[Richard M. Stern and Nelson Morgan]

Hearing Is Believing

[Biologically inspired methods for
robust automatic speech recognition]

© ISTOCKPHOTO.COM/SUCHOA LERTADIPAT

**FUNDAMENTAL TECHNOLOGIES
IN MODERN SPEECH RECOGNITION**

Many feature extraction methods that have been used for automatic speech recognition (ASR) have either been inspired by analogy to biological mechanisms, or at least have similar functional properties to biological or psychoacoustic properties for humans or other mammals. These methods have in many cases provided significant reductions in errors, particularly for degraded signals, and are currently experiencing a resurgence in community interest. Many of them have been quite successful, and others that are still in early stages of application still seem to hold great promise, given the existence proof of amazingly robust natural audio processing systems.

The common framework for state-of-the-art ASR systems has been fairly stable for about two decades now: a transformation of a short-term power spectral estimate is computed every 10 ms and then is used as an observation vector for Gaussian-mixture-based HMMs that have been trained on as much data as possible, augmented by prior probabilities for word sequences generated by smoothed counts from many examples. The most probable word sequence is chosen, taking into account both acoustic and language probabilities. While these systems have improved greatly over time, much of the improvement has arguably come from the availability of more data, more detailed models to take advantage of the greater amount of data, and larger computational resources to train and evaluate the models. Still other improvements have come from the increased use of discriminative training of the models. Additional gains have come from changes to the front end (e.g., normalization and compensation schemes), and/or from adaptation methods that could be implemented either in the stored models or through equivalent transformations of the incoming features. Even improvements obtained through discriminant training [such as with the minimum phone error (MPE) method] have been matched in practice by discriminant transformations of the features [such as feature MPE (fMPE)].

System performance is crucially dependent on advances in feature extraction, or on modeling methods that have their equivalents in feature transformation approaches. With degraded acoustical conditions (or, more generally, when there are mismatches between training and test conditions), it is even more important to generate features that are insensitive to nonlinguistic sources of variability. For instance, relative spectra (RASTA) processing and cepstral mean subtraction both have the primary effect of reducing sensitivity to linear channel effects. Significant problems of this variety remain—even as seemingly simple a task as the recognition of digit strings becomes extremely difficult when the utterances are corrupted by real-world noise and/or reverberation.

While machines struggle to cope with even modest amounts of acoustical variability, human beings can recognize speech remarkably well in similar conditions; a solution to the difficult problem of environmental robustness does indeed exist. While a number of fundamental attributes of auditory processing remain poorly understood, there are many instances in which analysis of psychoacoustical or physiological data can inspire signal processing research. This is particularly helpful because providing a particular conceptual framework for feature-extraction research makes feasible a search through a potentially very large set of possible techniques.

In many cases, researchers develop feature-extraction approaches without any conscious mimicking of any biological function. Even for such approaches, a post hoc examination of the method often reveals similar behavior.

In summary, the feature extraction stage of speech recognition is important historically and is the subject of much current research, particularly to promote robustness to acoustic disturbances such as additive noise and reverberation. Biologically inspired and biologically related approaches are an important subset of feature extraction methods for ASR.

BACKGROUND

STATE OF THE ART FOR ASR IN NOISE VERSUS HUMAN PERFORMANCE

Human beings have amazing capabilities for recognizing speech under conditions that still confound our machine implementations. In most cases ASR is still more errorful, even for speech signals with high signal-to-noise ratios (SNRs) [41]. More recent results show a reduced but still significant gap [47], [15]. For instance, the lowest word error rates on a standard conversational telephone recognition task are in the midteens, but interannotator error rates for humans listening to speech from this corpus have been reported to be around 4% for careful transcription. As Lippmann noted, for the 5,000-word *Wall Street Journal* task, human listeners' error rate was tiny and virtually the same for clean and noisy speech (for additive automotive noise at 10-dB SNR). Even the most effective current noise robustness strategies cannot approach this. There is, however, a caveat: human beings must pay close attention to the task to achieve these results.

Human hearing does much more than speech recognition. In addition, the nature of information processing by computers is inherently different from biological computation. Consequently, when implementing ASR mechanisms inspired by biology, we must take care to understand the function of each enhancement. Our goals as speech researchers are fundamentally different from those of biologists, whose aim is to create models that are functionally equivalent to the real thing.

What follows is a short summary of mammalian auditory processing, including the general response of the cochlea and the auditory nerve to sound, basic binaural analysis, some important attributes of feature extraction in the brain stem and primary auditory cortex, along with some relevant basic psychoacoustical results. We encourage the reader to consult texts and reviews such as [37] and [44] for a more complete exposition of these topics.

PERIPHERAL PROCESSING

PERIPHERAL FREQUENCY SELECTIVITY

The initial part of the auditory chain is a series of connected parts, moving signals forward. Time-varying air pressure associated with sound impinges on the ears, inducing small inward and outward motion of the tympanic membrane (eardrum). The eardrum is connected mechanically to the three bones in the middle ear, the malleus, incus, and stapes (or, more commonly, the hammer, anvil, and stirrup). The mechanical vibrations of the stapes induce wave motion in fluid in the spiral tube known as the cochlea. The basilar membrane is a structure that runs the length of the cochlea. It has a density and stiffness that vary along its length, causing its resonant frequency to vary as well. Affixed to the human basilar membrane are about 15,000 "hair cells," which innervate about 30,000 spiral ganglion cells whose axons form the individual fibers of the auditory nerve. Because of the spatially specific nature of this transduction, each fiber of the auditory nerve only responds to a narrow range of frequencies.

Sound at a particular frequency elicits vibration along a localized region of the basilar membrane, which in turn causes the initiation of neural impulses along fibers of the auditory nerve that are attached to that region of the basilar membrane. The frequency for which a fiber of the auditory nerve is the most sensitive is referred to as the characteristic frequency (CF) of that fiber. This portion of the auditory system is frequently modeled as a bank of bandpass filters (despite the many nonlinearities in the physiological processing), and the "bandwidth" of the filters appears to be approximately constant for fibers with CFs above 1 kHz when plotted as a function of log frequency. In other words, these physiological filters have a nominal bandwidth that is roughly proportional to center frequency. The bandwidth of the filters is roughly constant at lower frequencies. This frequency-based or "tonotopic" organization of individual parallel channels is generally maintained from the auditory nerve to higher centers of processing in the brain stem and the auditory cortex.

The basic description above is highly simplified, ignoring non-linearities in the cochlea and in the hair-cell response. In addition, there are actually two types of hair cells with systematic differences in response. The "inner" hair cells transduce and pass on the spectral representation of the signal produced by the basilar membrane to higher levels in the auditory system. The "outer" hair cells, which constitute the larger fraction of the total population, have a response that is affected in part by efferent feedback from higher centers of neural processing. They amplify the incoming signals nonlinearly, with low-level inputs amplified more than more intense ones, achieving a compression in dynamic range. Spikes generated by fibers of the auditory nerve occur stochastically, and hence the response of the nerve fibers must be characterized statistically.

THE RATE-INTENSITY FUNCTION

While the peripheral auditory system can roughly be modeled as a sequence of operations from the original arrival of sound pressure at the outer ear to its representation at the auditory nerve (ignoring the likely importance of feedback), these components are not linear. In particular, the rate-intensity function is roughly S-shaped, with a low and fairly constant rate of response for intensities below a threshold, a limited range of about 20 dB in which the response rate increases in roughly linear proportion to the signal intensity, and a saturation region for higher intensities. Since the range of intensities between the lowest detectable sounds and those that can cause pain (and damage) is about 120 dB, compression is a critical part of hearing. The small dynamic range of individual fibers is mitigated to some extent by the variations in fiber threshold intensities at each CF, as well as the use of a range of CFs to represent sounds.

SYNCHRONY OF TEMPORAL RESPONSE TO INCOMING SOUNDS

For low-frequency tones, e.g., below 5 kHz (in cats), auditory neurons are more likely to fire in phase with the stimulus, even though the exact firing times remain stochastic in nature. This results in a response that roughly follows the shape of the input signal on the average, at least when the signal amplitude is positive. This "phase locking" permits the auditory system to compare arrival times of signals to the two ears at low frequencies, a critical part of the function of binaural spatial localization. However, this property may also be important for robust processing of the signal from each individual ear. The extent to which the neural response at each CF is synchronized to the nearest harmonic of the fundamental frequency of a vowel, called the averaged localized synchronized rate (ALSR), is comparatively stable over a range of input intensities, while the mean rate of firing can vary significantly. This suggests that the timing information associated with the response to low-frequency components can be substantially more robust to intensity and other sources of signal variability than the mean rate of neural response. Note that the typical signal processing measures used in ASR are much more like mean-rate measures, and entirely ignore this timing-based information.

LATERAL SUPPRESSION

For more complex signals than pure tones, the response to signals at a given frequency may be suppressed or inhibited by energy at adjacent frequencies. The presence of a second tone over a range of frequencies surrounding the CF inhibits the response to the probe tone at CF, even for some intensities of the second tone that would be below threshold if it had been presented in isolation. This form of "lateral suppression" enhances the response to changes in the signal content with respect to frequency, just as overshoots and undershoots in the transient response have the effect of enhancing the response to changes in signal level over time.

As an example of the potential benefit that may be derived from such processing, Figure 1(a) depicts the spectrogram of an utterance from the TIMIT database for clean speech (left column) and speech in the presence of additive white noise at an SNR of 10 dB (right column). Figure 1(b) depicts a spectrogram of the same utterance derived from standard MFCC features. Finally, Figure 1(c) shows reconstructed spectrograms of the same utterance derived from a physiologically based model of the auditory-nerve response to sound proposed by [57] that incorporates the above phenomena. It can be seen that the formant trajectories are more sharply defined, and that the impact of the additive noise on the display is substantially reduced.

PROCESSING AT MORE CENTRAL LEVELS

SENSITIVITY TO INTERAURAL TIME DELAY AND INTENSITY DIFFERENCES

Two important cues for human localization of the direction of arrival of a sound are the interaural time difference (ITD) and interaural intensity difference (IID). ITDs are most useful at low frequencies and IIDs are only significant at higher frequencies for spatial aliasing and physical diffraction, respectively. Units in the superior olivary complex and the inferior colliculus appear to respond maximally to a single "characteristic" ITD, sometimes referred to as the characteristic delay (CD) of the unit. An ensemble of such units with a range of CFs and CDs can produce a display that represents the interaural cross-correlation of the signals to the two ears after the frequency-dependent and nonlinear processing of the auditory periphery.

SENSITIVITY TO AMPLITUDE MODULATION AND MODULATION FREQUENCY ANALYSIS

Physiological recordings in cochlear nuclei, the inferior colliculus, and the auditory cortex have revealed the presence of units that appear to be sensitive to the modulation frequencies of sinusoidally amplitude-modulated (SAM) tones (e.g., [22]). In some of these cases, response would be maximum at a particular modulation frequency, independently of the carrier frequency of the SAM tone complex, and some of these units are organized anatomically according to best modulation frequency. These results have led to speculation that the so-called modulation spectrum may be a useful and consistent way to describe the
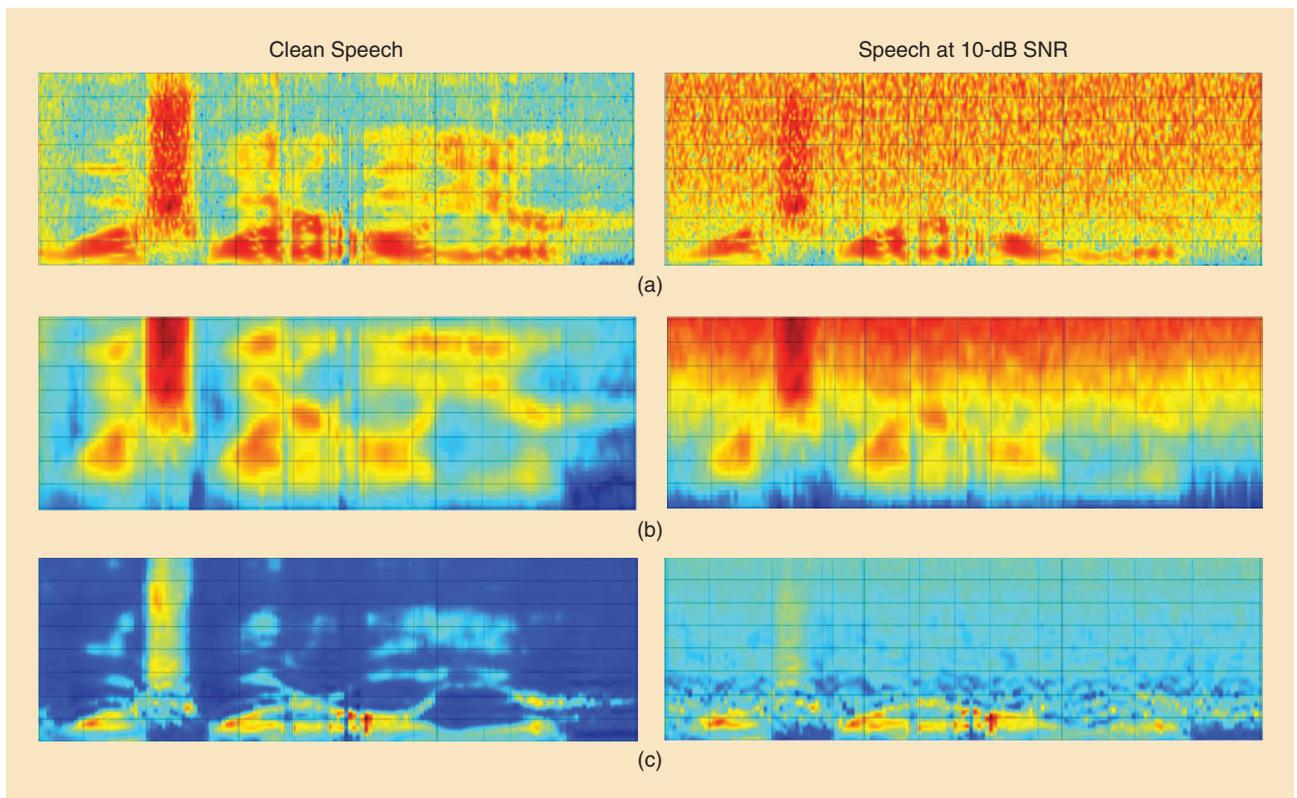
[FIG1] (a) Original spectrogram: MATLAB spectrogram of speech in the presence of additive white noise at an SNR of 10 dB. (b) Mel-freqency cepstral coefficients (MFCCs) spectrogram: Reconstructed spectrogram of the utterance after traditional MFCC processing. (c) Peripheral auditory spectrogram: Reconstructed spectrogram of the utterance after peripheral auditory processing based on the model by Zhang et al. [57]. The columns represent responses to clean speech (left) and speech in white noise at an SNR of 10 dB (right).

dynamic temporal characteristics of complex signals like speech after the peripheral frequency analysis. However, psychoacoustic results show that lower modulation frequencies appear to have greater significance for phonetic identification.

SPECTROTEMPORAL RECEPTIVE FIELDS

Finally, the firing patterns of neurons in the A1 cortical spiking in ferrets show that neurons in A1 (primary auditory cortex) are highly tuned to specific spectral and temporal modulations (as well as being tonotopically organized by frequency, as in the auditory nerve) [9]. The sensitivity patterns of these neurons are often referred to as spectrotemporal receptive fields (STRFs) and are often illustrated as color temperature patterns on an image of the time-frequency plane.

PSYCHOPHYSICAL PHENOMENA
THAT HAVE MOTIVATED AUDITORY MODELS

Psychoacoustical as well as physiological results have enlightened us about the functioning of the auditory system. In fact, interesting auditory phenomena are frequently first revealed through psychoacoustical experimentation, with the probable physiological mechanism underlying the perceptual observation identified at a later date. We briefly discuss several sets of basic psychoacoustic observations that have played a major role in auditory modeling.

AUDITORY FREQUENCY RESOLUTION

We have noted above that many physiological results suggest that the auditory system performs a frequency analysis, which is typically approximated in auditory modeling by a bank of linear filters. Beginning with the pioneering efforts of Harvey Fletcher and colleagues in the 1940s auditory researchers have attempted to determine the shape of these "filters" and their effective bandwidth (frequently referred to as the critical band) through the use of many clever psychoacoustical experiments.

Three distinct frequency scales have emerged from these experiments that describe the putative bandwidths of the auditory filters. The Bark scale (named after Heinrich Barkhausen, who proposed the first subjective measurements of loudness) is based on the results of traditional masking experiments [58]. The Mel scale (referring to the word "melody") is based on pitch comparisons [54]. The equivalent rectangular bandwidth (ERB) scale was developed to describe the results of several experiments using differing procedures [38]. Despite the differences in how they were developed, the Bark, Mel, and ERB scales describe a very similar dependence of auditory bandwidth on frequency, implying constant-bandwidth filters at low frequencies and constant-Q filters at higher frequencies, consistent with the physiological data. All common models of auditory processing begin with a bank of filters whose center frequencies and bandwidths are based on one of these three frequency scales.

THE PSYCHOACOUSTICAL TRANSFER FUNCTION

The original psychoacousticians were physicists and philosophers of the 19th century who sought to develop mathematical functions that related sensation and perception, such as the dependence of the subjective loudness of a sound on its physical intensity. Two different "psychophysical" scales for intensity have emerged over the years. The first, developed by Gustav Fechner, was based on the 19th-century empirical observations of Weber, who observed that the just-noticeable increment in intensity was a constant fraction of the reference intensity level. MFCCs and other common features use a logarithmic transformation that is consistent with this observation as well as the assumption that equal increments of perceived intensity should be marked by equal intervals along the perceptual scale.

Many years after Weber's observations, Stevens [53] proposed an alternate loudness scale, which implicitly assumes that perceived ratios in intensity should represent equal ratios (rather than increments) on the perceptual scale. This suggests a nonlinearity in which the incoming signal intensity is raised to a power to approximate the perceived loudness. This approach is supported by psychophysical experiments in which subjects directly estimate perceived intensity in many sensory modalities, with an exponent of approximately 0.33 fitting the results for experiments in hearing. This compressive power-law nonlinearity has been incorporated into PLP processing and other feature extraction schemes.

AUDITORY THRESHOLDS AND PERCEIVED LOUDNESS

The human threshold of hearing varies with frequency, as the auditory system achieves the greatest sensitivity between about 1 and 4 kHz with a minimum threshold of about –5 dB SPL [11]. The threshold of hearing increases for both lower and higher frequencies of stimulation. The human response to acoustical stimulation undergoes a transition from hearing to pain at a level of very roughly 110 dB SPL for most frequencies. The frequency dependence of the threshold of hearing combined with the relative frequency independence of the threshold of pain causes the perceived loudness of a narrowband sound to depend on both its intensity and frequency.

NONSIMULTANEOUS MASKING

Nonsimultaneous masking occurs when the presence of a masker elevates the threshold intensity for a target that precedes or follows it. Forward masking refers to inhibition of the perception of a target after the masker is switched off. When a masker follows the probe in time, the effect is called backward masking. Masking effects decrease as the time between masker and probe increases, but can persist for 100 ms or more [37].

THE PRECEDENCE EFFECT

Another important attribute of human binaural hearing is that binaural localization is dominated by the first-arriving components of a complex sound. This phenomenon, which is referred to as the precedence effect, is clearly helpful in enabling the perceived location of a source in a reverberant environment to remain constant, as it is dominated by the characteristics of the components of the sound which arrive directly from the sound source while suppressing the potential impact of later-arriving reflected components from other directions. In addition to its role in maintaining perceived constancy of direction of arrival in reverberation, the precedence effect is also believed by some to improve speech intelligibility in reverberant environments, although it is difficult to separate the potential impact of the precedence effect from that of conventional binaural unmasking.

BIOLOGICALLY RELATED METHODS IN CONVENTIONAL FEATURE EXTRACTION FOR ASR

The overwhelming majority of speech recognition systems today make use of features that are based on either MFCCs [7] or features based on perceptual linear predictive (PLP) analysis of speech [16]. In this section, we discuss how MFCC and PLP coefficients are already heavily influenced by knowledge of biological signal processing. Both MFCC and PLP features make explicit or implicit use of the frequency warping implied by psychoacoustical experimentation [the Mel scale for MFCC parameters and the Bark scale for PLP analysis as well as psychoacoustically motivated amplitude compression (the log scale for MFCC, and a power-law compression for PLP coefficients)]. The PLP features include additional attributes of auditory processing including more detailed modeling of the asymmetries in frequency response implied by psychoacoustical measurements of auditory frequency selectivity [48], pre-emphasis based on the loudness contours of Fletcher and Munson [11], among other phenomena. Finally, both MFCC and PLP features have mechanisms to reduce the sensitivity to changes in the long-term average log power spectra, through the use of cepstral mean normalization or RASTA processing [18].

While the primary focus of this article is on ASR algorithms that are inspired by mammalian recognition (especially human), we would be remiss in not mentioning a now-standard method that is based on a simplified model of speech production. What is commonly called vocal tract length normalization (VTLN) is a further scaling of the frequency axis based on the notion that vocal tract resonances are higher for shorter vocal tracts. In practice, vocal tract measurements are not available, and a true customization of the resonance structure for an individual would be quite complicated. This is typically implemented using a simplified frequency scaling function, which is most commonly a piecewise-linear approach warping function to account for edge effects as suggested by Cohen et al. [6], obtaining the best values of the free parameters statistically.

## FEATURE EXTRACTION BASED ON MODELS OF THE AUDITORY SYSTEM

The trends and results obtained over three decades involving feature extraction based on computational models of the auditory system are reviewed and discussed in this section. In citing specific research results, we have attempted to adopt a broad perspective that includes representative work from most of the relevant conceptual categories. Nevertheless, we recognize that in a

brief review such as this it is necessary to omit some important contributors. We apologize in advance to the many researchers whose relevant work is not included here.

"CLASSICAL" AUDITORY MODELS OF THE 1980s

Much work in feature extraction based on physiology is based on three seminal auditory models developed in the 1980s by Seneff [49], Ghitza [14], and Lyon [31]. All of these models included a description of processing at the cochlea and the auditory nerve that is far more physiologically accurate and detailed than the processing used in MFCC and PLP feature extraction, including more realistic auditory filtering, more realistic nonlinearities relating stimulus intensity to auditory response rate, synchrony extraction at low frequencies, lateral suppression (in Lyon's model), and higher-order processing through the use of either cross-correlation processing or autocorrelation processing. The early stages of all three models are depicted in a general sense by the blocks in Figure 2(a). Specifically, Seneff's model proposed a generalized synchrony detector (GSD) that implicitly provided the autocorrelation value at lags equal to the reciprocal of the center frequency of each processing channel. Ghitza proposed ensemble-interval histogram (EIH) processing, which developed a spectral representation by computing times between a set of amplitude-threshold crossings of the incoming signal after peripheral processing. Lyon's model [31] included many of the same components found in the Seneff and Ghitza models, producing a display referred to as a cochleagram, which serves as a more physiologically accurate alternative to the familiar spectrogram. Lyon also proposed a correlogram representation based on the running autocorrelation function at each frequency of the incoming signal, again after peripheral processing, as well as the use of interaural cross correlation to provide the separation of incoming signals by direction of arrival [32], building on earlier theories by Jeffress [21] and Licklider [29].

There was very little quantitative evaluation of the three auditory models in the 1980s, in part because they were computationally costly for that time. In general, these approaches provided no benefit in recognizing clean speech compared to MFCC/PLP representations, but they did improve recognition accuracy when the input was degraded by noise and/or reverberation (e.g., [14] and [40]). In general, work on auditory modeling in the 1980s failed to gain traction, not only because of the computational cost, but also because of a poor match between the statistics of the features and the statistical assumptions built into standard ASR modeling. In addition, there were more pressing fundamental shortcomings in speech recognition technology that first needed to be resolved. Despite this general trend, there was at least one significant instantiation of an auditory model in a major large vocabulary system in the 1980s, specifically, IBM's Tangora [5]. This front end incorporated, in addition to most of the other properties above, a form of short-term adaptation inspired by the auditory system. Other important contemporary work included the auditory models of Deng and Geisler [8], Payton [43], and Patterson et al. [42].
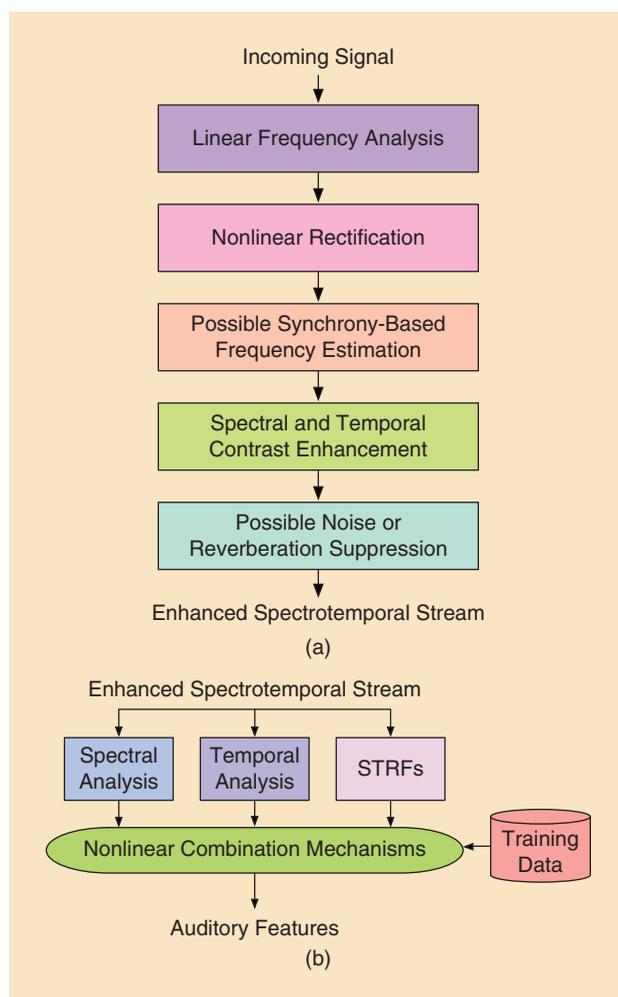

[FIG2] (a) Generalized organization of contemporary feature extraction procedures that are analogous to auditory periphery function. Not all blocks are present in all feature extraction approaches, and the organization may vary. (b) Processing of the spectrotemporal feature extraction by spectral analysis, temporal analysis, or the more general case of (possibly many) STRFs. These can incorporate long temporal support, can comprise a few or many streams, and can be combined simply or with discriminative mechanisms incorporating labeled training data.

CONTEMPORARY ENABLING TECHNOLOGIES AND CURRENT TRENDS IN FEATURE EXTRACTION

While the classical models described in the previous section are all more or less complete attempts to model auditory processing at least at the auditory-nerve level to varying degrees of abstraction, there have been a number of other current trends that have been motivated directly or indirectly by auditory processing that have influenced feature extraction in a more general fashion, even for features that are not characterized specifically as representing auditory models.

MULTISTREAM PROCESSING

The articulation index model of speech perception, which was suggested by Fletcher [10] and French and Steinberg [12], and revived by Allen [2], modeled phonetic speech recognition as
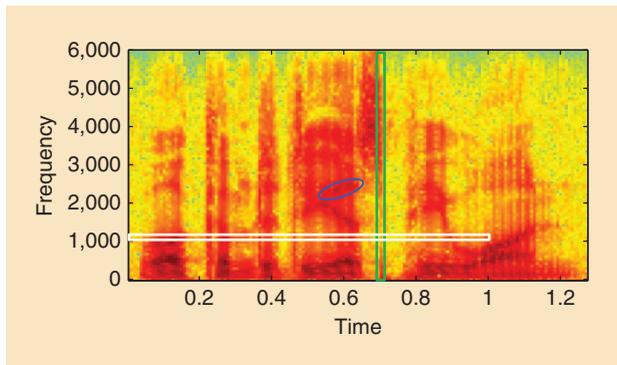
[FIG3] Comparison of the standard analysis window used in MFCC, PLP, and other feature sets, which computes features over a vertical slice of the time-frequency plane (green box) with the horizontal analysis window used in TRAPS analysis (white box), and the oblique ellipse that represents a possible STRF for detecting an oblique formant trajectory (blue ellipse).

arising from independent estimators for critical bands. This led to a great deal of interest initially in the development of multiband systems based on this view of independent detectors per critical band that were developed to improve robustness of speech recognition, particularly for narrowband noise (e.g., [20]). This approach in turn can be generalized to the consideration of fusion of information from parallel detectors that are presumed to provide complementary information about the incoming speech. This information can be combined at the input (feature) level [39], at the level at which the HMM search takes place [33], or at the output level by merging hypothesis lattices [34]. The incorporation of multiple streams with different modulation properties can be done in a number of ways, many of which require nonlinear processing. This integration is depicted in a general sense in Figure 2(b).

LONG-TIME TEMPORAL EVOLUTION

An important parallel trend has been the development of features that are based on the temporal evolution of the envelopes of the outputs of the bandpass filters that are part of any description of the auditory system. Human sensitivity to overall temporal modulation has also been documented with perceptual experiments. As noted earlier, spectrotemporal receptive fields have been observed in animal cortex, and these fields are sometimes much more extended in time than the typical short-term spectral analysis used in calculating MFCCs or PLP.

Initially, information about temporal evolution has been used to implement features based on frequency components of these temporal envelopes, which is referred to as the modulation spectrum [27]. Subsequently, various groups have characterized these patterns using nonparametric models as in the temporal patterns (TRAPS) and hidden activation TRAPS (HATS) methods (e.g., [19]) or using parametric all-pole models such as frequency-domain linear prediction (FDLP) (e.g., [3]). It is worth noting that RASTA, mentioned earlier, was developed to emphasize the critical temporal modulations, and in so doing emphasize transitions (as was suggested in

perceptual studies such as Furui [13]), and reduce sensitivity to irrelevant steady state convolutional factors. More recently, temporal modulation in subbands was normalized to improve ASR in reverberant environments [30].

SPECTROTEMPORAL RECEPTIVE FIELDS

Two-dimensional (2-D) Gabor filters are obtained by multiplying a 2-D sinewave by a 2-D Gaussian. Their frequency response can be used to model the spectrotemporal receptive fields of A1 neurons. They also have the attractive properties of being self-similar, and they can be generated from a mother wavelet by dilation and rotation. Mesgarani et al. [36] have used 2-D Gabor filters successfully in implementing features for speech/nonspeech discrimination and similar approaches have also been used to extract features for ASR by multiple researchers (e.g., [28]). In many of these cases, multilayer perceptrons (MLPs) were used to transform the filter outputs into a form that is more amenable to use by Gaussian mixture-based HMMs, typically using the tandem approach [17].

Figure 3 compares typical analysis regions in time and frequency that are examined in 1) conventional frame-based spectral analysis, 2) the long-term temporal analysis employed by TRAPS processing, and 3) the generalized spectrotemporal analysis that can be obtained using STRFs. The central column of Figure 4 illustrates a pair of typical model STRFs obtained from 2-D Gabor filters. As can be seen, each of these filters elicit responses from different aspects of the input spectrogram, resulting in the parallel representations seen in the column on the right.

FEATURE EXTRACTION WITH TWO OR MORE "EARS"

All of the attributes of auditory processing cited above are essentially single-channel in nature. It is well known that human listeners compare information from the two ears to localize sources in space and separate sound sources that are arriving from different directions, a process generally known as binaural hearing. Source localization and segregation is largely accomplished by estimating the ITD and the IID of the signals arriving at the two ears as reviewed in [51]. In addition, the precedence effect, which refers to emphasis placed on the early-arriving components of the ITD and IID, can substantially improve sound localization and speech understanding in reverberant and other environments.

Feature extraction systems based on binaural processing have taken on several forms. The most common application of binaural processing is through the use of systems that provide selective reconstruction of spatialized signals that have been degraded by noise and/or reverberation by selecting those spectrotemporal components after short-time Fourier analysis that are believed to be dominated by the desired sound source (e.g., [46], [1], and [23]). These systems typically develop binary or continuous masks in the time-frequency plane for separating the signals, using the ITD or interaural phase difference (IPD) of the signals as the basis for the separation. Correlation-based emphasis is a second approach that is based on binaural
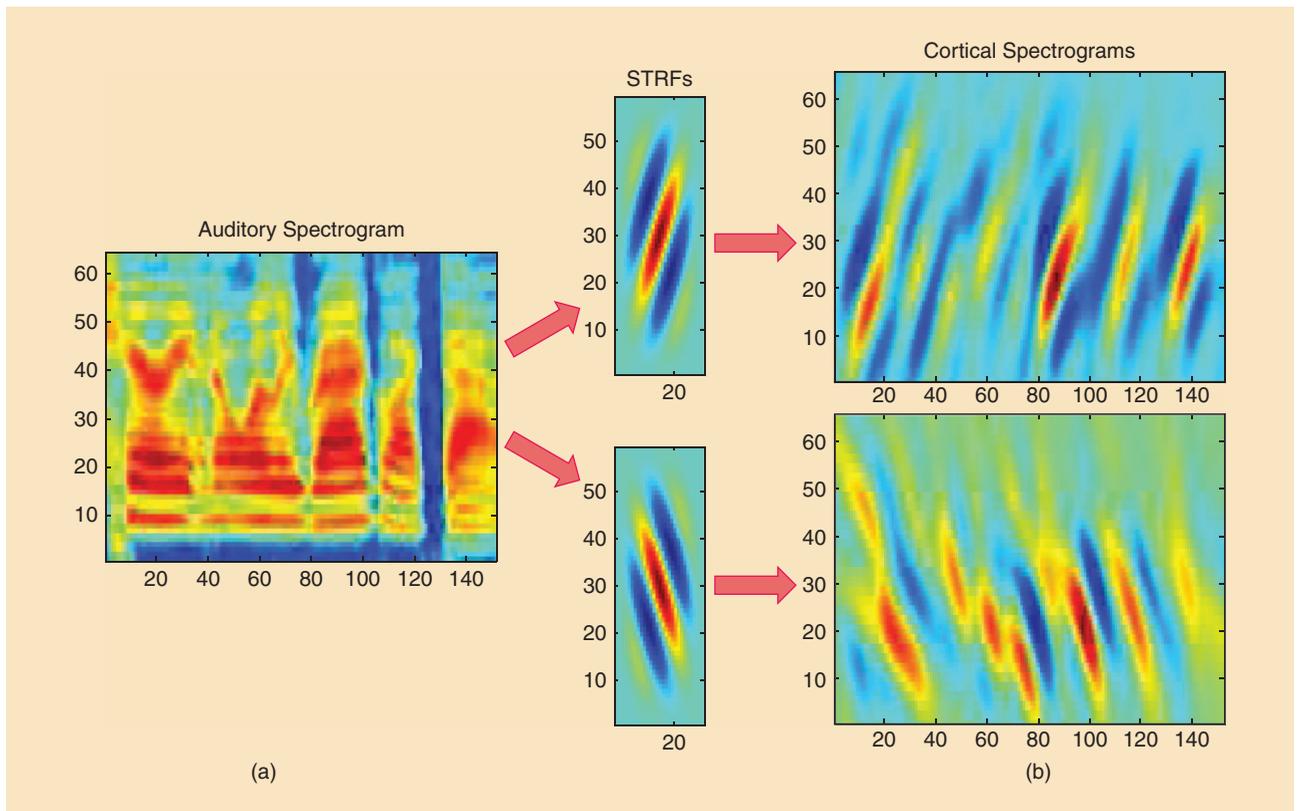
[FIG4] Two decompositions of spectrograms using differing STRFs. The spectrogram in (a) is correlated with STRFs of differing rate and scale, producing the parallel representations as in (b).

hearing that operates in a fashion similar to multimicrophone beamforming, but with additional nonlinear enhancement of the desired signal based on interaural cross correlation of signals after a model of peripheral auditory processing [52]. Finally, several groups have demonstrated that processing based on the precedence effect, both at the monaural level (e.g., [35]) and at the binaural level (e.g., [59]), typically through the use of enhancement of the leading edge of envelopes of the outputs of the auditory filters. This type of processing has been shown to be particularly effective in reverberant environments.

CONTEMPORARY AUDITORY MODELS

Auditory modeling enjoyed a renaissance during the 1990s and beyond for several reasons. First, the cost of computation became much less of a factor because of continued developments in computer hardware and systems software. Similarly, the development of efficient techniques to learn the parameters of Gaussian mixture models for observed feature distributions mitigated the problem of mismatch between the statistics of the incoming data and the assumptions underlying the stored models, and techniques for discriminative training potentially provide effective ways to incorporate features with very different statistical properties. We briefly cite a small number of representative complete models of the peripheral auditory system as examples of how the concepts discussed above can be incorporated into complete feature extraction sys-

tems. These feature extraction approaches were selected because they span most of the auditory processing phenomena cited above; again, we remind the reader that this list is far from exhaustive.

TCHORZ AND KOLLMEIER

Tchorz and Kollmeier [55] developed an early "modern" physiologically motivated feature extraction system, updating the classical auditory modeling elements. In addition to the basic stages of filtering and nonlinear rectification, their model also included adaptive compression loops that provided enhancement of transients, some forward and backward masking, and short-term adaptation. These elements also had the effect of imparting a specific filter for the modulation spectrum, with maximal response to modulations in the spectral envelope occurring in the neighborhood of 6 Hz. This model was initially evaluated using a database of isolated German digits with additive noise of different types, and achieved significant error rate reductions in comparison to MFCCs for all of the noisy conditions. It has also maintained good performance in a variety of other evaluations.

KIM, LEE, AND KIL

A number of researchers have proposed ways to obtain an improved spectral representation based on neural timing information. One example is the zero-crossing peak analysis (ZCPA) proposed by [26], which can be considered to be a

simplification of the level-crossing methods developed by Ghitza. The ZCPA approach develops a spectral estimate from a histogram of the inverse time between successive zero crossings in each channel after peripheral bandpass filtering, weighted by the amplitude of the peak between those crossings. This approach demonstrates improved robustness for ASR of isolated Korean words in the presence of various types of additive noise. The ZCPA approach is functionally similar to aspects of Sheikhzadeh and Deng's earlier model [50], which also develops a histogram of the interpeak intervals of the putative instantaneous firing rates of auditory-nerve fibers, and weights the histogram according to the amplitude of the initial peak in the interval. Their model makes use of the very detailed composite auditory model proposed by Deng and Geisler [8].

KIM AND STERN
Kim and Stern [25] described a feature set called power normalized cepstral coefficients (PNCCs), incorporating relevant physiological phenomena in a computationally efficient fashion. PNCC processing includes 1) traditional pre-emphasis and short-time Fourier transformation, 2) integration of the squared energy of the STFT outputs using gammatone frequency weighting, 3) "medium-time" nonlinear processing that suppresses the effects of additive noise and room reverberation, 4) a power-function nonlinearity with exponent 1/15, and 5) generation of cepstral-like coefficients using a discrete cosine transform (DCT) and mean normalization. For the most part, noise and reverberation suppression is accomplished by a nonlinear series of operations that accomplish running noise suppression and temporal contrast enhancement, working in a "medium-time" context with analysis intervals on the order of 50–150 ms. PNCC processing has been found by the CMU group and independent researchers to outperform baseline processing as well as several systems developed specifically for noise robustness such as the European Telecommunications Standards Institute advanced front end. This approach with minor modifications is also quite effective in reverberation [23].

CHI, RU, AND SHAMMA
In a seminal paper, Chi et al. [4] presented a new abstract model of the putative representation of sound at both the peripheral level and in the auditory cortex, based on the research by Shamma's group and others. The model describes a representation with three independent variables: auditory frequency, "rate" (characterizing temporal modulation), and "scale" (characterizing spectral modulations), as would be performed by successive stages of wavelet processing. The model relates this representation to feature extraction at the level of the brain stem and the cortex, including detectors based on STRFs, incorporating a cochlear filterbank at the input to the STRF filtering. Chi et al. also generated speech from the model outputs.

RAVURI
In his 2011 thesis, Ravuri [45] describes a range of experiments incorporating over 100 2-D Gabor filters, each implementing a single STRF, and each with its own discriminatively trained neural network to generate noise-insensitive features for ASR. The STRF parameters were chosen to span a range of useful values for rate and scale, as determined by many experiments, and then were applied separately to each critical band. The system thus incorporated multiple streams comprising discriminative transformations of STRFs, many of which also were focused on long temporal evolution. The effectiveness of the representation was demonstrated for Aurora2 noisy digits and for a noisy version of the Numbers 95 data set.

SUMMARY
Feature extraction methods based on an understanding of both auditory physiology and psychoacoustics have been incorporated into ASR systems for decades. In recent years, there has been a renewed interest in the development of signal processing procedures based on much more detailed characterization of hearing by humans and other mammals. It is becoming increasingly apparent that the careful implementation of physiologically based and perceptually based signal processing can provide substantially increased robustness in situations in which speech signals are degraded by interfering noise of all types, channel effects, room reverberation, and other sources of distortion. And the fact that humans can hear and understand speech, even under conditions that confound our current machine recognizers, makes us believe that there is more to be gained through a greater understanding of human speech recognition—hearing is believing.

ACKNOWLEDGMENTS

AUTHORS
Richard M. Stern (rms@cs.cmu.edu) is currently a professor in the Department of Electrical and Computer Engineering and in the Language Technologies Institute as well as a lecturer in the School of Music at CMU, where he has been since 1977. He is well known for work in robust speech recognition and binaural perception, and his group has developed algorithms for statistical compensation for noise and filtering, missing-feature recognition, multimicrophone techniques, and physiologically motivated feature extraction. He is a Fellow of the Acoustical Society of America and the International Speech Communication Association (ISCA). He was the ISCA Distinguished Lecturer in 2008–2009 and the general chair of Interspeech 2006.

Nelson Morgan (morgan@icsi.berkeley.edu) has led speech research at ICSI since 1988 and is a professor in residence at the

University of California, Berkeley. He has over 200 publications, including a speech and audio textbook coauthored with Ben Gold (with a new edition prepared with Dan Ellis of Columbia). He was formerly coeditor-in-chief of *Speech Communication*. He received the 1997 IEEE Signal Processing Magazine Best Paper Award. He jointly developed RASTA processing with Hynek Hermansky, and hybrid HMM/MLP systems with H. Bourlard. He is a Fellow of the IEEE and ISCA.

REFERENCES

[1] P. Aarabi and G. Shi "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Syst., Man, Cybern., B*, vol. 34, no. 4, pp. 1763–1773, 2004.

[2] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio*, vol. 2, no. 4, pp. 567–577, 1994.

[3] M. Athineos and D. Ellis, "Frequency-domain linear prediction for temporal features," in *Proc. IEEE ASRU Workshop*, 2003, pp. 261–266.

[4] T. Chi, P. Ru, and S. A. Shamma "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 887–906, 2005.

[5] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, vol. 85, no. 6, pp. 2623–2629, 1989.

[6] J. R. Cohen, T. Kamm, and A.G. Andreou, "Vocal track normalization in speech recognition: Compensating for systematic speaker variability," *J. Acoust. Soc. Amer.*, vol. 97, no. 5, pp. 3246–3247, 1995.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[8] L. Deng and D. C. Geisler, "A composite auditory model for processing speech sounds," *J. Acoust. Soc. Amer.*, vol. 82, no. 6, pp. 2001–2012, 1987.

[9] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 85, no. 3, pp. 1220–1234, 2001.

[10] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, no. 1, pp. 47–65, 1940.

[11] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement, and calculation," *J. Acoust. Soc. Amer.*, vol. 5, no. 2, pp. 82–108, 1933.

[12] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.* vol. 19, no. 1, pp. 90–119, 1947.

[13] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, vol. 80, no. 4, pp. 1016–1025, 1986.

[14] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, no. 2, pp. 109–130, 1986.

[15] M. Glenn, S. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, "Transcription methods for consistency, volume and efficiency," in *Proc. LREC 2010*, Valletta, Malta.

[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[17] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, Istanbul, Turkey, 2000, pp. 1635–1638.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[19] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. IEEE ICASSP*, Phoenix, AZ, 1999, pp. 1255–1258.

[20] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on partially corrupted speech," in *Proc. ICSLP*, Philadelphia, PA, 1996, vol. 1, pp. 462–465.

[21] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, no. 1, pp. 35–39, 1948.

[22] P. X. Joris, C. E. Schreiner, and A. Rees, "Neural processing of amplitude-modulated sounds," *Physiol. Rev.*, vol. 84, no. 2, pp. 541–577, 2004.

[23] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *Proc. Interspeech*, Makuhari, Japan, 2010.

[24] C. Kim, R. M. Stern, K. Eom, and J. Lee, "Automatic selection of thresholds for signal separation algorithms based on interaural delay," in *Proc Interspeech*, Makuhari, Japan, 2010.

[25] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Processing*, to be published.

[26] D. Kim, S. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real world noisy environments," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 1, pp. 55–69, 1999.

[27] B. Kingsbury, "Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments," Ph.D. dissertation, Univ. California, Berkeley, Dec. 1998.

[28] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proc. Eurospeech*, 2003, pp. 2573–2576.

[29] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, no. 4, pp. 128–134, 1951.

[30] X. Lu, M. Unoki , and S. Nakamura. "Subband temporal modulation spectrum normalization for automatic speech recognition in reverberant environments," in *Proc. Interspeech 2009*.

[31] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proc. ICASSP*, Paris, France, 1982, pp. 1282–1285.

[32] R. F. Lyon, "A computational model of binaural localization and separation," in *Proc. ICASSP*, Boston, MA, 1983, pp. 1148–1151.

[33] C. Ma, K.-K. J. Kuo, H. Soltau, X. Cui, U. Chaudhari, L. Mangu, and C.-H. Lee, "A comparative study on system combination schemes for LVCSR," in *Proc. ICASSP 2010*, Dallas, TX, pp. 4394–4397.

[34] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition; word error minimization and other applications of confusion networks," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 373–400, 2000.

[35] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE ASSP Workshop Applications of Signal Processing to Audio and Acoustics*, 1997.

[36] N. Mesgarani, M. Slaney, and S. Shamma. "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 14, no. 3, pp. 920–929, 2006.

[37] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. London: Academic Press, 2003.

[38] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 750–753, 1983.

[39] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Processing* (Special Issue on Deep Learning), vol. 20, no. 1, pp. 7–13, Jan. 2012.

[40] Y. Ohshima and R. M. Stern, "Environmental robustness in automatic speech recognition using physiologically motivated signal processing," in *Proc. ICSLP 1994*, pp. 1619–1622.

[41] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, no. 1, pp. 1–15, 1997.

[42] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception (Proc. 9th Int. Symp. Hearing)*, Y Cazals, L. Demany, and K. Horner, Eds. Oxford: Pergamon, 1992, 429–446.

[43] K. L. Payton, "Vowel processing by a model of the auditory periphery: A comparison to eighth-nerve responses," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 145–162, 1988.

[44] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 3rd ed. New York: Academic Press, 2008.

[45] S. Ravuri, "On the use of spectro-temporal features in noise-additive speech," M.S. thesis, Dept. Elect. Eng. Comput. Sciences, Univ. California, Berkeley, Spring 2011.

[46] N. Roman, DeL. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[47] O. Scharenborg and M. Cooke, "Comparing human and machine recognition performance on a VCV corpus," in *Proc. Workshop Speech Analysis and Processing for Knowledge Discovery*, Aalborg, 2008.

[48] M. R. Schroeder, "Recognition of complex acoustic signals," in *Life Sciences Research Report 5*, T. H. Bullock, Ed. Abakon Verlag, 1977.

[49] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonet.*, vol. 15, no. 1, pp. 55–76, 1988.

[50] H. Sheikhzadeh and L. Deng, "Speech analysis and recognition using interval statistics generated from a composite auditory model," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 1, pp. 50–54, 1998.

[51] R. M. Stern, G. J. Brown, and D. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis*, D. Wang and G. J. Brown, Eds. New York: IEEE Press, 2006, pp. 147–185.

[52] R. M. Stern, E. Gouvêa, and G. Thattai, "'Polyaural' array processing for automatic speech recognition in degraded environments," in *Proc. Interspeech 2007*.

[53] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, pp. 153–181, 1957.

[54] S. S. Stevens, J. Volkman, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, 1937.

[55] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 2040–2060, 1999.

[56] T. Virtanen, R. Singh, and B. Raj, Eds., *Noise Robust Techniques for Automatic Speech Recognition*. Hoboken, NJ: Wiley, 2012.

[57] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the response of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *J. Acoust. Soc. Amer.*, vol. 109, no. 2, pp. 648–670, 2001.

[58] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoustic. Soc. Amer.*, vol. 33, no. 248, , p. 248, 1961.

[59] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibltion: I Simulation of lateralizaton for stationary signals," *J. Acoustic. Soc. Amer.*, vol. 80, no. 6, pp. 1608–1622.

[SP]