Multi-stream approach in acoustic modeling. *LVCSR-Hub5 Workshop*, Baltimore

Valente F, Hermansky H 2006 Discriminant linear processing of time-frequency plane. *ICSLP'98*, Pittsburgh

van Vuuren S, Hermansky H 1997 Data-driven design of RASTA-like filters. *Eurospeech'97*, ESCA, Rhodes, Greece

van Vuuren S, Hermansky H 1998 On the importance of components of the modulation spectrum for speaker verification. *ICSLP'98*, Sydney

von Helmholtz A 1863 Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik (On the sensations of tone as a physiological basis for the theory of music) Trans. Ellis. Kaufmann, London: Longmans, Green, and Co., 1875