8

Features Based on Auditory Physiology and Perception

Richard M. Stern and Nelson Morgan

Carnegie Mellon University
and *The International Computer Science Institute and the University of California at Berkeley*

8.1 Introduction

It is well known that human speech processing capabilities far surpass the capabilities of current automatic speech recognition and related technologies, despite very intensive research in automated speech technologies in recent decades. Indeed, since the early 1980's, this observation has motivated the development of speech recognition feature extraction approaches that are inspired by auditory processing and perception, but it is only relatively recently that these approaches have become effective in their application to computer speech processing. The goal of this chapter is to review some of the major ways in which feature extraction schemes based on auditory processing have facilitated greater speech recognition accuracy in recent years, as well as to provide some insight into the nature of current trends and future directions in this area.

We begin this chapter with a brief review of some of the major physiological and perceptual phenomena that have motivated feature extraction algorithms based on auditory processing. We continue with a review and discussion of three seminal 'classical' auditory models of the 1980s that have had a major impact on the approaches taken by more recent contributors to this field. Finally, we turn our attention to selected more recent topics of interest in auditory feature analysis, along with some of the feature extraction approaches that have been based on them. We conclude with a discussion of the attributes of auditory models that appear to be most effective in improving speech recognition accuracy in difficult acoustic environments.

8.2 Some Attributes of Auditory Physiology and Perception

In this section we very briefly review and discuss a selected set of attributes of auditory physiology that historically or currently have been the object of attention by developers of

Techniques for Noise Robustness in Automatic Speech Recognition Virtanen, Singh, and Raj (eds)

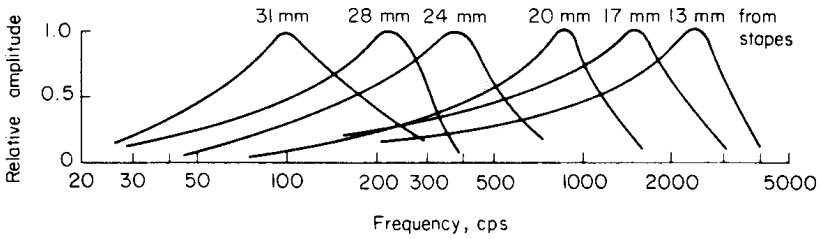

Figure 8.1: The response of the basilar membrane as a function of frequency, measured at six different distances from the stapes. As the frequency axis is plotted on a logarithmic scale, it can be easily seen that the effective bandwidth is proportional to center frequency at higher frequencies; effective bandwidth is roughly constant at lower frequencies. From [113].

auditory-based features. This discussion has been simplified for clarity's sake at the expense of other interesting phenomena that have received less attention in constructing models, at least to date, and it is far from comprehensive, even with respect to the auditory response to speech sounds. Furthermore, the physiological response to speech sounds is the object of much current attention, so that any report on current progress will inevitably be quickly out of date. The reader is referred to standard texts and reviews in physiology and psychoacoustics (*e.g.*[77, 89, 118]) for more comprehensive descriptions of general auditory physiology as well as the psychoacoustical response to speech sounds. The physiological response of the auditory system to speech and speech-like sounds is described in [86], among other places.

8.2.1 Peripheral processing

Mechanical response to sound. The peripheral auditory response to sound has been well documented over a period of many years. Very briefly, small increases and decreases in pressure of the air that impinges on the two ears induce small inward and outward motion on the part 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 *cochlea* is the organ that converts the mechanical vibrations in the middle ear to neural impulses that can be processed by the brainstem and brain. The cochlea can be thought of as a fluid-filled spiral tube, and the mechanical vibrations of the structures of the middle ear induce wave motion of the fluid in the cochlea. The *basilar membrane* is a structure that runs the length of the cochlea. As one moves from the basal end of the cochlea (closest to the stapes) to the apical end (away from the stapes), the stiffness of the basilar membrane decreases, causing its fundamental resonant frequency to decrease as well. Figure 8.1 illustrates some classical measurements of cochlear motion by Georg von Békésy [113] which were obtained using stroboscopic techniques in the 1940s. These curves show that the membrane responds to high-frequency tones primarily at the basal end, while low-frequency signals elicit maximal vibration at the apical end, although the response to low-frequency sounds is more asymmetric and distributed more broadly along the membrane.

Affixed to the human basilar membrane are about 15,000 hair cells, which enervate about 30,000 individual fibers of the auditory nerve. Through an electrochemical mechanism, the

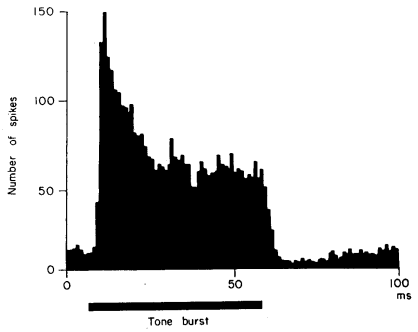

Figure 8.2: PST histograms in response to tone bursts. From [48].

mechanical motion of the hair cells elicits the generation of a brief transient or 'spike' in the voltage inside the cell wall. This transient is then propagated along the nerve fibers and beyond to the cochlear nucleus and subsequently to higher centers in the brainstem and the brain. The most important attribute of these spikes is the time at which they take place. Because each nerve fiber takes input from a relatively small number of fibers that in turn move in response to vibration over only a limited length along the basilar membrane, and because a given location along the basilar membrane is most sensitive to a narrow range of frequencies, each fiber of the auditory nerve also only responds to a similar range of frequencies.

It should be borne in mind that the basic description above is highly simplified, ignoring nonlinearities in the cochlea and in the hair-cell response. In addition, the there are actually two different types of hair cells with systematic differences in response. The *inner hair cells*, which constitute about 90% of the total population, transduce and pass on information from the basilar membrane to higher levels of analysis in the auditory system. The remaining *outer hair cells* have a response that is affected in part by efferent feedback from higher centers of neural processing. These cells appear to amplify the incoming signals nonlinearly, with low-level inputs amplified more than more intense signal components, hence achieving a compression in dynamic range. We describe some of the simple attributes of the auditory-nerve response to simple signals in the sections below, focussing on those attributes that are most commonly incorporated into feature extraction for automatic speech recognition.

Transient response of auditory-nerve fibers. The spikes that are generated by fibers of the auditory nerve occur at random times, and hence the response of the nerve fibers must be characterized statistically. Figure 8.2 is a 'post-stimulus-time' (PST) histogram of the rate of firing in response to tone bursts as a function of the time after the initiation of the burst, averaged over many presentations of the tone burst. It can be seen there is a low level of spontaneous activity before the tone burst is gated on. When the tone is suddenly turned on, there is an 'overshoot' in the amount of activity, which eventually settles down to about 50 spikes per second. Similarly, when the tone is gated off, the response drops below the spontaneous rate before rising to it. These results illustrate the property that the auditory system tends to emphasize transients in the response to the signals, with less response from the steady-state portions being more suppressed.

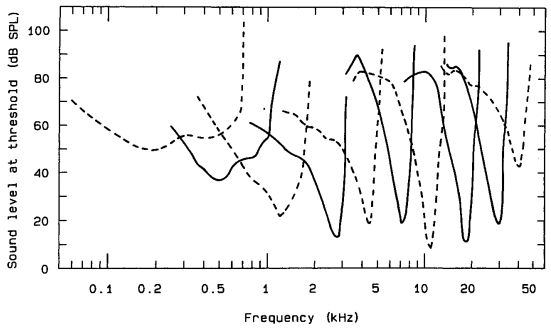

Figure 8.3: Tuning curves indicating the response threshold for pure tones for an assortment of auditory-nerve fibers with different CFs. From [87] as redrawn by [77].

Frequency resolution of the auditory system. As was noted above, different frequency components of an incoming signal elicit maximal vibrations of the basilar membrane at different locations along the membrane. Because each of the roughly 30,000 fibers of the auditory nerve is connected to a particular location along the membrane, the response of each of these fibers is frequency specific as well, as illustrated by the curves of Fig. 8.3. Each of the curves in this figure represents the intensity at a given frequency that is needed to cause the mean rate of firing from a particular auditory-nerve fiber in response to a sine tone to increase a pre-determined percentage above the spontaneous average firing rate for that fiber. Each curve in the figure represents the response of a different fiber. It can be seen that for each of the fibers responds only over a relatively narrow range of frequency and there is a specific frequency of the incoming signal at which the fiber is the most sensitive, called 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 we note that 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. This means that these physiological filters could be considered to be 'constant-Q' in that the nominal bandwidth is roughly proportional to center frequency. The bandwidth of the filters is roughly constant at lower frequencies, although this is less obvious from the curves in Fig. 8.3. This frequency-specific or 'tonotopic' organization of individual parallel channels is generally maintained as we move up from the auditory nerve to higher centers of processing in the brainstem and the auditory cortex.

Rate-level responses. We have previously stated that many aspects of auditory processing are nonlinear in nature. This is illustrated rather directly in Fig. 8.4, which shows the manner in which the rate of response increases as a function of signal intensity. (The spontaneous rate of firing for each fiber has been subtracted from the curves.) As can be seen, the rate-intensity function is roughly S-shaped, with a relatively flat portion corresponding to intensities below the threshold intensity for the fiber, 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 in which the response is again essentially independent of the incoming signal intensity. (There are some exceptions to this, such as the fiber with CF 1.6 kHz in the figure.) The fact

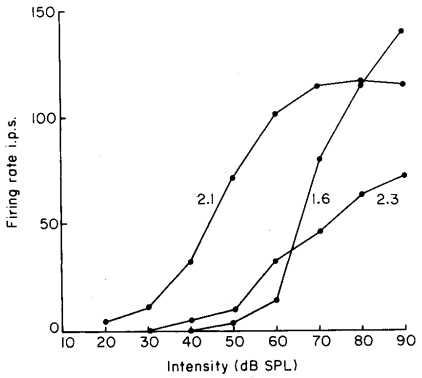

Figure 8.4: Rate of spike discharges as a function of intensity for three auditory-nerve fibers with different CFs (as indicated), after subtracting the spontaneous rate of firing. From [90].

that each individual fiber is limited to approximately 20 dB of active response implies that psychophysical phenomena such as loudness perception must be mediated by the combined response of a number of fibers over a range of frequencies.

Synchronized response to low-frequency tones. When the excitation for a particular auditory-nerve fiber is below the threshold intensity level for that fiber, spikes will be generated randomly, following a Poisson interval distribution with a refractory interval of no response for about 4 ms after each spike. Figure 8.5 is a 'post zero-crossing histogram' (PZC histogram) which describes the firing rate that is observed as a function of the phase of the input signal. We note that the response roughly follows the shape of the input signal at least when the signal amplitude is positive (which actually corresponds to times at which the instantaneous pressure is lower than the baseline level). This 'phase-locking' behavior enables the auditory system to compare arrival times of signals to the two ears at low frequencies, which is the basis for the spatial localization of a sound source at these frequencies. While the auditory system loses the ability to respond in synchrony to the fine structure of higher-frequency components of the input signal, its response is synchronized to the *envelopes* of these signal components (*e.g.* [21]). The frequency at which the auditory system loses its ability to track the fine structure of the incoming signal in this fashion is approximately the frequency at which such timing information becomes useless because that information becomes ambiguous for localization purposes, which strongly suggests that the primary biological purpose for low-frequency synchrony is indeed sound localization.

While temporal coding is clearly important for binaural sound localization, it may also play a role in the robust interpretation of the signals from each individual ear as well. For example, the upper panel of Fig. 8.6 depicts the mean rate of response to a synthetic vowel sound by an ensemble of auditory-nerve fibers as a function of the CF of the fibers, with the intensity of the vowel sound varied over a wide range of input intensities, as described by Sachs and Young [97]. The lower panel of that figure depicts the derived *averaged localized synchronized rate* (or *ALSR*) to the same signals [120], which describes the extent to which the neural response at a given CF is synchronized to the nearest harmonic of the fundamental frequency of the

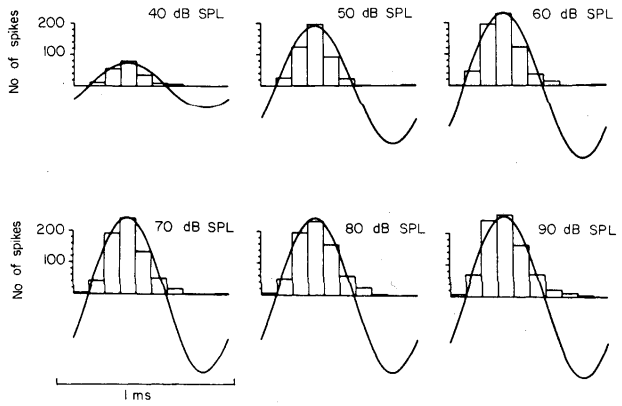

Figure 8.5: Period histograms in response to a 1100-Hz pure tone at various signal intensities. From [95].

vowel. It can be easily seen that the mean rate of response varies dramatically as the input intensity changes, while the ALSR remains substantially unchanged. These results suggest that the timing information associated with the response to low-frequency components of a signal can be substantially more robust to variations in intensity (and potentially various other types of signal variability and/or degradation) than the mean rate of the neural response. Most conventional feature extraction schemes (such as MFCC and PLP coefficients) are based on short-time energy in each frequency band, which is more directly related to mean rate than temporal synchrony in the physiological responses.

Lateral suppression. The response of auditory-nerve fibers to more complex signals also depends on the nature of the spectral content of the signals, as the response to signals at a given frequency may be suppressed or inhibited by energy at adjacent frequencies (*e.g.* [96, 5]). For example, Fig. 8.7 summarizes some aspects of the response to a pairs of tones. The signal in this case is a pair of tones, a 'probe tone' that is 10 dB above threshold at the CF (indicated by the open triangle in the figure) plus a second tone presented at various different frequencies and intensities. The cross-hatched regions indicate the frequencies and intensities for which the response to the two tones combined is less than the response to the probe tone at the CF alone. The open circles outline the tuning curve for the fiber that describes the threshold intensity for the probe tone alone as a function of frequency. It can be seen that the presence of the second tone over a range of frequencies surrounding the CF inhibits the response to the probe tone at CF, even when the second tone is presented at intensities that would be below threshold if it had been presented in isolation. This form of 'lateral suppression' has the effect of enhancing the response to changes in the signal content with respect to frequency, just as the overshoots and undershoots in the transient response have the effect of enhancing the response to changes in signal level over time.

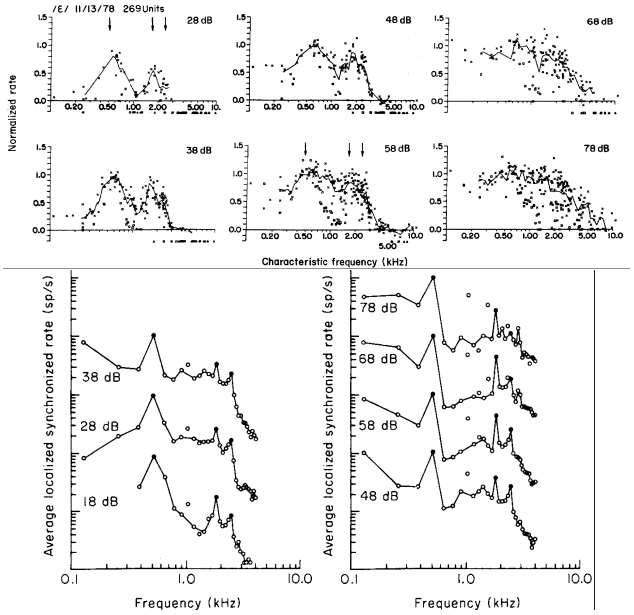

Figure 8.6: Comparison of auditory-nerve responses to a computer-simulated vowel sound at various intensities based on mean rate (upper panel) and synchrony to the signal (see text). Redrawn from [97] and [120].

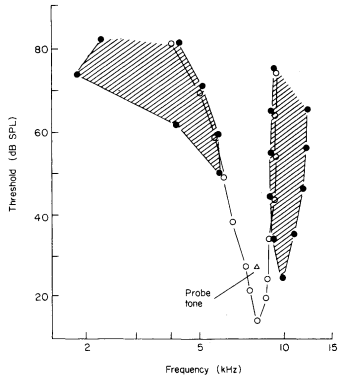

Figure 8.7: The shaded portions of the figure indicate combinations of intensities and frequencies at which the presence of a second tone suppresses the auditory-nerve response to a tone at a fiber's CF presented 10 dB above threshold. From [5].

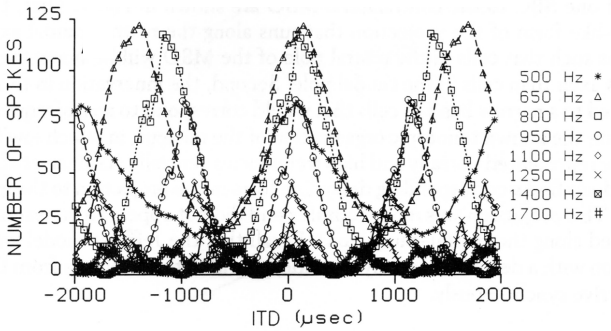

Figure 8.8: Response of a unit in the MSO to pure tones at various frequencies plotted as a function of the stimulus ITD. From [117].

8.2.2 Processing at more central levels

While the phenomena described above all observed at the level of the cochlea or the auditory nerve, substantial processing takes place at the level of the pre-cortical centers of the brainstem as well as in the auditory cortex itself. We note here three sets of more central phenomena that also play significant roles in auditory modeling.

Sensitivity to interaural time delay and other binaural phenomena. It is well known that 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) (*e.g.* [24]). As first noted by Rayleigh [109], ITDs are most useful at low frequencies and IIDs are only present at higher frequencies for reasons related to spatial aliasing and physical diffraction, respectively. Physiologists have observed units in the superior olivary complex and the inferior colliculus that appear to respond maximally to a single 'characteristic' ITD (*e.g.* [94, 117]). As an example, Fig. 8.8 depicts the response of a unit in the superior olivary complex in the brainstem of a cat which responds in excitatory fashion when signals are presented binaurally. The figure plots the relative number of spikes in response to tones at various frequencies as a function of the ITD with which the signals are presented to the two ears. We note that this unit exhibits a maximum in response when the tones are presented with an ITD of approximately 33 μs for frequencies ranging from 500 to 1700 Hz. In other words, the function of this unit appears to be the detection of a specific ITD, and that ITD of best response is 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. Over the years many theories have been developed that describe how a display of this sort can be used to describe and predict a wide range of binaural phenomena as reviewed by Stern and Trahiotis [104] and Stern *et al.* [106]). Units have also been described that appear to record the IIDs of a stimulus (*e.g.* [94]).

Another important attribute of human binaural hearing is that localization is dominated by the first-arriving components of a complex sound [114]. This phenomenon, which is referred to as the *precedence effect*, is clearly helpful in causing the perceived location of a source in a reverberant environment to remain constant, as it is dominated by the characteristics of the direct field (which arrives straight 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.

Sensitivity to amplitude and frequency modulation. Physiological recordings in the cochlear nucleus, 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.* [47]). 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 [58]. Similar responses have been observed in the cochlear nucleus to sinusoidal frequency modulations [76], with modulation frequencies of 50 to 300 Hz providing maximal response. These results have lead to speculation that the so-called *modulation spectrum* may be a useful and consistent way to describe the dynamic temporal characteristics of complex signals like speech after the peripheral frequency analysis. Nevertheless, the extent to which the physiological representation of amplitude modulation is preserved and remains invariant at higher levels remains an open issue at present. For example, Drullman *et al.* [22, 23] conducted a series of experiments in which they measured the perception of speech that had been analyzed and resynthesized with modified temporal envelopes in each frequency band, concluding that modulation spectrum components between 4 Hz and 16 Hz are critical for speech intelligibility.

Feature detection at higher levels: spectro-temporal response fields. There is a rich variety of feature-detection mechanisms that have been observed at the higher levels of the brainstem and in the auditory cortex as reviewed by Palmer and Shamma [86]. A characterization that has proved useful is that if the *spectro-temporal response field* or *STRF*, which can in principle be used to describe sensitivity to amplitude modulation, frequency modulation, as well as a more general sensitivity to sweeps in frequency over time, as might be useful in detecting formant transitions in speech. As an example, researchers at the University of Maryland and elsewhere have used used dynamic 'ripple' stimuli, with drifting sinusoidal spectral envelopes, to develop the STRF patterns in the responses of units of the primary auditory cortex (A1) in ferrets [20, 56]. They reported units with a variety of types of response patterns, including sensitivity to upward and downward ripples, as well as a range of best frequencies, bandwidths, asymmetries in response with respect to change in frequency, temporal dynamics, and related characteristics. It is frequently convenient to illustrate the observed STRS as color temperature patterns in the time-frequency plane.

8.2.3 Psychoacoustical correlates of physiological observations

All of the phenomena cited above have perceptual counterparts, which are observed by carefully-designed psychoacoustical experiments. The results of these experiments give us direct insight into the characteristics and limitations of auditory perception, although we must infer the mechanism underlying the experimental results. (In contrast, physiological experiments provide direct measurements of the internal response to sound, but the perceptual significance of a given observation must be inferred.) 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 two sets of basic psychoacoustic observations that have played a major role in

auditory modeling.

The psychoacoustical transfer function. The original psychoacousticians were physicists and philosophers of the nineteenth century who had the goal of developing mathematical functions that related sensation and perception, such as the dependence of the subjective loudness of a sound on its physical intensity. As can be expected, the nature of the relationships will depend on the temporal and spectral properties of the signals, as well as how the scales are constructed. The original psychophysical scales for intensities were based on the empirical observations of Weber [116], who observed that the increment in intensity needed to just barely perceive that a simple sound (such as a tone) was louder than another was a constant fraction of the reference intensity level. This type of dependence of the *just-noticeable difference* or *JND* of intensity on reference intensity is observed in other sensory modalities as well, such as the perception of the brightness of light or the weight of a mass. Fechner [26] proposed that a psychophysical scale that describes perceived intensity as a function of the intensity of the physical stimulus could be constructed by combining Weber's empirical observation with the assumption that JNDs in intensity should represent equal *intervals* on the perceptual scale. It is easy to show that this assumption implies a logarithmic perceptual scale

$$\Psi = C \log(\Phi) \tag{8.1}$$

where Φ in the above equation represents physical intensity and Ψ represents its perceptual correlate (presumably loudness in hearing). The logarithmic scale for intensity perception, of course, motivated the decibel scale, and it is partially supported by the fact that there is typically a linear relation between the neural rate of response and intensity in dB for intermediate intensity levels, as in intermediate range of the curves in Fig. 8.4. Many years later Stevens proposed an alternate loudness scale, which implicitly assumes that JNDs in intensity should represent equal *ratios* on the perceptual scale. This gives rise to the power law relationship

$$\Psi = K_1 \Phi^{K_2} \tag{8.2}$$

where Φ and Ψ are as in Eq. (8.1). The Stevens power law is supported by the results of many *magnitude estimation* experiments in which subjects are asked to apply a subjective numerical label to the perceived intensity of a signal. While the value of the exponent K_2 depends to some extent on the nature of the signal and how the experiment is conducted, it is typically on the order of 0.33 when physical intensity is expressed in terms of stimulus amplitude [107]. More extensive discussion of these theories and their derivation are available in texts by Gescheider [31] and Baird and Noma [10], among many other sources.

Auditory frequency resolution. As noted above, the individual parallel channels of the auditory system are frequency selective in their response to sound. It is generally assumed that the first stage of auditory processing may be modeled as a bank of bandpass filters, and all modern theories of auditory perception are attentive to the impact of processing by the peripheral auditory system has on the representation of sound. For example, the detection of a tonal signal in a broadband masker is commonly assumed to be mediated by the signal-to-noise ratio at the output of the auditory filter that contains the target tone. Auditory frequency resolution was first studied psychophysically in the 1930s by Fletcher and colleagues at Bell Laboratories [28], which preceded Békésy's physiological measurements of cochlear mechanics in the 1940s [113] as well as subsequent descriptions of the frequency-specific physiological response to sound at the level of the fibers of the auditory nerve and at higher

centers (*e.g.* [48]). There are a number of ways of measuring auditory frequency selectivity (*cf.* [78] and Chapter 3 of [77]), and to some extent the estimated bandwidth of the auditory channels (commonly referred to as the 'critical band' associated with each channel) depends on the assumed filter shape and the way in which bandwidth is measured. In general the estimated channel bandwidth increases with increasing center frequency of the channel, and at higher frequencies the filter bandwidth tends to be roughly proportional to center frequency, as was observed in Fig. 8.3.

From these experiments, three distinct frequency scales have emerged that describe the bandwidths of the auditory filters and, correspondingly, the center frequencies of the filters that are needed to ensure that the filters are separated by a constant number of critical bands at all frequencies. The *Bark scale* (named after Heinrich Barkausen), based on estimates of the critical band from traditional masking experiments, was first proposed by Zwicker [124], quantified by Zwicker and Terhardt [125], and subsequently represented in simplified form by Traunmüller [111] as

$$Bark(f) = [26.8/(1 + (1960/f))] - 0.53 \tag{8.3}$$

where the frequency f is in Hz.

The *mel scale* (which refers to the word 'melody') was proposed by Stevens *et al.* [108] and is based on pitch comparisons; it is approximated by the formula [85]

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{8.4}$$

The original critical band estimates of Fletcher were based on the simplifying assumption that the auditory filters were rectangular in shape. The shape of the auditory filters has been estimated in several ways, frequently making use of notch-shaped maskers (*e.g.* [88]). A popular scale proposed by Moore and Glasberg [79] called the *ERB* scale describes the *equivalent rectangular bandwidth* of these filters. The number of ERBs as a function of frequency is approximated by the formula

$$ERB_N(f) = 21.4 \log_{10}(1 + 4.37f/1000) \tag{8.5}$$

where again f is in Hz. For example, at 1 kHz this function is equal to about 130 Hz, which means that an increase of frequency of 130 Hz centered about 1 kHz would constitute one ERB.

Figure 8.9 compares the Bark, Mel, and ERB scales from the equations above after multiplying each curve by a constant that was chosen to minimize the squared difference between the curves. It can be seen that despite the differences in how the frequency scales were formulated, they all look similar, reflecting the fact that the perceptual scale is expanded with respect to frequency at low frequencies and compressed at higher frequencies. All common models of auditory processing begin with a bank of filters whose center frequencies and bandwidths are based on one of the three frequency scales depicted in this figure.

Loudness matching and auditory thresholds. A final set of results that have had an impact on feature extraction and auditory models is the set of *equal loudness contours* depicted in Fig. 8.10 after measurements by Fletcher and Munson [29]. Each curve depicts the intensity of a tone at an arbitrary frequency that matches the loudness of a tone of a

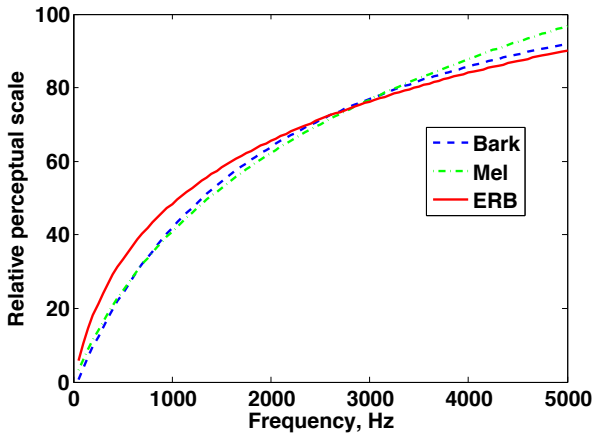

Figure 8.9: Comparison of frequency scales derived from the Bark, mel, and ERB scales.

specified intensity at 1 kHz, which is defined to be the loudness of that tone in *phons*. These curves indicate that threshold intensities (the lowest curve) vary with frequency, with the ear being the most sensitive between frequencies of about 1000 and 4000 Hz. The upper limit of hearing is much less sensitive to 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 [77].

8.2.4 The impact of auditory processing on conventional feature extraction

The overwhelming majority of speech recognition systems today make use of features that are based on either *Mel-Frequency Cepstral Coefficients* (*MFCCs*) first proposed by Davis and Mermelstein in 1980 [19] or features based on *perceptual linear predictive (PLP)* analysis of speech [36], proposed by Hermansky in 1990. We briefly discuss MFCC and PLP processing in this section. The major functional blocks used in these procedures are summarized in Fig. 8.11.

As is well known, MFCC analysis consists of (1) short-time Fourier analysis using Hamming windows, (2) weighting of the short-time magnitude spectrum by a series of triangularly-shaped functions with peaks that are equally spaced in frequency according to the Mel scale, (3) computation of the log of the total energy in the weighted spectrum, and (4) computation of a relatively small number of coefficients of the inverse discrete-cosine transform (DCT) of the log power coefficients from each channel. These steps are summarized in the left column of Fig. 8.11. Expressed in terms of the principles of auditory processing, the triangular weighting functions serve as a crude form of auditory filtering, the

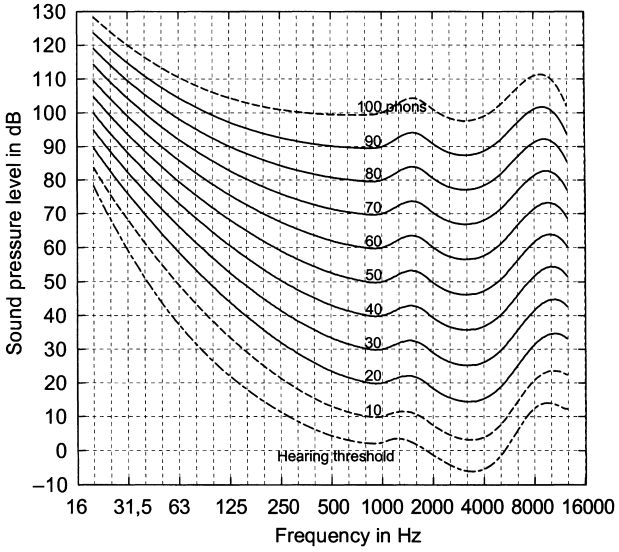

Figure 8.10: Equal-loudness matching curves (after Fletcher and Munson, [29]).

log transformation mimics Fechner's psychophysical transfer function for intensity, and the inverse DCT can be thought of as providing a lowpass Fourier series representation of the frequency-warped log spectrum. The cepstral computation can also be thought of as a means to separate the effects of the excitation and frequency-shaping components of the familiar source-filter model of speech production (*e.g.* [91]).

The computation of the PLP coefficients is based on somewhat different implementation of similar principles. PLP processing consists of (1) short-time Fourier analysis using Hamming windows (as in MFCC processing), (2) weighting of the power spectrum by a set of asymmetrical functions that are spaced according to the Bark scale, and that are based on the auditory masking curves of [98], (3) pre-emphasis to simulate the equal-loudness curve suggested by Makhoul and Cosell [66] to model the loudness contours of Fletcher and Munson (as in Fig. 8.10), (4) a power-law nonlinearity with exponent 0.33 as suggested by Stevens *et al.* [108] to describe the intensity transfer function, (5) a smoothed approximation to the frequency response obtained by all-pole modeling, and (6) application of a linear recursion that converts the coefficients of the all-pole model to cepstral coefficients.

indexRelative spectral analysis (RASTA) PLP processing is also frequently used in conjunction with Hermansky and Morgan's RASTA algorithm [37], a contraction of *relative spectral analysis*. RASTA processing in effect applies a bandpass filter to the compressed spectral amplitudes that emerge between Steps (3) and (4) of the PLP processing above. RASTA processing also models the tendency of the auditory periphery to emphasize the transient portions of incoming signals, as noted in Sec. 8.2.1 above. In practice, the bandpass nature of the filter causes the mean values of the spectral coefficients to equal zero, which effects a normalization that is similar to the normalization provided by cepstral mean

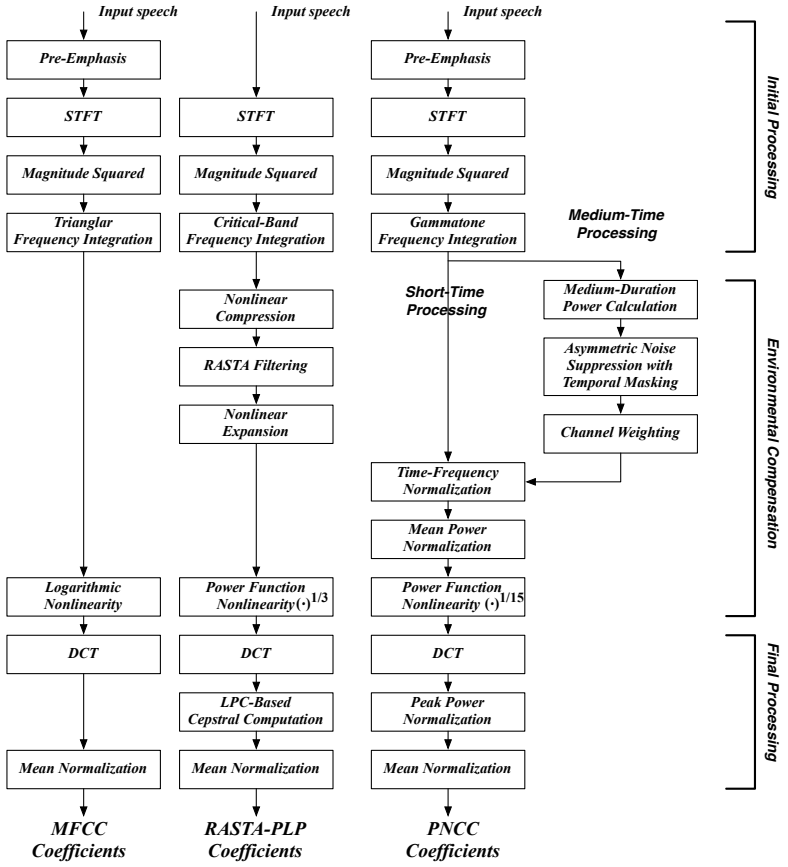

Figure 8.11: Comparison of major functional blocks of the MFCC, PLP-RASTA, and PNCC processing methods. (PNCC processing is discussed in Sec. 8.4 below.)

normalization (CMN) that is commonly used in conjunction with MFCC processing. Both the RASTA filter and CMN are effective in compensating for the effects of unknown linear filtering in cases for which the impulse response of the filter is shorter than the duration of the analysis window used for processing. Hermansky and Morgan [37] also propose an extension to RASTA processing, called *J-RASTA* processing, which provides similar compensation for additive noise at low signal levels.

In summary, PLP feature extraction is an attempt to model several perceptual attributes of the auditory system more exactly than MFCC processing, including the use of the Zwicker filters to represent peripheral frequency selectivity and the pre-emphasis to characterize the dependence of loudness on frequency. In addition, it replaces the mel scale by the Bark scale, the log relation for intensity by a power function, and it uses auto-regressive modeling

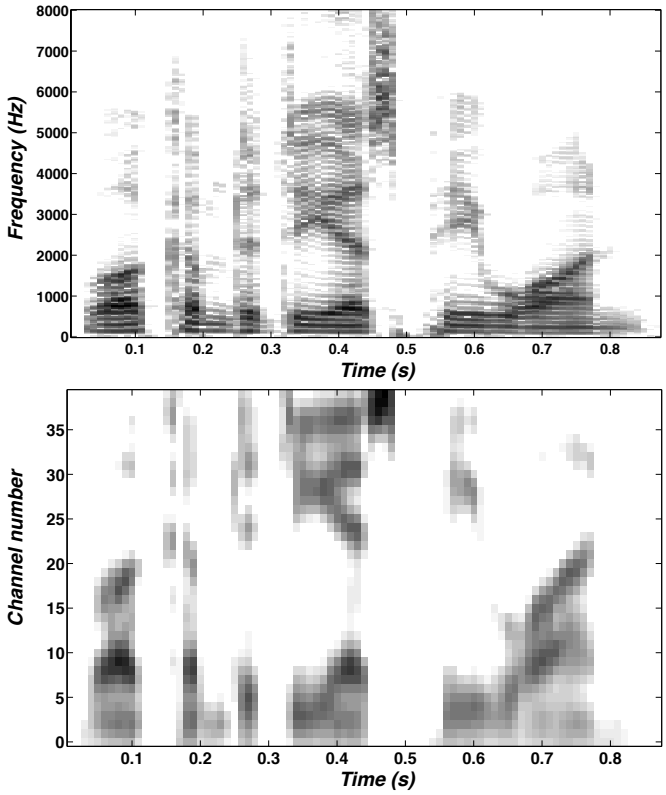

Figure 8.12: Upper panel: Wideband spectrogram of a sample utterance. Lower panel: Reconstruction of the spectrogram after MFCC analysis.

of a low order (rather than truncation of a Fourier-based expansion) to obtain a smoothed representation of the spectrum.

8.2.5 Summary

We have described a number of physiological phenomena that have motivated the development of auditory modeling for automatic speech recognition. These phenomena include frequency analysis in parallel channels, a limited dynamic range of response within each channel, preservation of temporal fine structure, enhancement of temporal contrast at signal onsets and offsets, enhancement of spectral contrast (at adjacent frequencies), and preservation of temporal fine structure (at least at low frequencies). Most of these phenomena also have psychoacoustical correlates. Conventional feature extraction procedures (such as MFCC and PLP coefficients) preserve some of these attributes (such as basic frequency

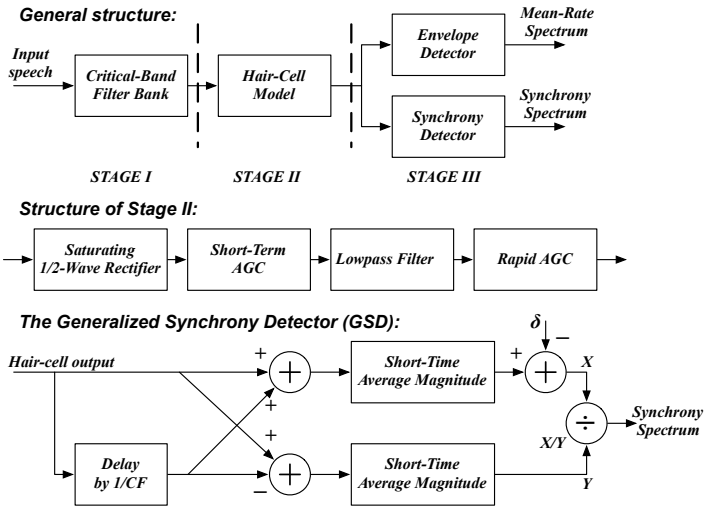

Figure 8.13: Upper panel: General structure of the Seneff model. Central panel: Block diagram of the Seneff hair cell model. Lower panel: Block diagram of of Seneff's generalized synchrony detector. After [100].

selectivity and spectral bandwidth) but omit others (such as temporal and spectral enhancement and detailed timing structure). As an example, Fig. 8.12 compares a high-resolution spectrogram in response to a short utterance to a spectrogram reconstructed from MFCC coefficients computed for the same utterance. In addition to the frequency warping that is intrinsic to MFCC (and PLP) processing, it is clear that substantial detail is lost in the MFCC representation, some of which is sacrificed deliberately to remove pitch information. One of the goals of the auditory representations is to restore some of this lost information about the signal in a useful and efficient fashion.

8.3 'Classic' Auditory Representations

The first significant attempts to develop models of the peripheral auditory system for use as front ends to ASR systems occurred in the 1980s with the models of Seneff [100], Ghitza [32], and Lyon [61, 62], which we summarize in this section. The Seneff model, in particular, has been the basis for many subsequent studies, in part because of it was described in great detail and it is easily available in MATLAB form as part of the Auditory Toolbox developed and distributed by [103]. This very useful resource, which also includes the Lyon model, is based on earlier work by Lyon and Slaney using Mathematica.

Seneff's auditory model. Seneff's auditory model [100] is summarized in block diagram form in Fig. 8.13. The first of three stages of the model consisted of 40 recursive linear filters implemented in cascade form to mimic the nominal auditory-nerve frequency responses as described by Kiang *et al.* [48] and other contemporary physiologists.

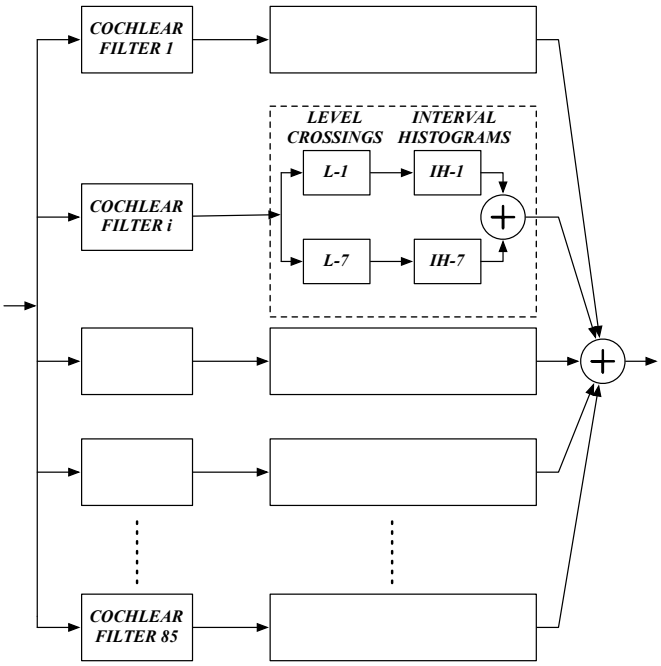

Figure 8.14: General structure of the Ghitza model. After [32].

Substantial effort was devoted in Stage II to describing the nonlinear transduction from the motion of the basilar membrane to the mean rate of of auditory-nerve spike discharges. As indicated in the central panel of Fig. 8.13, this 'inner-hair-cell model' included four stages: (1) nonlinear half-wave rectification using an inverse tangent function for positive inputs and an exponential function for negative inputs, (2) short-term adaptation that modeled the release of transmitter in the synapse, (3) a lowpass filter with cutoff frequency of approximately 1 kHz to suppress synchronous response at higher input frequencies, and (4) a rapid AGC stage to maintain an approximately-constant response rate at higher input intensities when an auditory-nerve fiber is nominally in saturation.

Stage III consisted of two parallel operations on the hair-cell model outputs. The first of these was an envelope detector, which produced a statistic intended to model the instantaneous mean rate of response of a given fiber. The second operation was called a *generalized synchrony detector (GSD)*, and was motivated by the ALSR measure of [120]. The GSD is summarized in the lower panel of Fig. 8.13. The hair-cell output is compared to itself delayed by the reciprocal of the center frequency of the filter in each channel, and the short-time averages of the sums and differences of these two functions are divided by one another. The threshold δ is introduced to suppress response to low-intensity signals and the resulting quotient is passed through a saturating half-wave rectifier to limit the magnitude of the predicted synchrony.

Ghitza's EIH model. A second classic auditory model developed by Ghitza [32] is called the *Ensemble Interval Histogram* (EIH) model and is summarized in Fig. 8.14. Ghitza makes use of the peripheral auditory model proposed by Allen [3] to describe the transformation of sound pressure into the neural rate of firing and focussed on the mechanism used to interpret the neural firing rates. The most interesting aspect of the EIH model is its use of timing information to develop a spectral representation of the incoming sound. Specifically, the EIH model records in each frequency channel the times at which the outputs of the auditory model crosses a set of seven thresholds that are logarithmically spaced over the dynamic range of each channel. Histograms are compiled of the reciprocals of the times between the threshold crossings of each threshold in each channel, and these histograms are summed over all thresholds and channels, producing an estimate of the internal spectral response to the incoming sound.

The EIH model was the only one of the three original auditory models for which the developer included speech recognition evaluations with the original description of the model. Ghitza obtained these results using a contemporary DTW recognizer [32]. He observed that while the use of the auditory model provided no advantage in clean speech (and in some cases degraded performance compared to baseline MFCC processing), improvements were noted in noise and reverberation.

Lyon's model. The third major model of the 1980s was described initially by Lyon [61, 62]. Lyon's model for auditory-nerve activity [61] included many of the same elements as the models of Seneff and Ghitza (such as bandpass filtering, nonlinear rectification and compression, along with several types of short-time temporal adaptation), as well as a mechanism for lateral suppression, which was unique among the classical models. Lyon was particularly concerned with the nature and shape of the filters used to model peripheral analysis and a longitudinal of his perspective on this subject may be found in [64]. In an extension of this work Lyon proposed a 'correlogram' display [63] that is derived from the short-time autocorrelation of the outputs of each channel that was believed to be useful for pitch detection and timbre identification. In 1983 Lyon described a computational binaural model based on cross-correlation of the corresponding outputs from the monaural processors. This model has the ability to separate signals based on differences in time of arrival of the signals to the two ears, and is similar in concept to the classic mechanism for extracting interaural time delays (ITDs) first suggested by Jeffress [46].

Performance of early auditory models. The classic models included a number of attributes of auditory processing beyond MFCC/PLP feature extraction: more realistic auditory filtering, more realistic auditory nonlinearity, and in some cases synchrony extraction, lateral suppression, and interaural correlation. Unsurprisingly, each system developer had his or her own idea about which attribute of auditory processing was the most important for robust speech recognition.

While the EIH model was the only one of the original three to be evaluated quantitatively for speech recognition accuracy at the time of its introduction, a number of early studies compared the recognition accuracy of auditory-based front ends with conventional representations (*e.g.* [32, 43, 70, 45]). It was generally observed that while conventional feature extraction in some cases provided best accuracy when recognizing clean speech, auditory-based processing would provide superior results when speech was degraded by added noise. Early work in the CMU Robust Speech Group (*e.g.* [84], [105]) confirmed these trends for reverberation as well as for additive noise in an analysis of the performance

of the Seneff model. We also noted, disappointingly, that the application of conventional engineering-approaches such as *codeword-dependent cepstral normalization* (CDCN, [1]) provided recognition accuracy that was as good as or better than the accuracy obtained using auditory-based features in degraded acoustical environments. In a more recent analysis of the Seneff model we observed that the saturating half-wave nonlinearity in Stage II of the Seneff model is the functional element that appears to provide the greatest improvement in recognition accuracy compared to baseline MFCC processing [17].

One auditory model of the late 1980s that was successful was developed by Cohen [18], and it exhibited a number of the physiological and psychoacoustical properties of hearing described in Sec. 8.2. Cohen's model included a bank of filters that modeled critical-band filtering, an empirical intensity scaling to describe equal loudness according to the curves of Fletcher and Munson [29], a cube-root power-law compressive nonlinearity to describe loudness scaling after Stevens [107]. The final stage of the model was a simple differential equation that models the time-varying release of neural transmitter based on the model of Schroeder and Hall [99]. This stage provided the type of transient overshoots observed in Fig. 8.2. Feature extraction based on Cohen's auditory model provided consistently better recognition accuracy than features that approximated cepstral coefficients derived from a similar bank of bandpass filters for a variety of speakers and microphones in relatively quiet rooms. On the basis of these results, Cohen's auditory model was adopted as the feature extraction procedure for the IBM Tangora system and was used routinely for about a decade.

Despite the adoption of Cohen's feature extraction in Tangora and interesting demonstrations using the outputs of the models of Seneff, Ghitza, and Lyon, interest in the use of auditory models generally diminished for a period of time around the late 1980s. As noted above, the auditory models generally failed to provide superior performance when processing clean speech, which was the emphasis for much of the research community at this time. In part this may well have been a consequence of the typical assumption in the speech recognition systems of the day that the probability densities of the features were normally distributed. In contrast, the actual outputs of the auditory models were distinctly non-Gaussian in nature. For example, Chigier and Leung [16] noted that the accuracy of speech recognition systems that used features based on the Seneff model was greatly improved when a multilayer perceptron (which learns the shape of the feature distributions without *a priori* assumptions) is used instead of a Bayesian classifier that assumed the use of unimodal Gaussian densities. The classical auditory models fared even worse when computation was taken into account. Ohshima [83], for example, observed that the Seneff model requires about 40 times as many multiplies and 33 times as many additions compared to MFCC or PLP feature extraction. And in all cases, the desire to improve robustness in speech recognition in those years was secondary to the need to resolve far more basic issues in acoustic modeling, large-vocabulary search, etc.

8.4 Current Trends in Auditory Feature Analysis

By the late 1990s physiologically-motivated and perceptually-motivated feature extraction methods began to flourish once again for several reasons. Computational capabilities had advanced over the decade to a significant degree, and front-end signal processing came to consume a relatively small fraction of the computational demands of large-vocabulary speech recognition compared to score evaluation, graph search, etc. The development of

fully-continuous hidden Markov models using Gaussian mixture densities as probabilities for the features, along with the development of efficient techniques to train the parameters of these acoustic models, meant that the non-Gaussian form of the output densities of the auditory models was no longer a factor that limited their performance.

In this section we describe some of these trends in auditory processing that have become important for feature extraction beginning in the 1990s. These trends include closer attention to the details of the physiology, a reconsideration of mechanisms of synchrony extraction, more effective and mature approaches to information fusion, serious attention to the temporal evolution of the outputs of the auditory filters, the development of models based on spectro-temporal response fields, concern for dealing with the effects of room reverberation as well as additive noise, and the use of two or more microphones motivated by binaural processing (which we do not discuss in this chapter).

In the sections below, with some exceptions, we characterize the performance of the systems considered only indirectly. This is because it is almost impossible to meaningfully compare recognition results across different research sites and different experimental paradigms. For example, the baseline level of recognition accuracy will depend on many factors including the types of acoustical models employed and the degree of constraint imposed by language modeling. The type of additive noise used typically affects the degree of improvement to be expected from robust signal processing approaches: for example, it is relatively easy to ameliorate the effects of additive white noise, but effective compensation for the effects of background music is far more difficult to achieve. As the amount of available acoustic training increases, the degree of improvement observed by advanced feature extraction or signal processing techniques diminishes because the initial acoustical models become intrinsically more robust. While most of the results in robust speech recognition that are reported in the literature are based on training on clean speech, the amount of improvement provided by signal processing also diminishes when an ASR system is trained in a variety of acoustical environments ('multi-style training') or when the acoustical conditions of the training and testing data are matched.

We begin with peripheral phenomena and continue with more central phenomena.

Speech recognition based on detailed physiological models. In addition to the 'practical' abstractions proposed by speech researchers including the classical representations discussed in Sec. 8.3, auditory physiologists have also proposed models of their own that describe and predict the functioning of the auditory periphery in detail. For example, the model of Meddis and his colleagues (*e.g.* [68, 69]) is a relatively early formulation that has been quite influential in speech processing. The Meddis model characterizes the rate of spikes in terms of a mechanism based on the dynamics of the flow of neurotransmitter from inner hair cells into the synaptic cleft, followed by its subsequent uptake once again by the hair cell. Its initial formulation, which has been refined over the years, was able to predict a number of the physiological phenomena described in Sec. 8.2.1 including the nonlinear rate-intensity curve, the transient behavior of envelopes of tone bursts, synchronous response to low-frequency inputs, the interspike interval histogram, and other phenomena. Hewitt and Meddis reviewed the physiological mechanisms underlying seven contemporary models of auditory transduction, and compared their ability to describe a range of physiological data, concluding that their own formulation described the largest set of physiological phenomena most accurately [42].

The Carney group (*e.g.* Zhang *et al.* [121]; Heinz *et al.* [35]; Zilarny *et al.* [123]) has also

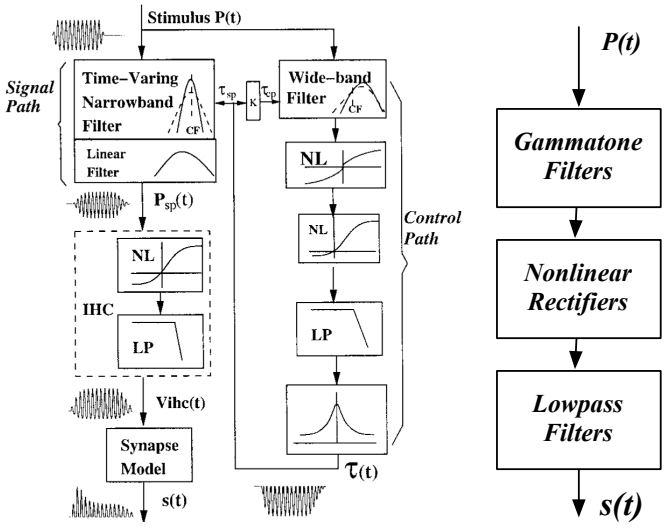

Figure 8.15: Left panel: block diagram of Zhang-Carney model (from Zhang *et al.*, 2001). Right panel: block diagram of a much simpler computational model of auditory processing.

developed a series of physiologically-based models of auditory-nerve activity over the years. The original goal of the work of Carney and her colleagues had been to develop a model that can account describe the response to more complex signals such as noise-masked signals and speech, primarily through the inclusion into the model of the compressive nonlinearity of the cochlear amplifier in the inner ear. A diagram of most of the functional blocks of the model of Zhang *et al.* is depicted in the left panel of Fig. 8.15. As can be seen in the diagram, the model includes a *signal path* that has many of the attributes of the basic phenomenological models introduced in Sec. 8.3, with a time-varying nonlinear narrowband peripheral filter that is followed by a linear filter. Both of these filters are based on gammatone filters. The time constant that determines the gain and bandwidth of the nonlinear filter in the signal path is controlled by the output of the wideband *control path* that is depicted on the right side of the panel. The level-dependent gain and bandwidth of the control path enable the model to describe phenomena such as two-tone suppression within a single auditory-nerve channel, without needing to depend on inhibitory inputs from fiber at adjacent frequencies, as in Lyon's model [61].

A few years ago Kim *et al.* [52] from the CMU group presented some initial speech recognition results that developed simple measures of mean rate and synchrony from the outputs of the model of Zhang *et al.* Fig. 8.16 compares the recognition accuracy for speech in white noise using feature extraction procedures that were based on the putative mean rate of auditory-nerve response [52]. The CMU Sphinx-3 ASR system was trained using clean speech for these experiments. The curves in Fig. 8.16 describe the recognition accuracy obtained using three types of feature extraction: (1) features derived from the mean rate

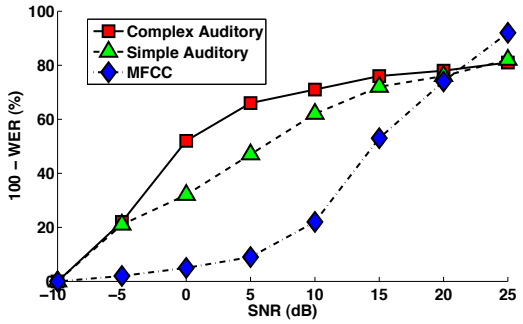

Figure 8.16: Comparison of speech recognition accuracy obtained using features derived from the Zhang-Carney model (squares), features obtained from the much simpler model in the right panel of Fig. 8.15 (triangles), and conventional MFCC coefficients (diamonds). Data were obtained using sentences from the DARPA Resource Management corpus corrupted by additive white noise. The language model is detuned, which increases the absolute word error rate from the best possible value. Replotted from [52].

response based on the complete model of Zhang *et al.* [121]; (2) features derived from the extremely simplified model in the right panel of Fig. 8.15 (triangles) which contains only a bandpass filter, a nonlinear rectifier, and a lowpass filter in each channel; and (3) baseline MFCC processing as described in [19] (diamonds). As can be seen, for this set of conditions the full auditory model provides about 15 dB of effective improvement in SNR compared to the baseline MFCC processing, while the highly simplified model provides about a 10-dB improvement. Unfortunately, the computational cost of features based on the complete model of Zhang *et al.* is on the order of 250 times the computational cost incurred by the baseline MFCC processing. In contrast, the simplified auditory processing consumes only about twice the computation of the baseline MFCC processing. We note that ASR performance in small tasks including the DARPA Resource Management task used for these comparisons can easily become dominated by the impact of a strong language model. In obtaining the results for this figure, as well as for Figs. 8.17 and 8.18, we deliberately manipulated the language weight parameter to reduce the impact of the language model in order to emphasize differences in recognition accuracy that were due to changes in feature extraction procedures. As a consequence, the absolute recognition accuracy is not as good as it would have been had we optimized all system parameters.

Power-normalized cepstral coefficients (PNCC processing). The extreme computational costs associated with the implementation of a complete physiological model such as that of Zhang *et al.* [121] have motivated many researchers to develop simplified models that capture the essentials of auditory processing that are believed to be most relevant for speech perception. The development of *power-normalized cepstral coefficients* (PNCC, [50, 49, 51]) is a convenient example of computationally-efficient 'pragmatic' physiologically-motivated feature extraction. PNCC processing was developed with the goal of obtaining features that incorporate some of the relevant physiological phenomena in a computationally efficient

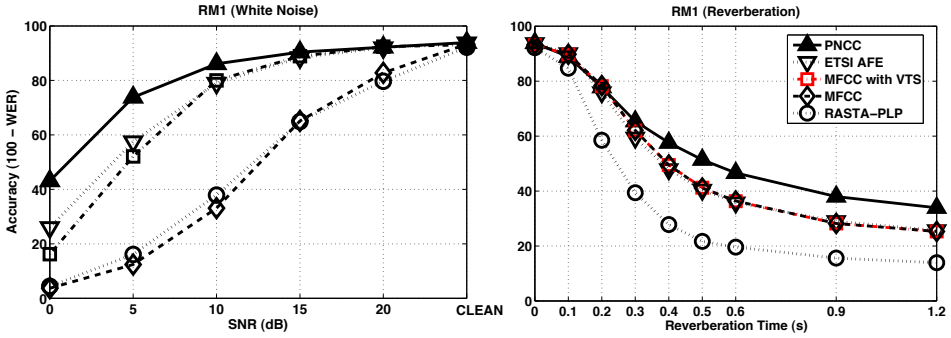

Figure 8.17: Comparison of recognition accuracy on the DARPA Resource Management RM1 database, obtained using PNCC processing with processing using MFCC features, RASTA-PLP features, the ETSI AFE, and MFCC features augmented by VTS processing. From [51].

fashion. A summary of the major functional blocks of PNCC processing is provided in the right column of Fig. 8.11. 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. The power law, rather than the more common logarithmic transformation, was adopted because it provides reduced variability at very low signal intensities, and the exponent of $1/15$ was selected because it provides a best fit to the onset portion of the rate-intensity curve developed by the model of Heinz *et al.* [35]. The power-function nonlinearity has the additional advantage of preserving ratios of responses that are independent of input amplitude.

For the most part, noise and reverberation suppression is introduced to PNCC processing through the system blocks labeled 'medium-time processing' in the far right column of Fig. 8.11 [51]. Medium-time processing operates on segments of the waveform on the order of 50-150 ms duration (as do other waveform-based compensation approaches) in contrast to compensation algorithms such as Vector Taylor Series (VTS, [80]) that manipulate cepstral coefficients derived from from analysis windows on the order of 20-30 ms.

Figure 8.17 compares the recognition accuracy obtained using PNCC processing with the accuracy obtained using baseline MFCC processing (Davis and Mermelstein, 1980), PLP-RASTA processing (Hermansky and Morgan, 1994), MFCC with VTS [80], and the 'Advanced Front End' (AFE), a newer feature extraction scheme developed as a standard for the European Telecommunications Standards Institute (ETSI), which also has noise-robustness capabilities (ETSI, 2007). It can be seen from the panels of Fig. 4 that the

recognition accuracy obtained using features derived with PNCC processing is substantially better than baseline processing using either MFCC or RASTA-PLP features, MFCC features augmented by the VTS noise-reduction algorithms, or the ETSI Advanced Front End for speech that had been degraded by additive white noise and simulated reverberation. In considering these comparisons, it must be borne in mind that neither MFCC nor RASTA-PLP coefficients were developed with the goal of robustness in degraded acoustic environments. A version of RASTA-PLP known as J-RASTA [37] is far more effective in the presence of additive noise. A much more thorough discussion of PNCC processing, including recognition results in the presence of a number of other types of degradations, may be found in [51]. PNCC processing is only about 30% more computationally costly than MFCC processing, and comparable in computational cost to RASTA-PLP. All of these methods require substantially less computation than either the ETSI Advanced Front End or the VTS approach to noise robustness.

Spectral profiles based on synchrony information. Since the 1980s, the approaches of Seneff and Ghitza for developing a spectral representation from the temporal patterns of auditory-nerve firings (rather than simply their mean rate) have been elaborated upon, and other techniques have been introduced as well. We summarize some of these approaches in this section.

Ali *et al.* [2] proposed a simple but useful extension of the Seneff GSD model that develops a synchrony spectrum by simply averaging the responses of several GSDs tuned to the same frequency using inputs from bandpass filters with CFs in a small neighborhood about a central frequency. As described by Ali *et al.*, this approach, referred to as *average localized synchrony detection* (ALSD), produces a synchrony spectrum with smaller spurious peaks than are obtained using either Seneff's original GSD detector, mean-rate-based spectral estimates, or the synchrony spectrum produced by the lateral inhibitory network (LIN) of Shamma [101], and it provides the best recognition results of the methods considered for a small vowel-classification task in white noise.

D. Kim *et al.* [53] proposed a type of processing called *zero-crossing peak analysis* (ZCPA) that could be considered to be an elaboration of Ghitza's EIH processing, but without the complication of the multiple thresholds that are part of the EIH model. ZCPA is also different from other approaches in that there is no nonlinear rectification that is associated with most auditory models, including the EIH model. Instead, positive-going zero crossings are recorded directly from the outputs of each of the auditory filters, and the times of these zero crossings are recorded on a channel by channel basis. A histogram is generated of the reciprocal of the intervals between the zero crossings (a measure of instantaneous frequency), weighted by the amplitude of the peak between the zero crossings. While quantitative analysis of zero crossings of a random process is always difficult, the authors argue that setting the threshold for marking an event to zero will minimize the variance of the observations. Kim *et al.* [53] compared the recognition accuracy in a small isolated word task using ZCPA with similar results obtained using LPC-based features, features from the EIH model, and features obtained using zero crossings without the weighting by the peak amplitude. The ZCPA approach provided the greatest accuracy in all cases, especially at low SNRs. Ghulam *et al.* [33, 34] augmented the ZCPA procedure by adding auditory masking, Wiener filtering, and a weighting of the frequency histograms to emphasize components that are close to harmonics of the fundamental frequency.

C. Kim *et al.* [52] implemented a synchrony-based estimation of spectral contours using

a third method: direct Fourier transformation of the phase-locked temporal envelopes of the outputs of the critical-band filters. This produces a high-resolution spectral representation at low frequencies for which the auditory nerve is synchronized to the input up to about 2.2 kHz, and which includes the effects of all of the nonlinearities of the peripheral processing. The use of the synchrony processing at low frequencies provided only a modest improvement compared to the auditory model with mean-rate processing as shown in Fig. 8.16, although it was a large improvement compared to baseline MFCC processing.

Multi-stream processing. The *articulation index* model of speech perception, which was suggested by Fletcher [28] and French and Steinberg [30], and revived by Allen [4], modeled phonetic speech recognition as arising from independent estimators for critical bands. This initially led to a great deal of interest 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.* [11, 40, 75]). 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 [82, 81], at the level at which the HMM search takes place, which is sometimes referred to as 'state combination' [44, 65], or at the output level by merging hypothesis lattices [27, 67, 102]. In a systematic comparison of all of these approaches, Li [59] observed that state combination provides the best recognition accuracy by a small margin.

The *Tandem* approach, first proposed by Hermansky, Ellis, and Sharma [41], has been particular successful in facilitating the combination of multiple information streams at the feature level. Typically, the Tandem method is applied by expressing the outputs of a multilayer perceptron (MLP) as probabilities, which can be combined linearly or nonlinearly across the streams. These combined probabilities are then in turn (after some simple transformations, such as the logarithm followed by principal components analysis) used as features to a conventional hidden Markov model classifier. If the linear stream weights can be determined dynamically, there is at least the potential for robustness to time-varying environmental conditions. The MLP training is quite robust to the nature of the input distribution, and in particular can easily be used to handle acoustic inputs covering a large temporal context. Over the years the Tandem approach has proven to be a very useful way of combining rather diverse sets of features.

Long-time temporal evolution. An additional major 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. As noted in Sec. 8.2.2, some units in the brainstem of various mammals exhibit a sensitivity to amplitude modulation, with maximal responses at a particular modulation frequency independently of the carrier frequency. Psychoacoustical results also indicate that humans are sensitive to modulation frequency [112, 119]), with temporal modulation transfer functions indicating greatest sensitivity to temporal modulations at approximately the same frequencies as in the physiological data, despite the obvious species differences.

Initially this information has been used to implement features based on frequency components of these temporal envelopes, which (as noted in Sec. 8.2.2) are referred to by Kingsbury and others as the *modulation spectrum* [55]. Specifically, Kingsbury *et al.* [54] obtained lowpass and bandpass representations of the envelopes of the outputs of the critical-band filters by passing the filter outputs through a square-root nonlinearity, followed

by a lowpass filter with a 16-Hz cutoff and a bandpass filter with passband from 2 to 16 Hz (in parallel), and two subsequent AGC stages. The modulation spectrum is obtained by expressing these signals as a function of the center frequencies of the critical-band filters. This is a useful representation because speech signals typically exhibit temporal modulations with modulation frequencies in the range that is passed by this processing, while noise components often exhibit frequencies of amplitude modulation outside this range. Tchorz and Kollmeier [110] also developed an influential physiologically-motivated feature extraction system at about the same time that included the usual stages of filtering, rectification, transient enhancement. They were also concerned about the impact of modulation spectra, noting that their model provided the greatest output for temporal modulations around 6 Hz, and that in general lowpass filtering the envelopes of the outputs of the auditory model in each channel reduced the variability introduced by background noise.

Other researchers have subsequently characterized the temporal patterns more explicitly. In general these procedures operate on the time-varying envelope or log energy of a long temporal segment that is the output of a single critical-band filter. These representations effectively slice a spectrographic representation into horizontal 'slices' rather than the vertical slices isolated by the conventional windowing procedure, which is brief and time and broad and frequency. As an example, Hermansky and Sharma [38] developed the *TRAPS* representation (for *TempoRAl PatternS*), which operates on 1-second segments of the log spectral energies that emerge from each of 15 critical-band filters. In the original implementation, these outputs were classified directly by a multi-layer perceptron (MLP). This work was extended by Chen *et al.* [13] who developed *HATS* (for *Hidden Activation TRAPS*), which use the trained internal representation (input to hidden weights for each MLP) of a separate network for each critical band filter to provide a set of basis functions optimized to maximize the discriminability of the data to be classified.

Athineos and Ellis [6, 9, 7] have developed *frequency-domain linear prediction*, or *FDLP*. In this process, the temporal envelopes of the outputs of critical band filters are represented by linear prediction. Much as linear-predictive parameters computed from the time-domain signal within a short analysis window (*e.g.* 25 ms) represent the envelopes of the short-time spectrum within a slice of time, the FDLP parameters represent the Hilbert envelope of the *temporal sequence* within a slice of spectrum. This method was further incorporated into a method called *LP-TRAPs* [8], in which the FDLP-derived Hilbert envelopes were used as input to MLPs that learned phonetically-relevant transformations for later use in speech recognition. LP-TRAPS can be considered to be a parametric estimation approach to characterizing the trajectories of the temporal envelopes, while traditional TRAPS is non-parametric in nature.

It is also worth restating that RASTA processing, described in Sec. 8.2.4, was developed to emphasize the critical temporal modulations (and in so doing emphasizes transitions, roughly models forward masking, and reduces sensitivity to irrelevant steady state convolutional factors). More recently, temporal modulation in subbands was normalized to improve ASR in reverberant environments [60].

Joint feature representation in frequency, rate, and scale. There has been substantial interest in recent years in developing computational models of speech processing based on the spectro-temporal response functions (STRFs) that were described in Sec. 8.2.2. In an influential set of studies, Chi *et al.* [14] conducted a series of psychoacoustical experiments that measured the spectro-temporal modulation transfer functions (MTF) in

response to moving ripple signals such as those used to develop physiological STRFs, arguing that the results were consistent the physiologically-measured STRFs, and that the spectro-temporal MTFs are separable into the product of a spectral MTF and a temporal MTF. Subsequently this enabled Chi *et al.* to propose a model of central auditory processing with three independent variables: auditory frequency, 'rate' (characterizing temporal modulation), and 'scale' (characterizing spectral modulations), with receptive fields of varying extent as would be obtained by successive stages of wavelet processing [15]. The model relates this representation to feature extraction at the level of the brainstem and the cortex, including detectors based on STRFs, incorporating an auditory model similar to those described above that provides the input to the STRF filtering. Chi *et al.* also generated speech from the model outputs and compared the intelligibility of the reconstructed speech to the degree of spectral and temporal modulation in the signal.

A number of researchers have found it convenient to use two-dimensional Gabor filters as a reasonable and computationally-tractable approximation to the STRFs of A1 neurons. This representation was used by Mesgarani *et al.* to implement features for speech/nonspeech discrimination [71] and similar approaches were used to extract features for ASR by multiple researchers (*e.g.* [57, 39, 122, 73]). In many of these cases, 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 described above [41]. The filters can either be used as part of a single modulation filter bank that either does or does not incorporate a final MLP, or the filters can be split into multiple streams, each with their own MLP, as described in the multi-stream section above. The choice of filters can either be data-driven (as in [57]) or chosen to span the space of interest, *i.e.*, to cover the range of temporal and spectral modulations that are significant components of speech (*e.g.*[72, 74]).

In one recent study, Ravuri [92] developed a complex model that incorporates hundreds of 2-dimensional Gabor filters, each with their own discriminatively-trained neural network to generate noise-insensitive features for ASR. As an example, as shown in Fig. 8.18, Ravuri and Morgan [93] describe the recognition accuracy that is obtained by incorporating feature streams developed by modeling a range of STRFs of mel spectra using Gabor filters. The resulting MLP posterior probability outputs are linearly combined across streams (the 'Selection/Combination' block in the left panel of Fig. 8.18), where each stream is weighted by inverse entropy of the posterior distribution for each stream's MLP. The combined stream is further processed with a log function to roughly Gaussianize it, and with a Karhunen-Loeve transformation to orthogonalize the features; both steps are taken provide a better statistical match of the features to systems based on Gaussian mixtures. The system was trained on the high-SNR Numbers95 database and tested on an independent Numbers95 set with noises added from the RSG-10 database, which comprises a range of noises including speech babble, factory noise, etc. The results labeled 'multistream Gabor' were obtained using 174 feature streams, each of which included a single spectro-temporal filter followed by an MLP trained for phonetic discrimination.

8.5 Summary

In this chapter we have reviewed a number of signal processing concepts that have been abstracted from several decades of concentrated study of how the auditory system responds to sound. The results of these studies have provided insight into the development of more

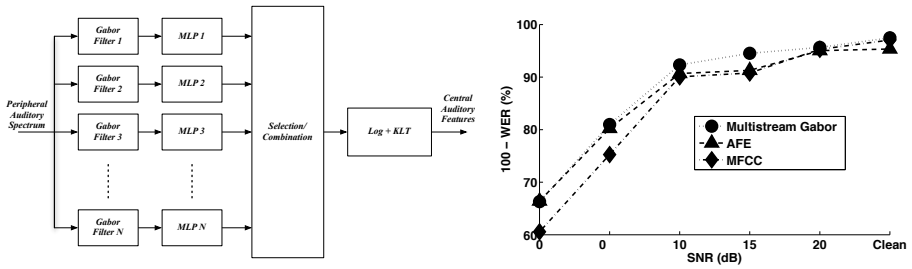

Figure 8.18: (left) Generation of multistream Gabor filter features. See text for details. (right) ASR word accuracy on the Numbers95 test set in the presence of a range of real-world noise sources using a system trained on clean speech. Results shown use the Gabor-based features to augment an MFCC feature vector, using SRI's Decipher system. From [93].

environmentally-robust approaches to feature extraction for automatic speech recognition. While we have only explicitly included a limited number of experimental ASR results from our own groups, many physiologically-motivated feature extraction procedures have demonstrated recognition accuracy that is as good as or better than the recognition accuracy provided by conventional signal processing, at least in degraded acoustical environments.

Although there remains no universally-accepted theory about which aspects of auditory processing are the most relevant to robust speech recognition, we may speculate with some confidence about some of the reasons for the general success of auditory models. The increasing bandwidth of the auditory analysis filters with increasing center frequency enables good spectral resolution at low CFs (which is useful for tracking formant frequencies precisely) and better temporal resolution at higher CFs (which is helpful in marking the precise time structure of consonant bursts). The nonlinear nature of the auditory rate-intensity function tends to suppress feature variability caused by additive low-level noise, and an appropriate shape of the nonlinearity can provide normalization to absolute amplitude as well. The short-time temporal suppression and lateral frequency suppression provides an ongoing enhancement of change with respect to running time and analysis frequency. As has been noted by Wang and Shamma [115] and others, the tendency of the auditory system to enhance local spectro-temporal contrast while averaging the incoming signals over a broader range of time and frequency enables the system to provide a degree of suppression to the effects of noise and reverberation. Bandpass filtering of the modulation spectrum between 2 and 16 Hz will help to separate the responses to speech and noise, as many disturbances produce modulations outside that range. In many respects, the more central representations of speech at the level of the brainstem and the cortex are based primarily on the dynamic aspects of the speech signal, and perceptual results from classical auditory scene analysis [12] confirm the importance of many of these cues in segregating individual sound sources in cluttered acoustical environments.

The good success of relatively simple feature extraction procedures such as PNCC processing, RASTA and similar approaches suggests that the potential benefits from the use of auditory processing are widespread. While our understanding of how to harness the potential of more central representations such as the spectro-temporal response functions

is presently in its infancy, we expect that we will be able to continue to improve the robustness and overall utility of our representations for speech as we continue to deepen our understanding of how speech is processed by the auditory system.

Acknowledgements

This research was supported by NSF (Grants IIS-0420866 and IIS-0916918) at CMU, and Cisco, Microsoft, and Intel Corporations and internal funds at ICSI. The authors are grateful to Chanwoo Kim and Yu-Hsiang (Bosco) Chiu for sharing their data, along with Mark Harvilla, Kshitiz Kumar, Bhiksha Raj, and Rita Singh at CMU, as well as Suman Ravuri, Bernd Meyer, and Sherry Zhao at ICSI for many helpful discussions.

References

[1] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 1990, pp. 849–852.

[2] A. M. A. Ali, J. V. der Spiegel, and P. Mueller, "Robust auditory-based speech processing using the average localized synchrony detection," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 279–292, 1999.

[3] J. B. Allen, "Cochlear modeling," *IEEE ASSP Magazine*, vol. 1, pp. 3–29, 1985.

[4] J. B. Allen, "How do humans process and recognize speech?" *IEEE Trans. on Speech and Audio*, vol. 2, pp. 567–577, 1994.

[5] R. M. Arthur, R. R. Pfeiffer, and N. Suga, "Properties of two-tone inhibition in primary auditory neurons," *J. Physiol.*, vol. 212, pp. 593–609, 1971.

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

[7] M. Athineos and D. P. W. Ellis, "Autoregressive modeling of termporal envelopes," *IEEE Trans. on Signal Processing*, vol. 15, pp. 5237–5245, 2007.

[8] M. Athineos, H. Hermansky, and D. P. W. Ellis, "LP-TRAP: Linear predictive temporal patterns," in *Proc. Int. Conf. Spoken Language Processing*, 2004, pp. 949–952.

[9] M. Athineos, H. Hermansky, and D. P. W. Ellis, "PLP^2: Autoregressive modeling of auditory-like 2-D spectro-temporal patterns," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, 2004, pp. 25–30.

[10] J. C. Baird and E. Noma, *Fundamentals of Scaling and Psychophysics*. Wiley, 1978.

[11] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards sub-band-based speech recognition," in *Proc. European Signal Processing Conference*, 1996, pp. 1579–1582.

[12] A. S. Bregman, *Auditory scene analysis*. Cambridge, MA: MIT Press, 1990.

[13] B. Chen, S. Chang, and S. Sivadas, "Learning discriminative temporal patterns in speech: Development of novel TRAPS-like classifiers," in *Proc. Eurospeech*, 2003.

[14] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. A. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 106, pp. 719–732, 1999.

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

[16] B. Chigier and H. C. Leung, "The effects of signal representations, phonetic classification techniques, and the telephone network," in *Proceedings of the International Conference of Spoken Language Processing*, 1992, pp. 97–100.

[17] Y.-H. Chiu and R. M. Stern, "Analysis of physiologically-motivated signal processing for robust speech recognition," in *Proc. Interspeech*, 2008.

[18] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, vol. 85, pp. 2623 – 2629, 1989.

[19] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.

[20] 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, pp. 1220–1234, 2001.

[21] A. Dreyer and B. Delgutte, "Phase locking of auditory-nerve fibers to the envelopes of high-frequency sounds: implications for sound localization," *J. Neurophysiol.*, vol. 96, no. 5, pp. 2327–2341, 2006.

[22] R. Drullman, J. M. Festen, and R. Plomp, "Effects of temporal envelope smearing on speech reception," *Journal of the Acoustical Society of America*, vol. 95, pp. 1053–1064, 1994.

[23] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of the Acoustical Society of America*, vol. 95, pp. 2670–2680, 1994.

[24] N. I. Durlach and H. S. Colburn, "Binaural phenomena," in *Hearing*, ser. Handbook of Perception, E. C. Carterette and M. P. Friedman, Eds. Academic Press, New York, 1978, vol. IV, ch. 10, pp. 365–466.

[25] European Telecommunications Standards Institute, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep. ETSI ES 202 050, Rev. 1.1.5, January 2007.

[26] G. T. Fechner, *Element der Psychophysik*. Breitkopf & Härterl; (English translation by H. E. Adler, Holt, Rinehart and Winston, 1966), 1860.

[27] J. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *Proc. IEEE ASRU Workshop*, 1997, pp. 347–354.

[28] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–65, 1940.

[29] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *J. Acoustic. Soc. Amer.*, vol. 5, pp. 82–108, 1933.

[30] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *Journal of the Acoustical Society of America*, vol. 19, pp. 90–119, 1947.

[31] G. A. Gescheider, *Psychophysics: The Fundamentals*. Psychology Press, 1997.

[32] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and Language*, vol. 1, pp. 109–130, 1986.

[33] M. Ghulam, T. Fukuda, J. Horikawa, and T. Niita, "Pitch-synchronous $zcpa$ ($ps-zcpa$)-based feature extraction with auditory masking," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2005.

[34] M. Ghulam, T. Fukuda, K. Katsurada, J. Horikawa, and T. Niita, "Ps-zcpa based feature extraction with auditoy masking, modulation enhancement and noise reduction," *IECE Trans. Information and Systems*, vol. E89-D, pp. 1015–1023, 2006.

[35] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory-nerve model for predicting performance limits of normal and impaired listeners," *Acoustics Research Letters Online*, vol. 2, pp. 91–96, July 2001.

[36] H. Hermansky, "Perceptual linear predictive (PLP) anlysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[37] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[38] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1999.

[39] H. Hermansky and F. Valente, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2008, pp. 4165–4168.

[40] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards asr on partially corrupted speech," in *Proc. Int. Conf. Spoken Language Processing*, vol. 1, 1996, pp. 462 – 465.

[41] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. IEEE ICASSP*, 2000, pp. 1635–1638.

[42] M. J. Hewitt and R. Meddis, "An evaluation of eight computer modles of mammalian inner hair-cell function," *J. Acoust. Soc. Amer.*, vol. 90, pp. 904–917, 1991.

[43] M. J. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 1989, pp. 262–265.

[44] A. Janin, D. P. W. Ellis, and N. Morgan, "Multi-stream speech recognition: Ready for prime time?" *Proc. Eurospeech*, pp. 591–594, 1999.

[45] C. R. Jankowski, H.-D. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 286–293, 1995.

[46] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psych.*, vol. 41, pp. 35–39, 1948.

[47] P. X. Joris, C. E. Schreiner, and A. Rees, "Neural processing of amplitude-modulated sounds," *Physiol. Rev.*, vol. 84, pp. 541–577, 2004.

[48] N. Y.-S. Kiang, T. Watanabe, W. C. Thomas, and L. F. Clark, *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*. MIT Press, 1966.

[49] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing Conf. on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4574–4577.

[50] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech*, September 2009, pp. 28–31.

[51] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Proc.* (accepted for pubication), 2012.

[52] C. Kim, Y.-H. Chiu, and R. M. Stern, "Physiologically-motivated synchrony-based processing for robust automatic speech recognition," in *Proc. Interspeech*, 2006, pp. 1975–1978.

[53] D.-S. Kim, S.-Y. Lee, and R. Kil, "Auditory proccessing of speech signals for robust speech recognition in real world noisy environments," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 55–59, 1999.

[54] B. E. D. Kingsbury, "Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments," Ph.D. dissertation, University of California, Berkeley, 1998.

[55] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1–3, pp. 117–132, 1998.

[56] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust spectro-temporal reverse correlation for the auditory system: Opttimizing stimulus design," *J. Comp. Neurosci.*, vol. 9, pp. 85–111, 2000.

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

[58] G. Langner and C. E. Schreiner, "Periodicity coding in the inferior colliculus of the cat. I. neuronal mechanisms," *J. Neurophysiol.*, vol. 60, pp. 1799–1822, 1988.

[59] X. Li, "Combination and generation of parallel feature streams for improved speech recognition," Ph.D. dissertation, Carnegie Mellon University, 2005.

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

[61] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, May 1982, pp. 1282–1285.

[62] R. F. Lyon, "A computational model of binaural localization and separation," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1983, pp. 1148–1151.

[63] R. F. Lyon, "Computational models of neural auditory processing," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing of the International Conference on Acoustics, Speech and Signal Processing*, 1984, pp. 36.1.1–36.1.4.

[64] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Procc. IEEE. Int. Symposium on Circuits and Systems*, 2010, pp. 3809–3812.

[65] 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. IEEE ICASSP*, 2010, pp. 4394–4397.

[66] J. Makhoul and L. Cosell, "LPCW: An LPC vocoder with linear predictive spectral warping," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1976, pp. 466–469.

[67] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition; word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, pp. 373–400, 2000.

[68] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Amer.*, vol. 79, pp. 702–711, 1986.

[69] R. Meddis, "Simulation of auditory-neural transduction: further studies," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1056–1063, 1988.

[70] H. Meng and V. W. Zue, "A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons," in *Proc. Int. Conf. Spoken Language Processing*, 1990, pp. 1053–1056.

[71] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, pp. 920–929, 2006.

[72] B. T. Meyer and B. Kollmeier, "Robustness of spectrotemporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, pp. 753–767, 2011.

[73] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition as a function of speech-intrinsic variabilities," *J. Acoust. Soc. Amer.*, vol. 128, pp. 3126–3141, 2010.

[74] B. T. Meyer, S. V. Ravuri, M. R. Schaedler, and N. Morgan, "Comparing different flavors of spectro-temporal features for asr," in *Proc. Interspeech*, 2011.

[75] N. Mirghafori, "A multi-band approach to automatic speech recognition," Ph.D. dissertation, University of California, Berkeley, Berkeley CA, January 1999.

[76] A. R. Møller, "Coding of amplitude and frequency modulated sounds in the cochlear nucleus," *Acustica*, vol. 31, pp. 292–299, 1974.

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

[78] B. C. J. Moore, "Frequency analysis and masking," in *Hearing*, 2nd ed., ser. Handbook of Perception and Cognition, B. C. J. Moore, Ed. Academic Press, 1995, ch. 5, pp. 161–205.

[79] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, pp. 750–731, 1983.

[80] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, May 1996, pp. 733–736.

[81] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 20, pp. 7–13, 2012.

[82] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cetin, H. Bourlard, and M. Athineos, "Pushing the envelope – aside," *IEEE Signal Processing Magazine*, vol. 22, pp. 81–88, 2005.

[83] Y. Ohshima, "Environmental robustness in speech recognition using physiological-motivted signal processing," Ph.D. dissertation, Carnegie Mellon University, December 1993.

[84] Y. Ohshima and R. M. Stern, "Environmental robustness in automatic speech recognition using physiologically-motivated signal processing," in *Proceedings of the International Conference of Spoken Language Processing*, 1994.

[85] D. O'Shaughnessy, *Speech Communication: Human and Machine*, 2nd ed. Wiley-IEEE Press, 2000.

[86] A. Palmer and S. Shamma, "Physiological representations of speech," in *Speech Processing in the Auditory System*, ser. Springer Handbook of Auditory Research, S. Greenberg, A. N. Popper, and R. R. Fay, Eds. Springer-Verlag, 2004, ch. 4.

[87] A. R. Palmer, "Physiology of the cochlear nerve and cochlear nucleus," in *Hearing*, M. P. Haggard and E. F. Evans, Eds. Churchill Livingstone, Edinburgh, 1987.

[88] R. D. Patterson and I. Nimmo-Smith, "Off-frequency listening and auditory filter asymmetry," *J. Acoustic. Soc. Amer.*, vol. 67, no. 1, pp. 229–245, 1980.

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

[90] J. O. Pickles, "The neurophysiological basis of frequency selectivity," in *Frequency Selectivity in Hearing*, B. C. J. Moore, Ed. Plenum, NY, 1986.

[91] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Prentice-Hall, 2010.

[92] S. Ravuri, "On the use of spectro-temporal features in noise-additive speech," Master's thesis, University of California, Berkeley, 2011.

[93] S. Ravuri and N. Morgan, "Easy does it: Many-stream ASR without fine tuning streams," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2012.

[94] J. E. Rose, N. B. Gross, C. D. Geisler, and J. E. Hind, "Some neural mechanisms in the inferior colliculus of the cat which may be relevant to localization of a sound source," *J. Neurophysiol.*, vol. 29, pp. 288–314, 1966.

[95] J. E. Rose, J. E. Hind, D. J. Anderson, and J. F. Brugge, "Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey," *J. Neurophysiol.*, vol. 34, pp. 685–699, 1971.

[96] M. B. Sachs and N. Y.-S. Kiang, "Two-tone inhibition in auditory-nerve fibers," *J. Acoustic. Soc. Amer.*, vol. 43, pp. 1120–1128, 1968.

[97] M. B. Sachs and E. D. Young, "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," *J. Acoustic. Soc. Amer.*, vol. 55, pp. 470–479, 1979.

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

[99] M. R. Schroeder and J. L. Hall, "A model for mechanical to neural transwduction in the auditory receptor," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1055–1060, 1974.

[100] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 15, pp. 55–76, 1988.

[101] S. A. Shamma, "The acoustic features of speech sounds in a model of auditory processing: Vowels and voiceless fricatives," *J. Phonetics*, vol. 16, pp. 77–91, 1988.

[102] R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, "Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2001, pp. 273–276.

[103] M. Slaney, *Auditory Toolbox (V.2)*, 1998, http://www.slaney.org/malcolm/pubs.html. [Online]. Available: http://www.slaney.org/malcolm/pubs.html

[104] R. M. Stern and C. Trahiotis, "Models of binaural interaction," in *Hearing*, 2nd ed., ser. Handbook of Perception and Cognition, B. C. J. Moore, Ed. Academic (New York), 1995, ch. 10, pp. 347–386.

[105] R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Speech Recognition*, C.-H. Lee and F. Soong, Eds. Kluwer Academic Publishers, 1996, ch. 14, pp. 351–378.

[106] R. M. Stern, D. Wang, and G. J. Brown, "Binaural sound localization," in *Computational Auditory Scene Analysis*, D. Wang and G. J. Brown, Eds. Wiley-IEEE Press, 2006, ch. 5.

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

[108] S. S. Stevens, J. Volkman, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, pp. 185–190, March 1937.

[109] Strutt JW, Third Baron of Rayleigh, "On our perception of sound direction," *Philosoph. Mag.*, vol. 13, pp. 214–232, 1907.

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

[111] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *J. Acoustic. Soc. Amer.*, vol. 88, pp. 97–100, 1990.

[112] N. Viemeister, "Temporal modulation transfer function based on modulation thresholds," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1364–1380, 1979.

[113] G. von Békésy, *Experiments in Hearing*. McGraw Hill; reprinted by the Acoustical Society of America, 1989, 1960.

[114] H. W. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *American Journal of Psychology*, vol. 62, pp. 315–337, 1949.

[115] K. Wang and S. A. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 421–435, 1994.

[116] E. H. Weber, *De pulsu, resorpitione, auditu et tactu: Annotations anatomicae et physiologicae.* Leipzig: Koehlor, 1834.

[117] T. C. T. Yin and J. C. K. Chan, "Interaural time sensitivity in medial superior olive of cat," *J. Neurophysiol.*, vol. 64, pp. 465–474, 1990.

[118] W. A. Yost, *Fundamentals of Hearing: An Introduction*, 5th ed. Emerald Group Publishing, 2006.

[119] W. A. Yost and M. J. Moore, "Temporal changes in a complex spectral profile," *J. Acoust. Soc. Amer.*, vol. 81, pp. 1896–1905, 1987.

[120] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the emporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoustic. Soc. Amer.*, vol. 66, pp. 1381–1403, 1979.

[121] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenologial model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppresion," *Journal of the Acoustical Society of America*, vol. 109, pp. 648–670, 2001.

[122] S. Y. Zhao, S. Ravuri, and N. Morgan, "Multi-stream to many-stream: Using spectgro-temporal features in ASR," in *Proc. Interspeech*, 2009.

[123] M. S. A. Zilarny, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomonological model of the synapse between the inner hair cell and auditory nerve: mlong-term adaptation with power-law dynamics," *J. Acoust. Soc. Amer.*, vol. 126, pp. 2390–2412, 2009.

[124] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoustic. Soc. Amer.*, vol. 33, p. 248, February 1961.

[125] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoustic. Soc. Amer.*, vol. 68, pp. 1523–1525, November 1980.