INVITED PAPER

Perceptual Properties of Current Speech Recognition Technology

Modern automatic speech recognition systems are reviewed in this paper, with a focus on comparing the technical approaches to aspects of the human auditory system.

By Hynek Hermansky, *Fellow IEEE*, Jordan R. Cohen, *Senior Member IEEE*, and Richard M. Stern, *Senior Member IEEE*

ABSTRACT | In recent years, a number of feature extraction procedures for automatic speech recognition (ASR) systems have been based on models of human auditory processing, and one often hears arguments in favor of implementing knowledge of human auditory perception and cognition into machines for ASR. This paper takes a reverse route, and argues that the engineering techniques for automatic recognition of speech that are already in widespread use are often consistent with some well-known properties of the human auditory system.

KEYWORDS | Auditory perception; feature extraction; speech recognition

I. INTRODUCTION

Automatic speech recognition (ASR) by machines attempts to emulate the part of the human speech communication chain that recovers the linguistic message from the speech signal. By some estimates, this involves reducing the information rate of the speech signal by about three orders of magnitude [1].

ASR is currently dominated by stochastic approaches, as outlined by Jelinek [2] in which the string of recognized words

$$\hat{W} = \arg\max_W \{p(x|W)p(W)\} \tag{1}$$

where x represents a series of measurements describing the speech signal and W refers to strings of words generated from the given stochastic model. The architecture of the stochastic model that generates W and the type of data x that describe the given speech segment need to be specified by the designer. Leaving aside the architecture of the model, we will discuss what the data x (usually referred to as "speech features") might be.

A typical process that executes (1) is shown in Fig. 1.

A. Features That Describe Speech

An ASR front end attempts to derive message-relevant information from the speech signal. The feature extraction module supports this goal. The signal itself contains extraneous information which has little to do with the words of the sentence (a message), such as information about who is speaking, information about the acoustic environment in which the speech was produced, information about the communication channel through which the speech was processed, etc. Ideally much of this unnecessary information would be eliminated so the features x would carry only information about the message.

The feature extraction module is a critical part of a speech recognizer. The useful information which is not passed to the ASR system is lost forever. On the other hand, irrelevant information which is not removed has to be dealt with by the ASR system, often at significant expense.

Manuscript received July 30, 2012; revised December 10, 2012; accepted December 20, 2012.. Date of publication July 23, 2013; date of current version August 16, 2013. This work was supported in part by the National Science Foundation under Grant IIS-I0916918, by the Cisco Research Center under Grant 570877, and by the Center of Excellence in Human Language Technologies at The Johns Hopkins University.

H. Hermansky is with the Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: hynek@jhu.edu).

J. R. Cohen is with Spelamode, Inc., Kure Beach, NC 28449 USA (e-mail: Jordan@spelamode.com).

R. M. Stern is with the Department of Electrical and Computer Engineering and the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: rms@cmu.edu).

Digital Object Identifier: 10.1109/JPROC.2013.2252316

0018-9219 © 2013 IEEE

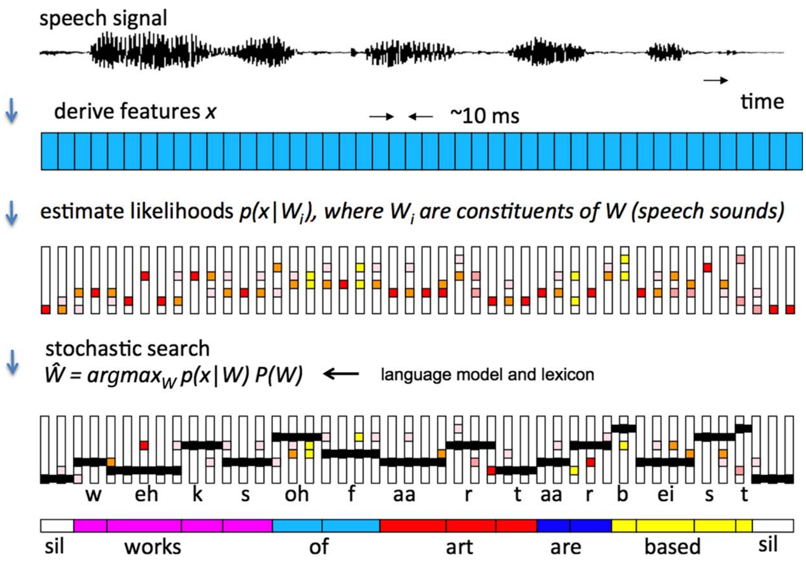

speech signal
time
derive features x
~10 ms
estimate likelihoods p(x|W_i), where W_i are constituents of W (speech sounds)
stochastic search
Ŵ = argmax_W p(x|W) P(W)
language model and lexicon
w eh k s oh f aa r t aa r b ei s t
sil works of art are based sil

Fig. 1. ***Speech features x are derived in regular intervals of about 10 ms from segments of the speech signal. The blue rectangles under the waveform represent the initial feature vectors x. The values of the features are used in the pattern classification module (which could be based on a Gaussian mixture model or an artificial neural network) to derive vectors of likelihoods of subword units $p(\mathbf{x}|\mathbf{W}_i)$, which are presumed to be related to the sounds of speech, and which are represented by the next row rectangles below (higher likelihoods are indicated by warmer colors). A stochastic search (typically a Viterbi search) yields the best sequence of subword units representing the recognized utterance W. The remaining part of the information used in the search, represented by $P(\mathbf{W})$, is supplied by a language model that is typically derived from text data, together with a lexicon that specifies which words are expected to occur and how they are pronounced.***

B. Perceptual Approaches to ASR

Human listeners solve the problem of speech recognition daily, and seemingly effortlessly. It is very likely that speech evolved under forces of nature over millennia, optimizing the use of the human auditory system. Thus, it is likely that important information-carrying elements in speech are easily heard and some of the elements which do not carry information are suppressed by the hearing system. Understanding human auditory perception and of the ways to emulate it in engineering systems should be useful for ASR.

Unfortunately, the situation is not so simple. Listening for the message in speech is not the only task that human auditory perception must accomplish. Knowing what and what not to emulate when recognizing the message in speech is important. We suggest that one way to proceed is to focus on successful and well-accepted ASR solutions and compare their properties with what we know about the perception of signals, and of speech in particular. Often, as we show in this paper, the engineering solution turns out to be a reflection of particular characteristics of hearing.

Hence, we suggest that clues from many engineering techniques for speech feature extraction can be understood as reflections of some more general characteristics of the hearing systems of humans. The important characteristics of perception can be sorted from the less important ones by assessing our current solutions.

II. COMPENSATING FOR DIFFERENT TALKERS

The first issue that we address is how ASR can accomplish its task independently of the identity of the talker.

A. The Problem of Speaker Variability

When different speakers produce the same message, the speech signal can be very different. The differences come from different anatomies and different speaking habits of the speakers. This introduces differences in the spectral composition and in the temporal structure of the signal. In spite of these differences, the message can be easily decoded by a human listener.

Differences in speaking rate induce differences in the temporal length of phones, words, and phrases but also differences in their spectral properties. Some ways of handling the differences in speaking rate are discussed in Section III; here we will focus on differences induced by speaker anatomy.

1) Effects of Gender: Since the time of the first ASR attempts to develop a voice-operated typewriter [4], the short-time power spectrum with its short-time phase eliminated was used as features for speech recognition. The most obvious spectral characteristic in the speech of different speakers is the fundamental frequency created by the periodic excitation of the vocal tract. Male speakers generally have lower frequencies of vocal cord vibrations in the production of voiced sounds than do female speakers. These differences are obvious in the detailed structure of the short-time spectrum of speech, seen in Fig. 2. This structure, heard as "pitch," does not interfere with conversations among humans.

Many of the remaining differences in spectral envelopes between speakers are due to differences in the lengths of their vocal tracts. These differences are easily seen in comparisons of spectral envelopes (formant structure) of the short-time spectrum of speech. More subtle differences in spectral composition of speech come from speaker differences such as accents or dialects. They are often observed in spectral properties of sonorants such as vowels. These do present some difficulties in speech communication, but a human listener adapts relatively quickly.

2) Speech of Children: The easily understood speech of small children is produced by small vocal tracts which could be as short as half the length of adult speakers and have fewer resonant modes (formants) than adult speakers do. Small children can also have extremely high fundamental frequency. As such, the speech of children presents a significant challenge to any spectrum-based feature extraction scheme. The left part of Fig. 2 (adopted from [5]) shows spectrograms of voiced utterances with identical messages produced by an adult male and by a four-year-old child. Substantial differences are obvious.

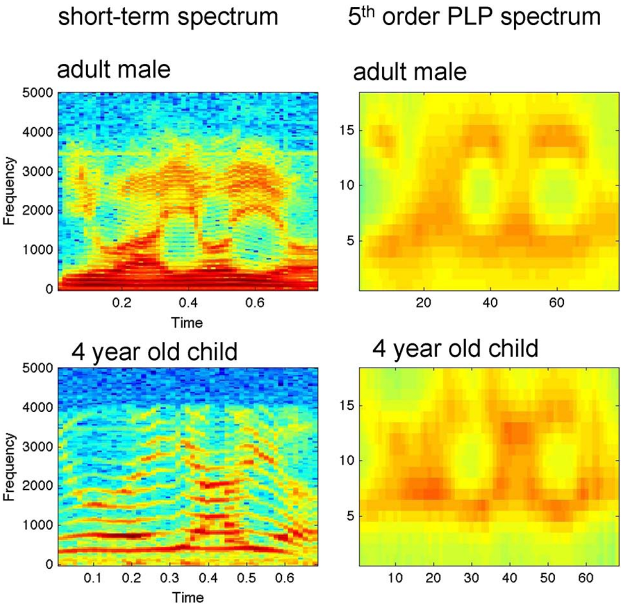


Fig. 2. ***Spectrograms of the speech of adults and children. Note differences in both the fine structures as well as in the overall spectral shapes. Adopted from [3].***

B. Engineering Approaches to Speaker-Independent Features

1) Extracting Spectral Envelope: The differences in fine spectral structure of short-time speech spectrum of different speakers are so obvious that no practical ASR system uses the full speech spectra directly. There is always some kind of spectral smoothing done in the feature extraction module so that the overall spectral envelope is emphasized. Deriving the spectral envelopes of the short-time spectra compensates for the effects of differences in fine spectral structures. However, the differences in spectral envelopes still remain.

2) Use of a Nonlinear Frequency Axis: The decreasing spectral resolution with frequency was known and well accepted by some early speech engineers and attempts were made to include it in engineering designs. However, the spectrograph, developed during World War II, employed a linear frequency scale. The fact that human perception is sensitive to relative rather than to absolute changes in formants [1], thus implying use of the logarithmic rather than linear frequency scale, was known but largely ignored by ASR engineers until Bridle and Brown proposed to derive features from speech by taking the cosine transform of the output of a nonuniform filterbank whose bandwidths follow the mel scale [7]. This idea was advanced by Mermelstein [8] who implemented a mel-scale filterbank by triangularly weighting adjacent bands of the Fourier power spectra of speech. This process is now widely referred to as computing the mel cepstrum. At about the same time, Itahashi and Yokoyama in Japan [9] described mel-scale-based linear prediction coefficients (LPCs), where the spectrum of the autoregressive LPC model is warped to the mel scale and approximated by another LPC spectrum; Makhoul and Cosell subsequently proposed mel spectral warping in an LPC vocoder [10]. Strube [11] used an all-pass filter to warp the Fourier spectrum prior to its approximation by an all-pole autoregressive model. Many groups in the 1970s used nonlinear warping of the frequency spectrum in their definition of speech features, usually represented in a low-dimensional orthogonal space. This made speech recognition much less sensitive to the identity of the speaker(s).

One such technique, which has survived through time, is perceptual linear predictive (PLP) analysis, developed by Hermansky *et al.* [3], [12] in the 1980s. Unlike standard LPC analysis, which approximates the power spectrum of speech, the PLP method applies linear prediction in the modified spectral domain [13] using a cube-root transformation to compress the auditory-like spectrum[1] prior to its approximation by an all-pole model. PLP features, together

[1] A Fourier power spectrum integrated over critical bands and pre-emphasized by a simulated equal loudness curve.

with mel cepstral features, are currently the dominant feature sets used in ASR systems.

An auditory-like frequency scale also emerges from data-driven investigations. Biem and Katagiri minimized classification error in multispeaker phonetic classification by deriving a proper filterbank [14]. The optimized filterbank center frequencies followed an auditory-like frequency scale. Attempts were made to find a universal frequency-warping function that would make the formants of the steady parts of vowels of four adult speakers similar, yielded mel-like warping [15]. In further support of this finding, Kamm *et al.* [16] found mel-like warping optimal for the classification of vowels. Subsequently, linear discriminant analysis of short-time spectral vectors taken from hand-labeled vowels from about 30 min of telephone speech yielded spectral bases that exhibited spectral resolution that was consistent with the nonuniform resolution of human hearing [17]. This analysis was repeated on a much larger and more realistic machine-labeled database, which confirmed the earlier observations [6]. Spectral bases derived by linear discriminant analysis (LDA) techniques are shown in Fig. 3. Higher spectral resolution at low frequencies, consistent with critical-band spectral resolution of human hearing, is evident from closer spacing of their zero crossings at lower frequencies and was demonstrated experimentally [6], [17]. Paliwal *et al.* also derived auditory-like warping by equalizing the spectral energy of speech in different frequency bands [18].

The fact that an auditory-like spectral resolution can be derived by optimizing speech sound classification is important. It supports the optimality of the speech code with respect to human auditory perception.

One algorithm for minimizing the distance between the same sounds created by different speakers is vocal tract length normalization (VTLN), developed by Andreou *et al.* [19] at the 1994 Workshop on Speech Recognition at Rutgers University (Piscataway, NJ, USA). They discovered that the first-order effects of vocal tract length variation can be accounted for by a single parameter, and that setting this parameter in switchboard data halved the error rates of a state-of-the-art recognizer. The procedure is widely used today, and it seems to positively affect all other normalization schemes, although not always providing reductions in error rate of 50%.

3) The Broad Shape of the Speech Spectrum: The first practical speech recognizer based on two-parameter temporal trajectories used two frequency bands with a lower bandwidth of about 1 kHz and the upper bandwidth close to 3 kHz [20]. It outperformed the elaborate short-time-spectrum-based scheme of Galt [4].

Nevertheless, ASR research soon returned to more detailed spectral envelopes as the basis for ASR features. It was only some three decades later when it was shown that the ASR performance of a low spectral resolution two-peak PLP model from the fifth-order PLP representation is noticeably less speaker dependent than LPC models that approximate linear power spectra or higher order PLP models that model more spectral details of auditory-like spectra [3], [21]. Speech synthesized from the low-order model was reported to be intelligible in informal listening tests [22]. This is consistent with results of later, more formal experiments [23], which show high intelligibility of

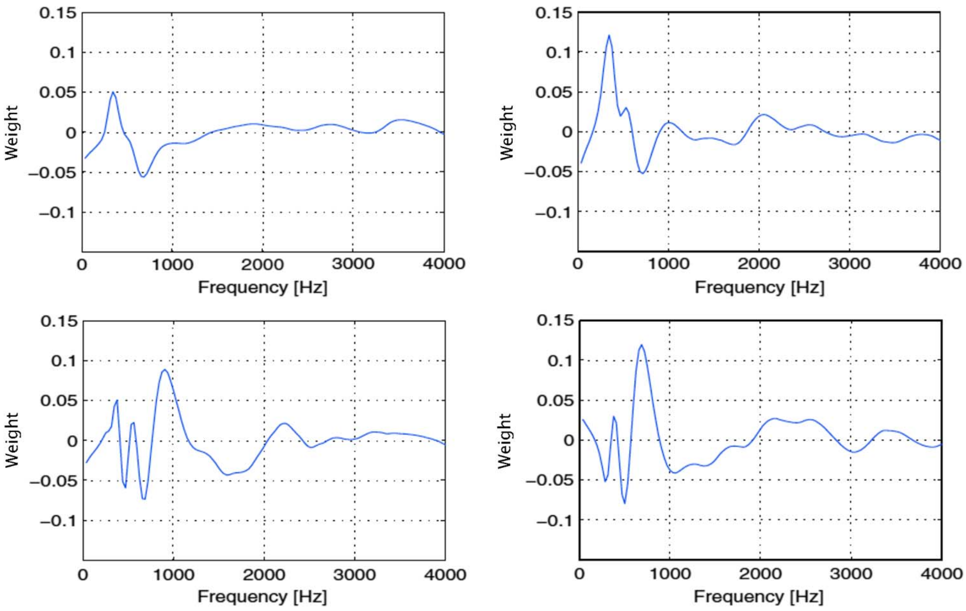


Fig. 3. ***Spectral bases derived by LDA analysis. Higher spectral resolution at low frequencies is evident from closer spacing of their zero crossings at lower frequencies. From [6], used with permission.***

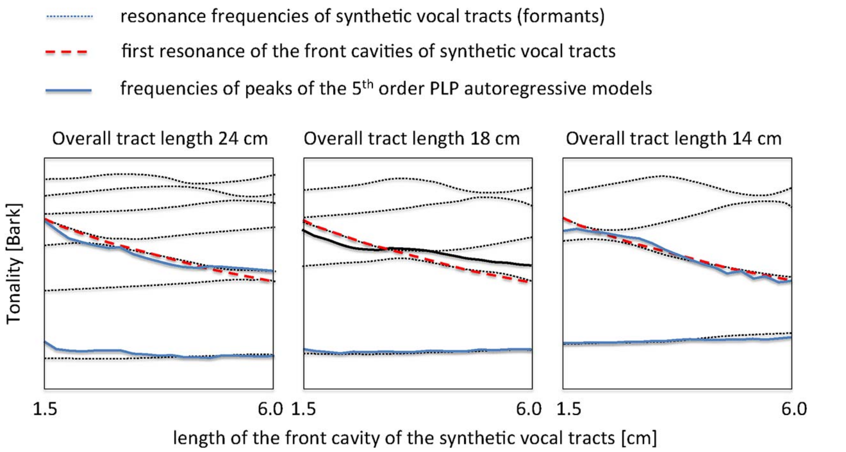


Fig. 4. ***Formants for three different simulated vocal tracts with different lengths are different but are always related to the resonance frequencies of the front cavity of the simulated vocal tract, which is always the same for all three tract lengths. The second peak from the fifth-order PLP model does not directly track individual formants but tracks the resonance frequency of the front cavity. Notice approximately linear relation between the cavity length and its resonance frequencies, first of which is indicated by dashed red line, resulting from the auditory-like Bark frequency scale. This relation would have been hyperbolic if the frequency was shown in hertz [30]. Adopted from [5].***

speech resynthesized from three to four bands of spectral energy.

The formants themselves must be speaker specific since they represent the shape of the whole vocal tract, which is, of course, speaker specific. However, smoothing out the formants emphasizes contributions of the underlying front cavity resonance and makes the spectral representation less speaker specific. The ability of the low-order PLP model to approximate the less speaker-dependent resonance frequency of the front part of a vocal tract was also demonstrated in [5]. Fig. 4 illustrates the key result.

Further support for the relative speaker independence of the model with low spectral resolution comes from experiments where speaker-specific linear regression models were used in generating male-like and female-like formants of voiced speech from the low-order PLP model [24]. Speaker-specific mapping between a high-order and low-order order model by a speaker-specific multilayer-perceptron neural network was successfully used in a speaker recognition experiment [25].

C. Physiological and Psychophysical Correlates

It is known from early experiments of von Békésy that the decreasing frequency resolution of hearing with decreasing frequency has its origin in resonance properties of the basilar membrane. The frequency channels that are formed in the periphery persist all the way to the auditory cortex (cf., [26]).

This decreasing spectral resolution with frequency is reflected in a number of psychophysical phenomena. Early experiments [27] revealed a logarithmic-like dependency of perceived pitch on frequency above about 800 Hz. Experiments in simultaneous masking indicate the existence of bandpass-like channels in human hearing. Above 800 Hz, the width of these bands increases approximately logarithmically with the band center frequency. Flanagan [28] demonstrated that human sensitivity to formant movements follows an approximately logarithmic scale, i.e., the just noticeable difference in the formant position depends on relative (rather than absolute) changes in formant frequency.

Fig. 5 from [5] shows tracings of the vocal tract of an adult and a child in the production of two vowels that indicate substantial speaker invariance of the front cavity shape, and the anatomical studies of Goldstein [29] support this notion. Experimental results with synthetic vocal tract shapes [30] show that formant positions in voiced speech are directly proportional to the frequency of

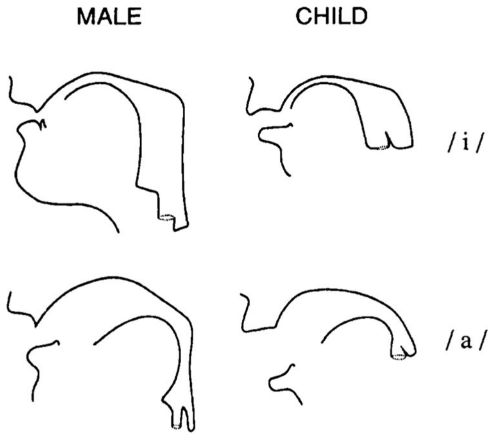


Fig. 5. ***Tracing of x-rays of the vocal tracts of adults and children in the production of vowels /aa/ and /iy/. Front parts of vocal tracts depend mainly on phonetic quality of the vowel, which the back part are highly speaker specific. (From [5], used with permission.)***

the nearest mode of the front cavity. Kuhn [31] experimentally demonstrated how the changes in higher formants of real voiced speech approximately follow changes in resonance modes of the vocal tract front cavity.

The relation between cavity resonance frequency and its length is hyperbolic. This has important implications because, as the front cavity changes its length and thus its resonance mode, it induces larger absolute frequency changes for higher formants and smaller changes for lower formants. Computing the logarithm of the frequency makes these relations linear. Thus, the logarithmic frequency scale is more appropriate for extracting the information coded in changes of the cavity shape and, subsequently, if one accepts the hypothesis pursued in [3] and [5], then the front cavity shape is the prime carrier of the speaker-independent phonetic information in speech, for extracting the message.

Even though perceptual experiments with nonspeech sounds suggest critical-band spectral resolution, there are indications that when it comes to speech perception, the processing of broader spectral spans than the critical band may be taking place. Newton heard vowel-sound sequences from the extreme back-rounded /*u*/ to the extreme front /*i*/ in gradually filled tall glass of beer [32]. Consistent with Newton, Helmholtz [33] discovered a single resonance frequency in back vowels and a combination of two resonances in front vowels. Histograms of frequencies of pure tones evoking different vowels in a number of listeners [34] also indicated such broad spectral features. Chistovich's perceptual experiments [35] indicate that human speech perception integrates spectral peaks that are closer than about three critical bands.

Pitch is a characteristic of human speech that is perceptually distinctive, and in tonal languages such as Mandarin, it carries phonetic information. Pitch is often used for source separation in systems based on computational auditory scene analysis (CASA, e.g., [36] and [37]), and it was a cue for speech separation in most systems participating in the 2011 PASCAL CHiME Challenge [38]. Pitch is occasionally used as part of the front end in ASR, and the addition of a pitch parameter, along with "toneme" phonetics, can often result in reductions in error rates for tonal languages on the order of 10%–15% [39]. An approximate pitch can be derived from MFCC parameters [40], and hence pitch is represented at least indirectly in the standard ASR front ends. Explicit pitch representation is not often found in ASR systems for western European languages, including English.

III. COMPENSATING FOR VARIATIONS IN THE RATE OF SPEECH PRODUCTION

It has been clear since Potter, Kopp, and Green's spectral displays of speech sounds that phonetic elements of speech with identical phonetic values can occur with varying lengths and variable timings, and that their spectral properties depend on other surrounding speech sounds [41]. This temporal variability presents one of the most fundamental difficulties in recognizing speech.

A. The Problem

The natural variations in timing among speakers, speaking conditions, and social interactions have little effect on the intelligibility of speech for human listeners. ASR's inability to compensate for temporal variation remains a substantial source of error in current systems. One surprising recent finding was the very low error rates of ASR systems observed when decoding Mandarin broadcast news, on the order of 2.5% character error [42]. Upon investigation, it was discovered that the Chinese government trains all newscasters to speak at the same tempo (they actually practice to a metronome) and thus identical sentences spoken by different announcers have very small temporal variability. Hidden Markov model (HMM) training captured this lack of variability, leading to the low error rates. Spontaneously produced speech, on the other hand, is not recognized nearly as well.

B. Engineering Approaches

Early attempts at large-vocabulary ASR [4] assumed initial segmentation of speech into classifiable subword units (phones), and decoding the resulting phonetic stream yielded the message. Attention was focused on finding the correct phonetic string underlying the speech; given the correct phonetic transcription, recovering the words is a relatively simple task. The underlying assumption of these segmentation efforts was that the speech can be represented by a "beads on a string" model, where one speech sound (phoneme) follows another (as in the upper panel of Fig. 6). This would inherently take care of the

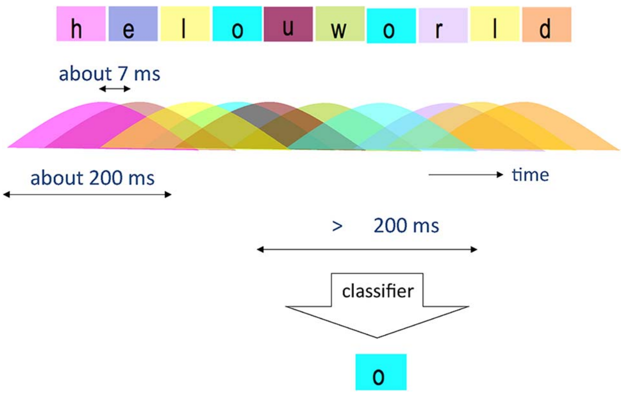


Fig. 6. ***The "beads on a string" model of speech (upper part of the figure) and the "eggs-passed-through-the-wringer" model of speech (middle part of the figure). The bottom part illustrates a suggestion discussed in Section III-B2 where information from a sufficiently long segment of the signal is used as an input to a neural net classifier to derive estimate of a posterior probability of the underlying coarticulated phoneme.***

variable speech production rates; all that would matter would be the order in which the speech sounds occur; the timing between them would be of no consequence for the decoding the message. Thus, phonetic-first analysis finessed the timing differences in the signal, but unfortunately, at that time, the assumed solution turned out to be impossible to implement. Pursuit of these techniques led to the influential paper "Whither speech recognition" by Pierce [43], which led to a substantial decrease in funding for speech recognition research for more than a decade.

After that fiasco, the "beads on a string" model yielded to a more sensible model: "eggs on a stretchy belt after having been passed through a wringer." Not only are the elements smeared, but they are distributed in unequal time intervals (as in the middle part of Fig. 6).

Current ASR solutions manage this problem by dynamically modifying the relative timing of either a model or of the speech to be recognized. Dynamic time warping (DTW) is a method that attempts to find the smallest error path through potential compression and expansion of a template and the input speech simultaneously, and it was independently developed at several research institutions [44]–[46]. This method allowed limited stretching or squeezing of one utterance with respect to a second, and it met with success in small- and medium-vocabulary speech systems.

DTW has been supplanted by hidden Markov modeling, developed in parallel at IDA and at the IBM Research Laboratories (e.g., [2]). In hidden Markov modeling, phones (or related segments) are represented as "states," where timing variability is associated with the dwell time in a state during decoding, and search algorithms automatically found the best matching temporal pattern. Despite substantial success, these models allow unrealistic spectral dynamics in the model matches, and imply exponential probability distributions for the segment lengths which do not match the actual duration distributions found in the data.

1) Dealing With Coarticulation: Modifying the temporal axis by techniques such as DTW or HMM would not be necessary if the speech sounds could be derived from coarticulated speech. More recent techniques based on multilayer perceptron (MLP) artificial neural networks aim at direct estimation of posterior probabilities of speech sounds, thus rejuvenating interests in the old "beads on a string" model. To avoid the pitfalls that were encountered earlier, the coarticulation of speech sounds would have to be addressed. There is some recent evidence that this is possible, as will be discussed below.

2) Estimating Phonemes From Coarticulated Speech: The posterior probabilities of phoneme classes can be estimated by trained MLP artificial neural network [47]. One of the important advantages of MLP-based classifiers is that they do not put too many restrictions on the type of input features used for the classification. Indeed, improvements in the MLP-based recognition of spelled English letters by including long temporal context (up to 500 ms) has been observed by Fanty *et al.* [48]. Such a time interval is consistent with data-driven discriminative RASTA filters derived by the LDA technique. These are Mexican-hat-like finite impulse response (FIR) filters with impulse responses on the order of 200 ms [6], as discussed in the next section. Taken to extreme, the TempoRAl Patterns (TRAPs) approach uses 1001-ms trajectories of spectral energies in critical bands of hearing as inputs to MLP classifiers that estimate at each frequency the posterior probability of the phoneme that is in the center of this long time interval. Such estimates are then fused by another MLP to yield the final phoneme posteriors. Later, the TRAP concept was extended by including spectral slope estimates with each spectral energy measurement [49]. The MRASTA approach [50] projects 1-s-long critical-band spectral trajectories onto a number of time–frequency bases as inputs to the MLP that estimates phoneme posteriors. A Hungarian phoneme recognizer [51] that is currently being used in a number of speech applications uses 310-ms temporal segments of speech data. Pinto *et al.* [52] use a hierarchy of two MLPs where the second MLP's inputs are syllable-length segments of phoneme posteriors estimated by the first MLP.

Relatively long segments of the signal are now often used as inputs to nonlinear classifiers. When a long segment of the signal that contains most of the coarticulation pattern of a given phoneme forms the input to a classifier, the classifier has enough information to classify the phoneme in the center of the pattern correctly, as in Fig. 7. A cartoon illustration of the situation is shown in the bottom panel of Fig. 6.

There has been some success in deriving time signatures of phonemes using matched filters. The task is

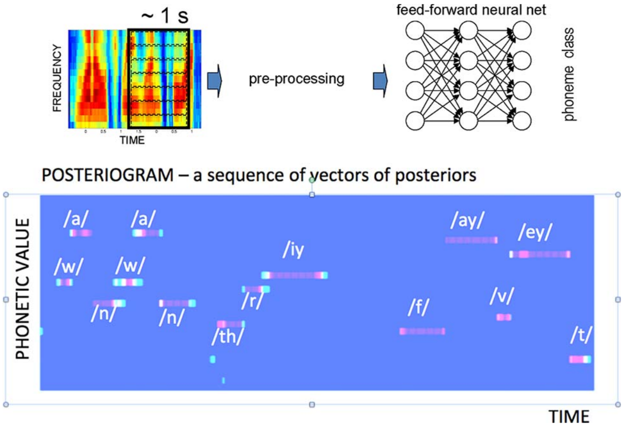


Fig. 7. ***Posterior probabilities of phonemes are indicated in the posteriogram by various colors, warmer colors indicating larger posteriors. Phoneme classes indicated by highest posteriors are indicated in the figure. These posterior probabilities were derived from the telephone-quality utterance "one-one-three- five-eight" by the MRASTA technique [50].***

to filter each phoneme probability trajectory with a matched filter that represents a typical probability trajectory in the vicinity of this phoneme [53]. The local maxima of such filtered trajectories indicate the centers of the phonemes. Such estimates form a sparse pattern in time that has been used successfully in extremely fast keyword searching [54], [55].

The MLP-derived posterior probabilities of phonemes could be scaled by phoneme priors to derive the scaled likelihoods that are appropriate for stochastic search for the best sound sequence in the HMM framework [56]. This technique is often used in estimation of phoneme strings in recognition of phonemes, where it can yield better than 80% phoneme string accuracies on well-articulated speech without the use of any word-level language knowledge (e.g., [57]). Some recent experiments [58] indicate that such accuracy is roughly in line with human performances on similar tasks.

In addition, the posterior probabilities of the decision classes form the smallest feature set for classification [59]. Posterior probabilities of phonemes could be used as an intermediate representation for deriving normally distributed and decorrelated features for current HMM-based LVCSR systems. The tandem approach [60], illustrated in Fig. 8, can be applied to turn posterior estimates into such features, and it is currently used in many state-of-the-art LVCSR systems. Results indicate that better posterior probability estimates yield more efficient features for ASR, independently of the crude temporal models used.

C. Psychophysical and Physiological Correlates

The identity of the basic unit with which humans recognize speech has still not been established, partly because there may be different answers for different situations. It is possible that frequently encountered words or whole phrases are recognized as single units in casual conversation, and that human listeners resort to recognition of subword units only when holistic recognition fails. When recognizing subword units, syllables are sometimes suggested as appropriate elements to be recognized [61], supported by the observation that coarticulation appears to be weaker across syllable boundaries than it is within a syllable. Nevertheless, reaction times when recognizing consonant–vowel syllables by human listeners show that the consonant is always recognized before the vowel [62]. This, in addition to the short time delay associated with speech shadowed by another person [62], provides evidence that is clearly inconsistent with syllabic-level holistic recognition. This validates the proposition by Kozhevnikov and Chistovich [62] that the units of human decision are phoneme-level speech sounds, but decisions are based on information collected from larger

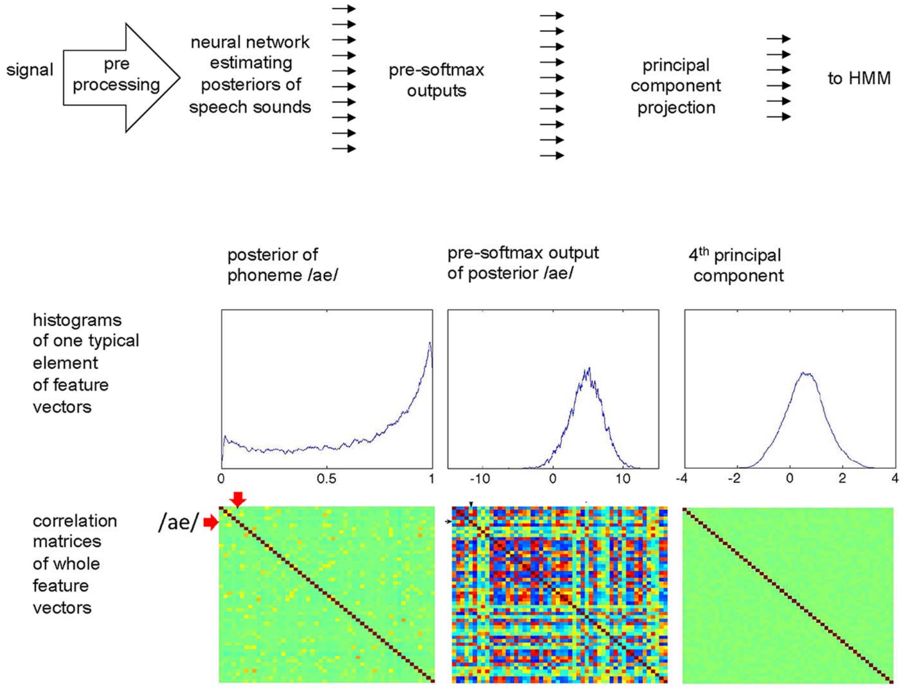

signal
pre processing
neural network estimating posteriors of speech sounds
pre-softmax outputs
principal component projection
to HMM
posterior of phoneme /ae/
pre-softmax output of posterior /ae/
4th principal component
histograms of one typical element of feature vectors
0
0.5
1
-10
0
10
-4
-2
0
2
4
correlation matrices of whole feature vectors
/ae/

Fig. 8. ***The tandem approach: Posterior probabilities of speech sounds, estimated by an artificial neural net, are first transformed by a static nonlinearity so that their distributions are closer to normal. They are then decorrelated by a Karhunen–Loeve transform and reduced in dimensionality to be used as features for a conventional Gaussian-mixture-model-based HMM ASR. The matrices below the feature distributions indicate the correlations among the features at each stage.***

contexts that typically extend over several neighboring phonemes.

The existence of such a temporal buffer for providing evidence about underlying speech sounds is supported by many phenomena in psychoacoustics as well as in the physiology of higher levels of neural processing and motor control (e.g., [63]). One example is the well-known phenomenon of temporal forward masking which suggests that components outside the critical time interval of about 200 ms do not contribute to detection of components inside this critical interval, implying that phones interact within 200 ms. The temporal component of a typical spectrotemporal receptive field in the auditory cortex often spans time intervals that would be consistent with a syllable-length temporal buffer. This does not, however, imply that human perception necessarily recognizes these relatively long speech segments (syllables) [61]. It merely implies that, due to coarticulation, these segments carry the information about elements (speech sounds) within them [62], [64]. Temporal interference using feedback [65] reinforces this finding. Many talkers, when listening to feedback consisting of their own speech, become unable to talk as the feedback delay is increased to about 200 ms. In addition, when reversing short segments of speech, there is no intelligibility loss when these segments are shorter than about 50 ms. As the segments get longer, the intelligibility of speech gradually decreases and the speech becomes completely unintelligible for the reversed segments longer than 200 ms [66].

IV. COMPENSATING FOR THE EFFECTS OF LINEAR DISTORTION

A. The Problem of Linear Distortion

Speech contains components over a broad range of frequencies, spanning many octaves. Many conditions including electrical equipment, filters, and room reverberation cause some of these frequencies to be amplified while others are decreased in amplitude. Such linear distortions are typically introduced through the recording or communication channels. Such distortions are typically perceived, but unless they are very severe they do not have significant effect on the accuracy of human speech recognition. However, when not anticipated and not observed in training data, or compensated for by the ASR system, linear channel distortion can dramatically degrade performance of machine ASR.

B. Engineering Solutions

It is quite straightforward to show that purely linear and steady distortions can be handled well by normalizing the magnitudes of components at various frequencies [67]. One of the early demonstrations of the (reversible) effect of linear distortion was the work of Stockham *et al.* [68]. He adjusted the long-term spectrum of the Caruso recordings to match that of the modern tenors, thus cancelling many of the effects of the fixed frequency characteristics of a recording horn used in early recordings. The first experimenter who approached the problem of linear distortion in ASR was (to our knowledge) Itakura [69], who divided the LPC-derived spectrum in each frame of an utterance by a spectrum of a two-pole autoregressive model derived from the whole utterance, thus subtracting an estimate of the long-term logarithmic spectrum of the utterance from the logarithmic spectrum at each frame. This process is functionally equivalent to cepstral mean subtraction, used almost everywhere today.

Cepstral mean subtraction of speech utterances removes fixed biases and apparently does no harm to speech information. However, the time span over which the mean is computed changes the character of this simple processing. For purely static distortions, the mean can be removed from the whole recording. When each utterance might be corrupted differently, utterance-level mean removal is often performed. When the length of short utterances changes dramatically, such utterance-level mean removal can introduce additional spectral distortions.

It is possible to approximate this utterance-level solution with calculations from a moving window. The most extreme is to compute temporal differences between neighboring short-time speech spectra, which effectively removes the mean of two neighboring frames. This solution was applied in [70] and later evolved into widely used and successful dynamic features [71] that describe local dynamics of logarithmic spectrum. Computation of dynamic features typically removes means from about 50–100-ms spans. Local adaptation to recent events in the band-limited spectrum yielded robust speech recognition results in Cohen's auditory model produced at IBM in 1985 [72], and was also important in the models of Seneff [73] and Ghitza [74]. These models showed promise in the 1980s, but were later superseded by processes which produced similar effect but at greatly reduced computational load [75].

RelAtive SpecTrAl (RASTA) processing introduces bandpass filtering of temporal trajectories of logarithmic critical-band spectral energies to emphasize typical rates of spectral changes in speech. The filter was designed experimentally to optimize recognition accuracy on telephone speech. The impulse response of the RASTA filter [76] implies mean removal from exponentially weighted past spectral values with a time constant on the order of 200 ms. This time span is longer than the time span implied by dynamic features but shorter than the length of a typical speech utterance.

RASTA processing can be very effective in handling linear distortions. Fig. 9 shows the original utterance and the utterance filtered in such a way that the vowel in the utterance has, in effect, a flat spectrum. While spectrograms of both utterances differ dramatically, the spectrograms derived by the RASTA–PLP technique are almost identical.

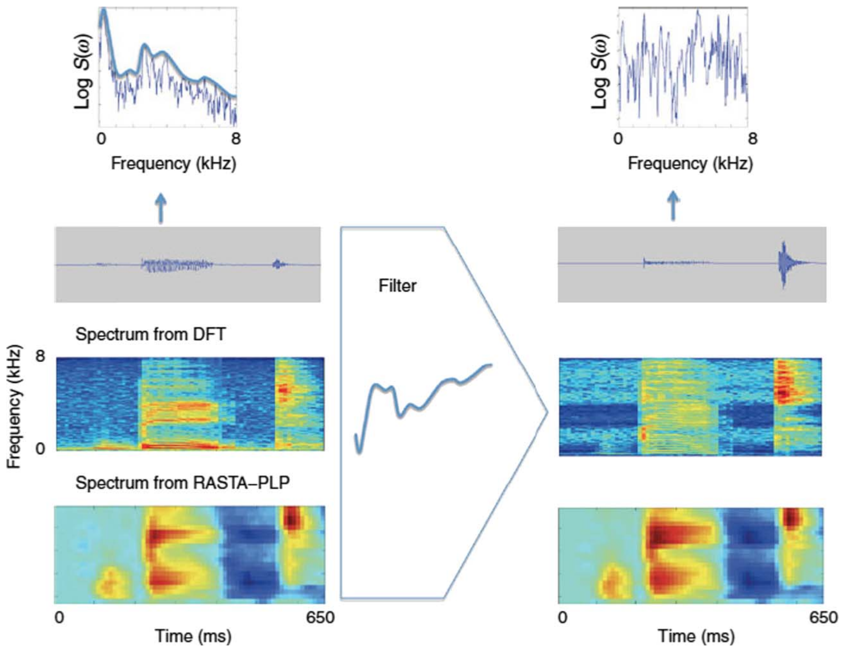


Fig. 9. *The left panel of the figure shows the time domain signal of the utterance "beet" (/b/ /ee/ /t/) together with its spectrogram computed by the conventional DFT analysis (as in the left middle part of the figure) and by the RASTA–PLP technique (left bottom part of the figure). Above the speech waveform, a single spectral slice from the spectrogram, extracted at the instant indicated by the arrow (spectrum of the vowel /ee/), is shown, together with its spectral envelope. The right part of the figure shows the speech waveform, the conventional spectrogram, the RASTA–PLP-derived spectrogram, and the spectral slice from the /ee/ vowel part after the speech waveform was filtered by the filter that has a frequency response that is the inverse of the spectral envelope of the vowel /ee/. The filtering flattens the spectral envelope of the vowel /ee/ but has only a negligible effect on the RASTA–PLP representation of speech.*

Data-driven discriminative RASTA filters, shown in Fig. 10, derived by the LDA technique [77], are Mexican-hat-like FIR filters with an impulse response with an effective length on the order of 200 ms (about one syllable) and have frequency responses that are consistent with the original RASTA filter and its temporal derivatives. This analysis was repeated on a much larger and more realistic machine-labeled database of continuous telephone speech, which confirmed the earlier observations [6]. An analysis of RASTA processing suggests that the most important processing step that alleviates both the effects of linear distortion and coarticulation is the subtraction of a running average of the logarithmic spectrum of neighboring speech segments within about 200 ms, from the spectrum of the current phoneme.

C. Relation to Human Auditory Perception

It is well accepted that perception is more sensitive to changes in the stimulus than it is to its steady components. Starting at the auditory nerve where the firing rate is largest at the onset of the stimulus, firing rates throughout the auditory system indicate greater sensitivity to changing stimuli than to steady state ones.

In a convincing demonstration that the channel has only minor effect on conveying the linguistic information in speech, Husein Yilmaz (personal communication, 1972) observed that vowels with a flat spectrum obtained by filtering the whole speech utterance through a filter that approximated the inverse of the spectral envelope of that particular vowel (shown in the upper part of Fig. 9) are clearly heard with their original vowel quality. It seems that, while the human listener senses modifications in the long-term spectrum of speech, these modifications do not affect their ability to correctly identify phonetic values of underlying speech sounds. Later, Watkins and Makin [78] demonstrated more formally that fixed linear distortions do not have significant effect on human judgments of phonetic quality of vowels. They designed a set of filters that could change a phonetic value of a particular vowel. The perceived vowel value changed when the filter was applied only to this vowel but not when it was applied to a whole utterance that contained the vowel.

Time constants that seem to be effective for the compensation of linear distortions in human hearing may suggest that this phenomenon is related to temporal forward masking. The nonlinear effect of temporal forward masking lasts about 200 ms, independent of the level of the masker [79]. The consistency of RASTA processing with forward temporal masking of human hearing has been observed and is discussed in [64].

More on this topic as well as on its relevance to modulation spectra of speech may be found in [80].

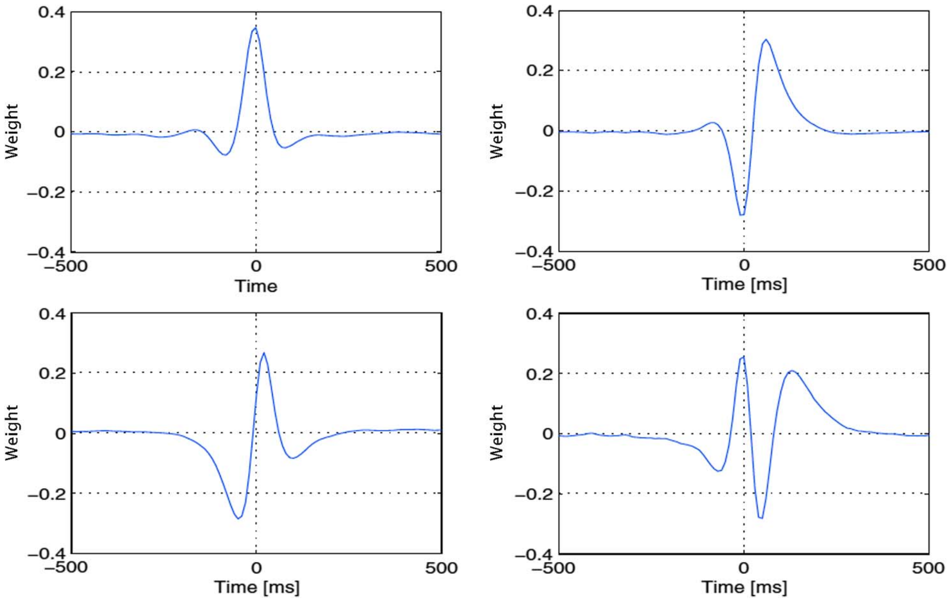


Fig. 10. ***Impulse responses of finite impulse response filters, derived as the first four linear discriminants by LDA technique. These discriminants suggest that the most effective temporal strategy is to average components within a phoneme and subtract components from the nearest neighboring phonemes. Frequency responses of the implied filters are not shown here but they are all bandpass, mainly attenuating modulation spectrum components below 1 Hz and above 10–15 Hz. (From [6], used with permission.)***

V. COMPENSATING FOR THE EFFECTS OF ADDITIVE NOISE

Additive noise can degrade the accuracy of ASR systems in almost every practical application environment, regardless of whether the source is acoustical, electrical, or electromagnetic in nature. Because the general topic of compensation for additive noise has received a great deal of attention over the past several decades, our treatment of the subject in this paper will be somewhat superficial, with an emphasis on identifying the major approaches to ameliorating the effects of the noise and relating these approaches to physiological processing. Readers seeking a more detailed discussion of these issues are encouraged to consult one of the many relevant reviews including [81]–[83], and especially the recent treatise edited by Virtanen *et al.* [84], among many other sources.

A. The Problem of Additive Noise

While the presence of noise is almost ubiquitous, the best approach to compensation will depend on the nature of the interference. It is relatively straightforward to compensate for additive quasi-stationary noise such as white noise, wind noise in a car, ambient air flow from a ventilation system, speech babble in a crowded cafeteria, or many types of continuous machinery. More difficult compensation challenges are posed by transient interference-like slamming doors, ringing telephones, percussive machinery (such as hammers and punches), and gunshots. Background speech and background music are doubly challenging because of their highly transient nature, and because many speech and music sources are easily confused with the desired speech signal. In addition to the acoustic noise sources, nonlinear channel distortion in the presence of noise creates a mixing between the wanted and unwanted signals that is difficult to disentangle. While the distortion-producing carbon button microphones in the traditional telephone network are now rarely encountered, nonlinear distortion remains commonplace in lossy coding for cellphones and voice transmission over the Internet, as well as in the modulation and demodulation processes in many practical point-to-point communication systems used by the military and in industry.

B. Engineering Approaches to Noise Compensation

1) Statistically-Based Approaches: The first viable approach to noise compensation for speech enhancement and ASR was the spectral subtraction method proposed by Boll [85]. Spectral subtraction is accomplished by estimating the magnitude spectrum of noise in the absence of speech, subtracting the noise estimate from the degraded speech on a frame-by-frame basis, and then reconstructing the signal by combining the noise-subtracted magnitude with the original phase of the degraded signal. Short-term differences in noise power can cause negative magnitudes to be computed for some spectra (a physical impossibility). Many of the dozens (if not hundreds) of extensions to this seminal approach that appeared over the ensuing years dealt with alternate methods to avoid the problems of oversubtraction of the noise estimate and its consequences.

Acero and Stern [86] noted that the effects of additive noise and linear filtering combine in a nonlinear fashion. By characterizing degraded speech as clean speech that is passed through an unknown linear filter and corrupted by unknown stationary noise, they developed several useful compensation algorithms such as codeword-dependent cepstral normalization (CDCN) which estimates the parameters that characterize the noise and the filter. Using the E–M algorithm, they obtained compensated cepstra using ML estimation. This approach is a generalization and unification of Boll's spectral subtraction combined with homomorphic deconvolution as proposed by [68].

The vector-Taylor series (VTS) approach [87] is a further generalization of this approach, using the same model of degradation but characterizing its effects in the cepstral domain, statistically using Gaussian mixtures to represent the cepstra of speech and a single Gaussian for the effects of noise. A very large number of variations of this approach have been proposed over the years (e.g., [88] and [89]).

2) Missing-Feature and Multistream Approaches: Algorithms such as VTS work reasonably well in the presence of quasi-stationary noise but are ineffective in the presence of transient interference such as background music (e.g., [90]). The use of "missing-feature" techniques (e.g., [91] and [92]) is useful for speech recognition in the presence of the type of transient interference that is not handled well by algorithms like VTS, as reviewed in [93]. Briefly, in missing-feature approaches, one attempts to determine which portions of speech are unreliable by mapping the signal to a spectrogram-like time–frequency display, and retaining only those components of the spectrogram that are deemed to be reliable. These approaches can be extremely effective when the "missing" spectrotemporal cells are correctly identified, but identifying the degraded elements is frequently difficult to accomplish without substantial *a priori* knowledge about the nature of the degradation, and success is critically dependent on the extent to which the missing cells can be identified correctly.

The multiband approach has similar goals and was developed contemporaneously [94]–[96]. In the original multiband method, speech is processed using multiple parallel classifiers in different frequency bands. The outputs of these channels are combined in a fashion that is intended to give greater weight to those channels that provide a more reliable representation of the incoming speech signal. This approach subsequently evolved into the more general multistream processing, in which classification is performed using a variety of complementary parallel techniques and the results from these different classification streams are combined in making the final decision. The different streams may use different projections of the incoming data [97], where some projections could be affected by a particular unexpected corruption less than other. Individual streams may also differ in their prior constraints (e.g., [98] and [99]), where one stream uses full strength of a language model and of global HMM search, while the other one classifies speech sounds only from local acoustics. Information from individual streams can be fused to yield the final output or compared to find sources of corruption. Efficient engineering techniques for accomplishing these goals remain an open problem [100]–[102], with more detail provided in [103].

3) Physiologically Motivated Auditory Models: Over the years a number of researchers have proposed signal processing schemes based on computational models of the auditory system, and three of the original models proposed in the 1980s [73], [74], [104] have been widely circulated in MATLAB implementations by Slaney [105]. Additional relevant computational auditory models include the work of Cohen [72], Tchorz and Kollmeier [106], Chi *et al.* [107], D.-S. Kim *et al.* [108], and C. Kim *et al.* [109]. All of these processing schemes (and others) typically produce recognition accuracy that is comparable to that which is obtained by MFCC or PLP processing in clean speech, and better recognition accuracy in the presence of additive noise.

As an example, Kim and Stern [110], [111] recently described a processing approach called power-normalized cepstral coefficients (PNCCs). PNCC processing represents an attempt to develop a pragmatic computationally efficient feature extraction procedure that is motivated by auditory processing in the spirit of PLP and (more abstractly) MFCC processing, but that also has built-in robustness with respect to additive noise. In addition to elements that are common to most of the approaches listed above, PNCC processing includes "medium-time" nonlinear processing that suppresses the effects of additive noise and room reverberation, along with a power-law nonlinearity. The "medium-time processing" in effect performs a nonlinear high-pass filtering of the cepstral coefficients that both suppresses noise by filtering in the modulation spectrum domain and reduces differences between the training and testing environments by the same processing. In addition, the use of the power-law nonlinearity renders PNCC insensitive to changes in input amplitude. PNCC has been shown to be as effective in reducing the impact of additive noise as VTS processing and the ETSI advanced front end (AFE), at a computational cost that is similar to that of MFCC and PLP processing [110], [111]. The success of relatively simple physiologically inspired feature extraction procedures such as PNCC suggests that the potential benefits from the use of auditory processing are widespread, and that we will continue to improve robustness in speech technologies as we deepen our understanding of the auditory processing of natural speech.

C. Physiological and Psychophysical Correlates

There are many physiological mechanisms that assist in suppression of noise and unwanted background information. Auditory processing, like many sensory systems, emphases *change* in the physical input in several dimensions. For example, the auditory system enhances the temporal onsets and offsets of the envelopes of signals in each frequency band (e.g., [112]), and emphasizes differences in frequency through lateral suppression (e.g., [113]). These processes result in suppression of many forms of slowly changing background noise, as has been noted by Wang and Shamma [114]. Modulation spectral analysis, as described in more detail in [80], also aids in the suppression of background noise, as the envelope modulation frequencies of speech in each analysis band are typically at different modulation frequencies than those of typical background noise. Units that are sensitive to envelope modulation frequencies, and which consequently may be used to implement modulation spectral analysis have been described by Langner and Schreiner [115] and others.

Another physiological mechanism which contributes to noise suppression is the nonlinear representation of signal intensity in each frequency band. The apparent response rate of auditory-nerve fibers is an S-shaped function of signal intensity, with no response to the incoming signal at low intensities (below threshold for a given fiber), a log-linear response in an intermediate range, and finally saturation at high intensities where there is relatively little change in response. As has been noted by Wang and Shamma [114] and Chiu and Stern [116], noise suppression would be accomplished if responses to speech were captured in the log-linear range, while the response to the background stimulation at a lower level were below the threshold of active response. More generally the auditory system suppresses weak stimuli when excited by stronger ones. (This property is frequently used in communications in such nonlinear processes as frequency modulation.)

Finally, the auditory system records not only the short-time average energy of the frequency components of the signal as a function of time, but also the temporal patterns of the signal as well. In fact, Licklider and Pollack [117] demonstrated that speech remained quite intelligible even if a speech signal is infinitely clipped, so that only zero crossings of the signal remain intact. It is well documented that the peripheral auditory systems preserve the cycle-by-cycle temporal variations of a signal at low frequencies, along with the cycle-by-cycle temporal variations of the low-frequency envelopes of higher frequency components. For example, Young and Sachs [118] have noted that the average localized synchrony rate (ALSR, a measure of the extent to which the temporal patterns of neural response are synchronized to the timing of the incoming signal) produces a response to steady-state vowels that is more robust (at least with respect to signal intensity) than the mean rate of response, which is more of a measure of short-time energy as a function of frequency. A number of speech researchers have incorporated the temporal response (as well as mean rate) into computational models of the auditory system, beginning with the DOMIN model of Blomberg *et al.* [119], continuing with the classical models of Seneff [73] and Ghitza [74], and followed by several more recent implementations (e.g., [108], [109], and [120]).

VI. EXPLOITING THE SPATIAL CHARACTERISTICS OF SIGNALS

All of the attributes of auditory processing cited above are essentially single channel in nature. It is well known that human listeners compare information from the two ears to localize sources in space and separate sound sources that are arriving from different directions, a process generally known as binaural hearing. The mechanisms of binaural hearing can be emulated in computational speech processing systems to improve recognition accuracy, signal separation, and speech enhancement.

A. Exploiting Spatial Information for Speech Recognition

Let us begin with a discussion of the type of information that is the basis for source separation based on source location in computational auditory scene analysis (CASA). As originally suggested by Lord Rayleigh [122], the human auditory system is able to identify the direction of arrival of incoming sound sources by estimating the difference in arrival time [interaural time delay (ITD)] at low frequencies and interaural intensity differences (IIDs) at high frequencies. [In some cases, ITDs are calculated as interaural phased differences (IPDs).] More recent studies indicate that we are able to make use of the ITDs of low-frequency envelopes of high-frequency signal components as well. Elevation angles and front–back ambiguities are most likely disambiguated by a combination of cues based on the spectral coloring of the outer ear and head-motion information. A number of studies have demonstrated that competing speech sources are more easily individually segregated and understood by humans when they are more spatially separated (e.g., [123] and [124]).

There are several ways in which signal processing approaches based on exploitation of ITDs and/or IIDs can be useful in automated speech processing. First and most obviously, this information can be used as cues to enable the separation of simultaneously presented sources in a complex acoustical field. Second, the use of two ears (and correspondingly, two microphones) has been demonstrated to improve the intelligibility of speech signals in reverberant environments (e.g., [125]). Finally, speech (and other) signals are much more easily detected in the presence of maskers when the interaural differences for the target signal are different from those of the masker. The first two of these phenomena have been demonstrated to be useful in automatic speech recognition, and are probably useful as well for techniques such as speaker identification and verification, language identification, keyword spotting, etc. The latter phenomenon is potentially useful for speech activity detection. We described some of the approaches that have been followed below.

B. Engineering Approaches to Binaural Enhancement

1) Separation According to Estimated ITD and IID: The first computational system for binaural processing of simultaneously presented speech sounds was developed by Lyon [126], based on the "correlogram," which describes the short-time cross correlation of paired outputs of his model for auditory nerve activity. The most common application of binaural processing is through the use of systems that provide selective reconstruction of spatialized signals that have been degraded by noise and/or reverberation by selecting those spectrotemporal components after short-time Fourier analysis that are believed to be dominated by the desired sound source. Modern systems that adopt this approach typically: 1) compute the short-time Fourier transforms of the incoming signal to the two microphones; 2) estimate the ITD (or IPD) of each spectrotemporal component by comparing the phases of the STFTs; 3) select the subset of spectrotemporal components that have ITDs that appear to correspond to the ITD corresponding to the direction of the desired speaker; 4) (optionally) fill in the missing spectrotemporal components using missing-feature techniques; and finally 5) either develop cepstral coefficients directly from the subset of spectrotemporal components that remain or resythesize an enhanced target waveform using the overlap–add method or a similar technique (e.g., [121] and [127]–[132]). As an example, the upper panel of Fig. 11 depicts the spectrogram of two concurrent speech signals, one coming from an azimuth on the perpendicular bisector of two microphones separated by 4 cm, and the other from an azimuth 45° to one side. The lower two panels depict the spectrograms of the two speech signals after separation using the phase difference channel weighting (PDCW) algorithm, which separates signals according to the ITDs of their spectrotemporal components. Good recognition accuracy is obtained using this procedure [121].

Some systems also explicit extract IIDs and use that information to further enhance those time–frequency components that exhibit a plausible combinations of ITD and IID that are associated with the target azimuth (e.g., [127], [129], and [130]).

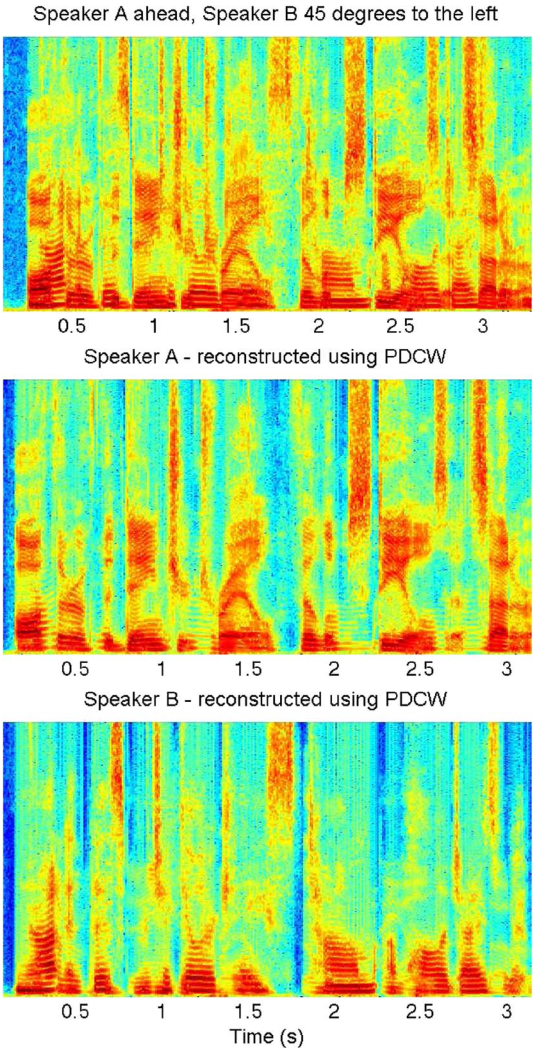


Fig. 11. *Upper panel: spectrograms of two concurrent speech signals arriving at azimuths separated by 45°. Central and lower panels: spectrograms of the speech signals following separation using the PDCW algorithm [121], which is based on the difference of arrival time of the spectrotemporal components to the two microphones.*

2) Processing for Robustness to Reverberation: Some CASA systems intended for use in reverberant environments have also incorporated into their processing the precedence effect, which is the observation that localization is dominated by the first arriving components of a complex sound [133]. The precedence effect is clearly helpful in enabling the perceived location of a source in a reverberant environment to remain constant, as it is dominated by the characteristics of the 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.

Several groups have incorporated processing based on the precedence effect, typically through the use of enhancement of the leading edge of envelopes of the outputs of the bandpass filters that are part of all feature extraction systems for speech. This emphasis can be implemented both at the monaural level (e.g., [134] and [135]) and at the binaural level (e.g., [129], [136], and [137]), and it has been shown to be particularly effective in reverberant environments in both cases [129], [135].

C. Physiological and Psychophysical Correlates

There is extensive neurophysiological evidence in the brainstem and the cortex that supports the type of computational processing that is developed in this section. For example, 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., [138] and [139]). 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 described in recent reviews (e.g., [140] and [141]). Units have also been described that appear to record the IIDs of a stimulus (e.g., [138]).

VII. CONCLUSION

In summary, we have seen that representations and modifications of the speech signal used in modern ASR often mimic the psychophysical or physiological processes found in mammalian auditory systems. Sometimes this happened when researchers explicitly modeled some aspects of the system, and sometimes it occurred when optimizing processes were used to tune the performance of a classifier or of the entire system. In every case, however, the information-reduced representation used in the speech recognizer could be cast as modeling one or more characteristics of the auditory system.

Given this convergence, it is still true that speech recognition remains an underperforming discipline. Modern systems fail in noise and reverberation, except under very carefully controlled circumstances. They do not gracefully account for accents, new words, or distorted acoustics, although people have little difficulty in these situations. What we have shown, however, is that the best recognizers of today mimic nonlinear frequency representations, power-law scaling, longer temporal buffers, signal-adaptive responses, focused acoustic attention, and noise adaptation when the noise is particularly well behaved.

Other aspects of the perception of speech, however, are not accounted for in current systems. We do not generally take advantage of pitch, although the auditory system appears to be pitch synchronous up to 2 kHz or so. We do not generally account for syllabic structure, although even unschooled speakers of native languages can all count syllables and speak in rhythm. We separate different sources occurring simultaneously with difficulty, although people are very good at this process. Of course, we also do not have even reasonable models of language, although that is a discussion for another day.

It is our hope that this review has highlighted those physiological and psychophysical processes which we believe to be important, leaving better models and more insight to those working in the field. We do not mean to exclude yet other processes, although we have started here to sort the more likely speech analysis characteristics from the less likely, in terms of the performance of speech systems. We look forward to further insights as systems improve, and as our knowledge of biological processes expand.

Of course, there is more to speech recognition than the features and their modifications, but we feel that getting those right might allow us to avoid the constraints of "garbage-in, garbage-out" transformations. Good luck to us all. ■

REFERENCES

[1] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception,* 2nd ed. New York, NY, USA: Springer-Verlag, 1983.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, no. 4, pp. 532–536, Apr. 1976.

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

[4] G. Galt*Galt: Study of speech and hearing at Bell Telephone Laboratories: Correspondence files (1917–1933) and other internal reports 1917–1933*, CD-ROM published by the ASA compiled by C. M. Rankovic and J. B. Allen, Book 23651, p. 2.

[5] H. Hermansky and D. Broad, "The effective second formant F2' and the vocal tract front cavity," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1989, pp. 480–483.

[6] F. Valente and H. Hermansky, "Discriminant linear processing of time-frequency plane," in *Proc. Interspeech 2006*, pp. 349–352.

[7] J. S. Bridle and M. D. Brown, "An experimental automatic word-recognition system," Joint Speech Res. Unit (JSRU), Ruislip, U.K., Rep. 1003, 1974.

[8] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed. New York, NY, USA: Academic, 1976, pp. 374–388.

[9] S. Itahashi and S. Yokoyama, "Automatic formant extraction utilizing mel scale and equal loudness contour," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1976, pp. 310–313.

[10] J. Makhoul and L. Cosell, "LPCW: An LPC vocoder with linear predictive spectral warping," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1976, vol. 1, pp. 466–469.

[11] H. W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Amer.*, vol. 68, no. 4, pp. 1071–1076, 1980.

[12] H. Hermansky, B. A. Hanson, and H. Wakita, "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain," *Speech Commun.*, vol. 4, no. 1–3, pp. 181–187, 1985.

[13] H. Hermansky, H. Fujisaki, and Y. Sato, "Analysis and synthesis of speech based on spectral transform linear predictive method," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1983, vol. 8, pp. 777–780.

[14] A. Biem and S. Katagiri, "Filter bank design based on discriminative feature extraction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1994, pp. 485–488.

[15] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-warping in speech," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 414–417.

[16] T. Kamm, H. Hermansky, and A. G. Andreou, "Learning the mel-scale and optimal VTN mapping," Johns Hopkins Ctr. Lang. Speech Process., Baltimore, MD, USA, Tech. Rep., 1997.

[17] H. Hermansky and N. Malayath, "Spectral basis functions from discriminant analysis," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 1379–1382.

[18] K. Paliwal, B. Shannon, J. Lyons, and K. Wojcicki, "Speech-signal-based frequency warping," *IEEE Signal Process. Lett.*, vol. 16, no. 4, pp. 319–322, Apr. 2009.

[19] A. Andreou, J. Cohen, and T. Kamm, "An experiment in systematic speaker variability," in *Proc. DoD Workshop: Front. Speech Process. II*, 1994.

[20] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Amer.*, vol. 24, no. 6, pp. 637–642, 1952.

[21] H. Hermansky, K. Tsuga, S. Makino, and H. Wakita, "Perceptually based processing

in automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1986, vol. 11, pp. 1971–1974.

[22] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptually based linear predictive analysis of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1985, pp. 509–512.

[23] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[24] H. Hermansky and A. L. Cox, "Perceptual linear predictive (PLP) analysis-resynthesis technique," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1991, pp. 331–334.

[25] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification," *Digit. Signal Process.*, vol. 10, pp. 55–74, 2000.

[26] G. von Békésy, *Experiments in Hearing*. New York, NY, USA: McGraw Hill, 1960.

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

[28] J. L. Flanagan, "A difference limen for vowel formant frequency," *J. Acoust. Soc. Amer.*, vol. 27, no. 3, pp. 613–617, 1955.

[29] U. G. Goldstein, "An articulatory model for the vocal tracts of growing children," Sc.D. thesis, Dept. Electr. Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 1980.

[30] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA, USA: MIT Press, 1998.

[31] G. M. Kuhn, "On the front cavity resonance and its possible role in speech perception," *J. Acoust. Soc. Amer.*, vol. 58, no. 2, pp. 428–433, 1975.

[32] P. Ladefoged, *Three Areas of Experimental Phonetics: Stress and Respiratory Activity, the Nature of Vowel Quality, Units in the Perception and Production of Speech*. Oxford, U.K.: Oxford Univ. Press, 1967.

[33] H. Helmholtz, *On the Sensations of Tone*. New York, NY, USA: Dover, 1954.

[34] G. Fant, *Acoustic Theory of Speech Production*. Berlin, Germany: Moulton De Gruyter, 1970.

[35] L. Chistovich and V. V. Lublinskaya, "The 'center of gravity' effect in vowelspectra and criticaldistance between the formants: Psychoacoustical study of the perception of vowel-like stimuli," *Hearing Res.*, vol. 1, no. 3, pp. 185–195, 1979.

[36] A. Coy and J. Barker, "An automatic speech recognition system based on the scene analysis account of auditory perception," *Speech Commun.*, vol. 17, no. 3, pp. 384–401, 2007.

[37] Z. Jin and D. Wang, "Reverberant speech segregation based on multipitch tracking and classification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 8, pp. 2328–2337, Nov. 2011.

[38] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Speech Commun.*, vol. 27, no. 3, pp. 621–633, 2013.

[39] P. Ding and L. He, "Improve the implementation of pitch features for mandarin digit string recognition task," in *Proc. Interspeech 2012*, pp. 914–917.

[40] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 51, no. 1, pp. 24–33, Jan. 2007.

[41] R. K. Potter, G. A. Kopp, and H. C. G. Kopp, *Visible Speech*. New York, NY, USA: Dover, 1966.

[42] S. M. Chu, H.-K. Kuo, L. Mangu, Y. Yi, Y. Qin, Q. Shi, S. L. Zhang, and H. Aronowitz, "Recent advances in the IBM GALE Mandarin transcription system," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 4329–4332.

[43] J. R. Pierce, "Whither speech recognition?" *J. Acoust. Soc. Amer.*, vol. 46, no. 4B, pp. 1049–1051, 1969.

[44] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE. Trans. Acoust. Speech Signal Process.*, vol. ASSP-26, no. 1, pp. 43–49, Jan. 1978.

[45] T. Vintsyuk, "Speech discrimination by dynamic programming," *Kibernetika*, vol. 4, no. 2, pp. 81–88, 1968.

[46] M. R. Schroeder, "Similarity measure for automatic speech and speaker recognition," *J. Acoust. Soc. Amer.*, vol. 32, no. 2, pp. 375–377, 1967.

[47] H. Bourlard and C. J. Wellekens, "Links between Markov models and multilayer perceptrons," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 12, pp. 1167–1178, Dec. 1990.

[48] M. Fanty, R. A. Cole, and K. Roginski, "English alphabet recognition with telephone speech," *Advances in Neural Information Processing Systems 4*, San Mateo, CA, USA: Morgan Kaufmann, 1992, pp. 199–206.

[49] P. Jain and H. Hermansky, "Beyond a single critical-band in TRAP based ASR," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2003, pp. 437–440.

[50] H. Hermanksy and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proc. Interspeech 2005*, pp. 361–364.

[51] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proc. Interspeech 2005*, Geneva, Switzerland, 1990, pp. 2237–2240.

[52] J. Pinto and H. Hermansky, "Combining evidence from a generative and a discriminative model in phoneme recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 2414–2417.

[53] M. Lehtonen, P. Fousek, and H. Hermansky, "Hierarchical approach for spotting keywords," in *Proc. 2nd Workshop Multimodal Interaction Related Mach. Learn. Algorithms*, Edinburgh, U.K., 2005.

[54] K. Kintzley, A. Jansen, and H. Hermansky, "Event selection from phone posteriorgrams using matched filters," in *Proc. Interspeech 2011*, pp. 1905–1908.

[55] K. Kintzley, A. Jansen, and H. Hermansky, "MAP estimation of whole-word acoustic models with dictionary priors," in *Proc. Interspeech 2012,* Paper Tue.O4a.01.

[56] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer, 1993.

[57] S. G. S. V. S. and H. Hermansky, "Sparse multilayer perceptron for phoneme recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 23–29, Jan. 2012.

[58] W. Shen, J. Olive, and D. Jones, "Two protocols comparing human and machine phonetic recognition performance in conversational speech," in *Proc. Interspeech 2008*, pp. 1630–1633.

[59] K. Fukunaga, *Introduction to Statistical Pattern Classification*. New York, NY, USA: Academic, 1990.

[60] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2000, pp. 1635–1638.

[61] S. Greenberg, "Speaking in shorthand—A syllable-centric perspective for understanding," *Speech Commun.*, vol. 29, no. 2–4, pp. 159–176, 1999.

[62] V. A. Kozhevnikov and L. A. Chistovich *Speech: Articulation and perception*," Trans. U.S. Dept. Commerce, Clearing House Fed. Sci. Tech. Inf., 1967.

[63] N. Cowan, "On short and long auditory stores," *Psychol. Bull.*, vol. 96, no. 2, pp. 341–370, 1984.

[64] H. Hermansky, "Should recognizers have ears?" *Speech Commun.*, vol. 25, pp. 3–27, 1998.

[65] A. J. Yates, "Delayed auditory feedback," *Psychol. Bull.*, vol. 60, no. 3, pp. 213–232, 1963.

[66] K. Saberi and D. R. Perrott, "Cognitive restoration of reversed speech," *Nature*, vol. 398, p. 760, 1999.

[67] C. Avendano and H. Hermansky, "On the effects of short-term spectrum smoothing in channel normalization," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 4, pp. 372–374, Jul. 1997.

[68] J. T. G. Stockham, T. M. Cannon, and R. B. Ingrebretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, no. 4, pp. 678–692, Apr. 1975.

[69] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-23, no. 1, pp. 67–72, Jan. 1975.

[70] M. Mlouka and J.-S. Lienard, "Word recognition based on either stationary items or on transitions," *Speech Communication,* G. Fant Ed. Stockholm, Sweden: Almqvist & Wiksell, vol. 3, pp. 257–263, 1975.

[71] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

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

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

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

[75] R. M. Stern and N. Morgan, "Features based on auditory physiology and perception," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, B. Raj, and R. Singh, Eds. New York, NY, USA: Wiley, 2012.

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

[77] S. van Vuuren and H. Hermansky, "Data-driven design of RASTA-like filters," in *Proc. Eurospeech 1997*, pp. 409–412.

[78] A. J. Watkins and S. J. Makin, "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," *J. Acoust. Soc. Amer.*, vol. 99, pp. 3749–3757, 1996.

[79] W. Jestead, S. P. Bacon, and J. R. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *J. Acoust. Soc. Amer.*, vol. 71, pp. 950–962, 1982.

[80] H. Hermansky, "Speech recognition from spectral dynamics," *Sādhanā*, vol. 36, no. 5, pp. 729–744, 2011.

[81] B.-H. Juang, "Speech recognition in adverse environments," *Comput. Speech Lang.*, vol. 5, no. 3, pp. 275–294, 1991.

[82] R. Singh, R. M. Stern, and B. Raj, "Signal and feature compensation methods for robust speech recognition," in *Noise Reduction in Speech Applications*, G. M. Davis, Ed. Boca Raton, FL, USA: CRC Press, 2002.

[83] R. Singh, B. Raj, and R. M. Stern, "Model compensation and matched condition methods for robust speech recognition," in *Noise Reduction in Speech Applications*, G. M. Davis, Ed. Boca Raton, FL, USA: CRC Press, 2002.

[84] T. Virtanen, B. Raj, and R. Singh, *Noise-Robust Techniques for Automatic Speech Recognition*. New York, NY, USA: Wiley, 2012.

[85] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

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

[87] 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 Signal Process.*, May 1996, pp. 733–736.

[88] J. R. Hershey, S. J. Rennie, and J. LeRoux, "Factorial models for robust speech recognition," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. New York, NY, USA: Wiley, 2012.

[89] M. L. Seltzer, "Acoustic model training for robust speech recognition," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. New York, NY, USA: Wiley, 2012.

[90] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1997, vol. 2, pp. 851–854.

[91] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, 2001.

[92] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 275–296, 2004.

[93] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 101–115, Sep. 2005.

[94] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards subband-based speech recognition," in *Proc. Eur. Signal Process. Conf.*, 1996, pp. 1579–1582.

[95] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on partially corrupted speech," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 462–465.

[96] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and re-combination of partial frequency bands," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 426–429.

[97] S. Tibrewala and H. Hermansky, "Multi-stream approach in acoustic modeling," in *Proc. DARPA Large Vocabulary Continuous Speech Recognit. Hub 5 Workshop*, 1997.

[98] H. Ketabdar, M. Hannemann, and H. Hermansky, "Detection of out-of-vocabulary words in posterior based ASR," in *Proc. Interspeech 2007*, pp. 1757–1760.

[99] H. Ketabdar and H. Bourlard, "Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 4065–4068.

[100] F. Valente and H. Hermansky, "Data-driven extraction of spectral-dynamics based posteriors," in *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, J. Olive, C. Christianson, and J. McCary, Eds. New York, NY, USA: Springer-Verlag, 2011.

[101] N. Mesgarani, S. Thomas, and H. Hermansky, "Adaptive stream fusion in multistream recognition of speech," in *Proc. Interspeech 2011*, pp. 2329–2332.

[102] S. Badiezadegan and R. Rose, "A performance monitoring approach to fusing enhanced spectrogram channels in robust speech recognition," in *Proc. Interspeech 2011*, pp. 477–480.

[103] H. Hermansky, "Multistream recognition of speech: Dealing with unknown unknowns," *Proc. IEEE*, vol. 101, no. 5, pp. 1076–1088, May 2013.

[104] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Paris, France, May 1982, pp. 1282–1285.

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

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

[107] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Amer.*, vol. 118, pp. 887–906, 2005.

[108] D.-S. Kim, S.-Y. Lee, and R. Kil, "Auditory processing of speech signals for robust speech recognition in real world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55–59, Jan. 1999.

[109] 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.

[110] 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 Int. Conf. Acoust. Speech Signal Process.*, Mar. 2010, pp. 4574–4577.

[111] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, 2013.

[112] J. O. Pickles, "The neurophysiological basis of frequency selectivity," in *Frequency Selectivity in Hearing*, B. C. J. Moore, Ed. New York, NY, USA: Academic, 1986.

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

[114] K. Wang and S. A. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 3, pp. 421–435, Jul. 1994.

[115] 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.

[116] Y.-H. Chiu and R. M. Stern, "Analysis of physiologically-motivated signal processing for robust speech recognition," in *Proc. Interspeech 2008*, pp. 1000–1003.

[117] J. C. R. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping on the intelligibility of speech," *J. Acoust. Soc. Amer.*, vol. 20, no. 1, pp. 44–51, 1948.

[118] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 66, no. 5, pp. 1381–1403, Nov. 1979.

[119] M. Blomberg, R. Carlson, K. M. B. Elenius, R. Carlson, K. Elenius, and B. Granstrom, "Auditory models and isolated word recognition," *STL-QPRS*, vol. 24, no. 4, pp. 001–015, 1983.

[120] A. M. A. Ali, J. V. der Spiegel, and P. Mueller, "Robust auditory-based speech processing using the average localized synchrony detection," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 279–292, Jul. 1999.

[121] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain," in *Proc. Interspeech 2009*, pp. 18–21.

[122] "On our perception of sound direction," *Philosoph. Mag.*, vol. 13, pp. 214–232, 1907.

[123] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Amer.*, vol. 105, pp. 3436–3448, 1999.

[124] R. Drullman and A. W. Bronkhorst, "Multichannel speech intelligibility and speaker recognition using monaural, binaural and 3D auditory presentation," *J. Acoust. Soc. Amer.*, vol. 107, pp. 2224–2235, 2000.

[125] B. Shinn-Cunningham, "Speech intelligibility, spatial unmasking, and realism in reverberant spatial auditory displays," in *Proc. Int. Conf. Auditory Display*, Kyoto, Japan, 2002, pp. 183–186.

[126] R. F. Lyon, "A computational model of binaural localization and separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1983, pp. 1148–1151.

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

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

[129] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, no. 4, pp. 361–378, 2004.

[130] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.

[131] S. Srinivasan, M. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486–1501, 2006.

[132] H.-M. Park and R. M. Stern, "Spatial separation of speech signals using continuously-variable weighting factors estimated from comparisons of zero crossings," *Speech Commun.*, vol. 51, no. 1, pp. 15–25, 2009.

[133] H. W. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *Amer. J. Psychol.*, vol. 62, pp. 315–337, 1949.

[134] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Proc. IEEE Mohonk Workshop Appl. Signal Process. Acoust. Audio*, 1997, DOI: 10.1109/ASPAA.1997.625622.

[135] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *Proc. Interspeech 2010*, pp. 2058–2061.

[136] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1608–1622, 1986.

[137] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wavefront," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1623–1630, 1986.

[138] 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.

[139] 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.

[140] R. M. Stern and C. Trahiotis, "Models of binaural interaction," in *Hearing*, B. C. J. Moore, Ed. 2nd ed. New York, NY, USA: Academic, 1995, pp. 347–386.

[141] R. M. Stern, D. Wang, and G. J. Brown, "Binaural sound localization," in *Computational Auditory Scene Analysis*, D. Wang and G. J. Brown, Eds. New York, NY, USA: Wiley–IEEE Press, 2006.

ABOUT THE AUTHORS

Hynek Hermansky (Fellow, IEEE) received the Dipl.Ing. degree from Brno University of Technology, Brno, Czech Republic, in 1972 and the Dr.Eng. degree from the University of Tokyo, Tokyo, Japan, in 1985.

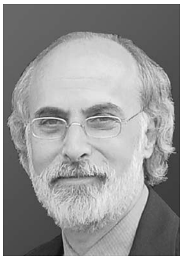
He is the Julian S. Smith Professor of Electrical Engineering and the Director of the Center for Language and Speech Processing at The Johns Hopkins University, Baltimore, MD, USA. He is also a Professor at the Brno University of Technology, and an External Fellow at the International Computer Science Institute at Berkeley, CA, USA. He has been working in speech processing for over 30 years, previously as a Director of Research at the IDIAP Research Institute, Martigny, Switzerland; a Titulary Professor at the Swiss Federal Institute of Technology, Lausanne, Switzerland; a Professor and Director of the Center for Information Processing, Oregon Health and Science University (OHSU), Portland, OR, USA; a Senior Member of Research Staff at U.S. WEST Advanced Technologies, Boulder, CO, USA; a Research Engineer at Panasonic Technologies, Santa Barbara, CA, USA; and a Research Fellow at the University of Tokyo. His main research interests are in acoustic processing for speech recognition.

Prof. Hermansky is a Fellow of the International Speech Communication Association, the General Chair of the 2013 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), and he was responsible for the plenary sessions at the 2011 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. He has also served as the Technical Chair at the 1998 ICASSP, Seattle, WA, USA, and as an Associate Editor for the IEEE Transaction on Speech and Audio. He is also a Member of the Editorial Board of *Speech Communication*.

Jordan R. Cohen (Senior Member, IEEE) received the B.S.E.E. degree from the University of Massachusetts, Amherst, MA, USA, in 1968, the M.S.E.E. degree from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 1970, and the M.S. and Ph.D. degrees in linguistics from the University of Connecticut, Storrs, CT, USA, in 1976 and 1982, respectively.

He is currently the Founder and Chief Technologist at Spelamode Consulting, Inc., Kure Beach, NC, USA, engaged in technical and intellectual property pursuits. He serves as the Co-CTO to Kextil, an emerging field service support company. He was previously the Principal Investigator for GALE at SRI International, and the CTO of Voice Signal Technologies. He was a research staff member at IDA, and at IBM Research, and worked at NSA as a research engineer. He served in the U.S. Air Force from 1970 to 1974. He is engaged in the application of speech recognition for practical systems, and in various aspects of intellectual property pursuit. He is a coauthor of 13 U.S. patents.

Dr. Cohen is a member of the Acoustical Society of America and of the International Speech Communication Association.



Richard M. Stern (Senior Member, IEEE) received the S.B. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1970, the M.S. degree from the University of California Berkeley, Berkeley, CA, USA, in 1972, and the Ph.D. degree from MIT in 1977, all in electrical engineering.

He has been on the faculty of the Carnegie Mellon University (CMU), Pittsburgh, PA, USA, since 1977, where he is presently a Professor in the Electrical and Computer Engineering and the Computer Science, and the Language Technologies Institute, and a Lecturer in the School of Music. Much of his current research is in spoken language systems, where he is particularly concerned with the development of techniques with which automatic speech recognition can be made more robust with respect to changes in environment and acoustical ambience. He has also developed sentence parsing and speaker adaptation algorithms for earlier CMU speech systems. In addition to his work in speech recognition, he also maintains an active research program in psychoacoustics, where he is best known for theoretical work in binaural perception.

Dr. Stern is a Fellow of the Acoustical Society of America and of the International Speech Communication Association (ISCA), the 2008–2009 Distinguished Lecturer of ISCA, a recipient of the Allen Newell Award for Research Excellence in 1992, and he served as General Chair of the 2006 Annual Conference of the International Speech Communication Association (Interspeech). He is also a member of the Audio Engineering Society.