Chapter 8
Robust Features in Deep Learning-Based Speech Recognition

Vikramjit Mitra, Horacio Franco, Richard Stern, Julien Van Hout, Luciana Ferrer, Martin Graciarena, Wen Wang, Dimitra Vergyri, Abeer Alwan, John H.L. Hansen

Abstract Recent progress in deep learning has revolutionized speech recognition research, with Deep Neural Networks (DNNs) becoming the new state of the art for acoustic modeling. DNNs offer significantly lower speech recognition error rates compared to those provided by the previously used Gaussian Mixture Models (GMMs). Unfortunately, DNNs are data sensitive, and unseen data conditions can deteriorate their performance. Acoustic distortionssuch as noise, reverberation, channel differences, etc.add variation to the speech signal, which in turn impact DNN acoustic model performance. A straightforward solution to this issue is training the DNN models with these types of variation, which typically provides quite impressive performance. However, anticipating such variation is not always possible; in these cases, DNN recognition performance can deteriorate quite sharply. To avoid subjecting acoustic models to such variation, robust features have traditionally been used to create an invariant representation of the acoustic space. Most commonly, robust feature-extraction strategies have explored three principal areas: (a) enhancing the speech signal, with a goal of improving the perceptual quality of speech; (b) reducing the distortion footprint, with signal-theoretic techniques used to learn the distortion characteristics and subsequently filter them out of the speech signal; and finally (c) leveraging knowledge from auditory neuroscience and psychoacoustics, by using robust features inspired by auditory perception.

In this chapter, we present prominent robust feature-extraction strategies explored by the speech recognition research community, and we discuss their relevance to coping with data-mismatch problems in DNN-based acoustic modeling. We present results demonstrating the efficacy of robust features in the new paradigm of DNN acoustic models. And we discuss future directions in feature design for mak-

Vikramjit Mitra
SRI International, STAR Lab, 333 Ravenswood Ave. Menlo Park, CA, 94532, USA, e-mail: vikramjit.mitra@sri.com

Horacio Franco
SRI International, STAR Lab, 333 Ravenswood Ave. Menlo Park, CA, 94532, USA, e-mail: horacio.franco@sri.com

ing speech recognition systems more robust to unseen acoustic conditions. Note that the approaches discussed in this chapter focus primarily on single channel data.

8.1 Introduction

Before the advent of deep learning, Gaussian Mixture Model (GMM)-based Hidden Markov Models (HMM) were the state-of-the-art acoustic models for automatic speech recognition (ASR) systems. However, GMM-HMM systems are susceptible to background noise and channel distortions, and a small mismatch between training and testing conditions can make speech recognition a futile effort. To counter this issue, the speech research community undertook significant efforts to reduce the mismatch between training and testing conditions by processing the speech signal, either through speech enhancement [115, 106] or by using robust signal-processing techniques [26, 62, 112, 77]. Studies also explored making acoustic models more robust by either using data augmentation or introducing a reliability mask [65, 29, 18].

The emergence of Deep Neural Network (DNN) architecture has significantly boosted speech recognition performance. Several studies [85, 102, 64] demonstrated significant improvement in speech recognition performance from DNNs compared to their GMM-HMM counterparts. Recent studies [103, 23] showed that DNNs work quite well for noisy speech and again significantly improve performance under these conditions compared to GMM-HMM systems. Given the versatility of DNN systems, it has been stated [120] that speaker normalization techniques, such as Vocal Tract Length Normalization (VTLN) [122] do not significantly improve speech recognition accuracy, as the DNN architectures rich multiple projections through multiple hidden layers enable it to learn a speaker-invariant data representation.

State-of-the-art DNN architectures also deviate from using traditional cepstral representation to instead employing simpler spectral representations. While GMM-HMM architectures necessitated uncorrelated observations due to their widely used diagonal covariance design (which in turn required that the observations undergo a decorrelation step using the popular Discrete Cosine Transform (DCT)), DNN architectures suffer no such requirement. Rather, the neural network architectures are known to benefit from cross-correlations [81] and hence demonstrate similar or better performance when using spectral features rather than their cepstral counterparts [103].

The Convolutional Neural Network (CNN) [2, 50] is often found to outperform fully connected DNN architectures [1]. CNNs are also expected to be noise robust [2], especially when noise or distortion is localized in the spectrum. Speaker-normalization techniques, such as VTLN [122], are also found to have less impact on speech recognition accuracy for CNNs as compared to for DNNs. With CNNs, the localized convolution filters across frequency tend to normalize the spectral variations in speech arising from vocal tract length differences, enabling the CNNs to learn speaker-invariant data representations. Recent results [84, 83, 80] confirm that CNNs are more robust to noise and channel degradations than DNNs. Typically

for speech recognition, a single layer of convolution filters is used on the input-contextualized feature space to create multiple feature maps that, in turn, are fed to fully connected DNNs. However, in [96], adding multiple convolution layers (usually up to two) was shown to improve performance of CNN systems beyond their single-layer counterparts. A recent study [74] observed that performing convolution across the time and frequency dimensions gives better performance than that provided by the CNN counterparts, especially for reverberated speech. Temporal processing through the use of Time-Delay Neural Networks (TDNNs) provided impressive results when dealing with reverberated speech [91].

DNN models can be quite sensitive to data mismatches, and changes in the background acoustic conditions can result in catastrophic failure of such models. Further, any unseen distortion introduced at the input layers of the DNN results in a chain reaction of distortion propagation through the DNN. Typically, deeper neural nets offer better speech recognition performance for seen data conditions than shallower neural nets; while the shallower nets are relatively robust to unseen data conditions [75]. This observation is a direct consequence of distortion propagation through the hidden layers of the neural nets, where deeper neural nets typically have more distorted information at their output-activation level compared to shallower ones, as shown by Bengio [11]. The literature reports that data augmentation to match the evaluation condition [90, 61] improves the robustness of DNN acoustic models and combats data mismatch. All such conditions assume that we have a priori knowledge about the kind of distortion that the model will see, which is often quite difficult, if not impossible, to achieve. For example, ASR systems deployed in the wild encounter unpredictable and highly dynamic acoustic conditions that are unique and hence difficult to augment.

A series of speech recognition challenges (MGB [9]; CHiME-3 [6]; ASpIRE [45]; REVERB-2014 [67]; and many more) revealed the vulnerability of DNN systems to realistic acoustic conditions and variations, and has resulted in innovative ways for making DNN-based acoustic models more robust to unseen data conditions. Typically, robust acoustic features are used to improve acoustic models when dealing with noisy and channel-degraded acoustic data [84, 83, 80]. A recent study [97] showed that instead of performing ad-hoc signal processing, as typically is done for robust feature generation, one can directly use the raw signal and employ a Long Short-Term Memory (LSTM) neural net to perform the signal processing for a DNN acoustic model where the feature-extraction step parameters are jointly optimized with the acoustic model parameters. Although such an approach is intriguing for future speech recognition research, unknown is both whether the limited training data would impact acoustic model behavior and how well such systems generalize to unseen data conditions, where feature transforms learned in a data-driven way may not generalize well for out-of-domain acoustic data.

Model adaptation is another alternative for dealing with unseen acoustic data. Several studies have explored novel ways of performing unsupervised adaptation of DNN acoustic models [118, 98, 88], where techniques based on Maximum Likelihood Linear Regression (MLLR) transforms, i-vectors, etc. have shown impressive performance gains over un-adapted models. Supervised adaptation with a limited

set of transcribed target domain data is typically found to be helpful [7], and such approaches mostly involve updating the DNN parameters with the supervised adaptation data with some regularization. The effectiveness of such approaches is usually proportional to the volume of available adaptation data; however, such systems are typically found to digress away from the original training data and to learn the details of the target adaptation data. A solution for coping with this issue was proposed in [121], where a Kullback-Leibler Divergence (KLD) regularization was proposed for DNN adaptation, which differs from the typically used L2 regularization [68] in the sense that it constrains the model parameters themselves rather than the output probabilities.

In the next sections, we first provide a brief historical background of acoustic features as used in ASR systems, present some of the prominent robust feature-extraction strategies that have been used in the literature, and discuss how some of those features have been used in the current DNN-based acoustic models.

8.2 Background

The study of speech technologies began during the second half of the 18th century, with an attempt to create machines that imitate the process of human speech production [59]. Acoustic-phonetics dominated the early years of modern speech recognition research, with analysis of acoustic realizations of phonetic elements in spoken utterances being the primary focus. Vocal-tract resonances or formant structures in speech at sustained vowel contexts was widely researched, and the vowel space with respect to formant frequency values was defined [19, 40].

In the late 1960s, Linear Predictive Coding (LPC) [3, 56] was introduced, which enabled estimating the vocal-tract response from speech waveforms. Introduction of LPC, in turn, enabled designing pattern-recognition methodologies that recognized speech using LPC-based information [55, 93]. In 1980, Davis and Mermelstein [20] first introduced Mel-Frequency Cepstral Coefficients (MFCCs), which have since served as the acoustic feature of choice across all speech applications. The steps involved in MFCC feature computation consist of (1) short-time Fourier analysis using Hamming windows; (2) weighting of the short-time magnitude spectrum by a series of triangularly shaped filterbanks with peaks that are equally spaced in frequency according to the mel scale; (3) computation of the log of the total energy in the weighted spectrum; and (4) computation of a relatively small number of coefficients of the inverse Discrete Cosine Transform (DCT) of the log-power coefficients for each channel. The mel-filterbank crudely mimics human auditory filtering; the log-compression mimics the nonlinear psychophysical transfer function for intensity; and the inverse DCT provides a low-pass Fourier series representation of the frequency-warped log spectrum, where the fine structure corresponding to source information is filtered out, retaining mostly the phonetic content of speech.

The Perceptual Linear Prediction (PLP) feature is somewhat different than the

MFCC feature, but the motivating principles behind both features are similar. The steps involved in PLP feature extraction are as follows: (1) short-time Fourier analysis using Hamming windows (as in MFCC processing); (2) weighting of the power spectrum by a set of asymmetrical functions that are spaced according to the Bark scale, and that are based on the auditory masking curves of [99]; (3) pre-emphasis to simulate the equal-loudness curve suggested by Makhoul and Cosell [70], to model the loudness contours of Fletcher and Munson (as in Fig. 8.10); (4) a power-law nonlinearity with exponent 0.33 as suggested by Stevens et al. [107] to describe the intensity transfer function; (5) a smoothed approximation to the frequency response obtained by all-pole modeling; and (6) application of a linear recursion that converts the coefficients of the all-pole model to cepstral coefficients.

In this section, we briefly discussed how speech science evolved and the motivation behind conventional feature-extraction techniques and described the steps involved. Next, we describe the various facets of speech-signal processing that have been explored to improve performance and robustness of automatic speech recognition systems.

8.3 Approaches

Since the introduction of Automatic Speech Recognition (ASR) systems, a tremendous effort has been made toward understanding the problem of speech recognition and making such systems more robust, with reliability of ASR systems under realistic background conditions being a critical research topic. Digitized audio signals serve as input to ASR systems, and therefore, signal-processing methodologies have been exhaustively investigated for coping with background conditions, with an aim of producing invariant speech representations that least impact ASR acoustic-model performance and thus speech recognition quality. The study of robust features explores different signal-processing techniques that produce reliable and invariant speech representations, where the phonetic classes are more easily recognizable, and the background distortions are minimized. In this section, we discuss robust feature-extraction techniques that have been investigated in the ASR research literature.

8.3.1 Speech Enhancement

Speech enhancement has received a tremendous amount of attention over the previous few decades. A detailed exploration of the different speech-enhancement techniques can be found in [10]. Most speech-enhancement techniques aim to modify the Short-Time Spectral Amplitude (STSA) of noisy speech signals. Subtractive-type speech-enhancement techniques assume that background noise is locally stationary, such that the noise characteristics can be estimated from the speech ab-

sent/pause regions. Since the introduction of the spectral subtraction algorithm [13], several variants/enhancements of subtractive algorithms have been proposed [8, 44]. In [115], a detailed analysis of the various subtraction parameters was explored, and a generalized spectral subtraction algorithm that adapts its parameters based on the masking properties of the human auditory system was presented.

In [31], robustness of ASR systems was investigated where speech in additive noise conditions was considered. The ETSI (European Telecommunications Standards Institute) basic [25] and advanced [26] frontends have been proposed for Distributed Speech Recognition (DSR). Such frontends performed speech enhancement to attenuate background noise, before extracting the spectral features for acoustic model training. The ETSI advanced has two stages, where the first stage consists of a Voice Activity Detection (VAD) to detect speech-absent regions for estimating the noise spectral characteristics necessary for speech enhancement, and the second stage performs speech enhancement followed by acoustic feature extraction. The ETSI advanced frontend is typically found to offer better performance than the ETSI-basic frontend for noisy conditions [31].

Auditory Scene Analysis (ASA) is usually considered to be a key factor behind the human ability to robustly perceive speech in varying acoustic environments [14]. ASA helps human listeners to organize the audio mixture into streams [14] that correspond to the different sound sources in the mixture. A feature-based Computational Auditory Scene Analysis (CASA) system was proposed in [105], which makes weak assumptions about the various sound sources in the mixture. In [116], an ideal binary time-frequency mask was proposed as a major computational goal of CASA, where the binary mask is constructed from a priori knowledge of target and interference. Using time-frequency masks is motivated by the phenomenon of auditory masking, where a weaker signal is masked by a stronger one within a critical band [86]. Soft-mask based approaches have been successfully applied to noise-robust ASR on small- and large-vocabulary tasks. In [113], a technique called Log-Spectral Enhancement (LSEN) was proposed, in which the variability caused by noise on the log-spectra is reduced while preserving the variability from the speech energy. First, an SNR-based soft-decision mask is computed in the mel-spectral domain as an indicator of speech presence. Then, the known time-frequency correlation of speech is exploited by treating this mask as an image and performing median filtering and blurring to remove the outliers and to smooth the decision regions. Finally, log-spectral flooring is applied on the lifted spectra of both clean and noisy speech, so as to match their respective dynamic ranges and to emphasize the information in the spectral peaks.

8.3.2 Signal-Theoretic Techniques

Signal-theoretic approaches use signal characteristics to perform filtering or transformation of the speech signal to generate robust feature representations that can improve the robustness of speech recognition systems against varying acoustic con-

ditions.

Typically, acoustic features are expected to demonstrate different distributions for different phonetic units. It is well known that noise, channel, reverberation, etc. result in significant deviation from the usual distributions for the different acoustic units. Such deviation typically results in acoustic condition mismatch, in which the training data and the test data statistics do not match, resulting in significant errors during test data decoding.

The most direct robustness techniques are based on normalization of various statistics of the features. The simplest such approach is Cepstral Mean Normalization (CMN), in which the mean of the cepstra in an utterance is subtracted frame by frame on a sentence-by-sentence basis, both in training and testing a speech recognition system. CMN has been so successful that it (or a similar technique) is used invariably in speech recognition. The success of CMN can be understood from two points of view. First, if a speech signal undergoes unknown linear filtering, showing that the filter imposes an additive shift to the cepstral coefficients is easy, provided the filters impulse response is briefer than the analysis windows duration. Hence, subtracting the mean cepstral values eliminates any effects introduced by stationary linear filtering. Second, and more prosaically, equating the features utterance-level means reduces variability between the features representing the training and testing data. This principle is easily extended to the features other attributes, such as in Mean Variance Normalization (MVN, in which the means are typically set to zero, and the variances set to one in training and testing) and Histogram Equalization (HEQ, in which the values of the features are warped monotonically to match a standard distribution of their values) [46]. MVN and HEQ typically provide some additional robustness to many types of distortion, again by reducing the statistical disparities between the training and testing samples.

8.3.3 Perceptually Motivated Features

It is well known that human speech-processing capabilities surpass the capabilities of current automatic speech recognition and related technologies. Since the early 1980s, this observation has motivated development of feature-extraction approaches for speech recognition systems that are based on auditory physiology and perception. Influential early examples include the auditory models of Seneff [104], Lyon [69], Ghitza [37], and Cohen [17]. Typically, these features provide little or no benefit for the recognition of clean speech, but they tend to be helpful in recognizing degraded speech.

Researchers have had differing opinions concerning which aspects of auditory processing are the most important to preserve in feature-extraction schemes. The most successful auditory modeling schemes have included some of the following components:

- Peripheral frequency selectivity, which typically includes a bank of filters that

mimic the shape of the frequency-selective response of individual fibers of the auditory nerve and more central structures. The gammatone filterbank [89] is frequently used to implement this stage of processing.

- Rate-level response, which typically takes the form of an S-shaped function (such as a sigmoid or inverse tangent) that relates signal intensity in a given frequency channel to output level, as opposed to the strictly logarithmic relationship between input and output in MFCC and similar representations.

- Synchrony to low-frequency fine structure, in the form of a mechanism that responds in synchrony to the fine structure of the low-frequency components of sound. (This component is believed by some to improve recognition accuracy in noisy environments.)

- Emphasis of onsets and suppression of steady state components, as in RASTA processing. In effect, this enhances temporal contrast and improves recognition accuracy in reverberant environments.

- Lateral suppression, which enhances contrast in signal content with respect to frequency. This is believed by some to be useful especially in distinguishing components of complex sound fields.

- Modulation-spectrum analysis, which can be useful for separating speech and non-speech components in noisy environments.

In recent years, advances in computation and statistical modeling of the features produced by auditory models have enabled much more practical use of physiologically and perceptually motivated features. Examples of successful systems include RASTA-PLP, TRAPS, PNCC, FDLP, MHEC, NMC, DOC, and many more.

Temporal processing plays a key role in human speech perception and ASR [28, 60]. For example, short-time spectral features, such as Mel-Frequency Cepstral Coefficients (MFCCs), are routinely concatenated with their first- and second-order temporal derivatives. The delta (δ) and double delta (δ^2) feature coefficients capture temporal dynamics of the acoustic features, and are overwhelmingly used in speech tasks such as speech recognition, speaker recognition, language identification, etc. Temporal information has also been incorporated through extracting temporal Amplitude Modulation (AM) of speech spectra or cepstra. The widely popular and frequently used modulation-based acoustic feature is the RelAltive SpecTrA (RASTA) processed PLP feature [48], which uses an Infinite-Impulse Response (IIR) filter that emphasizes AM frequencies between 1 and 12 Hz. The goal of RASTA processing is retaining the perceptually relevant modulation bands that correspond to linguistically meaningful information while filtering out extrinsic information [28, 60]. RASTA processing effectively applies a bandpass filter to the compressed spectral amplitudes in the intermediate stages of the PLP features, with the intention of modeling the emphasis in the transient portions of incoming signals, which is considered to be an attribute of human auditory processing.

PLP features [47] were developed with the intent of obtaining a representation that was similar to MFCC features, but implemented in a manner that was attentive to more detailed attributes of peripheral auditory physiology and perception. Details regarding PLP processing are provided in section 8.2. Many researchers have

obtained better recognition accuracy with PLP features than with MFCC features and PLP feature extraction is frequently combined with the RASTA algorithm to produce RATSA-PLP features.

Both physiological and psychophysical data suggest that mammalian auditory systems include units in the brainstem that are sensitive to the specific modulation frequencies to amplitude-modulated signals, independent of carrier frequency [58]. Similarly, psychoacoustical findings also indicate that humans are sensitive to modulation frequency [114, 119], with temporal modulation transfer functions indicating greatest sensitivity to temporal modulations at approximately the same frequencies as in the physiological data, despite the obvious species differences. This information has been used to implement features based on frequency components of the temporal envelopes of the bandpass-filtered components of speech signals, which Kingsbury and others referred to as the modulation spectrum [66]. Typically, the modulation spectrum is obtained by passing the speech signal through bandpass peripheral auditory filters, computing the envelopes of the filter outputs, and passing these envelopes through a second set of parallel bandpass modulation filters with center frequencies between 2 and 16 Hz. As a result, the modulation spectrum is a joint function of the center frequencies of the initial peripheral auditory filters, which span the range of useful speech frequencies, and the center frequencies of the modulation filters. This is a useful representation, because speech signals typically exhibit temporal modulations with modulation frequencies in the range that is passed by this processing, while noise components often exhibit frequencies of amplitude modulation outside this range. Tchorz and Kollmeier [108], among other researchers, observed the greatest amount of temporal modulation at modulation frequencies of approximately 6 Hz, and that low-pass filtering the envelopes of the outputs of each channel generally reduced the variability introduced by background noise.

8.3.3.1 TempoRAl PatternS (TRAPS)

Hermansky and Sharma [49] developed the TRAPS representation, which operates on one-second segments of the log-spectral energies that emerge from each of 15 critical-band filters. In the original implementation, these outputs were classified directly by a Multilayer Perceptron (MLP). This work was extended by Chen et al. [123], who developed HATS (for Hidden Activation TRAPS), which trains an additional MLP layer at the level of each critical band filter to provide a set of basic functions optimized to maximize the discriminability of the data to be classified. TRAP-DCT features were proposed in [100], which is a variation of the previously proposed TRAP features, where Discrete Cosine Transform (DCT) was applied on 310 ms-long segments of critical spectral energies. The TRAP-DCT features reduce word error rates (WERs) for ASR tasks in noisy conditions [100].

8.3.3.2 Frequency-Domain Linear Prediction (FDLP)

Athineos and Ellis [4] developed FDLP, where the temporal envelopes of the outputs of critical band filters are represented by linear prediction. Much as linear-predictive parameters computed from the time-domain signal within a short analysis window (e.g., 25 ms) represent the envelopes of the short-time spectrum within a slice of time, the FDLP parameters represent the Hilbert envelope of the temporal sequence within a slice of spectrum. This method was further incorporated into a method called LP-TRAPs [5], in which the FDLP-derived Hilbert envelopes were used as input to MLPs that learned phonetically relevant transformations for later use in speech recognition. LP-TRAPS can be considered to be a parametric estimation approach to characterizing the trajectories of the temporal envelopes, while traditional TRAPS is nonparametric in nature. Traditional FDLP [5, 33] features approximate the temporal Hilbert envelope within spectral sub-bands by linear prediction on one-second-long cosine-transformed audio segments [33]. The derived set of temporal subband envelopes forms a two-dimensional representation that is convolved with an integration window of 25 ms before resampling at a frame rate of 100 Hz. Further, the spectral bands are integrated by using a mel-filterbank, and cepstral coefficients are derived by applying a DCT. FDLP features demonstrate improved robustness against channel noise, additive noise, and room reverberation using a phoneme-recognition task with conversational telephone speech [33].

Mel-filterbanks have served as the state-of-the-art spectral analysis filters for speech-processing tasks since their introduction. Recently, with more availability of computing resources, better-precision filterbanks, such as the gammatone filterbanks, are used more frequently. The gammatone filterbanks address the limitations of the mel-filterbanks, where the former uses asymmetric filters to replace the computationally efficient triangular filters of the latter [39]. Gammatone filters are a linear approximation of the auditory filterbank found in the human ear. Gammatone Filterbank (GFBs) energies have been used for DNN acoustic model training [84]. For GFB feature extraction, the power of the bandlimited time signals within an analysis window of 26 ms is usually computed at a frame rate of 10 ms. The subband powers from 40 filters were then root compressed by using the 15th root.

8.3.3.3 Power-Normalized Cepstral Coefficients (PNCC)

PNCCs [62, 63] are a representative feature set that attempt to include many of the auditory processing attributes in a computationally efficient way. PNCC processing begins in traditional fashion with a short-time Fourier transform, with the outputs in each frame multiplied by gammatone frequency weighting, along a power-function nonlinearity, and generation of cepstral-like coefficients using DCT and mean normalization. For the most part, noise and reverberation suppression is accomplished by a nonlinear series of operations that perform running noise suppression and temporal contrast enhancement, respectively, working in a medium time context, with

analysis intervals on the order of 50150 ms. (The results of this longer-duration analysis are applied to signal representations extracted over traditional 20 to 35 ms analysis frames for speech recognition.) Multiple groups have found that PNCC processing provides effective noise robustness as well as suppression of reverberation effects, with minor modification, and the computation required is comparable to that used in MFCC and PLP feature extraction.

8.3.3.4 Modulation spectrum features

Modulation spectrum features incorporate low-pass filtering of critical bands with a cutoff frequency of 28 Hz and a subsequent AM band-pass filtering step [66]. The band-pass filter consists of a complex exponential function, which is windowed by a Hamming window and has its peak sensitivity at 4 Hz, matching the temporal characteristics of syllables. The filter emphasizes AM frequencies between approximately 0 and 8 Hz (i.e., the dominant AM range of speech) [36]. An approach to estimate the modulation spectrum of speech signals using the Hilbert envelopes in a nonparametric way was proposed in [112], where a modulation spectrum feature extracted from mel-filterbanks were used in an ASR task. That work showed that the logarithm of a particular mel-filterbanks Hilbert envelope over an analysis window of 100 ms produced a better amplitude modulation (AM) estimate of the subband signals compared to shorter window lengths. Lower DCT coefficients (in the range 0 to 25 Hz) of the AM signal were used as the acoustic feature, which was named as the fepstrum features. The fepstrum features performance was evaluated on the Conversational Telephony Speech (CTS) recognition experiments of the Switchboard (SWB) corpus, where the results indicated that such features in combination with short-term features, such as MFCCs, provided up to 2.5% absolute improvement in phone recognition accuracy and up to 2.5%3.5% absolute word recognition accuracy improvement on the 1.5 hour SWB test set [112].

8.3.3.5 Normalized Modulation Coefficient (NMC)

Studies [27, 38] have shown that amplitude modulation (AM) of the speech signal plays an important role in speech perception and recognition. Hence, recent studies [112, 92] have treated the speech signal as a sum of amplitude-modulated (AM) narrow-band signals. Demodulation of a narrow-band signal into its AM and frequency modulation (FM) components can be performed through the use of Discrete Energy Separation Algorithm (DESA) [71], which uses the nonlinear Teager Energy Operator (TEO) to perform the demodulation operation. TEO has been used in [57] to create mel-cepstral features that demonstrated robustness against car noise and improved ASR performance. The nonlinear DESA tracks the instantaneous AM energies quite reliably [71], which in turn provide better formant information [57] compared to conventional power spectrum-based approaches. Normalized Modulation Coefficient (NMC) was proposed in [77] that uses the DESA algorithm to ex-

tract instantaneous AM estimates for generating acoustic features. The significance of DESA is twofold: (a) it doesnt impose a linear model to analyze speech and (2) it tracks the frequency and amplitude variations at the sample level without imposing any stationary assumption as done by linear prediction or Fourier transform. For DESA to give good AM/FM estimates the input signal has to be sufficiently bandlimited [92]; for which a gammatone filter-bank was used in NMC feature extraction.

The TEO used in DESA was first introduced in [109] as a nonlinear energy operator, Ψ , that tracks the instantaneous energy of a signal, where the signals energy is defined to be a function of its amplitude and its frequency. Considering a discrete sinusoid $x[n]$, where A = const. amplitude, Ω = digital frequency, f = frequency of oscillation in hertz, f_s = sampling frequency in hertz, and θ = initial phase angle.

$$x[n] \;=\; A\cos[\,\Omega n+\theta\,];\; \Omega \;=\; 2\pi\left(\frac{f}{f_s}\right), \tag{8.1}$$

If $\Omega \leq \pi/4$ and sufficiently small, then Ψ takes the form

$$\Psi\{x[n]\} \;=\; x^2[n]-x[n-1]x[n+1] \;\approx\; A^2\Omega^2, \tag{8.2}$$

where, the maximum energy estimation error in Ψ will be 23%, if $\Psi \;\leq\; \pi/4$. DESA was formulated in [71], where Ψ was used to formulate demodulation algorithm that can instantaneously separate the AM/FM components of a narrow-band signal using the following sets of equations

$$\Omega[n] \;\approx\; cos^{-1}\left\{1-\frac{\Psi(x[n])+\Psi(x[N-1])}{4\Psi(x[n])}\right\} \tag{8.3}$$

$$|a_i[n]| \approx \sqrt{\frac{\Psi\{x[n]\}}{1-cos(\Omega_i[n])^2}} \tag{8.4}$$

Note that in 8.2, $x^2[n]-x[n-1]x[n+1]$ can be less than zero if $x^2[n] < x[n-1]x[n+1]$, while the right hand side is strictly non-negative, $A^2\Omega^2 \geq 0$, hence in [77] the TEO in 8.2 was modified to

$$\Psi\{x[n]\} \;=\; \left|x^2[n]-x[n-1]x[n+1]\right| \;\approx\; A^2\Omega^2, \tag{8.5}$$

which tracks the magnitude of energy changes. Also, the AM/FM signals computed from 8.3 and 8.4 may contain discontinuities (that substantially increase their dynamic range). To prevent such discontinuities the AM estimation equation 8.4 was modified in NMC feature extraction, as detailed in [77]. Fig. 8.1 shows the overlaying plot of the windowed narrow-band time signal and its corresponding AM magnitude.

At this point the question remains, how robust is the TEO to different noisy condition, to answer that we need to revisit 8.5 and consider a noisy bandlimited signal $s[n] = x[n] + v[n]$, the TEO, $\Psi s[n]$ is given as

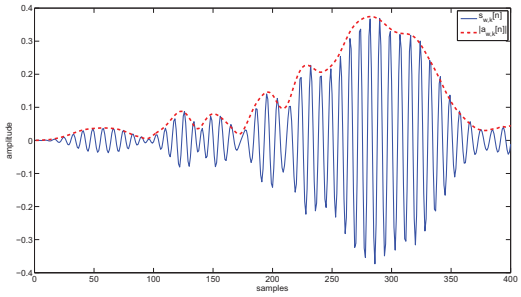

Fig. 8.1: A windowed narrow-band speech signal (in blue) and its corresponding AM signal (red) from the modified DESA algorithm.

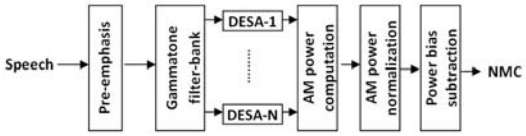

Fig. 8.2: Flow-diagram of NMC feature extraction from speech.

$$\Psi\{s[n]\} \;=\; \Psi\{x[n]\} + \Psi\{v[n]\} + \widetilde{\Psi}\{x[n]\,v[n]\} \tag{8.6}$$

where, $\widetilde{\Psi}\{x[n]v[n]\} = x[n]v[n] - (1/2)x[n-1]v[n+1] - (1/2)v[n-1]x[n+1]$ is the cross TEO of $x[n]$ and $v[n]$. If we assume the subband noise $v[n]$ to be zero mean and additive, then the expected value of the cross term $\widetilde{\Psi}\{x[n]v[n]\}$ is zero, resulting in

$$E\left[\Psi\{s[n]\}\right] = E\left[\Psi\{x[n]\}\right] + E\left[\Psi\{v[n]\}\right] \tag{8.7}$$

If we assume that the noise is hi-pass in every subband then using the low pass filtering results in $E[\Psi\{v[n]\}] << E[\Psi\{x[n]\}]$, and Ω^2 is almost constant for narrow band signals, results in

$$E[\Psi\{s[n]\}] \;\approx\; E[\Psi\{x[n]\}]\; hence,\; E[A_s^2] \;\approx\; E[A_x^2] \tag{8.8}$$

Where A_s represents the instantaneous amplitude of the noisy signal $s[n]$ and A_x represents the same for the clean signal $x[n]$. Thus 8.8 indicates how the estimated AM signals are robust to noise corruption.

The steps involved in obtaining the NMCC features are shown in Figure 8.2. At the onset, speech signal is pre-emphasized (using a pre-emphasis filter) and then analyzed using 26ms hamming window with 10ms frame rate. The windowed speech signal $s^w[n]$ is passed through a gamma-tone filter-bank having 40 channels between 200Hz and 7500Hz (for 16kHz signal). The AM time signals $a_{k,j}[n]$ for k^{th} chan-

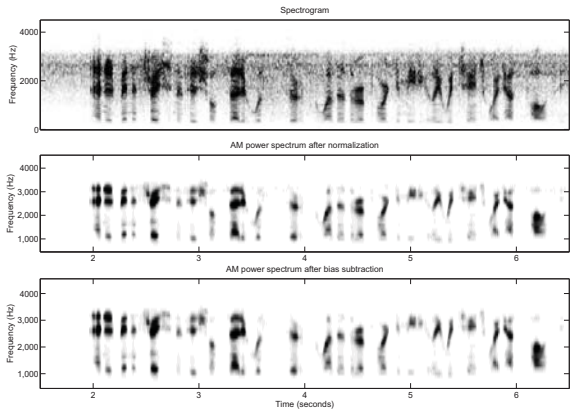

Fig. 8.3: Spectrogram of a noisy utterance corrupted with 15.6dB non-stationary noise and its normalized and bias subtracted amplitude modulation spectrum.

nel and j^{th} frame are then obtained for each of the 40 channels using the modified DESA algorithm.

The normalized AM powers were then bias subtracted using a similar approach as specified in [63]. Figure 8.3 shows the spectrogram of a noise corrupted signal, its normalized AM power spectrum and its corresponding bias subtracted version. $1/15^{th}$ root power compression is performed on the bias subtracted AM power spectrum and the resultant is typically used as the NMC feature set.

8.3.3.6 Modulation of Medium Duration Speech Amplitudes (MMeDuSA)

Note that, given 8.5 in section 8.3.3.6 a simpler approach to estimate the instantaneous AM signal can be devised by assuming that the instantaneous FM signal will be approximately equal to the center frequency of the analysis gammatone filterbank when the subband signals are sufficiently bandlimited, i.e.,

$$\Omega_i \approx f_c \tag{8.9}$$

Given 8.9, the estimation of the instantaneous AM signal from 8.5 becomes very simple

$$A_i \approx \sqrt{\frac{|x^2[n] - x[n-1]x[n+1]|}{\Omega_i^2}} \tag{8.10}$$

This simplification is essentially used in obtaining the MMeDuSA feature [78], as shown in Figure 8.4. In the MMeDuSA pipeline the speech signal is first pre-emphasized and then analyzed using a Hamming window of 51ms with a 10ms

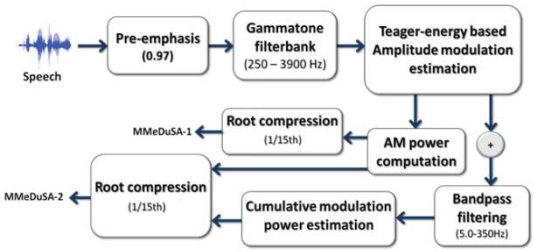

Fig. 8.4: MMeDuSA1 and MMeDuSA2 feature extraction pipeline [78].

frame rate. The windowed speech signal $s[n]$ is passed through a gammatone filterbank having 40 critical bands, with center frequencies spaced equally in the equivalent rectangular bandwidth (ERB) scale between 150 Hz and 7400 Hz. For each of these 40 subband signals, their AM signals are computed using 8.10. The power of the estimated AM signals was computed followed by non-linear dynamic range compression (using $1/15^{th}$ root). For a given analysis window, 40 power coefficients were obtained and these are identified as the MMeDuSA1 features. Note that the feature operates in medium duration as it uses an analysis window of size 52 ms compared to the traditionally used 10-25ms windows. In parallel, each of the 40 estimated AM signals (as shown in Figure 8.4) were band-pass filtered using DCT, retaining information only within 5 Hz to 350 Hz. These are the medium duration modulations (represented as: $a_{mod_{k,j}}[n]$), which were summed across the frequency scale to obtain a medium duration modulation summary

$$\overline{a_{mod_j}} = \sum_{k=1}^{N} a_{mod_{k,j}}[n] \tag{8.11}$$

The power signal of the medium duration modulation summary was obtained, followed by $1/15^{th}$ root compression. The resultant was transformed using DCT and the first n (typically 50%) coefficients were retained. These n coefficients were combined with the MMeDuSA1 feature, to produce a combined feature set named as the MMeDuSA2. Typically MMeDuSA2 is found to be more useful in ASR experiments and are usually termed as the MMeDuSA feature unless categorized specifically.

8.3.3.7 Two dimensional modulation extraction - Gabor features

Most of the approaches discussed so far extracted modulation information across time only, extracting such information across time and frequency scales was performed in [73]. A 2D Gabor filter was used in [73] to extract specific modulation frequencies from the spectro-temporal information of speech. The design of the Gabor features is motivated by Spectro-Temporal Receptive Fields (STRFs), which provide

an estimate of the stimulus that results in a high firing rate in isolated neurons [72]. It is observed that a significant proportion of the STRFs exhibit patterns that span durations of 200 ms, which is significantly longer than the traditionally used analysis durations in most speech features [73]. The Gabor/Tandem posterior features use a Multilayer Perceptron (MLP) to predict the monophone class posteriors of each frame by using the Gabor features as input; the posteriors were then Karhunen-Loeve transformed to 22 dimensions and appended with standard 39-dimensional MFCCs to yield 64-dimensional features. In [15], a convolutional neural network called the Gabor Convolutional Neural Network (GCNN) was proposed, which incorporated the Gabor functions into convolutional filter kernels. Features extracted using GCNNs showed significant performance improvement on noisy and channel-degraded speech over MFCC and other robust features such ETSI-AFE, PNCC, and Gabor-DNN posterior features.

8.3.3.8 Damped Oscillator Coefficients (DOC)

Studies have indicated that auditory hair cells exhibit damped oscillations in response to external stimuli [87] and such oscillations result in enhanced sensitivity and sharper frequency responses. To model such oscillations, a forced damped oscillator was propose in [76] to generate acoustic features for ASR systems. The simplest oscillator is a simple harmonic oscillator, which is neither driven nor damped and is defined by the following equation

$$m\frac{d^2x}{dt^2} + 2\zeta\omega_0 m\frac{dx}{dt} + \omega_0^2 mx = F_e(t) \tag{8.12}$$

where m is the mass of the oscillator; x is the position of the oscillator, ω_0 is the undamped angular frequency of the oscillator; and ζ is called the damping ratio. Assuming that the force can be represented as a sum of pulses, it can be shown that 8.12 can be written as

$$m\frac{d^2z(t)}{dt^2} + 2\zeta\omega_0 m\frac{dz(t)}{dt} + \omega_0^2 mz(t) = F_e e^{j\omega t} \tag{8.13}$$

where $z(t) = x(t) + jy(t)$ and represent $cos\omega t + jsin\omega t = e^{j\omega t}$. Equation 8.13 suggests that there exists a solution of the form $z(t) = z_0 e^{\zeta t}$, where $\frac{d^2z(t)}{dt^2} = \zeta^2 z_(0) e^{zetat}$ and $\frac{dz(t)}{dt} = \zeta z_0 e^{\zeta t}$

If $z(t)$ is a complex exponential with the same frequency as the applied force, then the displacement $x(t)$ will also vary as a sine or cosine with a frequency ω. It can be shown [76] that the amplitude of oscillation in response to a force at frequency ω is given as

$$|z_0| = \frac{\frac{F_e}{m}}{\sqrt{(\omega_0^2 - \omega^2)^2 + (2\zeta\omega_0\omega)^2}} \tag{8.14}$$

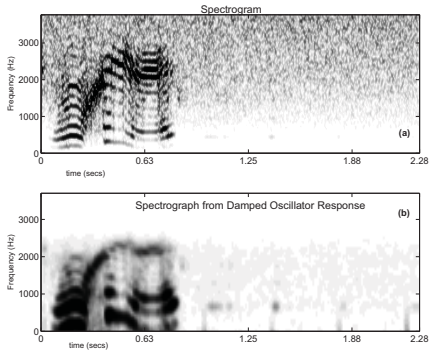

Fig. 8.5: (a) Spectrogram of signal corrupted with 3 dB noise and (b) Spectral representation of the damped oscillator response after gammatone filtering.

From 8.14, at resonance, i.e., $\omega_0 = \omega_t$, $|z_0|$ becomes

$$|z_0| = \frac{F_e}{2m\zeta\omega_0^2} \tag{8.15}$$

indicating that the bank of oscillators behave as a low pass filter, where it uses lower gains for high frequency bands and higher gains for the low frequency bands. To counter this effect we have selected m to be as follows

$$m = \frac{1}{2\zeta\omega_0^2} \tag{8.16}$$

Note that ω_0 and ζ can be user defined and for underdamped oscillation $\zeta < 1$. Modeling damped oscillator equation 8.12 in discrete time results in

$$x[n] = \frac{(2\zeta\Omega_0^2)F_e[n] + 2(1+\zeta\Omega_0)x[n-1] - x[n-2]}{(1+2\zeta\Omega_0+\Omega_0^2)} \tag{8.17}$$

where $\Omega_0 = \omega_0 T$ and $T = 1/f_s$.

The time response of the forced damped oscillators is obtained using 8.17 and their power over a hamming analysis window of 25.6 ms is computed. Figure 8.5 shows the spectrogram of a speech signal and the damped oscillator response; where the oscillator model is found to successfully retain the harmonic structure of speech while suppressing the background noise.

The DOC feature extraction block diagram is shown in Figure 8.6, where the damped oscillator response is computed using gammatone filterbank outputs as forcing functions. In DOC processing, speech signal is pre-emphasized and then analyzed using a 25.6 ms Hamming window with a 10 ms frame rate. The windowed speech signal is passed through a gammatone filterbank having 40 channels with cutoff frequencies at 200 Hz to 7000 Hz (for 16 kHz). The damped oscillator re-

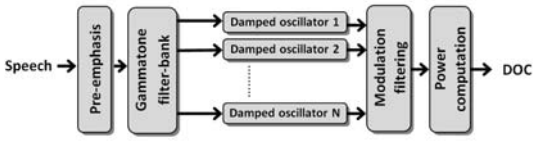

Fig. 8.6: Block diagram of the damped oscillator coefficient (DOC) feature extraction.

sponse is smoothed using a modulation filter with cutoff frequencies at 0.9 Hz and 100 Hz, which helps to reduce the background subband noise. The powers of the resulting time signals are computed which are then root compressed ($1/15^{th}$ root) and the resulting 40 dimensional feature is used as the DOC features.

8.3.4 Current Trends

Recent advances in deep learning technology have redefined the common strategies used in acoustic modeling for Automatic Speech Recognition (ASR) systems, where Gaussian Mixture Model (GMM)-based models have been replaced by Deep Neural Network (DNN)-based models. DNNs [85, 102, 64] have demonstrated significant improvement in speech recognition performance compared to their GMM-HMM counterparts. Given the versatility of the DNN systems, [120] stated that speaker-normalization techniques, such as Vocal Tract Length Normalization (VTLN) [122], do not significantly improve speech recognition accuracy, as the DNN architectures rich multiple projections through multiple hidden layers enable it to learn a speaker-invariant data representation. The current state-of-the-art architectures also deviate significantly from the traditional cepstral representation to simpler spectral representations, where MFCCs are usually replaced by Mel-Filterbank Energy (MFB) features. While the basic assumptions in GMM-HMM architectures necessitated uncorrelated features due to their widely used diagonal covariance design (which in turn forced the observation to undergo a decorrelation step using the widely popular Discrete Cosine Transform (DCT)), the current paradigm makes no such assumption. Rather, the neural network architectures are known to benefit from cross-correlations [81] and hence demonstrate better performance by using spectral features rather than their cepstral versions [103]. Recent studies [103, 23] demonstrated that DNNs work very well for noisy speech and improve performance significantly compared to GMM-HMM systems. Convolutional Neural Networks (CNNs) [2, 50] have been found to perform as well as or sometimes better than the fully connected DNN architectures [1]. CNNs are expected to be noise robust [2], especially in those cases where noise/distortion is localized in the spectrum. With CNNs, the localized convolution filters across frequency tend to normalize the spectral variations in speech arising from vocal tract length differences, enabling the CNNs to learn speaker-invariant data representations. Recent results [84, 83, 74] also showed that

CNNs are more robust to noise and channel degradations than DNNs.

In CNN/DNN-based ASR systems, speaker adaptation is usually done by using a generative framework that involves transforming features to a different space by using transforms such as Feature-Space Maximum Likelihood Linear Regression (fMLLR) [32] or by applying a speaker-dependent bias by appending features like i-vectors [21, 98]. However, using i-vectors is often problematic, especially in mismatched conditions [90], where careful pre-processing, such as segmentation and additional architectural enhancements, may be required [34]. The i-vector framework was first developed for speaker verification as a way of summarizing the information in a variable-length utterance into a single fixed-length vector.

Much research has been performed on exploring and advancing feature-space adaptation methods, such as Feature-Space Maximum Likelihood Linear Regression (fMLLR) approaches for GMM-HMM models. fMLLR applies a linear transform to the feature vectors for every frame, where the transform parameters are estimated by optimizing an auxiliary Q-function. DNN models are typically adapted by providing fMLLR transformed features as input. Use of fMLLR features has several advantages: firstly, it is efficient, as a few iterations of Expectation-Maximization (EM) usually suffice; secondly, estimation of the fMLLR transform is quite robust even with a very limited adaptation data; thirdly, it is quite versatile as it can be applied to both supervised setting, where it is more robust to transcription errors than using a discriminative criterion, and to unsupervised setting where reference transcription is not available.

Seide et al., [101] investigated the effectiveness of applying feature transforms developed for GMM-HMM, including HLDA, VTLN, and fMLLR, to Context-Dependent Deep Neural Network HMMs, or CD-DNN-HMMs. The authors observed that unsupervised speaker adaptation with discriminatively estimated fMLLR like transforms works nearly as well as fMLLR for GMM-HMMs. Rath et al., [94] explored various methods of providing higher-dimensional features to DNNs, while still applying speaker adaptation with fMLLR of low dimensionality. The best-observed features consist of the baseline 40-dimensional speaker-adapted features that have been spliced again, followed by de-correlation and dimensionality reduction using another Linear Discriminant Analysis (LDA). The authors believe that the whitening transform performed by LDA on the features will be favorable for DNN training, as the LDA would work as a pre-conditioner of the data, enabling setting higher learning rates, leading to faster learning (especially when pre-training is not used) [94]. Parthasarathi et al., [88] investigated fMLLR for DNN adaptation and proposed early fusion and late fusion to improve fMLLR performance, where early fusion with a bottleneck can act as a strong regularizer, and late fusion can provide significant robustness when fMLLR estimation is noisy.

In [43], Stacked Bottleneck (SBN) neural network architecture was proposed to cope with limited data from a target domain, where the SBN net was used as a feature extractor. The SBN system was used to deal with unseen languages in [43] and, in [61], was extended to cope with unseen reverberation conditions. Unseen data conditions can significantly impact the performance of DNN ASR systems, and either supervised or unsupervised data adaptation is typically used to overcome such

problems. In most such cases, labeled adaptation data is used to adapt the acoustic model (i.e., the DNN), with typically L2 [68] regularization employed. However, such approaches may veer the acoustic model away from the initial training acoustic conditions; hence in [121] a Kullback-Leibler divergence (KLD) regularization was proposed for DNN model parameter adaptation, which differs from the typically used L2 regularization in the sense that it constrains the model parameters themselves rather than the output probabilities. Using such an approach, the model learns new acoustic conditions without digressing from what it had learned from the initial training data.

Recent results [95] showed that a single DNN can be trained to learn both feature extraction and phonetic classification. [111] proposed directly using the raw time signal as input to a DNN, and several others [53, 97, 12] have explored different ways to process the raw waveform and to train an acoustic model from it. In [97], using the raw signal resulted in better recognition performance than from using conventional acoustic features. In a different study [12], conventional acoustic features were appended with DNN-generated features from the raw waveform, and the combination produced better performance than the conventional acoustic features did alone. While several research efforts have proposed different ways to learn data-driven feature-extraction processes through DNN training, an open question remains about how robust such approaches are to unseen data conditions.

8.4 Case Studies

8.4.1 Speech Processing for Noise- and Channel- Degraded Audio

Voice activity detection (VAD) is an essential stage of any ASR system. If a segment is not detected by VAD, it will not be processed by ASR, leading to word deletion errors. The performance of this stage greatly affects the quality of the final ASR hypothesis [35]. Robust features have been explored in recent years for the task of speech activity detection (VAD), in large part motivated by the challenges posed by noisy datasets [41]. Several robust features (such as PNCC, NMC etc.) were explored for VAD in a DNN-based framework in [42] and it was observed that the fusion of all these features gave the best performance across different conditions. In [110], FDLP, rate-scale features demonstrated significant robustness for VAD task on noise and channel degraded data.

DNN acoustic models using traditional MFB or MFCC features have been observed to suffer performance loss when the evaluation data is different than the training data [84]. Robust features are typically found to improve the performance of DNN/CNN acoustic models [84, 15]. In [74], baseline MFBs were compared with respect to MMeDuSA, NMC, and DOC features in a time-frequency CNN (TFCNN) acoustic model-based Aurora-4 noisy word-recognition task [51], and the

results demonstrated (see Table 8.1) a relative 5% reduction in WER compared to the baseline MFB features. Multiple robust features can be used in combination to provide a multi-view representation of the acoustic signal, and these combinations typically improve recognition performance [75, 79].

Table 8.1: WER (averaged across all conditions) on multi-conditioned training task of Aurora-4 (16 kHz) from using different features.

Features	Avg. WER (%)
MFB	9.4
NMC	9.0
DOC	8.9
MMeDuSA	9.2

Robust features are found to be quite useful for performing keyword spotting (KWS) in mismatched training-testing conditions. Table 8.2 shows the performance of a CNN-based keyword-agnostic KWS system for Farsi datasets from the DARPA-RATS KWS evaluation conditions. Performance is given in terms of average probability of false alarm [P(fa)] between 15% and 50% probability of miss [P(miss)]. As is evident from Table 8.2, the robust features demonstrated much better performance than the MFB features. Beyond the good individual performance of the robust features for KWS, the availability of multiple features enables creating systems that potentially capture complementary information, which in turn can be leveraged to provide even better results through system fusion [30].

Table 8.2: Performance from different feature sets for Farsi KWS system from SRIs DARPA RATS submission.

	P(fa)	
Features	P(miss) = 15% to 50%	P(miss) = 15%
MFB	0.060	0.675
NMC	0.057	0.474
DOC	**0.054**	0.413
MMeDuSA	0.057	**0.389**

Besides the good individual performance of the robust features in KWS, the availability of multiple features opens up the opportunity to create multiple systems that can potentially capture complementary information which in turn can be leveraged to provide even better results through system fusion [30].

8.4.2 Speech Processing under Reverberated Conditions

Robust acoustic features resistant to reverberation artifacts have shown significant promise in DNN acoustic models. Reverberation introduces mostly temporal distortion, where temporal smearing of information occurs whose duration is dependent on the impulse response of the room where the speech is recorded. The REVERB-2014 challenge [67] presented results from several research groups that used inverse filtering, NMF, modulation based features, i-vectors, and several other methods to greatly improve performance of DNN-based acoustic models under reverberated conditions [22, 117]. The results from REVERB-2014 indicate that sufficiently augmenting the training data with reverberation conditions similar to the evaluation conditions significantly improves ASR performance (e.g., [22]) showed an average relative reduction of 20% in WER through data augmentation).

The impact of training-testing data-condition mismatch was investigated in the ASpIRE [45] evaluation, where the training data consisted of the full Fisher training data [16], and the evaluation data contained reverberated speech recorded by far-field microphones. ASpIRE was a Large-Vocabulary Continuous Speech Recognition (LVCSR) evaluation, where speech recognition robustness was investigated in a variety of acoustic environments and recording scenarios without having any knowledge about such conditions in the training and development data [45]. In ASpIRE, the participants were allowed to augment the training data by artificially introducing reverberation and/or noise into the training data. Data augmentation was found to be useful across all systems submitted to the challenge [54, 90, 82]. Speech enhancement using maximum kurtosis dereverberation reduced the WER by 2.3% absolute [24]. A Time-Delay Neural Network (TDNN) using mel-cepstral features and i-vectors was presented in [90], where longer temporal information processing through the time-delay layer was found to be crucial for dealing with reverberation. A Time-Frequency CNN (TFCNN) was presented in [82], which gave better performance than traditional CNN and DNNs, with the use of robust features found to be useful. Table reftab:3 shows the results from baseline Gammatone Filterbank (GFB) features and DOC features, where DOC features, owing to their long temporal memory, were found to be robust against reverberation corruption.

Table 8.3: WER from the ASpIRE dev set using GFB and DOC features, with different acoustic models

Features	Acoustic Models	Avg. WER (/%)
GFB	DNN	47.3
DOC	DNN	42.6
DOC	CNN	41.4
DOC	TFCNN	40.7

In [54, 61], an autoencoder-based enhancement approach was proposed, where the role of the autoencoder was to de-noise and de-reverberate the degraded speech.

Table 8.4: WER from DNN acoustic models trained with WDAS beamformed signals using baseline and noise-robust features for Chime-3 real evaluation data

Features	Real-Test WER (%)
MFB	20.17
DOC	18.53
MMeDuSA	18.27
DOC+fMLLR	15.28
MMeDuSA+fMLLR	14.96

In addition, FDLP and stacked-bottleneck features were also used in [54] along with DNN adaptation and data augmentation, which resulted in significant improvement compared to the baseline system.

In a recent study [52], robust features were used on top of multi-microphone beamforming-based dereverberation in the Chime-3 challenge, where significant reduction in error rate was observed compared to using mel-filterbank features. Beamforming techniques such as Weighted Delay-And-Sum (WDAS) and Minimum Variance Distortionless Response (MVDR) are popular methods for leveraging multi-microphone data for coping with reverberation artifacts, and studies [24, 22] have shown impressive ASR performance under reverberated conditions when beamforming is used. The gain from robust features after beamforming in [52] was quite encouraging (see Table 8.4), as the results indicate that further performance improvement can be achieved from multi-microphone beamforming-based solutions when robust features are used.

8.5 Conclusion

Use of robust features have helped in improving acoustic model performance under mismatched training testing conditions across different flavors of deep learning architectures. In recent speech recognition evaluations it has been overwhelmingly witnessed that while the DNN models produce state-of-the-art results under matched training-testing condition; they are susceptible to performance degradation when the testing conditions are grossly mismatched from the training conditions. Traditional approaches such as data augmentation, adaptation has been found to be quite useful in data mismatched cases, enabling the models to deal with unseen data conditions. Robust features typically aim to create an invariant representation of speech, such that data perturbation has minimal effect on its feature space, hence providing a reliable feature representation to the acoustic models. Use of robust features has been beneficial on top of data augmentation and adaptation steps. The design of the signal processing steps in acoustic feature engineering has been largely motivated by signal theoretic approaches or speech perception studies, where several different realizations of speech signal processing have been explored and evaluated. The human auditory processing is a complex interaction of several nonlinear processes, such as

auditory attention, temporal filtering, masking etc., and on top of that it allows information to flow in both bottom-up and top-down direction, providing human listeners the capability to deal with varying acoustic conditions and extremely quick adaptation skills. Researchers in auditory neuroscience and psychoacoustics have been actively investigating the different mechanisms of human auditory perception and their mutual interactions; such observations may provide promising future directions for speech feature engineering which can potentially lead to more versatile and robust acoustic features.

The surge of raw-signal processing in recent years have revolutionized the way speech scientists and technologists used to think about speech systems. The current trend replaces the signal processing frontend as an ad-hoc step to an integrated acoustic modeling step where the neural networks are made learn both signal decomposition and phonetic discrimination all in one step using common objective criteria. Raw-signal based approaches are usually found to be data-hungry where several hundred (preferably a thousand or more) hours of training is necessary to reliably learn the frontend transformation. The draw-back of such system is the requirement of computational resources as the traditional acoustic models are no longer using encoded/compressed feature forms, but are using information that are 5 to 10 times or more in size. Also learning the frontend in a data driven way may result in over-fitting the model to the training acoustic conditions, hence one may require a significant amount of diverse acoustic data to train acoustic models that can generalize well across unseen acoustic conditions. Given the recent impressive results from raw-signal based systems, more and more researchers are investigating alternative models for raw-signal based acoustic modeling which renders the optimism that some of the drawbacks of raw-signal processing systems may be addressed in the near future making such systems integral part of our ASR systems.

References

1. Abdel-Hamid, O., Deng, L., Yu, D.: Exploring convolutional neural network structures and optimization techniques for speech recognition. In: INTERSPEECH, pp. 3366–3370 (2013)
2. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Penn, G.: Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4277–4280. IEEE (2012)
3. Atal, B.S., Hanauer, S.L.: Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America **50**(2B), 637–655 (1971)
4. Athineos, M., Ellis, D.P.: Frequency-domain linear prediction for temporal features. In: Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on, pp. 261–266. IEEE (2003)
5. Athineos, M., Hermansky, H., Ellis, D.P.: Lp-trap: Linear predictive temporal patterns. Tech. rep., IDIAP (2004)
6. Barker, J., Marxer, R., Vincent, E., Watanabe, S.: The thirdchime'speech separation and recognition challenge: Dataset, task and baselines. In: 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015) (2015)

7. Bartels, C., Wang, W., Mitra, V., Richey, C., Kathol, A., Vergyri, D., Bratt, H., Hung, C.: Toward human-assisted lexical unit discoverywithout text resources. In: SLT (2016)
8. Beh, J., Ko, H.: A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, vol. 1, pp. I–648. IEEE (2003)
9. Bell, P., Gales, M., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M., et al.: The mgb challenge: Evaluating multi-genre broadcast media recognition. Proc. of ASRU, Arizona, USA (2015)
10. Benesty, J., Makino, S.: Speech enhancement. Springer Science & Business Media (2005)
11. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. Unsupervised and Transfer Learning Challenges in Machine Learning **7**, 19 (2012)
12. Bhargava, M., Rose, R.: Architectures for deep neural network based acoustic models defined over windowed speech waveforms. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
13. Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. Acoustics, Speech and Signal Processing, IEEE Transactions on **27**(2), 113–120 (1979)
14. Bregman, A.S.: Auditory scene analysis: The perceptual organization of sound. MIT press (1994)
15. Chang, S.Y., Morgan, N.: Robust cnn-based speech recognition with gabor filter kernels. In: INTERSPEECH, pp. 905–909 (2014)
16. Cieri, C., Miller, D., Walker, K.: The fisher corpus: a resource for the next generations of speech-to-text. In: LREC, vol. 4, pp. 69–71 (2004)
17. Cohen, J.R.: Application of an auditory model to speech recognition. The Journal of the Acoustical Society of America **85**(6), 2623–2629 (1989)
18. Cooke, M., Green, P., Josifovski, L., Vizinho, A.: Robust automatic speech recognition with missing and unreliable acoustic data. Speech communication **34**(3), 267–285 (2001)
19. Davis, K., Biddulph, R., Balashek, S.: Automatic recognition of spoken digits. The Journal of the Acoustical Society of America **24**(6), 637–642 (1952)
20. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech and Signal Processing, IEEE Transactions on **28**(4), 357–366 (1980)
21. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on **19**(4), 788–798 (2011)
22. Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Hori, T., Nakatani, T., et al.: Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge. In: Proc. REVERB Workshop (2014)
23. Deng, L., Hinton, G., Kingsbury, B.: New types of deep neural network learning for speech recognition and related applications: An overview. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8599–8603. IEEE (2013)
24. Dennis, J., Dat, T.H.: Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2r's system description for the aspire challenge. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 518–524. IEEE (2015)
25. Doc, E.S.: speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms. ETSI ES **201**(108), ver 1.1.3 (2003)
26. Doc, E.S.: speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced frontend feature extraction algorithm; compression algorithms. ETSI ES **202**(050), ver 1.1.5 (2007)
27. Drullman, R., Festen, J.M., Plomp, R.: Effect of reducing slow temporal modulations on speech reception. The Journal of the Acoustical Society of America **95**(5), 2670–2680 (1994)

28. Elliott, T.M., Theunissen, F.E.: The modulation transfer function for speech intelligibility. PLoS comput biol **5**(3), e1000,302 (2009)
29. Fine, S., Saon, G., Gopinath, R.A.: Digit recognition in noisy environments via a sequential gmm/svm system. In: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, vol. 1, pp. I–49. IEEE (2002)
30. Fiscus, J.G.: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In: Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on, pp. 347–354. IEEE (1997)
31. Flynn, R., Jones, E.: Combined speech enhancement and auditory modelling for robust distributed speech recognition. Speech Communication **50**(10), 797–809 (2008)
32. Gales, M.J., Woodland, P.C.: Mean and variance adaptation within the mllr framework. Computer Speech & Language **10**(4), 249–264 (1996)
33. Ganapathy, S., Thomas, S., Hermansky, H.: Temporal envelope compensation for robust phoneme recognition using modulation spectrum. The Journal of the Acoustical Society of America **128**(6), 3769–3780 (2010)
34. Garimella, S., Mandal, A., Strom, N., Hoffmeister, B., Matsoukas, S., Parthasarathi, S.H.K.: Robust i-vector based adaptation of dnn acoustic model for speech recognition. In Proceedings of Interspeech (2015)
35. Gelly, G., Gauvain, J.L.: Minimum word error training of rnn-based voice activity detection. ISCA Interspeech (submitted) (2015)
36. Gemmeke, J.F., Virtanen, T.: Noise robust exemplar-based connected digit recognition. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp. 4546–4549. IEEE (2010)
37. Ghitza, O.: Auditory nerve representation as a front-end for speech recognition in a noisy environment. Computer Speech & Language **1**(2), 109–130 (1986)
38. Ghitza, O.: On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. The Journal of the Acoustical Society of America **110**(3), 1628–1640 (2001)
39. Gibson, J., Van Segbroeck, M., Narayanan, S.S.: Comparing time-frequency representations for directional derivative features. In: INTERSPEECH, pp. 612–615 (2014)
40. Giegerich, H.J.: English phonology: An introduction. Cambridge University Press (1992)
41. Graciarena, M., Alwan, A., Ellis, D., Franco, H., Ferrer, L., Hansen, J.H., Janin, A., Lee, B.S., Lei, Y., Mitra, V., et al.: All for one: feature combination for highly channel-degraded speech activity detection. In: INTERSPEECH, pp. 709–713. Citeseer (2013)
42. Graciarena, M., Ferrer, L., Mitra, V.: The sri system for the nist opensad 2015 speech activity detection evaluation. In: in review (2016)
43. Grezl, F., Egorova, E., Karafiát, M.: Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure. In: Spoken Language Technology Workshop (SLT), 2014 IEEE, pp. 48–53. IEEE (2014)
44. Gustafsson, H., Nordholm, S.E., Claesson, I.: Spectral subtraction using reduced delay convolution and adaptive averaging. Speech and Audio Processing, IEEE Transactions on **9**(8), 799–807 (2001)
45. Harper, M.: The automatic speech recognition in reverberant environments (aspire) challenge. ASRU (2015)
46. Harvilla, M.J., Stern, R.M.: Histogram-based subband powerwarping and spectral averaging for robust speech recognition under matched and multistyle training. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4697–4700. IEEE (2012)
47. Hermansky, H.: Perceptual linear predictive (plp) analysis of speech. the Journal of the Acoustical Society of America **87**(4), 1738–1752 (1990)
48. Hermansky, H., Morgan, N.: Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on **2**(4), 578–589 (1994)
49. Hermansky, H., Sharma, S.: Temporal patterns (traps) in asr of noisy speech. In: Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, vol. 1, pp. 289–292. IEEE (1999)

50. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE **29**(6), 82–97 (2012)
51. Hirsch, G.: Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task. ETSI STQ Aurora DSR Working Group (2002)
52. Hori, T., Chen, Z., Erdogan, H., Hershey, J.R., Roux, J., Mitra, V., Watanabe, S.: The merl/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition. In: Proc. IEEE ASRU (2015)
53. Hoshen, Y., Weiss, R.J., Wilson, K.W.: Speech acoustic modeling from raw multichannel waveforms. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 4624–4628. IEEE (2015)
54. Hsiao, R., Ma, J., Hartmann, W., Karafiat, M., Grézl, F., Burget, L., Szoke, I., Cernocky, J., Watanabe, S., Chen, Z., et al.: Robust speech recognition in unknown reverberant and noisy conditions. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (2015)
55. Itakura, F.: Minimum prediction residual principle applied to speech recognition. Acoustics, Speech and Signal Processing, IEEE Transactions on **23**(1), 67–72 (1975)
56. Itakura, F., Saito, S.: Statistical method for estimation of speech spectral density and formant frequencies. Electronics & Communications in Japan **53**(1), 36 (1970)
57. Jabloun, F., Cetin, A.E., Erzin, E.: Teager energy based feature parameters for speech recognition in car noise. Signal Processing Letters, IEEE **6**(10), 259–261 (1999)
58. Joris, P., Schreiner, C., Rees, A.: Neural processing of amplitude-modulated sounds. Physiological reviews **84**(2), 541–577 (2004)
59. Juang, B.H., Rabiner, L.R.: Automatic speech recognition–a brief history of the technology development. Encyclopedia of Language and Linguistics (2005)
60. Kanedera, N., Arai, T., Hermansky, H., Pavel, M.: On the importance of various modulation frequencies for speech recognition. In: Fifth European Conference on Speech Communication and Technology (1997)
61. Karafiát, M., Grézl, F., Burget, L., Szöke, I., Černockỳ, J.: Three ways to adapt a cts recognizer to unseen reverberated speech in but system for the aspire challenge. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
62. Kim, C., Stern, R.M.: Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In: ICASSP, pp. 4574–4577 (2010)
63. Kim, C., Stern, R.M.: Power-normalized cepstral coefficients (pncc) for robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4101–4104. IEEE (2012)
64. Kingsbury, B., Sainath, T.N., Soltau, H.: Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
65. Kingsbury, B., Saon, G., Mangu, L., Padmanabhan, M., Sarikaya, R.: Robust speech recognition in noisy environments: The 2001 ibm spine evaluation system. In: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, vol. 1, pp. I–53. IEEE (2002)
66. Kingsbury, B.E., Morgan, N., Greenberg, S.: Robust speech recognition using the modulation spectrogram. Speech communication **25**(1), 117–132 (1998)
67. Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R.: The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, pp. 1–4. IEEE (2013)
68. Li, X., Bilmes, J.: Regularized adaptation of discriminative classifiers. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, vol. 1, pp. I–I. IEEE (2006)

69. Lyon, R.F.: A computational model of filtering, detection, and compression in the cochlea. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82., vol. 7, pp. 1282–1285. IEEE (1982)
70. Makhoul, J., Cosell, L.: Lpcw: An lpc vocoder with linear predictive spectral warping. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76., vol. 1, pp. 466–469. IEEE (1976)
71. Maragos, P., Kaiser, J.F., Quatieri, T.F.: Energy separation in signal modulations with application to speech analysis. Signal Processing, IEEE Transactions on **41**(10), 3024–3051 (1993)
72. Mesgarani, N., David, S., Shamma, S.: Representation of phonemes in primary auditory cortex: how the brain analyzes speech. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol. 4, pp. IV–765. IEEE (2007)
73. Meyer, B.T., Ravuri, S.V., Schädler, M.R., Morgan, N.: Comparing different flavors of spectro-temporal features for asr. In: INTERSPEECH, pp. 1269–1272 (2011)
74. Mitra, V., Franco, H.: Time-frequency convolutional networks for robust speech recognition. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 317–323. IEEE (2015)
75. Mitra, V., Franco, H.: Coping with unseen data conditions: Investigating neural net architectures, robust features, and information fusion for robust speech recognition. In: in review (2016)
76. Mitra, V., Franco, H., Graciarena, M.: Damped oscillator cepstral coefficients for robust speech recognition. In: INTERSPEECH, pp. 886–890 (2013)
77. Mitra, V., Franco, H., Graciarena, M., Mandal, A.: Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4117–4120. IEEE (2012)
78. Mitra, V., Franco, H., Graciarena, M., Vergyri, D.: Medium-duration modulation cepstral feature for robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 1749–1753. IEEE (2014)
79. Mitra, V., van Hout, J., Wang, W., Bartels, C., Franco, H., Vergyri, D., et al.: Fusion strategies for robust speech recognition and keyword spotting for channel- and noise-degraded speech. In: Interspeech, 2016 (2016)
80. Mitra, V., Hout, J.V., McLaren, M., Wang, W., Graciarena, M., Vergyri, D., Franco, H.: Combating reverberation in large vocabulary continuous speech recognition. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
81. Mitra, V., Nam, H., Espy-Wilson, C.Y., Saltzman, E., Goldstein, L.: Retrieving tract variables from acoustics: a comparison of different machine learning strategies. Selected Topics in Signal Processing, IEEE Journal of **4**(6), 1027–1045 (2010)
82. Mitra, V., Van Hout, J., Wang, W., Graciarena, M., McLaren, M., Franco, H., Vergyri, D.: Improving robustness against reverberation for automatic speech recognition. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 525–532. IEEE (2015)
83. Mitra, V., Wang, W., Franco, H.: Deep convolutional nets and robust features for reverberation-robust speech recognition. In: Spoken Language Technology Workshop (SLT), 2014 IEEE, pp. 548–553. IEEE (2014)
84. Mitra, V., Wang, W., Franco, H., Lei, Y., Bartels, C., Graciarena, M.: Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions. In: INTERSPEECH, pp. 895–899 (2014)
85. Mohamed, A.r., Dahl, G.E., Hinton, G.: Acoustic modeling using deep belief networks. Audio, Speech, and Language Processing, IEEE Transactions on **20**(1), 14–22 (2012)
86. Moore, B.: An introduction to the psychology of hearing (1989)
87. Neiman, A.B., Dierkes, K., Lindner, B., Han, L., Shilnikov, A.L., et al.: Spontaneous voltage oscillations and response dynamics of a hodgkin-huxley type model of sensory hair cells. Journal of mathematical neuroscience **1**(11), 11 (2011)

88. Parthasarathi, S.H.K., Hoffmeister, B., Matsoukas, S., Mandal, A., Strom, N., Garimella, S.: fmllr based feature-space speaker adaptation of dnn acoustic models. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
89. Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M.: Complex sounds and auditory images. Auditory physiology and perception **83**, 429–446 (1992)
90. Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., Khudanpur, S.: Jhu aspire system: Robust lvcsr with tdnns, i-vector adaptation, and rnn-lms. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (2015)
91. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of INTERSPEECH (2015)
92. Potamianos, A., Maragos, P.: Time-frequency distributions for automatic speech recognition. IEEE Transactions on Speech and Audio Processing **9**(3), 196–200 (2001)
93. Rabiner, L.R., Levinson, S.E., Rosenberg, A.E., Wilpon, J.G.: Speaker-independent recognition of isolated words using clustering techniques. Acoustics, Speech and Signal Processing, IEEE Transactions on **27**(4), 336–349 (1979)
94. Rath, S.P., Povey, D., Veselỳ, K., Cernockỳ, J.: Improved feature processing for deep neural networks. In: INTERSPEECH, pp. 109–113 (2013)
95. Sainath, T.N., Kingsbury, B., Mohamed, A.r., Ramabhadran, B.: Learning filter banks within a deep neural network framework. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pp. 297–302. IEEE (2013)
96. Sainath, T.N., Mohamed, A.r., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for lvcsr. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8614–8618. IEEE (2013)
97. Sainath, T.N., Weiss, R.J., Senior, A., Wilson, K.W., Vinyals, O.: Learning the speech front-end with raw waveform cldnns. In: Proc. Interspeech (2015)
98. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker adaptation of neural network acoustic models using i-vectors. In: ASRU, pp. 55–59 (2013)
99. Schroeder, M.R.: Recognition of complex acoustic signals. Life Sciences Research Report **5**(324), 130 (1977)
100. Schwarz, P.: Phoneme recognition based on long temporal context (2009)
101. Seide, F., Li, G., Chen, X., Yu, D.: Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pp. 24–29. IEEE (2011)
102. Seide, F., Li, G., Yu, D.: Conversational speech transcription using context-dependent deep neural networks. In: Interspeech, pp. 437–440 (2011)
103. Seltzer, M.L., Yu, D., Wang, Y.: An investigation of deep neural networks for noise robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7398–7402. IEEE (2013)
104. Seneff, S.: A joint synchrony/mean-rate model of auditory speech processing. In: Readings in speech recognition, pp. 101–111. Morgan Kaufmann Publishers Inc. (1990)
105. Shao, Y., Srinivasan, S., Jin, Z., Wang, D.: A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech & Language **24**(1), 77–93 (2010)
106. Srinivasan, S., Wang, D.: Transforming binary uncertainties for robust speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on **15**(7), 2130–2140 (2007)
107. Stevens, S.S., Volkmann, J., Newman, E.B.: A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America **8**(3), 185–190 (1937)
108. Tchorz, J., Kollmeier, B.: A model of auditory perception as front end for automatic speech recognition. The Journal of the Acoustical Society of America **106**(4), 2040–2050 (1999)
109. Teager, H.M.: Some observations on oral air flow during phonation. Acoustics, Speech and Signal Processing, IEEE Transactions on **28**(5), 599–601 (1980)

110. Thomas, S., Saon, G., Van Segbroeck, M., Narayanan, S.S.: Improvements to the ibm speech activity detection system for the darpa rats program. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp. 4500–4504. IEEE (2015)
111. Tüske, Z., Golik, P., Schlüter, R., Ney, H.: Acoustic modeling with deep neural networks using raw time signal for lvcsr. In: INTERSPEECH, pp. 890–894 (2014)
112. Tyagi, V.: Fepstrum features: Design and application to conversational speech recognition. Tech. rep., IBM Research Report (2011)
113. Van Hout, J.: Low complexity spectral imputation for noise robust speech recognition (2012)
114. Viemeister, N.F.: Temporal modulation transfer functions based upon modulation thresholds. The Journal of the Acoustical Society of America **66**(5), 1364–1380 (1979)
115. Virag, N.: Single channel speech enhancement based on masking properties of the human auditory system. Speech and Audio Processing, IEEE Transactions on **7**(2), 126–137 (1999)
116. Wang, D.: On ideal binary mask as the computational goal of auditory scene analysis. In: Speech separation by humans and machines, pp. 181–197. Springer (2005)
117. Weninger, F., Watanabe, S., Le Roux, J., Hershey, J., Tachioka, Y., Geiger, J., Schuller, B., Rigoll, G.: The merl/melco/tum system for the reverb challenge using deep recurrent neural network feature enhancement. In: Proc. REVERB Workshop (2014)
118. Yoshioka, T., Ragni, A., Gales, M.J.: Investigation of unsupervised adaptation of dnn acoustic models with filter bank input. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 6344–6348. IEEE (2014)
119. Yost, W.A., Moore, M.: Temporal changes in a complex spectral profile. The Journal of the Acoustical Society of America **81**(6), 1896–1905 (1987)
120. Yu, D., Seltzer, M.L., Li, J., Huang, J.T., Seide, F.: Feature learning in deep neural networks-studies on speech recognition tasks. arXiv preprint arXiv:1301.3605 (2013)
121. Yu, D., Yao, K., Su, H., Li, G., Seide, F.: Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7893–7897. IEEE (2013)
122. Zhan, P., Waibel, A.: Vocal tract length normalization for lvcsr. Tech. rep., Tech. Rep. CMU-LTI-97-150. Carnegie Mellon University (1997)
123. Zhu, Q., Stolcke, A., Chen, B.Y., Morgan, N.: Incorporating tandem/hats mlp features into sris conversational speech recognition system. In: Proc. DARPA Rich Transcription Workshop (2004)