Long time gives the general idea, pertaining to the signal
Short time gives the detail of the signal. Would also get sutff which would die out over long time.

INVITED PAPER

Acoustic Analysis for Automatic Speech Recognition

Some terms:
Nyquist frequency
Shanon sampling theorem
Short time analysis and Long time analysis

This paper presents the theory and practice for methods of speech analysis, as used for automatic speech recognition.

By Douglas O'Shaughnessy, *Fellow IEEE*

ABSTRACT | As a pattern recognition application, automatic speech recognition (ASR) requires the extraction of useful features from its input signal, speech. To help determine relevance, human speech production and acoustic aspects of speech perception are reviewed, to identify acoustic elements likely to be most important for ASR. Common methods of estimating useful aspects of speech spectral envelopes are reviewed, from the point of view of efficiency and reliability in mismatched conditions. Because many speech inputs for ASR have noise and channel degradations, ways to improve robustness in speech parameterization are analyzed. While the main focus in ASR is to obtain spectral envelope measures, human speech communication efficiently exploits the manipulation of one's vocal-cord vibration rate [fundamental frequency (FO)], and so FO extraction and its integration into ASR are also reviewed. For the acoustic analysis reviewed here for ASR, this work presents modern methods as well as future perspectives on important aspects of speech information processing.

KEYWORDS | Automatic speech recognition; digital signal processing; pattern recognition; spectral analysis; speech analysis; time-frequency representation

I. INTRODUCTION

The conversion of a speech signal into a useful message (e.g., its corresponding text) is called automatic speech recognition (ASR). As a practical example of the general area of pattern recognition, ASR needs the efficient compression of input speech data to a small set of parameters, to be able to accurately classify portions of the data as phonemes (and eventually as text). Direct speech-to-text classification without data reduction is not feasible owing to the massive number of potential input audio signals and the lack of a reliable way to link all such signals to their corresponding texts (or to a decision that the audio is not speech).

Speech consists of a sequence of uttered sounds (phonemes), at an average rate of approximately 12 per second. As a nonstationary random process, speech is usually segmented into successive sets of samples for suitable processing, to help accomplish data reduction on each "windowed" set or "frame." (Within each of these sets, one usually assumes quasi-stationarity, which allows averages to be representative statistics.) If such segments are too long, analysis will yield unrepresentative results that are smoothed across sequential phonemes; such would cause errors in ASR classification. If too short, noise may dominate. A common choice is a window of approximately 25 ms, shifting the window by an offset of 10 ms at a time. This update rate of 100 frames/s is common in many speech applications (coding, recognition, and synthesis) as a suitable compromise between too short (leading to excessive calculations and unreliable estimates in noisy signal conditions) or long (smearing estimates across phonemes). Phonemes have widely varying durations, which hinders a simple choice of window length.

Based on what is known of speech production and perception, it is widely assumed that some form of spectral analysis is needed as part of ASR [1]–[3]. Direct analysis of speech in the time domain (i.e., without using a spectrum), as occurs in basic coding systems [e.g., log-pulse-code modulation (PCM) and adaptive differential pulse-code modulation (ADPCM)], would be heavily affected by phase and ignore many aspects of human communication, e.g., the relative lack of control of phase in speech production and a similar insensitivity to phase in perception. Except at very low bit rates, speech coders try to preserve phase, as coder output is reconstructed speech destined for human

Manuscript received May 24, 2012; revised August 30, 2012; accepted December 27, 2012. Date of publication April 5, 2013; date of current version April 17, 2013.
The author is with the Institut national de la recherche scientifique (INRS), University of Quebec, Montreal, QC H5A 1K6 Canada (e-mail: dougo@emt.inrs.ca).

Digital Object Identifier: 10.1109/JPROC.2013.2251592

ears. ASR's output, on the other hand, is text (or ideas) and thus allows discarding of phase. Nonetheless, some systems try to exploit phase, e.g., [4], which used a crude statistical model of speech waveforms. Some recent methods allow networks to use deep learning to extract ASR features automatically and directly from the temporal speech signal, thus allowing phase, if so judged as irrelevant, to be eliminated by low-level learning without supervised feature processing. (For speaker verification, on the other hand, phase may be an important aspect to discriminate similar speakers.) Besides phase, a time-domain-only analysis would ignore: 1) that speakers tend to focus their vocal-tract (VT) movements to accomplish positions for phonemes that correspond to desired spectral resonances (formants); and 2) that inner-ear audition focuses on spectra via the frequency-sensitive basilar membrane in the cochlea. As a result, almost all ASR features involve spectral analysis.

For speech over telephone links (and thus limited to the 300–3200-Hz band), coders typically send at rates in the range of 2–64 kb/s (1–3 kb/s for low-rate applications, 8–11 kb/s for cell phones, and 64 kb/s on the usual network). Assuming the usual telephone bandpass limitations, 8000 samples/s are used in analog/digital conversion. Direct microphone ASR applications use higher rates of 10–16 kilosamples/s. (Yet higher rates, as found on CDs, are rarely needed for ASR, as speech information above 7 kHz is highly redundant, adding mostly aspects of "speech quality" rather than helping intelligibility.) When comparing even the lowest coding rates of 1–2 kb/s against the text information rate of speech, a big gap exists and can be exploited in data reduction for ASR. If we ignore other aspects of speech (e.g., emotions, personality, stress, and gender of the speaker) and only consider text (the usual ASR output), one has a rate of perhaps 60 b/s: typically 12 phonemes/s and assuming a language of 32 phonemes ($\log 32 = 5$). ASR does not need to reduce speech data to its minimal amount, and any reduction must be carefully designed to minimize loss of relevant information.

Proper data reduction greatly aids the ASR process. Data compression for pattern recognition not only reduces cost (e.g., for storing and manipulating data), but also focuses on useful aspects of the data, by eliminating irrelevant aspects. Consider a data sequence of N speech samples, as an initial input vector of data to classify, say, as a portion of one of 32 phonemes in a language, e.g., English. Each such sequence can be viewed as a point in N-dimensional space. If data exemplars for each phoneme (from many speakers and contexts) cluster tightly in such a space, while also spreading apart for different phonemes, then a linear ASR classifier could partition the space well via hyperplanes. Recognition errors would result from cases where the data points for a given phoneme spread too much and overlap with the distributions of other phonemes. ASR approaches typically use Gaussian probability densities (and mixtures thereof) to model phonemes, and errors result when these probability density functions (pdfs) overlap. Given the advent of recent nonlinear classifiers [e.g., deep neural networks (DNNs)], tight clusters for similar sounds are not essential; nonetheless, a major challenge for ASR is to select feature spaces that allow good data reduction.

For ASR, a more fundamental problem than labeling speech is simply determining where, in an audio signal, speech occurs. Such voice-activity detection (VAD) is a prerequisite to (a minority of) ASR systems that ask speakers to pause between words (to simplify the segmentation problem) [5]. VAD is usually accomplished by applying an energy threshold (to distinguish speech from background noise), but some spectral analysis is useful as noise usually has different characteristics from speech [6].

A recent special issue of the IEEE Signal Processing Magazine (SPM) dealt with ASR methods [7], [8]: 1) large-vocabulary continuous speech recognition (LVCSR)—recent techniques of model adaptation; 2) a history of biologically inspired auditory models (see Section III); 3) pronunciation modeling—the use of speech units smaller than words (essential for LVCSR) and use of graphical models (machine learning); 4) discriminative training (learning techniques), which usually outperforms basic maximum likelihood (ML) for ASR acoustic model training (both of model parameters and structures); 5) DNNs that can outperform Gaussian mixture models (GMMs) for ASR; 6) exemplar-based ASR models (exploiting sparse coding or compressive sensing techniques found in signal processing and machine learning); 7) distant-talking speech (corrupted by interfering sounds and reverberation); and 8) microphone array techniques. In contrast, this paper focuses on acoustical features; readers are thus invited to see the SPM issue for more detail on other ASR issues. As well, of course, this current Proceedings of the IEEE issue has several papers on ASR information processing, using the features discussed herein.

II. SPEECH PRODUCTION BY HUMAN SPEAKERS

Data reduction should exploit pertinent characteristics of the signal to be recognized. To understand which relevant features to extract from speech, we briefly examine how humans speak. At a fundamental level, speakers utter a sequence of phonemes by moving tongue and lips to achieve VT positions in a succession suitable for each phoneme, while exciting the VT (which acts as a filter) via air from the lungs. Thus, each short segment of speech has: 1) a spectral envelope owing to the phoneme's VT shape; and 2) a spectral fine structure owing to the nature of the excitation. ASR analysis concentrates largely on the envelope, as the excitation varies widely as a function of many factors orthogonal to phoneme recognition. Also, short-time analysis over limited time windows is very suitable to

generate features for practical ASR systems. Many features of VT excitation, on the other hand, extend well beyond individual phonemes.

A. Segmentals

Each phoneme can be viewed as associated with a target VT shape and corresponding spectral envelope. The latter could theoretically be obtained by artificially exciting the VT with white noise. In actual speech, of course, the excitation is more complex and unknown *a priori*. It is often assumed that noisy speech (obstruent sounds, i.e., stops and fricatives) derives from a noise excitation in the VT that has a flat spectrum over audio frequencies (i.e., 0–10 kHz), and, thus, can be modeled by white noise. Sonorant sounds (vowels, liquids, glides, and nasals), un-, are "voiced," having a quasi-periodic excitation (puffs of air exiting the lungs) owing to the vibration of the vocal cords.

Meaning?

Relevant features to help classify "phones" (the physical realizations of phonemes) likely include: energy, spectral shape, and periodicity. Normal (nonwhispered) vowels are the strongest speech sounds, with high energy (usually decreasing with frequency) in the lower formants, whose spectral envelope weights the harmonics, which are spaced at intervals of F0. Nonvowel sonorants are less intense, sometimes having spectral zeros, but otherwise appear similar to vowels; among the former, the nasals and /l/, having abrupt changes in articulation at their onset and offset (owing to velic closure, or to tongue tip contact with the roof of the mouth), show sudden spectral changes, which do not occur in vowel sequences. Fricatives are composed of noise (random spectra across mostly high frequencies, owing to the use of a shortened VT), and each stop consists of a weak-energy interval followed by frication. (Linguists distinguish frication, which is generated by air passing through a narrow constriction usually high in the VT, and aspiration, made instead at the glottis; while these are quite similar spectrally, frication is limited to higher frequencies, as it only excites the shorter VT above the constriction—the lower portion only contributing spectral zeros.) In English, obstruents (i.e., stops and fricatives) can be either voiced or unvoiced; voicing appears mostly only as a "voice bar" at extremely low frequencies (i.e., F0 and its first harmonic), owing to the modulation of the excitation noise by the vocal cord vibration. In stops, an additional cue is the longer voicing onset time (VOT) for unvoiced stops, as voicing resumes within 20 ms in voiced stops after the release of the VT closure.

Such "manner of articulation," e.g., obstruent versus sonorant, can be more easily classified in ASR than "place of articulation," as the acoustic evidence for manner is easier to discern, even in degraded speech. Distinguishing place changes, e.g., among /p,t,k/ (unvoiced stops) or /i,I,e/ (high front vowels), is much more subtle, and has much higher error rates in both human and automatic speech recognition.

For ASR, the focus is on estimating relevant details of a speech segment's spectral envelope, mostly dealing with its peaks, as speakers tend to control VT shape to achieve specific positioning of spectral peaks. (This can be confirmed in cases where a VT has interference, e.g., food, gum, cigarettes, disease, or for ventriloquists.)

The VT is often modeled as a hard-walled circular acoustic tube of varying cross-sectional area. Basic models have resonances spaced 1 kHz on average, assuming a 17-cm length from glottis to lips (typical of an adult man). As the glottal end is largely closed (the vocal cords forming a narrow slit for voicing), while the lips are usually open, a quarter-wavelength resonator is a basic model for a uniform VT, such as occurs for schwa vowels (and for filled pauses, such as "uhh") [6]. Assuming 340 m/s as the speech of sound, resonances typically appear at: 500, 1500, 2500, 3500, ... Hz in such a model, which corresponds well to actual schwa sounds. In such ideal conditions, F1 = 500, F2 = 1500, etc., where F*i* means the *i*th formant.

For other VT shapes, the resonances move to other frequencies, e.g., for the vowel /i/, F1 is about 280 Hz (owing largely to a raising of the tongue), while F2 lies above 2 kHz (owing mostly to the tongue positioned more forward in the mouth). The traditional "vowel triangle" describes standard F1 and F2 values for all the vowels, and has useful general interpretations as to tongue height and lateral positioning. When plotted in a 4-D space of F0–F1–F2–F3, individual vowels (spoken in isolation, and not in normal speech context) for individual speakers often cluster fairly tightly, with only small amounts of overlap. Such is relevant for phoneme classification, as this would allow partitioning the four-feature space neatly with hyperplanes, where recognition error would be due to regions of overlap. Unfortunately, it is difficult to estimate formants accurately [10]–[12], and coarticulation in no induces much overlap. A direct mapping from s shape could be a step in ASR, but is very difficu

What is the difference between acoustic and phonological features?

A distinction may be noted between acoustic and phonological features, both being numerical outputs of speech analysis. Acoustic features range from simple direct mathematical formulations, e.g., Fourier transform (FT) and energy, to more advanced versions that require several transformations, e.g., linear prediction (LP) coefficients and mel-frequency cepstral coefficients (MFCCs). The simple ones come from elementary mathematical (continuous, but often nonlinear) transformations; some are invertible (e.g., FT), but most are lossy, as the data reduction tries to eliminate less relevant speech information. Phonological features require an additional level of classification (or quantization), and result from major compression of speech data, but at the cost of possible errors in classification. Examples are F0, formants, voicing, and phonetic labels such as "high vowel." ASR focuses on using acoustic features, to avoid phonological classification errors early in its decision process. To better understand ASR, however, phonological features are discussed as well.

In perception, as noted in Section III, humans pay most attention to the presence of stimuli, and tend to ignore absences; i.e., they see, hear, feel, and smell things that occur, rather than noting missing aspects. So, in speech, one focuses on resonances, e.g., their center frequencies, bandwidths, and amplitudes (and usually much less for phase). The formants with the most energy are most salient: spectral detail above 3 kHz in sonorants is far less relevant than the values for F1–F3, as energy in high frequencies is much reduced. Obstruents, on the other hand, have most energy at high frequencies, but listeners pay much less attention to specific detail in such sounds.

It is usually assumed that speakers control their VT so as to achieve (or approach) a target position associated with each phoneme, which is realized in terms of a desired distribution of energy across the spectrum. Aspects that are much less controlled appear to be: spectral slope and bandwidths. Resonance bandwidths tend to be larger for nasal sounds than nonnasals, but there seems to be little evidence otherwise of speaker control of bandwidths. Sonorant spectra generally decrease with frequency, owing mostly to the low-pass nature of glottal excitation (puffs of air from the lungs). While whispered speech has a greater slope and shouted speech a lesser slope, such sounds average about −6 dB/octave. As with bandwidths, speakers do not exercise control here to accomplish aspects of speech communication, i.e., the main focus is on resonance position.

English has approximately 11 vowels, distinguished by tongue height and lateral position, as well as by lip rounding and tongue tip retroflexion. In adult men, such variation causes F1 to have an approximate range from 250 to 800 Hz, and F2 from 800 to 2200 Hz. (F3 is very low for /r/, and high for /i/. Few studies describe the effects of F4 and above, as these seem to be far less relevant.) As a result, differences between vowels correspond to changes in one or more formants of about 100 Hz or so. Speakers appear to manipulate their VT to accomplish such resonance positioning.

B. Intonation

In addition to the spectral details of phones ("segments"), another pertinent aspect of speech production is intonation (also called prosodics or "suprasegmentals," because their effects exceed the temporal range of the phones), i.e., amplitude, duration, and fundamental frequency (F0). Intonation varies phonemically (e.g., lower vowels have higher energy, and F0 tends to be higher in unvoiced contexts), but intonation mostly cues nonphonemic aspects, such as syntax, semantics, and emotions. As discussed later in this paper, use of intonation in ASR has been problematic. Nonetheless, human speakers accomplish many communicative objectives via intonation: they group words together in syntactic units, use lengthening to indicate pause locations, and signal emphasis (important, new-information words versus less critical words to be recognized) as well as emotions. (In tone languages, F0 distinguishes phonemes as well.)

Some examples of how English uses F0 for syntactic purposes are: many major phrases start with a rise in F0 and end with a fall, phrases that are not sentence final tend to end with a small "continuation" rise (signaling to listeners that more speech follows immediately), vocatives (addressing a person) show similar F0, and yes/no questions end with an F0 rise to a high value. As an example, consider the sentence "The good flies quickly past/passed." F0 during both "good" and "flies" varies greatly depending on whether the final word is "past" or "passed."

Semantically, a speaker places larger rises or falls in F0 on words that are considered important, so as to highlight them to the listener. Such words are usually a subset of the "content words" (nouns, verbs, adjectives, adverbs) in an utterance, with "old information" words (repeated or anaphoric) being destressed. Speakers may emphasize words that are not normally stressed, e.g., in "Bob fainted, but came to after a few minutes," the word "to" has a large F0 change, as the speaker notes its prominence, being an adverb in this case, and not the (usual) low-stress preposition.

Amplitude is usually viewed as the least pertinent of the prosodics, in terms of its communicative uses, as it readily varies with: phonemics (e.g., vowels stronger than consonants), distance from a microphone (or attenuation in transmission), and speaker effort. Nonetheless, in each stressed word, the syllable that is denoted as having lexical stress is usually more intense than other syllables. For ASR, overall amplitude is often deemphasized for the reasons just noted. The distribution of energy across frequency is critical to discriminate phones, but the overall level is often associated with transmission and other effects not pertinent for ASR.

C. Variability, Coarticulation, and Speech Dynamics

All aspects of speech production are subject to variability, and for many reasons. Speakers have a great deal of freedom in articulating speech, owing to differences in effort, acoustic conditions and channels, audience, message content, etc. Such is evident in casual versus precise speech, in slow versus fast speech, in talking to friends versus strangers, and in reading versus spontaneous speech. Free variation is presumably constrained by the speaker's desire to avoid confusions by listeners, which occur if the speaker does not exercise sufficient control to correctly communicate.

Major challenges for ASR are differences in speech signals across speakers, contexts, communication channels, and languages. As a pattern recognition problem, ASR uses training sessions on available data, where system designers utilize much data across many conditions. However, one cannot anticipate all variations, and there is often a "mismatch" between training and testing data. A major challenge for ASR is to understand and model speech so as

to be able to handle such mismatches, without necessarily requiring access to excessive amounts of training data.

Another confounding factor for ASR is coarticulation: modifications in how phonemes are uttered owing to the interaction of successive sounds. VT positioning (and thus spectra) for individual phones can be greatly affected by neighboring phonemes, e.g., in the word "strew," one rounds one's lips throughout the initial /str/, in anticipation of the ensuing vowel /u/. With the exception of pauses, which divide utterances into sections of speech that are often several words (and seconds in duration) each, phones follow immediately one after another, greatly affecting VT positioning for all. ASR cannot simply model each phoneme by a supposed target shape and its related spectrum. A common approach is to use "triphone" models, where each phoneme is represented in terms of the context of its immediately preceding and ensuing phoneme neighbors (since most languages have more than 30 phonemes, this leads to many thousands of triphones).

Coarticulation comes from both phonological and phonetic effects [14]. Phonology refers to the discrete classification of aspects of phonemes, e.g., voicing, place, and manner of articulation, whereas phonetics deals with the variability in VT positioning of phonemes. The structure of a syllable is an example of phonological coarticulation: in most languages, all obstruent consonants within a cluster (e.g., [sp] in "split" or "lisp") have the same voicing; semivowels occur adjacent to vowels, and obstruents are the farthest from the vowel in each syllable. Phonological features owing to phoneme proximity have direct effects on timing (e.g., shorter consonants in clusters), and thus affect ASR. Some phonological coarticulation is language specific, e.g., English nasal+plosive clusters share the same place of articulation if the plosive is unvoiced, but not necessarily if voiced. Thus, the nasals in "limp/lint/link" are all short, while those in "beamed/beaned" are long.

A simple approach to ASR would treat each phone separately, assuming that suitable analysis of speech data within each phone would suffice to classify it. The contextual effects of variability and coarticulation do not allow this. Most ASR does indeed analyze data within each successive frame of data, usually spaced every 10 ms with a window spread of 20–30 ms, thus achieving a sequence of N frames of data per phone (where N varies greatly, as discussed below). Such frames are not easily grouped into sets for each phoneme, as phone boundaries are very hard to determine reliably based solely on such data analysis.

The earlier discussion on intonation said little about duration. Features from spectra and F0 can be readily obtained within each successive frame of speech data, but (the difficult task of) grouping frames into phones is usually postponed to a later level of ASR processing. The current standard for ASR modeling is the hidden Markov model (HMM), which partitions speech into phones indirectly, as part of the overall recognition process. The issue of timing and duration is complex and important, and is discussed in Section IV-F.

III. SPEECH PERCEPTION

Section II reviewed human speech production, because ASR needs to interpret its output. This section examines how humans perceive speech, as some ASR approaches emulate human audition [8]. It is not useful to extract speech features beyond a speaker's ability to control them, nor to use precision in feature estimation that exceeds a listener's ability to discriminate sounds. Normal data conversion of speech for ASR employs rates of 8 kilosamples/s (or higher), in order to preserve relevant bandwidth, but the choice of quantizer amplitude precision (e.g., 8 or 16 b) should be guided by levels of human difference limens [just-noticeable differences (JNDs)]. For speech, humans cannot distinguish changes below these approximate values: 1 dB in amplitude, 3 Hz in F0, 5% in formant center frequencies, and 20% in formant bandwidths [6]. As ASR rarely directly estimates formants (or even F0), these are mostly general guidelines, although the choice of bits/feature is relevant. Log PCM (as in basic telephone service) uses 8 b/sample, which is acceptable, although many ASR systems prefer to use 16-b precision, given the low cost of computer memory.

Direct interpretation of what information the brain exploits for speech is difficult, as our understanding of audition is much more limited than our knowledge of speech production. The outer ear and middle ear mostly act as an amplifying filter, enhancing the middle auditory range of 1–3 kHz, thus presenting the cochlea with enhanced information in the range of F1–F3. The inner ear transforms mechanical vibration of the basilar membrane, distributed in frequency as a series of "critical band" bandpass filters, sending neural firings from 12 000 hair cells in each ear along the auditory nerve to the brain. These cells respond to a wide spectral range (high frequencies at the near, basal end of the cochlea; low at the inner apex). Such action is modeled with 24 critical bands, of widths approximately 100 Hz at low frequencies, and increasing in width above 1 kHz. This nonlinear "Bark" or "mel" scale is widely used in ASR, as relevant speech information decreases with increasing frequency. (In general, log scales represent perception better than linear scales.) Some auditory models for ASR include the use of timing or interval statistics of neural firings [9].

In human hearing, audio signals are transformed in complex ways by numerous levels in the auditory network. On the other hand, artificial neural network (ANN) models [often called multilayer perceptrons (MLPs) and used for many pattern recognition applications, including ASR] are far less complex than human audition, and, until recently, were mostly limited to three levels. Three levels allow arbitrary shaping of decision surfaces in an N-dimensional signal representation space; a big challenge

for ANNs is to determine their many weighting factors automatically from speech training data. It is very difficult for a three-level ANN to directly handle ASR, owing to huge timing variations in speech [15]; however, ANNs can be excellent for static pattern recognition. Recent ANN work with more layers, e.g., DNNs (see below), have done a better job, but HMMs still dominate ASR classification.

Certain aspects of audio signals are well preserved in the neural routing from the ear to the brain. The initial neural firings in the cochlea occur stochastically when the basilar membrane bends enough (in one direction). Firings are more likely at times of intense audio amplitude, although the mapping is nonlinear. As each cochlear neuron is associated with a specific "characteristic frequency" (the center frequency of its critical band), neural information at the initial stage of the auditory system has both timing and frequency aspects. At low frequencies, firings can be synchronized with excursions of the audio waveform, but latency issues prevent this above 1 kHz. In general, the brain receives data in complex form about both timing and frequency. Further interpretation of audition is usually accomplished via psychological experiments using listeners.

Listeners tend to focus on the most salient aspects of sounds, e.g., their strongest frequencies. In sonorants, these are usually the harmonics of F1, followed by those of F2 and then F3 (owing to the negative spectral slope). The positions of F1–F3 appear to be highly relevant to distinguish sonorant sounds. The relevance of bandwidths is far less, with listeners judging nasals as needing wider bandwidths. Otherwise, bandwidths and spectral slope are not main foci of perception in speech.

Relative overall audio amplitude is used by listeners to help classify sonorants versus obstruents. Speech generally alternates between such strong and weak sounds, and grouping into syllables (each having one vowel or diphthong) can be done largely based on amplitude, and is very reliable even in noisy conditions. Further classification of the manner of sonorants is likely done in terms of perceived positions of F1–F3, although other gross spectral aspects may be salient. There is little evidence that audition employs "formant detectors" as such, rather than exploiting other possible ways to represent spectral peaks. Many attempts to track formants explicitly in degraded speech have yielded accuracy that is inadequate for ASR [11], [12], [15]–[17]. A major difficulty is that formants differ from VT resonances [14]; acoustic formants are peaks of smoothed spectra in narrow frequency regions, while resonances are underlying properties of the VT. The former vary greatly with VT excitation; the latter come directly from VT shape. Especially during consonants, resonances can be weakly excited or distorted by VT antiresonances (zeros).

Compared to sonorants, the precision needed to classify obstruent sounds, which have much more high-frequency energy, is far less. Listeners readily distinguish vowels with changes of a few tens of hertz in F1, but classify fricatives as being the same even with large variations in spectral shape. Listeners may primarily pay attention to the location of the cutoff frequency in the high-pass spectra of fricatives. In stops, the resonances present during the VOT period may help in aiding to distinguish place information. In general, perception of place in stops may be the most complicated to analyze. Stop place is the easiest to misunderstand in noisy conditions. Place for strident fricatives (/s,z,sh,zh/) is well signaled by steady-state spectra (except in limited telephone bandwidths). For all other cases of obstruents, place is largely cued by spectral transitions in adjacent sonorants, which have far stronger audio. The same is true for nasals: in isolation, /m, n, ng/ sound very similar; their place is heard primarily via transitions in sonorants adjacent to the nasal.

As in many aspects of speech communication, context is important. Speakers utter speech by moving their articulators, while listening to their own speech. They use auditory feedback to judge how well they are uttering their message. Listeners interpret aspects of what they hear in terms of what they are expecting to hear. As speech comes from a wide range of sources (e.g., small VTs for children, large VTs for adults), listeners must normalize auditory representations in order to properly interpret the cues. A formant (or F0, or intensity) is only high or low in the context of its behavior in the rest of an utterance. Such normalization occurs across speakers, speaking rates, channel conditions, noisy environments, etc. Being a mental process, the nature of normalization is hard to judge. It cannot be a linear transformation: for adults, a woman's VT is about 15% shorter than a man's, while her F0 is about twice as high. As a child grows, the pharynx changes much more than the upper VT; so resonance position changes are nonlinear with growth. ASR often uses a form of "vocal-tract normalization" to try to adjust models to reflect differences in spectral features.

There have been many attempts to integrate aspects of audition into ASR methods, most notably the use of the mel-scale, e.g., in MFCC [18]–[22]. Perceptual linear prediction (PLP) [23] includes aspects of loudness, for example. It is rare, however, to see latency or masking directly applied in ASR, despite evidence of their relevance in human perception. Exceptions include "relative spectral" RASTA–PLP and the use of auditory masking to separate multispeaker speech into individual streams (ideally of one speaker each), prior to applying ASR [24]. Few ASR tasks try to handle such "cocktail party" speech, even though humans are often successful; approaches such as computational auditory scene analysis are needed [25].

IV. METHODS OF ESTIMATING SPEECH FEATURES

For ASR, the primary features to be estimated from an input speech signal concern its spectral envelope, because of the strong correlation for each phoneme of VT shape

and its spectra. Overall energy is the simplest feature to obtain, and is very useful for speech activity detection and low-rate coding applications. Its use for ASR is less direct, as energy levels are affected by transmission conditions; e.g., average amplitude can be removed via cepstral mean normalization (see Section V-A), but local amplitude changes should be exploited to help discriminate strong versus weak phonemes.

A. Discrete Fourier Transform (DFT)

To classify phonemes, more details about speech spectra are needed than simple energy. Most often, one uses a version of the traditional short-time DFT, with input being a windowed section (frame) of speech. This operation is repeated periodically by shifting the window for successive frames, so that the entire signal is transformed into a series of DFTs. Here, one must choose two design aspects: the duration N of the DFT and the shift interval L. Typically, N corresponds to about 20–30 ms, while L is 10 ms. Both relate to temporal variations of VT articulation and are compromises to maximize accuracy while minimizing cost. The signal within the windowed frame should be approximately stationary, so as to obtain reliable features describing the speech in the frame. If N is too low (too little data), results may be unreliable, especially in noisy environments. Update rates $(1/L)$ need to be chosen to follow VT movements. As typical speech has 12 phonemes/s, 100 frames/s is a common compromise: even a short 10-ms frame could smear spectral results across a sudden stop release, while long vowels could tolerate much longer windows.

Assume that a 256-point DFT (25.6 ms) is used for speech at a sampling rate of 10 000/s. This would give frequency resolution of almost 40 Hz between output samples—a somewhat coarse quantization to track resonance changes in the formants. (One could instead set $N = 512$, and pad with zero-valued samples in the DFT [2].) If each frame is simply a set of N sequential speech samples, the window is thus rectangular, i.e., equal weighting for all samples. More often, a Hamming window (one period of a cosine wave, raised to be purely positive in value) is used, so that any specific choice of positioning of the window has limited effects. (The wide diversity of pitch periods, relative to window size, can be a major analysis problem—sometimes alleviated by synchronizing analysis to pitch periods.)

Windows are low-pass filters, because one wishes to obtain an average representation and to be able to decimate their output, for efficient use in ASR. Speech itself must be sampled at the Nyquist rate (e.g., 10 000/s to preserve the 0–5 kHz range of the speech, assuming that higher frequencies may be sacrificed for lower cost), but ASR features reflect VT positions, and thus are only needed at much lower (e.g., frame) rates. The DFT itself is not useful for data compression, but rather to transform temporal speech samples into spectra, which then better allow compression. The simplest reduction could set N very low, e.g., $N = 16$, but the corresponding 625 Hz spacing is too coarse for ASR. Instead, one could employ a set of M bandpass filters, whose bandwidths reflect the mel scale, i.e., narrow at perceptually important low frequencies, with increasing width (decreasing resolution) above 1 kHz. (A similar way is to smooth the DFT values, treating the DFT as a signal passing through a low-pass filter, then decimate down to M values.) This approach was used in early ASR systems, and still occurs in channel vocoders ($M = 10$–16) and sub-band speech coders ($M = 4$–8) [6]. A version of such simple smoothing occurs also in MFCC analysis (Section IV-C).

Normally, ASR uses a vector of M parameters to represent each successive frame of analyzed speech data. M can range from 5 to 6 if PLP or formants are used to 39 in many MFCC-based systems.

Data reduction of the DFT takes many forms—the best way to extract optimal features from a speech spectrum remains controversial. Simple smoothing, as in a filterbank approach, while possibly exploiting the auditory mel scale, ignores the fact that listeners focus on spectral peaks (and speakers focus on controlling the resonances of their speech). The common ways of LPC and MFCC are examined next.

Another alternative to the DFT is wavelets [26], [27]. A major DFT limitation is the need to select a specific length N, which fixes both the window duration and frequency resolution. Hearing uses a nonlinear (mel) scale, and a fixed N for analysis ignores that fact. For wavelets, the analysis window varies, being short at high frequencies (where spectral resolution is poor in perception) and long for low frequencies (where corresponding spectral resolution is good). As this better matches human speech analysis, wavelets have been tested for many speech applications. Unfortunately, their nonlinearity and difficulty of interpretation have limited their success for ASR.

B. Linear Predictive (LP) Analysis

A common way for large data reduction for speech applications is the linear predictive coding (LPC) approach, which imposes a speech-specific model of speech production [28], [29]. (LPC remains a major aspect of ACELP speech coding in cellphone use, but for ASR has largely been replaced by MFCC.) For a discrete-time speech signal $s(n)$, LPC presumes that each sample may be approximated as a linear combination of its preceding K samples $(s(n-k), k = 1 \ldots K)$. The weighting factors in this combination (a_k) are called LP coefficients, and are the multipliers in the LP "inverse filter": if one passes $s(n)$ through such a filter, one gets a residual or "error" signal that has minimal energy, and largely consists of impulses at times when the speaker's vocal cords close. Such closures cannot be predicted using a short-time (e.g., $K = 10$ samples) window, but much of the remaining spectral detail (e.g., four to five resonances) is reasonably well modeled with $K = M$ parameters.

Using a spectral matching criterion of minimum mean-square error (MMSE), efficient algorithms exist to determine the a_k parameters or equivalent versions [reflection coefficients and line-spectral frequencies (LSFs)] that are more efficient for coding and stable synthesis filters [2]. This usually involves forming an autocorrelation (or autocovariance) matrix $R(k)$ from windowed speech samples $s(n)$ and inverting the matrix. In multiplying $s(n)$ by $s(n-k)$ and summing over the analysis window, R retains pertinent time–frequency data of the speech. R is equivalent to the convolution of speech with its inverted version $(s(-n))$, which yields $|S(k)|^2$ via a DFT, i.e., the squared amplitude of the speech spectrum. (As noted earlier, speech analysis for ASR invariably discards phase.) For ASR, an important aspect of LPC is that it models well the spectral details around spectral peaks, because the MMSE criterion emphasizes peaks. (Bandwidths and spectral valleys are much less well modeled.)

A weakness of LPC is that it treats all frequencies equally; one applies Parseval's theorem in either time or frequency to minimize modeling error, treating each time or spectral sample the same. Other coders instead try to exploit the perceptual nonuniformity of hearing, e.g., the mel scale, to improve performance. Another issue is that one must, *a priori*, choose the order K of the LPC model. Often, $K = 10$ for telephone speech at 8000 samples/s, in order to obtain a spectral model having four resonances (each resonance needing two poles, and represented by two real parameters). In principle, K is proportional to the speech bandwidth, but wideband speech is rarely represented by LPC and formants above F4 have little energy. A choice for K that is higher than needed to model the formants having major energy can lead to spectral model peaks that correspond to harmonics (such can occur for high-pitched speakers as well), and thus not formants. Too low a value for K smooths out the spectrum, obscuring formants. When used for speech coding, such deviations are less important, as listeners tolerate small resynthesis distortions. For ASR, however, models that deviate from tracking the resonance peaks properly can cause lower recognition accuracy.

In many applications, the basic LP representation (the set of multiplying factors in the synthesis filter) is transformed into a set of reflection coefficients, which are more efficient for transmission and yield stable synthesizers. These have potential benefit for ASR as well, as they correspond to a VT model of K uniform cross-sectional areas that can generate a version of the original speech. However, such a model is not unique (e.g., many VT shapes can yield the same speech spectrum, as ventriloquists often exploit). One extension that may be more useful for ASR is the set of LSFs, which are even more efficient for transmission than reflection coefficients. LSFs map the poles of the LP representation onto the unit circle in the z-plane, which allows easier interpretation of the LSFs in terms of the actual speech formants. Other versions of LP exist [30], [31].

Both the LP spectrum and the LSFs allow easier interpretation (than the DFT) in terms of the resonances of the VT, as they greatly simplify the detail found in the DFT, but neither is efficient enough for most ASR purposes, and thus ASR generally uses another mechanism, i.e., the cepstrum [32], which is described next. Some recent ASR uses frequency-domain perceptual linear prediction (FDPLP), where the LP model is computed from a cosine transform of speech rather than from the signal itself [33].

C. Mel-Frequency Cepstral Coefficients

During the 1980s, a new speech analysis method [22] called MFCC overtook LP, and remains today's *de facto* method for ASR analysis. First, speech spectral power is obtained, usually via an N-point DFT, but sometimes via LP. This set of N spectral powers (indexed on frequency) is then multiplied by a series of (isosceles) triangular functions, which approximate the Bark or mel scale, to convert the N samples into typically 20–24 energy values, as in the number of critical bands. Then, a logarithm applies to convert to decibels. The final step to get the MFCC is an inverse cosine (DCT) transform.

This processing is justified partly by the fact that the last step orthogonalizes the spectrum, yielding a set of uncorrelated parameters that are approximately Gaussian. Its use of a log scale in both time and amplitude follows relevant aspects of perception, and as in most ASR analysis, the phase is discarded. The primary reason for its dominance in ASR is that its use has led to better ASR accuracy than LP, although serious comparison tests have not been reported in the literature. As LP cannot easily integrate the logarithmic aspects (which appear to be relevant in both production and perception of speech), its modeling exploited less well these factors than MFCC did.

Intuitively, the cepstrum "deconvolves" speech. Natural speech results from the VT filtering an excitation; in general, it is difficult to separate these two components. Their combination, e.g., the VT spectral envelope modulating the set of harmonics in voiced speech, confounds the information that ASR needs to decode speech into text. The task is easier if one accesses reliable estimates of envelope and excitation separately. (The smoothing of the speech spectrum that LP does, for example, is one common attempt to eliminate the effects of excitation and estimate the VT effects alone.) The log step in calculating the cepstrum converts a spectral multiplication into an addition (i.e., $\log AB = \log A + \log B$). In the time domain, speech is the convolution of the excitation (e.g., noise, or pulses at the F0 rate) with the VT impulse response. Spectrally, this is the product of their two DFTs. The log step of the cepstrum renders this as a sum, and it is thus easier to filter to yield its components.

For sonorants, the duration of a VT's impulse response is about 20 ms, roughly inversely proportional to the bandwidth of F1 (usually the strongest formant, with the smallest bandwidth). As F0 averages in the range 100–200 Hz,

speech has mostly overlapping (and, thus, confounding to ASR analysis) time responses to the individual pulse excitations from the vocal cord vibrations. The log step of the cepstrum, however, has the effect of compressing a speech cycle into a much briefer time range. Typically, ASR uses only about 13 cepstral values (i.e., the first 1–2 ms of the cepstrum). The log has no effect on the spacing of the periodic F0, and so the vocal cord closures still appear in the cepstrum spaced at periods of 1/F0. The separation of VT filter and excitation is not complete, however, as the model is only approximate.

The MFCC are very difficult to interpret intuitively, however. The final cepstral step (the inverse spectral transform) multiplies the mel-log spectral amplitudes by a series of increasing-frequency sinusoids, and then averages. The initial (zeroth) output coefficient C_0 is simply log energy (i.e., using a zero-frequency cosine wave). For C_1, the next coefficient, the one-period cosine wave weights the lower half of the mel frequency scale positively, and the upper half negatively. Thus, C_1 has high values for sonorants, which are dominated by low-frequency energy, and has negative value for obstruents; this reflects a basic phonetic distinction, and may thus be directly useful, both practically and intuitively.

For all the other MFCC, however, the sinusoidal weightings serve primarily a mathematical purpose, and have little direct phonetic value. The higher MFCC are needed for ASR to provide (very indirect) representation of greater spectral detail: C_0 and C_1 are very coarse speech measures. Typically, ASR uses 13 coefficients to cover the 0–4 kHz telephone bandwidth. Thus, the most precision comes from C_12, where each of 12 cosine periods covers an average of 333 Hz (if we ignore the mel scale). The effective 166-Hz precision is enough to help discriminate among the 11 vowels of English, where such deviations in F1 or F2 cause a shift from, e.g., /i/ to /I/. Other than the decorrelation of parameters, which has value, the motivation for using an inverse DCT is unclear (except, of course, from the practical point of view that MFCCs give good ASR results). Weighting spectral ranges of odd index positively and adjacent even-indexed ranges negatively has no inherent value or interpretation, other than the need to have precision about spectral detail to a level of 166 Hz. Even for C_1, there is little phonetic reason to weight the 1-kHz region much more heavily than the 500- or 1500-Hz regions, as occurs with the cosine weighting. MFCC today remains common in ASR, but modifications have been proposed [34]. Recent work has noted that DNNs may be obviating the cosine transform, falling back to filter banks or FTs, with improved ASR accuracy [7].

D. Two-Dimensional (Time–Frequency) Analysis

Most ASR analysis occurs frame by frame, where a short-time window (often Hamming, of duration about 25 ms) is repeatedly used to provide the speech data for successive analyses. The resulting parameter vectors can be placed in matrix form, e.g., a version of a spectrogram, having the results as a function of time, with the other axis being either frequency or some related dimension (e.g., LPC order, MFCC index). Detailed correlations across the two axes are often difficult to exploit, but appear to be very relevant. For example, on a wideband spectrogram, formant tracks are often well correlated to phoneme production and resulting speech perception ([35] provides a hand-labeled formant database). Several recent techniques have examined how the time-frequency plane may be more well exploited for ASR [36], [37]. Examples are sub-band temporal modulation envelopes [38] and modulation spectra [39]–[41].

E. Effect of Intonation on Spectral Analysis

Variations in F0 often interfere with spectral analysis. The objective of the latter is a smooth representation of the VT spectrum, without "interference" from its excitation. LP and MFCC, among other methods, do a reasonable job at this. LP focuses on broad, major spectral patterns, effectively grouping local strong harmonics (as occur in formants) together, by assigning one pair of poles per resonance, if the LP order is well chosen. In MFCC, the triangular weighting functions smooth out the effects of harmonics, thus focusing attention on broad aspects of the spectrum.

Ignoring or eliminating intonation from consideration in ASR is not ideal, however. F0 and other aspects of intonation have strong communicative relevance, and are heavily relied upon by both speakers and listeners. Future ASR systems will have to exploit this better to help achieve accuracy that will approach that of human listeners.

F. Dynamic Features and DNNs

Since the mid-1980s, basic methodology for ASR has employed the first-order HMM architecture, typically using MFCC as inputs (e.g., an MFCC vector of 13 static coefficients, plus often the delta and delta–delta versions, which indirectly characterize the VT velocity and acceleration, while the static parameters model VT position). Most ASR features are derived every 10–20 ms, using short-time windows of 20–30 ms. Recently, there has been much interest in extending analysis to much larger time spans. HMMs often now use an analysis vector where the short-term spectrum-based MFCC are complemented with "dynamic" features that reflect spectral envelope trends from speech segments up to 100 ms. For example, the modulation spectrum of speech is usually defined as the DFT of a logarithm of the sum of a time series of speech spectra within a time window (corresponding to auditory critical bands), with its mean removed [41]. Like RASTA filtering [42], this is intended to exploit the fact that human hearing is most sensitive to relatively slow modulations (e.g., in the range of 2–8 Hz, roughly corresponding to the rates of syllables in speech). This corresponds to shapes of temporal trajectories of elements of speech

spectral envelopes, which themselves reflect VT movements. The related TempoRAl Pattern (TRAP) methodology uses 1-s trajectories of speech spectral power in critical bands (derived from PLP analysis) with their means removed [33].

Recent research for ASR includes use of DNNs [43], [44]. DNNs try to extract better ASR features at lower network layers (those near the input) and then to perform classification at higher layers. They transform basic MFCC or PLP features through a multilayer neural network, yielding outputs (posterior-based features) that are pdfs. Sometimes such networks have a hidden layer of relatively few nodes (bottleneck architecture). They may have separate parallel paths trained independently and have several hidden levels, each possibly trained sequentially. A main idea here is to allow integrating speech information over longer temporal spans than a few 10-ms frames. Even with delta–deltas, standard frame-based HMMs rarely can use information beyond 50–80 ms, yet coarticulation can have a much wider range. DNNs convert basic acoustic features to density features, through significant nonlinear transformations. Such outputs are usually more decorrelated than the original acoustic features, thus allowing easier use of diagonal covariance matrices in the GMMs.

Conventional HMMs have a single layer of nonlinear feature transformations, whereas DNNs have several. Linear (or nonlinear) dynamical systems, conditional random fields (CRFs), maximum entropy models, support vector machines (SVMs), logistic regression, kernel regression, and MLP all function with a single hidden layer [45]. Such shallow learning models have a simple architecture to transform the basic features into a problem-specific feature space. Making the problem more difficult is that such a space is often unobservable. Shallow architectures can be effective for simple or well-constrained problems, but less so for more complicated natural signals such as human speech. DNNs use a greedy, layer-by-layer learning algorithm that optimizes network weights efficiently. They start with a restricted Boltzmann machine (RBM), a type of Markov random field, with one layer of stochastic hidden units and one layer of stochastic observable units [46].

A problem with basic ML training of HMMs is that maximizing model likelihood is only weakly related to minimizing ASR error. Hence, much of recent ASR research has gone toward discriminatory training methods, which focus the models on distinguishing similar phonetic categories, where the discriminative objective function used for training is closely related to ASR error rate. Such optimization has helped raise ASR accuracy significantly in recent years [47]. For ASR, discriminative learning uses three major discriminative learning objective functions: maximum mutual information (MMI), minimum classification error (MCE), and minimum phone/word error (MPE/MWE), to determine classifier parameters. To optimize discriminative training, ASR typically uses gradient-based methods, e.g., Quickprop (a batch-mode, second-order method that approximates Newton's optimization) or Rprop. Other optimization methods use reestimation, e.g., the expectation–maximization (EM) and EBW algorithms. Such growth transformations ensure rigorous monotone growth of the value of the objective functions iteratively and give useful monotone convergence in ASR model training.

Accuracy can be improved by augmenting (or concatenating) basic input features (such as MFCC) with "tandem" or bottleneck features generated using neural networks [48]. DNNs address a serious weakness of basic HMMs—their often mediocre modeling for speech data. Speech production of phonemes uses very few dimensions (e.g., some articulatory models for speech synthesis use as few as seven VT parameters), while phonemic HMMs often use hundreds of parameters. Recent advances in machine learning [49] have provided efficient methods for training DNNs with many layers of nonlinear hidden units and a very large output layer. The latter handles the many triphone HMM states, many of which can be tied or shared, owing to similarity of phonetic context. One advantage of this new approach is the use of an RBM as a unit for DNN, which acts as a "product of experts" versus a "sum of experts" for HMM mixture models [7]. The many components in GMMs are inefficient because each parameter applies to only a small portion of the data.

V. ROBUSTNESS IN ASR FEATURE EXTRACTION

In many practical environments, speech is often degraded before it reaches the ASR input microphone, e.g., by channel distortions and/or background noise. The success of the analysis methods discussed above varies in such conditions, and ASR using methods that are less robust to degradations suffers lower accuracy [50], [51]. Modifying ASR to handle the wide range of variations observed in different speech contexts is a major research issue. Options include: 1) speech enhancement: enhance or transform a received degraded speech signal prior to analysis (to more closely match its original undegraded version, using models trained on "clean" speech); 2) feature compensation: revise the features obtained by analysis of the degraded speech, to try to match standard stored models (the latter having been trained on clean speech); 3) robust features: try to find analysis methods that could resist expected degradations; and 4) model compensation: revise stored models to better match the distorted nature of the incoming speech.

A. Enhancement

Many ways have been suggested to improve the quality of degraded speech, prior to doing ASR. Most focus on attenuating the distortion as much as possible, while trying to preserve speech aspects that may be most useful in phonetic classification. ASR can employ general speech

enhancement techniques, which tend to focus on lessening listener fatigue, or may design methods that focus primarily on classification. For speech input from a single microphone, enhancement techniques have had great difficulty improving intelligibility, only getting better "quality." (Multimicrophone methods, using array beamforming, are far more successful, but we will assume that ASR has access to only one microphone, as is mostly the case, although some recent cellphones have two.)

Noise-robust ASR often assumes that the mismatch between training and testing is due to a transformation that may be modeled by suitable analysis and data, where speech features or acoustic models are modified using an estimate of the transformation. System parameters are optimized following a suitable criterion, given the adaptation data collected in noisy conditions [52]. Speech distortions are often modeled as a time-varying filter (e.g., transmission channel) plus added noise. Enhancement methods may try to estimate both the filter and the level and nature of the noise, so that the speech may be passed through an inverse filter of the estimated channel, the goal being to attenuate the noise. Common methods include spectral subtraction (SS) and Wiener filtering, both of which attenuate the input signal over frequency ranges where the estimated SNR is low [24], [53], [54]; e.g., RASTA includes "on-the-fly" syllable-length log-spectral subtraction. One assumes here that low-SNR ranges have been corrupted more than others and that their attenuation (or even complete removal) improves speech quality. In cases where unprocessed degradations cause listener fatigue, "enhanced" speech is often easier to listen to, although it may be further distorted.

In SS for human listeners, one does a DFT, subtracts from the amplitude spectrum an estimate of the short-term noise (constrained by a "floor" minimum to avoid "oversubtraction"), and then does an inverse DFT (inserting the original untreated phase). (For ASR applications, this final resynthesis step is not needed.) Wiener filtering is similar, but designs a filter (rather than a DFT for subtraction) to suppress the noise in regions of low SNR. In both cases, the noise is estimated during portions of the input signal when the amplitude is relatively low, on the assumption that such portions are less likely to have speech content.

Variable noise is only one of the reasons for a mismatch between actual input speech to ASR and its models. Differences among speakers lead to much lower accuracy for speaker-independent ASR than with speaker-dependent models. Sometimes robust ASR deals with both speaker and environmental variability simultaneously [55]. Varying channel degradations are another major source of difficulty. One common signal enhancement method is normalization via cepstral mean subtraction (CMS) or cepstral mean and variance normalization (CMVN): for each parameter in every analyzed speech frame, one subtracts the parameter's average (or normalizes for the variance) over a significant section of the current utterance (e.g., a few seconds) [56], [57]. This entails a processing delay and functions poorly for short utterances, such as one-word responses in some ASR applications. Removing the short-term average from each parameter is a form of differential encoding, which is very common in speech coding. Ideally, each section of speech for which this is done would have sounds that distribute well over all possible phones, with the net result being an elimination of stationary channel effects. In practice, the removed baseline may vary significantly, as speech rarely distributes its phonemes uniformly.

Given the predominance of the MFCC in commercial ASR, particular attention has been paid to the fact that added noise, after the log transformation of MFCC, is no longer a linear distortion, and thus not readily treated by filtering methods noted above. For this, approaches called vector Taylor series (VTS) seem to do well [58]. VTS approximates train/test mismatch with a first-order Taylor series expansion, which greatly simplifies the general nonlinear estimation problem. Piecewise-linear approximation methods, e.g., interacting multiple model (IMM) [59], stereo-based piecewise linear compensation for environments (SPLICE) [60], and noise-adaptive training [61], may also be used to handle nonlinear acoustic mismatches.

As audio distortion often perturbs only portions of speech that are limited in time and frequency, an approach called missing feature theory (MFT) can improve ASR noise robustness [62], [63]. Disturbed portions of the signal may be replaced by long-term averages (so-called data imputation). A technique called data marginalization modifies ASR models to handle partial feature vectors, using time–frequency masks.

Statistical or model-based methods that focus on major speech characteristics have found application in ASR [64]. A common way to enhance signals that have periodic components (e.g., voiced speech) is via averaging. Such speech has pitch periods that are mostly repetitions, while degradations are rarely synchronized to F0. Thus, overlaying copies of selected periods causes the speech aspects to reinforce, while other aspects do not [65]. The modulation spectrum, which is the power spectral density of the feature trajectories generated from a speech signal, describes relevant temporal structure [39], and has been used recently with some success. Last, the European Telecommunications Standards Institute has specified an enhancement method called the advanced front–end (AFE) for noise-robust ASR in distributed recognition systems [66].

B. Speaker Variations

While not degradations, speech variations owing to interspeaker and intraspeaker variability are issues to be handled, if mismatch between test and training data is to be reduced and accurate recognition to occur. Statistical models (e.g., averages of parameters, but often more complex Gaussian models) are common in ASR to handle at least intraspeaker variability, e.g., in speaker-dependent

ASR. In speaker-independent applications, as in many commercial telephone cases, variances are much larger, owing to large parameter differences across speakers. ASR errors (due to overlapping pdfs) greatly increase unless the mismatch between training and testing data is reduced. There are many ways to accomplish this, including having separate models for different classes of users, e.g., by gender or by region. This latter approach requires more training and risks errors in its preclassification, i.e., an ASR system needs to assign each user to the proper class, so as to use the corresponding models.

A major source of mismatch concerns variations in VT size and shape across different speakers, which greatly affect speech analysis features. Human listeners normalize for different speakers easily, but ASR needs to be so programmed. While there are many second-order effects, perhaps the biggest difference across speakers is the length of one's VT: typically 17 cm for adult men, about 15% shorter for women, and much less for children. In vocal tract-length normalization (VTLN) for ASR, one estimates a linear mapping of spectral features to adjust to the estimated overall size of each speaker's VT [67], [68]. Various other frequency warping methods have also been tried [69], [70].

A similar approach to adjust ASR model features is called maximum-likelihood linear regression (MLLR) [71]. VTLN, MLLR, and similar variants all typically seek an affine transformation (rotate and scale the parametric space of ASR models), on the assumption that the spectral effects of speaker differences can be largely modeled by a simple scale plus offset. Another similar method is RASTA filtering [42], which can reduce distortions from time-invariant short-term convolution channels, for distortions with short impulse responses. CMS and RASTA thus have difficulties for reverberated speech.

VI. FUNDAMENTAL FREQUENCY ESTIMATION

Despite its clear importance in human speech communication, F0 has had little exploitation in ASR so far. There are two major reasons: ASR avoids use of phonological features (in favor of acoustic features), and F0 has a communicative role that is much more complex than the other frame-based features commonly used for ASR. Acoustic features are not subject to classification errors, which occur in F0 (and formant) estimation. Any pattern recognition process that involves many sequential decisions risks high error rates if individual decisions are wrong and remain uncorrected. Many F0 estimators do well for clean speech signals, but tracking F0 is difficult during voicing transitions and in noisy conditions (in such cases, even human evaluators often disagree on F0 estimation).

Also problematic for ASR is that intonation affects speech over much larger time scales than individual frames [72]. Spectral features reflecting instantaneous VT shape are more easily handled frame by frame, which allows insertion of spectral detail readily into HMM-based ASR. Thus, many aspects of intonation are awkward for ASR, which almost always operates frame by frame. Amplitude, in the form of the zeroth cepstral coefficient (C_0), is frequently omitted from ASR vectors as being too linked to channel effects. Variations in phone durations are a major reason why HMMs replaced dynamic time warping (DTW) in the 1980s: some form of time normalization is needed to handle timing differences, but current methods treat duration as an aspect to be discarded or normalized, rather than exploited to improve ASR performance.

A. Techniques to Estimate F0 in the Time Domain

Looking at waveforms of typical voiced speech, it is reasonable that the simplest estimation method is a form of "peak picking." The primary excitation of the VT occurs at vocal cord closure, once per period. The signal at the VT output increases greatly 0.5 ms after each VC closure (owing to the delay for sound to travel 17 cm to the lips), and then declines, with each pitch period showing local temporal oscillations related to the formants. The first and largest excursion of the period is not always so clearly larger than ensuing excursions as to allow a simple thresholding to locate one such excursion per pitch period. Such is particularly true for cases where F1 and F2 are close, or when the bandwidth of F1 is small.

Spectral detail in the envelope (while greatly useful for ASR) is a hindrance for F0 estimation. Thus, estimators will often try to remove detail irrelevant to F0 estimation, such as phase and formants, which cause interference among the harmonics and thus rendering the time waveform less easy to peak-pick reliably. Speech is usually low-pass filtered to retain only F1 (e.g., up to 900 Hz), which then allows decimation for cheaper processing. Phase can be removed by using the autocorrelation of the filtered and decimated speech: recall that autocorrelation is the convolution of speech with a time-reversed version, which yields a zero-phase spectrum. For even better accuracy in F0 estimation, the signal can have its spectrum partly flattened (e.g., with an LP inverse filter, or more coarsely, with some nonlinear operation, each preserving the crucial harmonic spacing).

B. Techniques in the Frequency Domain

Evidence of F0 is clearly seen in voiced speech waveforms as the energy increases abruptly at the start of each period when the vocal cords close. Nonetheless, many F0 estimators prefer to examine a frequency display instead, e.g., DFT. For this, one multiplies the signal by an N-point window, takes the DFT, and usually discards the phase as not being useful. N must exceed at least one full period, so that some harmonic structure appears in the spectrum. Theoretically, an infinitely long periodic vowel would have a line spectrum with nonzero values only at multiples of F0. In practical cases, the N-point (finite) window smears the spectrum, but leaves enough periodic detail to allow

some form of peak-picking to estimate F0. Simply choosing the lowest frequency peak is inadequate as channels (e.g., the telephone) often suppress the actual fundamental. Similarly, choosing the highest peak is also usually wrong, as formant structure amplifies the spectrum, usually in the F1 region for sonorant sounds.

If one can reliably detect a series of energy peaks (i.e., harmonics) that are all multiples of a single value, then F0 is usually declared to be the highest such value that accounts for a (large) majority of the peaks. As interfering sounds (including other talkers) may insert peaks in the audio that are not harmonics of the intended sonorant under analysis, one should not require that all signal peaks be at such multiples.

One method estimates harmonic spacing to be the desired F0 estimate, via use of comb filters. Such a filter passes primarily energy spaced at equal frequencies. An easy way to do this is to apply a finite impulse response (FIR) filter whose time response consists of impulses spaced at possible period durations. The duration that reinforces periods is declared to be the actual pitch period. As voiced speech consists primarily of harmonics, some speech analyzers focus directly on them, as their locations are specified by a single number F0, which can lead to a simpler processing [45].

As the cepstrum is usually obtained for ASR, it could be used for F0 estimation as well. In voiced speech, there are usually clear spikes at multiples of the duration of the pitch period. Speech is often postprocessed after frame-by-frame F0 analysis, to render F0 estimates more reliable, eliminating rapid variations (errors) in estimates that cannot correspond to actual F0 changes.

This review highlights only a few approaches to F0 estimation. Recent work has focused on noisy speech and on tracking multiple speakers. Very effective approaches are found in the time–frequency domain [73].

C. Possible Use of Intonation for ASR

Among the three main aspects of intonation, amplitude is the easiest to exploit in modern HMM-based ASR. While some systems remove C_0 (energy) from among the other MFCC, others include it. Duration has long been explored through various modeling efforts, including explicit duration models, to replace the inherent Laplacian pdf of first-order HMMs. Inclusion of F0 as an additional feature, along with the normal set of MFCC, has not been helpful to ASR, as the relevant information of F0 is very different from that in the MFCC. The latter describe aspects of each speech frame's spectral envelope, in detail ranging from very coarse (C_1) to very detailed (C_12). F0, on the other hand, is a measure of how fast the vocal cords vibrate and is largely orthogonal to the phonemic information in the MFCC. The value of F0 lies instead in its pattern over several phones, e.g., whether F0 rises or falls on a syllable relative to its neighbors. ASR greatly benefits from contextual use of MFCC information, e.g., use of delta and delta–delta parameters. Such is inadequate for F0: a delta version of F0 could give an indication whether F0 were rising or falling, but its use frame by frame in standard HMMs is inappropriate to interpret the complex nature of F0 patterns in speech. Future use of F0 in ASR is more likely to arise from integration in alternative ASR, not directly using HMMs, but instead combining stochastic modeling with structured features.

VII. FUTURE PROSPECTS TO IMPROVE ASR FEATURES

One stumbling block against progress in ASR has been a general insistence on using a frame-based approach. HMMs do well but remain far from a final solution. Recent research on phonological feature detectors (e.g., locating nasals reliably in speech signals) may suggest ways to do ASR without the basis of a frame-by-frame approach. ASR has gone far with its recent methodology, but still remains far from being able to handle the full range of variations that human listeners can readily accommodate. One reason for this is that listeners exploit context and intonation quite well to interpret what they hear, using time spans well beyond the frame.

As noted earlier, it is not necessary for ASR to emulate human listeners, as modern computers have far greater serial computing power than humans do. However, ASR has tended to rely too much on the contextual assistance of language models to overcome weaknesses in its relatively coarse acoustic modeling (i.e., using HMMs and MFCC). Future major improvements in ASR may well need to find better ways to integrate more structural knowledge into the process. As noted above, there are many facets of human speech production and perception that are only weakly exploited in modern ASR. There are likely ways to model patterns of spectral time–frequency information more efficiently than MFCC vectors, frame by frame.

Speakers do not plan or produce (or understand) speech in 10-ms chunks. It makes more sense to try to utilize time–frequency patterns of formant-like resonances. Looking at spectrograms, as some expert human spectrogram readers do, one can often make very good observations about many aspects of speech recognition, and such an analysis has the potential to be very efficient, as essential elements of spectrograms can be displayed with very few numbers (e.g., versus hundreds of parameters in MFCC vectors) [74]. Exploiting such patterns linked to syllables should allow much easier integration with F0, as the domain of F0 is much closer to the syllable than the frame.

ASR has been dominated for decades by HMM approaches, having replaced earlier expert-system methods, which could not handle the immense amount of variation in speech situations. Nonetheless, it has been noted that use of statistical models may obscure valid phonetic detail in ASR comparisons. Some researchers suggest that

averaging analysis data, as usually done for Gaussian models, may not best utilize training data. Some alternative schemes are called exemplar-based methods [75], [76], e.g., SVMs, *k*-nearest neighbor (*k*NN) methods, and sparse representations. They store individual training examples and are thus computationally expensive. As they only represent individual occurrences, ASR must use many of them separately, much as DTW did for earlier ASR.

As noted just above, other recent ASR approaches are reexamining aspects of expert-system methodology, e.g., focusing on recognizing specific features, such as nasals [77]. There is no chance for successful ASR using a strict expert-system approach, as was common in the 1970s. However, it is highly likely that future major improvements in ASR will require better use of structural information than the traditional use of MFCC in HMMs. While successful in many limited applications to date, this approach has seen only incremental gains in recent years. The acoustical analysis information as used is likely too coarse and too diffuse to furnish enough discriminative data to render accurate decisions in many ambiguous cases. Language models help greatly to reduce ambiguity, but more focused analysis than the MFCC currently provide is likely to improve ASR accuracy. The repeated failure of formant detectors does not encourage that route; yet an analysis focused on gross details of the amplitude spectrum need not insist on locating three to four formants every frame, as many detectors have tried to do. Perceptual experiments suggest instead focusing on simpler measures of where the energy is in speech [78], [79]. The approaches to acoustic analysis of speech as described in this paper suggest ways to improve their use for ASR [80]–[82], and future measures may well be both more accurate for ASR as well as more efficient. ■

REFERENCES

[1] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. New York, NY, USA: Springer-Verlag, 1972.

[2] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York, NY, USA: Macmillan, 1993.

[3] L. R. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.

[4] H. Sheikhzadeh and L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 80–91, Jan. 1994.

[5] S. H. K. Parthasarathi, D. Gatica-Perez, H. Bourlard, and M. Magimai-Doss, "Privacy-sensitive audio features for speech/nonspeech detection," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 8, pp. 2538–2551, Aug. 2011.

[6] D. O'Shaughnessy, *Speech Communication: Human and Machine*, 2nd ed. Piscataway, NJ, USA: IEEE Press, 2000.

[7] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[8] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 34–43, Nov. 2012.

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

[10] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, Jan. 1998.

[11] T. Smit, F. Türckheim, and R. Mores, "Fast and robust formant detection from LP data," *Speech Commun.*, vol. 54, no. 7, pp. 893–902, Sep. 2012.

[12] L. Deng, H. Attias, L. Lee, and A. Acero, "Adaptive Kalman smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 13–23, Jan. 2007.

[13] H. Wakita, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 3, pp. 281–285, Jun. 1979.

[14] L. Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Commun.*, vol. 24, no. 4, pp. 299–323, 1998.

[15] J. Park, F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland, "The efficient incorporation of MLP features into automatic speech recognition systems," *Comput. Speech Lang.*, vol. 25, no. 3, pp. 519–534, Jul. 2011.

[16] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 435–444, Mar. 2006.

[17] M. Lee, J. van Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pt. 2, pp. 741–750, Sep. 2005.

[18] S. Chatterjee and W. B. Kleijn, "Auditory model-based design and optimization of feature vectors for automatic speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 6, pp. 1813–1825, Aug. 2011.

[19] W. Chu and B. Champagne, "A noise-robust FFT-based auditory spectrum with application in audio classification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 137–150, Jan. 2008.

[20] J. Woojay and B.-H. Juang, "Speech analysis in a model of the central auditory system," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp. 1802–1817, Aug. 2007.

[21] S. Dharanipragada, U. H. Yapanel, and B. D. Rao, "Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 224–234, Jan. 2007.

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

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

[24] K. Hu and D. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 6, pp. 1600–1609, Aug. 2011.

[25] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[26] G. Evangelista, "Pitch-synchronous wavelet representations of speech and music signals," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3313–3330, Dec. 1993.

[27] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Process. Mag.*, vol. 9, no. 2, pp. 21–67, Apr. 1992.

[28] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York, NY, USA: Springer-Verlag, 1976.

[29] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[30] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 65–73, Jan. 2008.

[31] A. Harma and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 579–588, Jul. 2001.

[32] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments," *Speech Commun.*, vol. 12, no. 3, pp. 277–288, Jul. 1993.

[33] S. Thomas, S. Ganapathy, and H. Hermansky, "Tandem representations of spectral envelope and modulation frequency features for ASR," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 2955–2958.

[34] C. Lee, D. Hyun, E. Choi, J. Go, and C. Lee, "Optimizing feature extraction for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 1, pp. 80–87, Jan. 2003.

[35] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing *Proc. Int. Conf. Acoust. Speech Signal Process.*, May 2006.

[36] T. T. Wang and T. F. Quatieri, "Two-dimensional speech-signal modeling," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 6, pp. 1843–1856, Aug. 2012.

[37] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Commun.*, vol. 53, no. 5, pp. 736–752, May–Jun. 2011.

[38] X. Lu, M. Unoki, and S. Nakamura, "Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments," *Comput. Speech Lang.*, vol. 25, no. 3, pp. 571–584, Jul. 2011.

[39] X. Xiao, E. Siong Chng, and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1662–1674, Nov. 2008.

[40] B. Kinsgbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, pp. 117–132, 1998.

[41] H. Hermansky, "Speech recognition from spectral dynamics," *Sadhana*, vol. 36, pt. 5, pp. 729–744, Oct. 2011.

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

[43] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[44] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[45] M. J. F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *Comput. Speech Lang.*, vol. 24, no. 4, pp. 648–662, 2010.

[46] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 5884–5887.

[47] X. He, L. Deng, and C. Wu, "Discriminative learning in sequential pattern recognition," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, Sep. 2008.

[48] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks *Int. Conf. Acoust. Speech Signal Process.*, Mar. 2012, pp. 4153–4156.

[49] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[50] B. Milner and J. Darch, "Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 2, pp. 338–347, Feb. 2011.

[51] G. Farahani, S. M. Ahadi, and M. M. Homayounpour, "Features based on filtering and spectral peaks in auto correlation domain for robust speech recognition," *Comput. Speech Lang.*, vol. 21, no. 1, pp. 187–205, Jan. 2007.

[52] H. K. Kim and R. C. Rose, "Cepstrum-domain model combination based on decomposition of speech and noise using MMSE-LSA for ASR in noisy environments," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 4, pp. 704–713, May 2009.

[53] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

[54] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[55] Y. Wang and M. J. F. Gales, "Speaker and noise factorization for robust speech recognition," *IEEE J. Sel. Top. Signal Process.*, vol. 20, no. 7, pp. 2149–2158, Jul. 2012.

[56] 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.

[57] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, pp. 133–147, Aug. 1998.

[58] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Int. Conf. Speech Lang. Process.*, Beijing, China, 2000, pp. 869–872.

[59] N. S. Kim, Y. J. Kim, and H. W. Kim, "Feature compensation based on soft decision," *IEEE Signal Process. Lett.*, vol. 11, no. 3, pp. 378–381, Mar. 2004.

[60] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of non-stationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 568–580, Nov. 2003.

[61] O. Kalinli, M. L. Seltzer, J. Droppo, and A. Acero, "Noise adaptive training for robust automatic speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 8, pp. 1889–1901, Nov. 2010.

[62] M. Van Segbroeck and H. Van Hamme, "Advances in missing feature techniques for robust large-vocabulary continuous speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 123–137, Jan. 2011.

[63] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 443–457, Jul. 2007.

[64] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[65] J. A. Morales-Cordovilla, A. M. Peinado, V. Sanchez, and J. A. Gonzalez, "Feature extraction based on pitch-synchronous averaging for robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 3, pp. 640–651, Mar. 2011.

[66] R. F. Astudillo, D. Kolossa, P. Mandelartz, and R. Orglmeister, "An uncertainty propagation approach to robust ASR using the ETSI advanced front-end," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 5, pp. 824–833, Oct. 2010.

[67] S. Panchapagesan and A. Alwan, "Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 42–64, Jan. 2009.

[68] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," presented at the CAIP Workshop: Frontiers in Speech Recognition II, Piscataway, NJ, USA, Spring 1994.

[69] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 49–60, Jan. 1998.

[70] S. Umesh, L. Cohen, N. Marinovic, and D. J. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 40–45, Jan. 1999.

[71] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[72] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 469–481, Oct. 1994.

[73] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 799–810, May 2011.

[74] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, and M. Athineos, "Pushing the envelope—Aside [speech recognition]," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 81–88, Sep. 2005.

[75] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: From TIMIT to LVCSR," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 8, pp. 2598–2613, Nov. 2011.

[76] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2067–2080, Sep. 2011.

[77] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Commun.*, vol. 43, no. 3, pp. 225–239, Aug. 2004.

[78] B. Lindblom, R. Diehl, and C. Creeger, "Do 'Dominant Frequencies' explain the listener's response to formant and spectrum shape variations?" *Speech Commun.*, vol. 51, no. 7, pp. 622–629, Jul. 2009.

[79] S. Dusan, "On the relevance of some spectral and temporal patterns for vowel classification," *Speech Commun.*, vol. 49, no. 1, pp. 71–82, Jan. 2007.

[80] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Commun.*, pp. 763–786, Oct.–Nov. 2007.

[81] 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, Sep. 2004.

[82] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 2, pp. 142–152, Feb. 2008.

ABOUT THE AUTHOR

Douglas O'Shaughnessy (Fellow, IEEE) received the Ph.D. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1976.

He has been a Professor at the Institut national de la recherche scientifique (INRS), Montreal, QC, Canada and an Adjunct Professor at McGill University, Montreal, QC, Canada, since 1977. He is the author of the textbook *Speech Communications: Human and Machine* (Reading, MA, USA: Addison-Wesley, 1986; revised Piscataway, NJ, USA: IEEE Press, 2000). In 2003, with L. Deng, he coauthored the book *Speech Processing: A Dynamic and Optimization-Oriented Approach* (Marcel Dekker). His research interests include all aspects of speech processing, focusing recently on automatic speech recognition.

Prof. O'Shaughnessy is a Fellow of the Acoustical Society of America. He served 12 years as an Associate Editor for the *Journal of the Acoustical Society of America*, and is the founding Editor-in-Chief of the *EURASIP Journal on Audio, Speech, and Music Processing*. He is now the Vice-President of the International Speech Communication Association (ISCA) and Chair of the IEEE Signal Processing Society's (SPS's) Speech and Language Technical Committee. He has served also as an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, on the SPS Board of Governors, and now on the IEEE Technical Activities Board (TAB) Periodicals Committee. He has presented tutorials on speech recognition at the 1996, 2001, and 2009 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the 2003 International Conference on Communications (ICC).