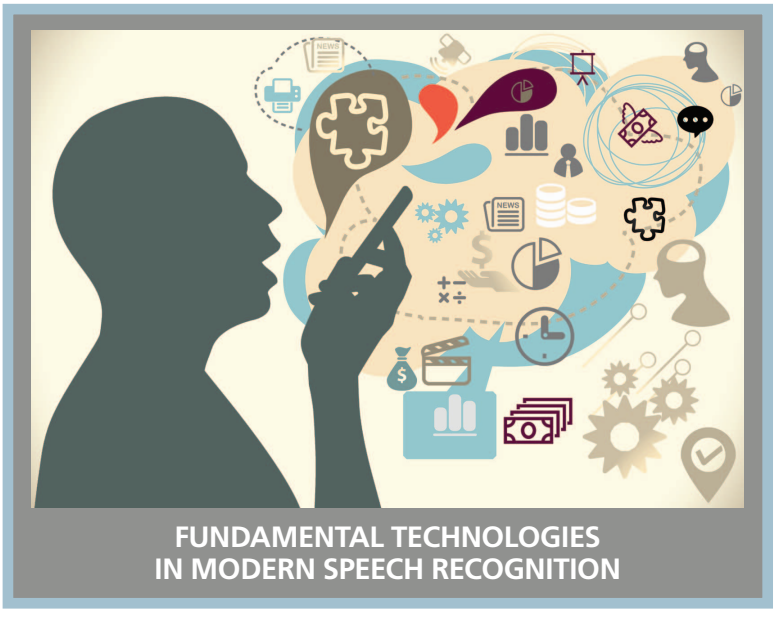
Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita,
Roland Maas, Tomohiro Nakatani, and Walter Kellermann

Making Machines Understand Us in Reverberant Rooms

Robustness against reverberation for automatic speech recognition



Speech recognition technology has left the research laboratory and is increasingly coming into practical use, enabling a wide spectrum of innovative and exciting voice-driven applications that are radically changing our way of accessing digital services and information. Most of today's applications still require a microphone located near the talker. However, almost all of these applications would benefit from distant-talking speech capturing, where talkers are able to speak at some distance from the microphones without the encumbrance of handheld or body-worn equipment [1]. For example, applications such as meeting speech recognition, automatic annotation of consumer-generated videos, speech-to-speech translation in teleconferencing, and hands-free interfaces for controlling consumer-products, like interactive TV, will greatly benefit from distant-talking operation. Furthermore, for a number of unexplored but important applications, distant

Digital Object Identifier 10.1109/MSP.2012.2205029
Date of publication: 15 October 2012

1053-5888/12/$31.00©2012IEEE

microphones are a prerequisite. This means that distant-talking speech recognition technology is essential for extending the availability of speech recognizers as well as enhancing the convenience of existing speech recognition applications.

The major problem in distant-talking speech recognition is the corruption of speech signals by both interfering sounds and reverberation caused by the large speaker-to-microphone distance. A range of successful techniques have been developed since the beginnings of speech recognition research to combat the additive and convolutional noise caused by interfering sounds, microphone mismatch, and the characteristics of transmission networks, e.g., [2]–[6]. The effects of those noise types are limited to a single frame of short-time signal analysis. In contrast, the effect of reverberation spans a number of consecutive time frames and thus requires dedicated approaches. Although several pioneering efforts were made, e.g., [7]–[10], compensating for such long-term distortion is very challenging and has not gained wide attention until recently. However, since a significant amount of reverberation is present in many practically relevant environments, overcoming the reverberation problem is paramount for realizing dependable distant-talking speech recognizers.

SINCE A SIGNIFICANT AMOUNT OF REVERBERATION IS PRESENT IN MANY PRACTICALLY RELEVANT ENVIRONMENTS, OVERCOMING THE REVERBERATION PROBLEM IS PARAMOUNT FOR REALIZING DEPENDABLE DISTANT-TALKING SPEECH RECOGNIZERS.

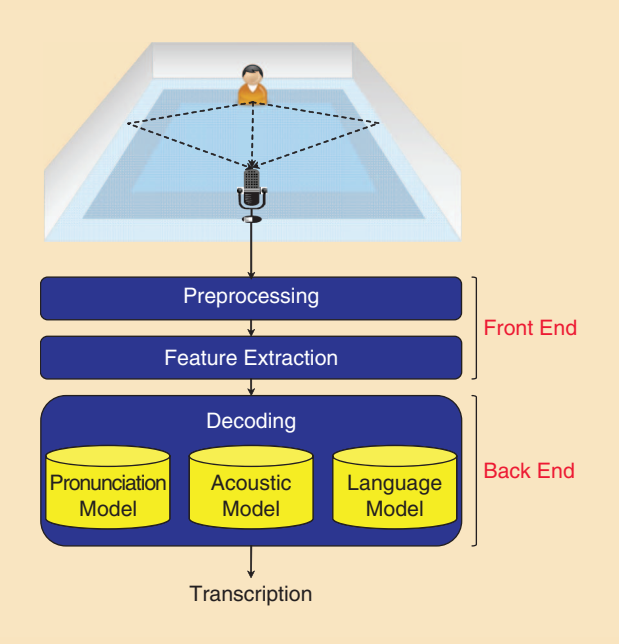
In recent years, research on reverberant speech processing has achieved significant progress in both the fields of audio processing and speech recognition, mainly driven by multidisciplinary approaches combining ideas from room acoustics, optimal filtering, machine learning, speech modeling, enhancement, and recognition. In audio processing, a number of dereverberation techniques have been developed, inspired by advanced signal processing techniques, such as blind deconvolution and nonnegative matrix factorization. In speech recognition, promising techniques have emerged, combining ideas from noise robustness methods and novel ways for modeling reverberant data. These recent studies are revealing the fundamental problem in reverberant speech recognition and are establishing promising lines of research using various approaches.

This article provides a systematic review of the state of the art of reverberant speech processing from the viewpoint of speech recognition. We endeavor to establish a taxonomy of approaches, thereby highlighting their similarities and differences. After briefly presenting basic principles of room acoustics and automatic speech recognition and discussing the fundamental problem in reverberant speech recognition, we provide a review of state-of-the-art techniques with some illustrative experimental results. Rather than generic robustness approaches, we focus on techniques dedicated to reverberation that exploit speech and reverberation models.

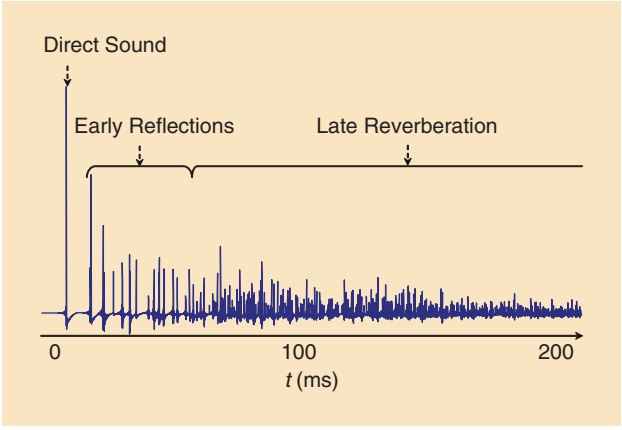
REVERBERATION AND ITS EFFECT ON SPEECH RECOGNIZERS

Figure 1 illustrates a speech recognition system in a reverberant room, where speech is captured with one or more distant microphones. While traveling from the speaker's mouth to the microphones, the wavefront of the speech is repeatedly reflected at the walls and other objects in the room. These reflections, perceived as reverberation, alter the acoustic characteristics of the original speech. The speech signals captured by the microphones are fed into the recognizer, whose processing steps are usually grouped into two broad units: the front end and the back end.

The goal of reverberant speech recognition is to correctly transcribe speech corrupted by reverberation regardless of the severity of the reverberation. The question that immediately arises is: "Why do we need fundamentally new techniques for handling reverberation?" To answer this question, this section describes the fundamental problem in reverberant speech recognition and discusses the properties of reverberation that can be leveraged to overcome the problem. Then, we explain why established speech recognition techniques cannot effectively cope with these properties, which shows the need for dedicated solutions. These solutions change the way that speech is processed in the front end and/or back end.



[FIG1] **Speech recognition system in reverberant environments. Front-end-based approaches remove reverberation from a speech signal or its features while back-end-based ones change an acoustic model or a decoding algorithm.**



[FIG2] **Room impulse response generated by the image method [12]. Room impulse responses actually measured in rooms are usually more noisy due to microphone characteristics, bandlimitation with analog-to-digital conversion, and measurement errors.**

ELEMENTS OF ROOM ACOUSTICS

Let us start our discussion by describing reverberation mathematically. The repeated sound reflections in a room create a sequence of numerous slowly decaying copies of the original sound, which is perceived as reverberation. The process of reverberating a speech signal can be represented as a linear convolution of the speech signal and a room impulse response [11]. The room impulse response represents the acoustic reaction of the room in response to an impulsive sound and describes the changes that the speech signal radiated from the speaker's mouth undergoes when traveling to the microphone in the room. Thus, when the clean speech signal, the reverberant speech signal, the room impulse response, and the additive background noise are denoted by $x(t)$, $y(t)$, $h(t)$, and $d(t)$, respectively, $y(t)$ is written as

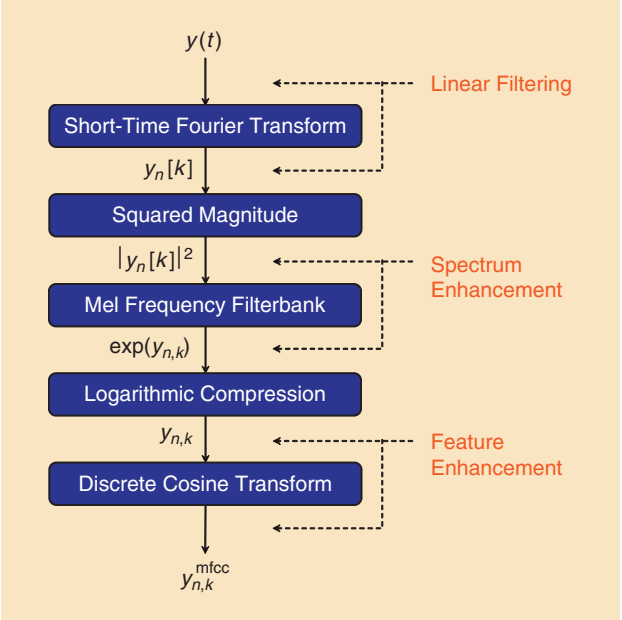
$$y(t) = \sum_{\tau=0}^{T_h} h(\tau)x(t-\tau) + d(t) = h(t) \circledast x(t) + d(t), \quad (1)$$

where $\circledast$ stands for a linear convolution and T_h is the length of the room impulse response. The room impulse response depends on the positions of the speaker and the microphone and thus changes when the speaker moves.

In some parts of this article, we neglect the additive noise $d(t)$ to focus on reverberation although we need to consider the presence of both additive noise and reverberation for actual distant-talking scenarios. Therefore, for most reverberation compensation approaches, we discuss extensions for dealing jointly with additive noise and reverberation.

The room impulse response can be divided into three portions as shown in Figure 2 [11]. After the arrival of the direct sound, several strong reflections, called early reflections, occur within 50 ms. After that comes a series of numerous indistinguishable reflections, called late reverberation. The characteristics of the early reflections depend strongly on the speaker and microphone positions. By contrast, the magnitude of the late reverberation decays approximately exponentially and the decay rate is independent of the positions. The time required for the late reverberation to decay by 60 dB relative to the level of direct sound is called the reverberation time T_{60}. For typical office and home environments, the reverberation time ranges from 200 to 1,000 ms. The insensitivity of the late reverberation magnitude to the speaker and microphone positions can be exploited to develop algorithms that are robust against speaker movement. All the approaches reviewed in this article except for linear filtering take advantage of this property because they do not rely on the phase information as we will see later.

THE INSENSITIVITY OF THE LATE REVERBERATION MAGNITUDE TO THE SPEAKER AND MICROPHONE POSITIONS CAN BE EXPLOITED TO DEVELOP ALGORITHMS THAT ARE ROBUST AGAINST SPEAKER MOVEMENT.

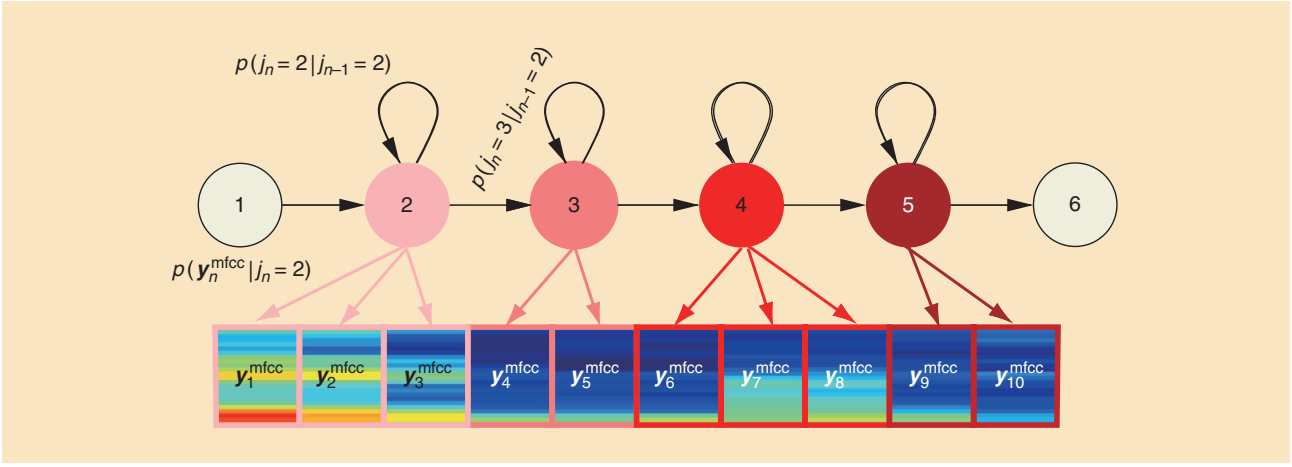


[FIG3] **Flow chart of front-end feature extraction and three front-end-based approaches.**

Therefore, it is useful to represent the early reflections and late reverberation separately. We denote the combined portion consisting of the direct sound and early reflections by $h_i(t)$ and the late reverberation by $h_l(t)$ so that we have

$$h_i(t) = \begin{cases} h(t) & \text{if } t < \Delta \\ 0 & \text{otherwise} \end{cases} \quad h_l(t) = \begin{cases} h(t+\Delta) & \text{if } t \geq 0 \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where Δ is the boundary between the early reflections and late reverberation, which is typically around 50 ms after the arrival of the direct sound.



[FIG4] **Example of an HMM.**

The energy ratio of the combined portion consisting of the direct sound and the early reflections to the late reverberation is measured by C_{50} [11], and is highly correlated with speech recognition performance [13]. This ratio mainly depends on the distance between the speaker and the microphone.

ELEMENTS OF AUTOMATIC SPEECH RECOGNITION

We briefly review basic principles of automatic speech recognition in the following. As shown in Figure 1, the front end of a recognizer converts the observed speech signal $y(t)$ to a sequence of feature vectors, $(\boldsymbol{y}_n^{\text{mfcc}})_{n\in\mathbb{T}}$, where $\mathbb{T}$ denotes an observation period. Figure 3 shows a block diagram for extracting Mel-frequency cepstral coefficients (MFCCs), which are typically used for speech recognition. The back end transcribes the feature vector sequence by searching the sentence ω^* that maximizes the posterior sentence probability, i.e.,

$$\omega^* = \operatorname*{argmax}_{\omega} p((\boldsymbol{y}_n^{\text{mfcc}})_{n\in\mathbb{T}}|\omega)p(\omega), \tag{3}$$

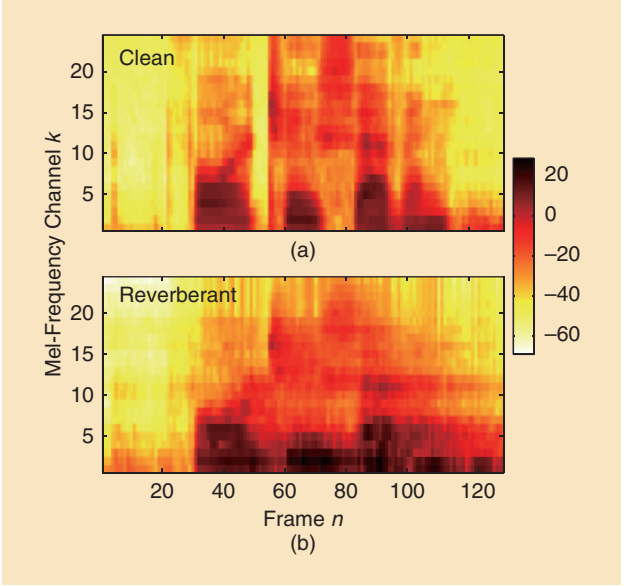
where $p(\omega)$, called the language model, defines a probability over sequences of words. On the other hand, $p((\boldsymbol{y}_n^{\text{mfcc}})_{n\in\mathbb{T}}|\omega)$, called the acoustic model, defines the conditional probability density function (pdf) of observing the feature vector sequence $(\boldsymbol{y}_n^{\text{mfcc}})_{n\in\mathbb{T}}$ given sentence ω is uttered. The acoustic model is usually realized by hidden Markov models (HMMs), which assume that there is an underlying sequence of discrete states, $(j_n)_{n\in\mathbb{T}}$, as shown in Figure 4. Using HMMs, $p((\boldsymbol{y}_n^{\text{mfcc}})_{n\in\mathbb{T}}|\omega)$ can be expressed as [14]

$$p((\boldsymbol{y}_n^{\text{mfcc}})_{n\in\mathbb{T}}|\omega) = \sum_{(j_n)_{n\in\mathbb{T}}} \prod_{n\in\mathbb{T}} p(\boldsymbol{y}_n^{\text{mfcc}}|j_n)p(j_n|j_{n-1},\omega). \tag{4}$$

Equation (4) is based on the conditional independence assumption and the first-order Markov assumption. As illustrated in Figure 4, the conditional independence assumption means that the current feature vector depends only on the current state while the first-order Markov assumption means that the current state depends only on the previous state. The pdf $p(\boldsymbol{y}_n^{\text{mfcc}}|j)$ and the probability $p(j|i,\omega)$ are called the emission pdf and the transition probability, respectively. The emission pdf is usually described by a Gaussian mixture model (GMM). In modern speech recognizers, sentence-level HMMs corresponding to $p((\boldsymbol{y}_n^{\text{mfcc}})_{n\in\mathbb{T}}|\omega)$ are created by concatenating phoneme-level HMMs during recognition. The parameters of the HMMs are trained using a speech corpus.

FUNDAMENTAL PROBLEM IN REVERBERANT SPEECH RECOGNITION

The effect of reverberation on feature vector sequences is illustrated in Figure 5, which compares clean and reverberant log Mel-frequency filterbank features extracted from an utterance "four, two, seven." Each log Mel-frequency filterbank feature is a frequency-warped and dimension-reduced version of each spectral bin obtained by short-time Fourier transform (STFT). The disparity between clean and reverberant data is obvious.



[FIG5] **Log Mel-frequency filterbank features corresponding to the utterance "four, two, seven" in dB color scale, extracted from (a) clean and (b) reverberant speech.**

Specifically, reverberation causes a temporal smearing of features as can be observed, e.g., around frame 55 in Figure 5, where the reverberation of the vowel /ao/ in "four" masks the short period of silence before the plosive /t/ in "two." This dispersive effect occurs because the room impulse response length T_h is much larger than the frame length of the short-time signal analysis used for feature extraction, and hence the time-domain convolution (1) cannot be described as multiplicative noise in the Mel-frequency filterbank domain.

REVERBERATION CAUSES A TEMPORAL SMEARING OF FEATURES.

To facilitate the explanation, let us recast the task of reverberant speech recognition as the more familiar task of recognizing speech convolved by a short impulse response and corrupted by additive noise. Using the notation of (2), the observed reverberant signal $y(t)$ is written as

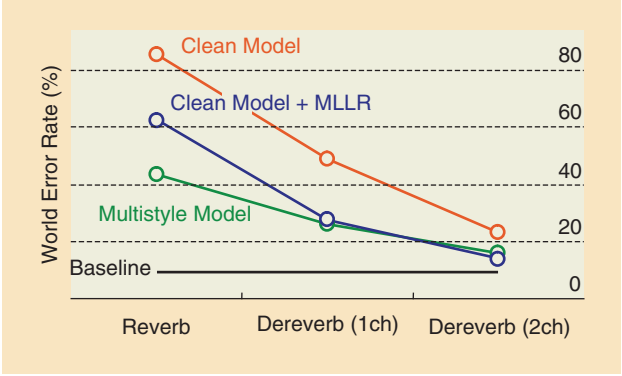
$$y(t) = h_i(t) \circledast x(t) + h_l(t) \circledast x(t - \Delta) = h_i(t) \circledast x(t) + r(t), \tag{5}$$

where $r(t)$ is the late reverberation component of $y(t)$. Some methods compensate only for the effect of the late reverberation instead of removing both the early reflections and late reverberation. This is because the effect of the early reflections, which is described as the convolution of a short impulse response $h_i(t)$, may be mitigated by standard techniques, such as cepstral mean normalization [14]. In many studies, an approximation is made where $r(t)$ is uncorrelated to $x(t)$, allowing us to consider the late reverberation $r(t)$ as additive noise. This approximation is made mainly for the sake of mathematical tractability, but it is partly justified by the fact that the autocorrelation coefficients of a clean speech signal are very small for time lags greater than 50 ms. Thus, the reverberant speech recognition task has been recast into the familiar problem of recognizing speech convolved by the short impulse response $h_i(t)$ and corrupted by additive noise $r(t)$.

The problem that makes reverberant speech recognition so challenging and that distinguishes it from the task of noisy speech recognition is the extreme nonstationarity of the late reverberation $r(t)$. The reason for these substantial time variations is that $r(t)$ results from filtering a delayed clean speech signal. This nonstationarity renders almost all noise robustness techniques, including parallel model combination (PMC) [4] and vector Taylor series (VTS) compensation [5], ineffective for combatting late reverberation because they usually assume stationary or slowly varying noise to make noise parameter estimation and compensation possible at a reasonable computational cost.

The main difference between reverberation and interfering speakers, which cause similarly nonstationary noise, is that the late reverberation $r(t)$ at time t can be predicted from past observations $(y(\tau))_{\tau \le t}$ of the reverberant signal, since both $r(t)$ and $y(t)$ are filtered copies of the clean signal $x(t)$. This implies the possibility of compensating for the late reverberation with high accuracy by leveraging the long-term relationship. Therefore, a reverberant speech recognizer must account for the long-term acoustic context, i.e., it has to exploit the sequence of reverberant feature vectors preceding the current feature vector.

The necessity of accounting for the long-term acoustic context implies that training-based approaches, such as matched training or multistyle training alone are insufficient to successfully cope with reverberation as shown by the experimental results in Figures 6 and 9. While matched training uses data recorded in the target room to estimate the parameters of the HMMs, multistyle training learns the HMM parameters from a vast amount of reverberant speech data collected in different rooms. Since environments identical or similar to the target room are expected to be included in the training environments, the mismatch between training and deployment environments is reduced. However, the HMMs used in speech recognition systems assume that a feature vector at the current time frame depends only on the current state and not on previous feature vectors. This conditional independence assumption between neighboring feature vectors prevents the HMMs from effectively modeling the statistical dependencies between reverberant feature vectors extending over several hundred milliseconds, though dynamic features [15] and extended feature vectors [16] may partly alleviate this drawback by increasing the temporal coverage of each feature vector. Capturing such long-term relationships with HMMs requires a correspondingly long left context for each HMM state. However, the left context provided by the commonly used triphone HMMs (see, e.g., [14]) is not sufficient for this purpose. Increasing the left context by using polyphones could alleviate this problem in principle. However, the enormous number of such polyphone models that would be needed to describe the reverberant data



[FIG6] **Word error rates (WERs) obtained with the long-term linear prediction [21] in a conference room with T_{60} = 0.78 s for the 20,000-word *The Wall Street Journal* (*WSJ*) task. The microphones were placed at 2 m from the speaker. While the "reverb" results are obtained by passing unprocessed reverberant speech to a recognizer, "dereverb (1ch)" and "dereverb (2ch)" mean that reverberant speech was dereverberated with one and two microphones, respectively, before recognition. MLLR adaptation was performed with unsupervised learning.**

would make the reliable training of the HMM parameters extremely challenging, if not impossible. Consequently, techniques are required that deal with the long-term acoustic context more directly and thereby compensate for the effect of reverberation more effectively.

REVERBERATION IS A SUPERPOSITION OF NUMEROUS TIME-SHIFTED AND ATTENUATED VERSIONS OF A CLEAN SIGNAL SO THAT BOTH THE AMPLITUDES AND PHASES ARE USEFUL FOR DEREVERBERATION.

In the subsequent sections, we review such techniques. We classify them into a hierarchy of approaches and present representative methods based on each approach. As illustrated in Figure 1, there are two classes of approaches at the root of the hierarchy: front-end-based and back-end-based approaches. The front-end-based approaches attempt to remove the effect of the entire or late reverberation from observed feature vectors by leveraging the long-term acoustic context. This process yields an estimate of a clean feature vector sequence, such as that shown in Figure 5(a), which can be directly fed into a standard back end. By contrast, the back-end-based approaches change the acoustic model and decoder to deal directly with reverberant feature vectors, such as those shown in Figure 5(b). Therefore, models that represent the long-term relationships between reverberant feature vectors efficiently are necessary.

FRONT-END-BASED APPROACHES

The front-end-based approaches, which aim at dereverberating corrupted feature vectors, can be categorized into three classes according to where enhancement takes place in the chain of processing steps for feature extraction (see Figure 3). The first approach, called *linear filtering*, dereverberates time-domain signals or STFT coefficients. The second approach, called *spectrum enhancement*, dereverberates corrupted power spectra while ignoring the signal phases. The third approach, called *feature enhancement*, attempts to remove the effect of reverberation directly from the corrupted feature vectors. Detailed description of various front-end-based methods can be found in [17].

Before surveying the three approaches in more detail, we summarize the notations. The STFT of reverberant signal $y(t)$ is denoted by $y_n[k]$, where n represents the index of a time frame. The log Mel-frequency filterbank feature elements and the corresponding feature vectors are denoted by $y_{n,k}$ and $\boldsymbol{y}_n$, respectively. Note that variable k is commonly used when referring to a frequency bin and when mentioning an element of a feature vector. The various representations of $x(t)$ are defined similarly. Furthermore, j is used as index for HMM states.

Note that tremendous effort has been made to find different speech representations that are more robust against reverberation than standard features, like MFCCs [7], [8], [18], [19]. Although we restrict our review to MFCC-based speech recognizers, the quest for robust features is an important thread of research.

LINEAR FILTERING

The linear filtering approach attempts to remove the effect of reverberation in the time or STFT domain taking consecutive reverberant observations into account. Then the dereverberated time- or STFT-domain speech signal is transformed into features used by the back end. In contrast to the other two approaches, linear filtering exploits both the amplitudes and phases of the signal, which is advantageous in terms of accuracy because reverberation is a superposition of numerous time-shifted and attenuated versions of a clean signal so that both the amplitudes and phases are useful for dereverberation. In addition, taking the signal phases into account enables us to effectively exploit the acoustical differences between multiple microphone positions [20]. This means that linear filter-based dereverberation methods can be coupled with beamforming techniques to jointly achieve noise reduction and dereverberation. Although, for conciseness, we explain the algorithms of this approach by assuming a single microphone and the STFT representation in the following, many of the algorithms can be extended to benefit from multiple microphones as discussed, e.g., in [17].

To represent the relationship between clean and reverberant STFT coefficients, $x_n[k]$ and $y_n[k]$, the following representation is often assumed in the literature [21], [22]:

$$y_n[k] \approx \sum_{\tau=0}^{T} h_\tau[k]^* x_{n-\tau}[k], \tag{6}$$

where the superscript $*$ stands for complex conjugation and T is the number of time frames over which reverberation continues to have an effect. The complex conjugate of $h_n[k]$ is used for consistency with the notation commonly accepted in the field of adaptive filtering [23]. Equation (6) means that the effect of reverberation may be represented as a one-dimensional convolution in each frequency bin, and therefore sequence $(h_n[k])_{0\le n\le T}$ can be viewed as an STFT-domain counterpart of the time-domain room impulse response. Our objective is to recover the corresponding clean STFT coefficients $(x_n[k])_{n\in\mathbb{T}}$ for each k, given a sequence of reverberant STFT coefficients $(y_n[k])_{n\in\mathbb{T}}$. Below, we omit the frequency bin index k for conciseness.

As the name suggests, linear filtering methods employ a linear filter to perform dereverberation according to

$$x_n = \sum_{\tau=T_\perp}^{T_\top} g_\tau^* y_{n-\tau}, \tag{7}$$

where $G = \{g_\tau\}_{T_\perp\le\tau\le T_\top}$ is a set of adjustable linear filter coefficients. Generally, $T_\perp \le 0$ and $T_\top > 0$. The clean STFT coefficient x_n is estimated based on $T_\perp + T_\top + 1$ consecutive reverberant frames, and thus the linear filtering methods naturally allow us to take the long-term acoustic context into account. Our goal is to find an optimal filter $G^\dagger$ that cancels the room impulse

response h_n. Denoting the convolution of the room impulse response and the linear filter by $f_n = \sum_{\tau=T_\perp}^{T_\top} g_\tau^* h_{n-\tau}$, our objective is to set G so that f_n is nonzero if $n = 0$ and zero otherwise. This problem is called blind deconvolution and has been studied extensively, especially in the field of digital communications [23]. Different blind deconvolution methods for speech signals are discussed in [24] and [25].

In the following, we look more closely at blind deconvolution based on long-term linear prediction [21], [26]–[28]. It leverages an explicit speech model to determine the filter G. In one exemplary concept [21], which has been applied to various speech recognition tasks including meeting recognition [29], the speech model defines the pdf of a clean STFT coefficient x_n and is assumed to be a normal distribution with zero mean and variance θ_n. The time-varying modeling, i.e., the dependence on frame index n, of the variance was shown to play a critical role in precise adjustment of the filter coefficients [21]. Since $\Theta = \{\theta_n\}_{n\in\mathbb{T}}$ is unknown in advance, the filter G is optimized jointly with Θ by using the method of maximum likelihood. Specifically, the likelihood of the combination of G and Θ given the sequence $Y = (y_n)_{n\in\mathbb{T}}$ of observed reverberant STFT coefficients is maximized according to

$$(\hat{G}, \hat{\Theta}) = \operatorname*{argmax}_{(G,\Theta)} \log p(Y|G, \Theta). \tag{8}$$

To facilitate the definition of the pdf $p(Y|G, \Theta)$, the concept of multistep prediction [28] is introduced. With multistep prediction, we assume in (7) that $g_0 = 1$ and that $g_n = 0$ when $T_\perp \le n < T_\delta$, where T_δ is a positive integer that approximately corresponds to the boundary Δ between early reflections and late reverberation. This approach is called multistep prediction because, with these assumptions, (7) can be rewritten in the form of long-term T_δ-step forward prediction of y_n as

$$y_n = x_n + \sum_{\tau=T_\delta}^{T_\top} g_\tau^* y_{n-\tau}, \tag{9}$$

representing the current reverberant observation y_n as the sum of the clean signal x_n and a signal predicted from past observations with filter $G = \{g_n\}_{T_\delta \le n \le T_\top}$. The sign of g_t has been inverted when deriving (9) from (7). Thanks to the predictive form of (9), $p(Y|G, \Theta)$ can be easily defined and the optimization problem in (8) is finally rewritten as the following minimization problem:

$$(\hat{G}, \hat{\Theta}) = \operatorname*{argmin}_{(G,\Theta)} \sum_{n\in\mathbb{T}} \left(\frac{\left| y_n - \sum_{\tau=T_\delta}^{T_\top} g_\tau^* y_{n-\tau} \right|^2}{\theta_n} + \log \theta_n \right), \tag{10}$$

which can be solved by an iterative algorithm updating estimates of G and Θ alternately [21]. If multiple microphones are available, (9) is modified so that the current reverberant observation at a microphone is predicted from past observations from all the microphones, i.e., (9) is rewritten in the form of multichannel prediction [28], [21].

Figure 6 summarizes the performance of the long-term linear prediction method [21]. The evaluation results were obtained using a slightly different version of the synthetically reverberated 20,000-word *WSJ* task described in [30]. The differences between the experimental conditions employed in [30] and in this article are summarized as follows. While [30] downsampled the signals to 8 kHz and conducts experiments only for a single microphone, here we kept the sampling rate at 16 kHz and added experimental results for the two microphone case. In addition, we added experimental results obtained with multistyle training. To generate the multistyle training data, the original training set was split into six groups, and the speech signals of the different groups were convolved with room impulse responses collected in different rooms with T_{60}'s ranging from 0 to 1.3 s. With an acoustic model trained from the clean (i.e., noise- and reverberation-free) speech data, which were recorded with close-talking microphones, the word error rate (WER) exceeded 80%, when directly passing the reverberant speech to the recognizer. Performing dereverberation prior to recognition considerably reduced the WER. The dereverberation effect was salient especially when two microphones were used for two-channel long-term linear prediction. Unsupervised MLLR adaptation and multistyle training improved the speech recognition performance in all cases. The multistyle acoustic models used for recognizing dereverberated test utterances were trained on a set of dereverberated utterances as in the noise adaptive training approaches employed for conventional noise robust speech recognition tasks [6]. It is worthwhile noting that the performance gains provided by dereverberation were significant even with the adapted and multistyle acoustic models. This indicates that conventional approaches are insufficient to totally make up for the degradation caused by reverberation and that their effects are complementary to that of dereverberation. The WER obtained when using two microphones and the adapted clean acoustic model was 14.2%, which was slightly higher than the baseline clean speech WER of 9.7% obtained by using a clean acoustic model adapted with MLLR.

Although the results presented here were obtained with artificial data, the long-term linear prediction method has been successfully applied to actual meeting data [29]. Furthermore, by extending the observation pdf $p(Y|G, \Theta)$, this method can be modified to deal jointly with multiple speakers, additive background noise, and reverberation as described, e.g., in [31].

SPECTRUM ENHANCEMENT

As an alternative to linear filtering, enhancement may be performed after taking the squared magnitudes of the STFT coefficients. The objective of the resulting spectrum enhancement methods is to restore the clean power spectrum coefficients $(|x_n[k]|^2)_{n\in\mathbb{T}}$, given a sequence of the corresponding reverberant power spectrum coefficients $(|y_n[k]|^2)_{n\in\mathbb{T}}$. The advantage of spectrum enhancement over linear filtering is its high robustness against speaker movement, which derives from the fact that the magnitude of the late reverberation is largely insensitive to changes in speaker and microphone positions. Furthermore, spectrum enhancement methods can be easily

combined with conventional additive noise reduction techniques, such as spectral subtraction, as shown in [32].

The spectrum enhancement methods can be categorized into two classes according to the estimator of the reverberation power spectrum: moving-average estimator and predictive estimator. The moving-average estimator is based on the power spectrum-domain reverberation model given by

$$|y_n[k]|^2 \approx \sum_{\tau=0}^{T} |h_\tau[k]|^2 |x_{n-\tau}[k]|^2, \tag{11}$$

which is derived from (6) by disregarding the cross-terms between different time frames. To estimate the power spectrum of late reverberation or clean speech with this model, we need to know the power spectrum-domain representation $(|h_n[k]|^2)_{0\le n\le T}$ of the room impulse response. This can be achieved by techniques such as correlation analysis [33], nonnegative matrix factorization [34], and an iterative least squares method [35].

The predictive reverberation estimator employs a much simpler model [36, 32]. Assuming a strict exponential decay of the late reverberation magnitude, the power spectrum $|r_n[k]|^2$ of the late reverberation at frame n can be predicted from the power spectrum $|y_{n-T_\delta}[k]|^2$ of the reverberant observation at frame $n - T_\delta$ via a scalar predictor $a[k]$ as

$$|r_n[k]|^2 = a[k]|y_{n-T_\delta}[k]|^2. \tag{12}$$

T_δ is set at a value corresponding to approximately 50 ms. The predicted late reverberation is removed from the reverberant power spectrum $|y_n[k]|^2$ with spectral subtraction. The predictor $a[k]$ is determined based on the knowledge of T_{60}.

FEATURE ENHANCEMENT

Several methods perform enhancement after the application of a logarithmic compression, i.e., they attempt to directly dereverberate features extracted from a reverberant signal. The objective is to estimate the corresponding clean feature vector sequence $(\boldsymbol{x}_n)_{n\in\mathbb{T}}$, given a sequence of reverberant feature vectors $(\boldsymbol{y}_n)_{n\in\mathbb{T}}$. Therefore feature enhancement approaches often exploit pretrained models of clean features, facilitating the reduction of the mismatch between reverberant observations and clean acoustic models used for speech recognition. The pretrained models also provide a straightforward way to compute confidence scores for uncertainty decoding techniques discussed at the end of this section.

To infer the clean features from the reverberant features, we need to hypothesize a feature-domain model of reverberation. By analogy with the power spectrum-domain model (11), the effect of reverberation on log Mel-frequency filterbank feature vectors may be modeled as

$$\boldsymbol{y}_n \approx \log\left(\sum_{\tau=0}^{T} \exp(\boldsymbol{h}_\tau + \boldsymbol{x}_{n-\tau})\right), \tag{13}$$

where each element of $\boldsymbol{h}_n$ is an approximate feature-domain representation of a room impulse response [22]. With the above model, the method proposed in [22] sequentially estimates a clean feature at each time frame by employing an extended Kalman filter, where the clean feature is regarded as the underlying unknown state. Alternatively, feature enhancement can be achieved by estimating the late reverberation in the time, STFT, or power spectrum domain, extracting features from the estimate, and employing a feature-domain additive noise model as in [37]. Note that both the methods of [22] and [37] can be extended to jointly compensate for additive noise and reverberation by including an additive noise term in (13).

BACK-END-BASED APPROACHES AIM AT ADJUSTING THE PARAMETERS OF THE ACOUSTIC MODEL TO THE STATISTICAL PROPERTIES OF REVERBERANT FEATURE VECTORS OR AT TAILORING THE DECODER TO THE REVERBERANT FEATURE VECTORS.

Finally, it is worth noting that uncertainty decoding and missing feature techniques, which are widely used for noise-robust speech recognition [38], [39], can also be employed for reverberation-robust speech recognition. The process of enhancing feature vectors is inevitably imperfect. The degree of imperfection varies from frame to frame and from dimension to dimension: The accuracy of the enhanced features is high in some portions of a feature stream while it is low elsewhere. With uncertainty and missing feature decoding, we take this time-varying degree of confidence into account during decoding instead of directly passing the enhanced features to a standard back end. The application of these techniques to reverberant data is discussed in [22], [30], and [40].

BACK-END-BASED APPROACHES

Back-end-based approaches aim at adjusting the parameters of the acoustic model to the statistical properties of reverberant feature vectors or at tailoring the decoder to the reverberant feature vectors. In this section, we first review different methods for adapting the parameters of HMMs to reverberation. These methods are called *HMM adaptation*. They adjust the HMM parameters once before recognition, and the adapted HMMs are used to transcribe reverberant utterances with a conventional Viterbi decoder. While these methods achieve significant performance gains compared with clean HMMs, they are not optimal as they cannot efficiently exploit the long-term relationships that are characteristic of reverberant feature vector sequences. Therefore, alternative approaches explicitly modeling the strong interframe relations between reverberant feature vectors are discussed in the second part of this section.

HMM ADAPTATION

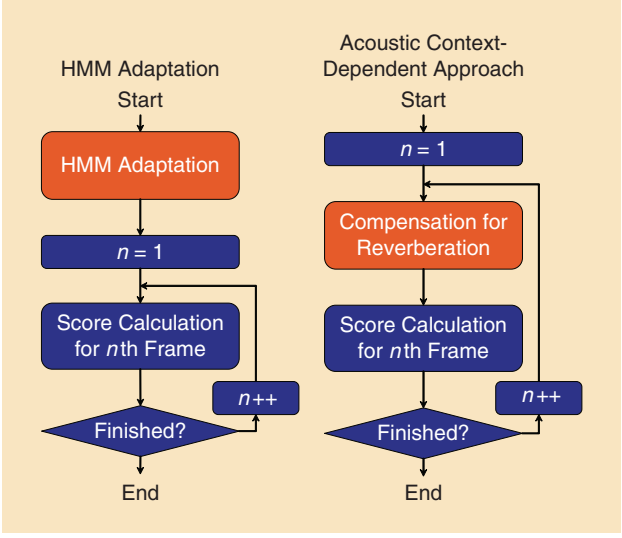
General adaptation schemes, developed for speaker or noise adaptation, including MLLR [3] and MAP adaptation [2], can

also be used to reduce the mismatch between clean HMMs and reverberant data. Although these techniques improve speech recognition performance in reverberant environments as shown in Figures 6 and 9, their performance gains are often insufficient. In addition, they require a relatively large amount of reverberant speech data to reliably adapt the HMM parameters to the reverberant environments.

By exploiting a reverberation model, the quantity of adaptation data can be reduced considerably as shown by the HMM adaptation schemes [41]–[43] tailored to reverberation. To simplify the discussion, we assume that the emission pdf of the clean HMMs is given by a single Gaussian distribution using static log Mel-frequency filterbank features, although these methods can employ GMMs for the emission pdfs and can handle dynamic features, such as delta cepstra. The methods [42], [43] propose adapting the mean vectors of the clean HMMs. Let $\boldsymbol{\mu}_j^X$ denote the clean HMM's mean vector for state j in the log Mel-frequency filterbank domain. The corresponding adapted mean vector $\boldsymbol{\mu}_j^Y$ is obtained by using the following state-level convolution:

$$\boldsymbol{\mu}_j^Y = \log\left(\sum_{\tau=0}^{j-1} \exp\left(\boldsymbol{\eta}_{j,j-\tau} + \boldsymbol{\mu}_{j-\tau}^X\right)\right), \tag{14}$$

where each element of $\boldsymbol{\eta}_{j,i}$ is a state-level reverberation representation describing the energy dispersion of state i to state j for the corresponding Mel-frequency channel. Equation (14) is the model-domain counterpart of the feature-domain reverberation model given by (13). While a maximum likelihood estimator based on a few calibration utterances with known transcriptions is used in [41] to determine $\boldsymbol{\eta}_{j,i}$, [42] assumes a strictly exponential energy decay for each channel and estimates the reverberation time T_{60} during recognition to obtain $\boldsymbol{\eta}_{j,i}$. If the adaptation approaches [41]–[43] are to be applied in noisy and reverberant environments, reverberation adaptation should be performed first, followed by noise adaptation, e.g., using PMC [4] or VTS [5]. Different HMM adaptation schemes tailored to noisy and reverberant data were proposed in [16] and [44] using extended feature vectors instead of the state-level convolution (14).



[FIG7] Comparison of HMM adaptation and acoustic context-dependent likelihood evaluation. "Compensation for reverberation" performs frame-by-frame adaptation [45] or optimization in (22) [46].

While the HMM adaptation methods achieve noticeable improvements in word accuracy compared to clean HMMs as shown in Figure 9, they are not optimal for reverberation-robust speech recognition for the same reason as matched and multistyle training—since the adapted HMMs still assume conditional independence between neighboring reverberant feature vectors, the long-term dependencies between consecutive reverberant feature vectors cannot be modeled appropriately.

This limitation can be stated more rigorously as follows. The heart of the Viterbi decoding algorithm used in conventional speech recognizers is the evaluation of the emission pdf $p_Y(\boldsymbol{y}_n|j)$, i.e., the likelihood of the observed feature vector $\boldsymbol{y}_n$ given state j. This specific form of the emission pdf implies that the state likelihood is evaluated independently of preceding reverberant feature vectors, explaining why HMMs cannot effectively account for the long-term acoustic context inherent in reverberant feature vector sequences. Capturing the long-term acoustic context requires an emission pdf of the form $p_Y(\boldsymbol{y}_n|j,(\boldsymbol{y}_\tau)_{\tau<n})$, where the dependency on preceding feature vectors is explicitly represented.

ACOUSTIC CONTEXT-DEPENDENT LIKELIHOOD EVALUATION

In this section, two representative methods for modeling the emission pdf $p_Y(\boldsymbol{y}_n|j,(\boldsymbol{y}_\tau)_{\tau<n})$ are discussed. The first method, frame-wise adaptation, modifies the HMM parameters on a frame-by-frame basis using past observations to model the dependency of the emission pdf on the acoustic context [45]. The second method, reverberation modeling for speech recognition (REMOS), decomposes a reverberant feature vector into contributions from the clean acoustic model and a reverberation model for each time frame and each state. Figure 7 contrasts the processing flows for the HMM adaptation schemes discussed in the previous section with that of the acoustic context-dependent likelihood evaluation methods discussed in the following.

The frame-wise model adaptation approach adjusts the means and covariance matrices of the HMMs at each frame depending on the preceding reverberant feature vectors [45]. Inspired by the time-domain representation of the early reflections and late reverberation of (2), the current reverberant feature vector $\boldsymbol{y}_n$ is approximated as

$$\boldsymbol{y}_n \approx \log\left(\exp(\boldsymbol{h} + \boldsymbol{x}_n) + \exp(\boldsymbol{r}_n)\right), \tag{15}$$

where $\boldsymbol{h}$ and $\boldsymbol{r}_n$ describe the early reflection portion of a room impulse response and the late reverberation component of reverberant speech, respectively, in the log

Mel-frequency filterbank feature domain. The early reflection parameter $\boldsymbol{h}$ is assumed to be a random variable with a normal distribution. Similarly to the predictive reverberation estimator (12), employed for front-end-based approaches, $\boldsymbol{r}_n$ is estimated by

$$\boldsymbol{r}_n \approx \boldsymbol{a} + \boldsymbol{y}_{n-1}, \tag{16}$$

where $\boldsymbol{a}$ predicts the current late reverberation vector from the previous observed feature vector. Based on the reverberation model (15), $p_Y(\boldsymbol{y}_n|j,(\boldsymbol{y}_\tau)_{\tau<n})$ is modeled as a normal distribution with adapted time-varying mean vector $\boldsymbol{\mu}^Y_{n,j}$ and covariance matrix $\sum^Y_{n,j}$, where these parameters are calculated by using the log-normal approximation [4] at each time frame. The unknown parameters, i.e., $\boldsymbol{a}$ and the mean vector and covariance matrix of $\boldsymbol{h}$, are determined from the observed data $(\boldsymbol{y}_n)_{n\in\mathbb{T}}$ using maximum likelihood estimation.

A major potential drawback of the above frame-wise adaptation method and also of many front-end-based approaches is the assumption of time-invariant reverberation model parameters. There are two main reasons why a time-varying reverberation model is preferable. First, although late reverberation is insensitive to changes in speaker and microphone positions (see the section "Elements of Room Acoustics"), the characteristics of the early reflections depend strongly on the speaker and microphone positions. Furthermore, there are time-varying errors in the approximations used to derive the reverberation models (13) and (15) even when the speaker stays at the same position (see [46] for a detailed analysis). A time-varying model can capture these approximation errors and therefore describe the effect of reverberation more precisely. This motivates the use of statistical reverberation models, from which the model parameters, specifically $\boldsymbol{h}$ and $\boldsymbol{a}$ in the case of (15), are sampled anew at each time frame.

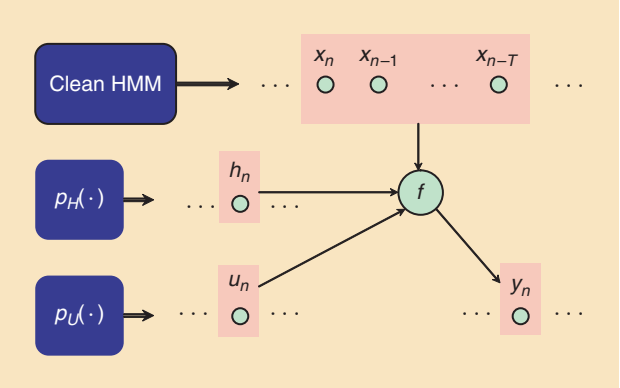
Inspired by the increased modeling accuracy, REMOS employs a time-varying version of (15) to define the emission pdf $p_Y(\boldsymbol{y}_n|j,(\boldsymbol{y}_\tau)_{\tau<n})$. Specifically, as illustrated in Figure 8, $\boldsymbol{y}_n$ is assumed to be produced according to

$$\begin{aligned}\boldsymbol{y}_n &= \boldsymbol{f}(\boldsymbol{x}_n,(\boldsymbol{x}_\tau)_{\tau<n},\boldsymbol{h}_n,\boldsymbol{u}_n)\\ &= \log(\exp(\boldsymbol{h}_n+\boldsymbol{x}_n)+\exp(\boldsymbol{r}_n+\boldsymbol{u}_n)),\end{aligned} \tag{17}$$

where $\boldsymbol{h}_n$ describes the early reflections at time frame n and is assumed to be drawn from a normal distribution $p_H(\boldsymbol{h})$. To estimate the late reverberation $\boldsymbol{r}_n$, [46] proposes replacing the predictive reverberation estimator (16) with a moving-average reverberation estimator similar to (13), which is defined as

$$\boldsymbol{r}_n = \log\left(\sum_{\tau=1}^{T}\exp(\boldsymbol{\mu}^H_\tau+\boldsymbol{x}_{n-\tau})\right). \tag{18}$$

The vector $\boldsymbol{u}_n$ in (17) accounts for the uncertainty and a potential bias in $\boldsymbol{r}_n$ and is assumed to obey a normal distribution $p_U(\boldsymbol{u})$. Each $\boldsymbol{\mu}^H_n$ is essentially equivalent to $\boldsymbol{h}_n$ in (14), but we changed the variable from $\boldsymbol{h}_n$ to $\boldsymbol{\mu}^H_n$ to distinguish it from the time-varying realization drawn from $p_H(\boldsymbol{h})$. The parameters of $p_H(\boldsymbol{h})$ and $p_U(\boldsymbol{u})$ as well as $(\boldsymbol{\mu}^H_n)_{1\le n\le T}$ are learned from a set of room impulse responses or reverberant utterances [46].

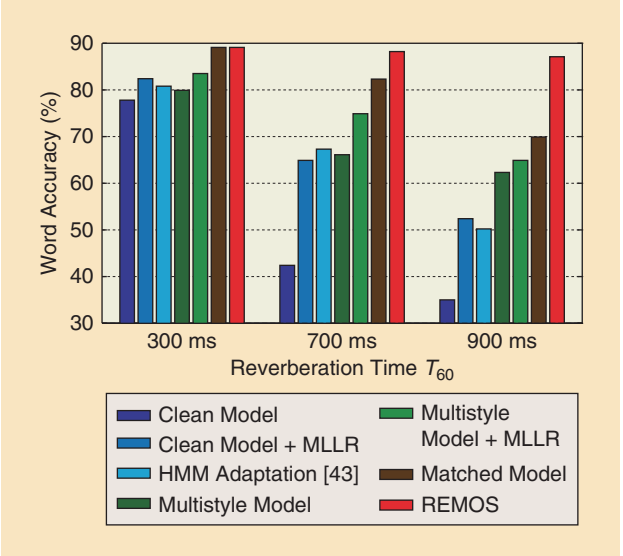


[FIG8] **Feature production model of REMOS.**

By combining those statistical models of the reverberation parameters and a clean HMM, the emission pdf $p_Y(\boldsymbol{y}_n|j,(\boldsymbol{y}_\tau)_{\tau<n})$ can be defined as

$$p_Y(\boldsymbol{y}_n|j,(\boldsymbol{y}_\tau)_{\tau<n}) = \int p_\eta(\boldsymbol{y}_n|\boldsymbol{x}_n,(\boldsymbol{y}_\tau)_{\tau<n})\,p_\lambda(\boldsymbol{x}_n|j)\,d\boldsymbol{x}_n, \tag{19}$$

where the second term on the right-hand side, $p_\lambda(\boldsymbol{x}_n|j)$, is the emission pdf of the clean HMM. The first term on the right-hand side, $p_\eta(\boldsymbol{y}_n|\boldsymbol{x}_n,(\boldsymbol{y}_\tau)_{\tau<n})$, describes the effect of reverberation on the clean vector $\boldsymbol{x}_n$ and is obtained by the following marginalization based on (17)



[FIG9] **Comparison of word accuracies in percentage for different conventional HMMs and REMOS in three different rooms for a connected digit recognition task using log Mel-frequency filterbank coefficients. The microphones were placed at 2.0 m, 4.1 m, and 4.0 m from the speaker. The baseline word accuracy achieved with clean data and clean HMMs is 95.3%.**

$$p_\eta(\boldsymbol{y}_n|\boldsymbol{x}_n,(\boldsymbol{y}_\tau)_{\tau<n}) = \iint p_H(\boldsymbol{h}_n)p_U(\boldsymbol{u}_n)\delta(\boldsymbol{y}_n - \boldsymbol{f}(\boldsymbol{x}_n,(\hat{\boldsymbol{x}}_\tau)_{\tau<n},\boldsymbol{h}_n,\boldsymbol{u}_n))d\boldsymbol{h}_n d\boldsymbol{u}_n, \tag{20}$$

where δ denotes the Dirac delta impulse, and the estimated clean speech vectors $\hat{\boldsymbol{x}}_\tau$ with $\tau < n$ are obtained during the optimization processes for the previous T frames. In practice, the integrals (19) and (20) are approximated by their values at the maximum of the integrand so that we have the following approximation:

$$p_Y(\boldsymbol{y}_n|j,(\boldsymbol{y}_\tau)_{\tau<n}) \approx p_H(\hat{\boldsymbol{h}}_n)p_U(\hat{\boldsymbol{u}}_n)p_\lambda(\hat{\boldsymbol{x}}_n|j), \tag{21}$$

where the values at the maximum are obtained as

$$(\hat{\boldsymbol{h}}_n,\hat{\boldsymbol{u}}_n,\hat{\boldsymbol{x}}_n) = \underset{(\boldsymbol{h}_n,\boldsymbol{u}_n,\boldsymbol{x}_n)}{\operatorname{argmax}} p_H(\boldsymbol{h}_n)p_U(\boldsymbol{u}_n)p_\lambda(\boldsymbol{x}_n|j) \quad \text{subject to } \boldsymbol{y}_n = \boldsymbol{f}(\boldsymbol{x}_n,(\hat{\boldsymbol{x}}_\tau)_{\tau<n},\boldsymbol{h}_n,\boldsymbol{u}_n). \tag{22}$$

Therefore, for each frame n and each state j, the Viterbi score is calculated by first solving the optimization problem (22) and then evaluating the emission pdf according to (21). Thereby, each reverberant feature vector $\boldsymbol{y}_n$ is decomposed into the contributions $\boldsymbol{x}_n$ from the clean HMM, the feature-domain effect of initial reflections $\boldsymbol{h}_n$, and the late reverberation component $\boldsymbol{r}_n$ depending on state j. Instead of obtaining a single estimate of the clean feature as with the front-end-based approaches, REMOS produces different clean feature estimates for different states by leveraging the statistical model of reverberation parameters, facilitating the discrimination between different phonemes and words. By adding a noise model to the REMOS framework, it is possible to handle background noise and reverberation jointly.

REMOS maintains high word accuracies even in severely reverberant environments, while the recognition performance of the HMM adaptation method [43] and many other conventional approaches deteriorate sharply with increasing reverberation as shown in Figure 9. It is also noteworthy that REMOS possesses a high flexibility to changes of speaker positions and even changes of the room [46]. These results indicate the potential of the approaches accounting for the long-term acoustic context over the conventional HMM adaptation approaches. However, it should be noted that REMOS currently requires a recognizer trained on static log Mel-frequency filterbank features and single Gaussian densities. This configuration does not represent the state of the art, and an extension of the method to MFCCs, GMMs, and dynamic features is still necessary so that REMOS can be applied to modern recognizers.

COMPARISON OF FRONT-END- AND BACK-END-BASED APPROACHES

Finally, we summarize the relative merits and demerits of the approaches presented above.

FRONT-END-BASED APPROACHES

A major advantage of the front-end-based approaches is that they do not require changes to the back-end processing steps, leading to two desirable properties. First, the computational complexity of the front-end-based approaches is independent of the acoustic model size. Since many front-end-based methods can be performed at a moderate computational cost, they can be easily employed for large vocabulary continuous speech recognition. Second, compared to the back-end-based approaches, these approaches could be relatively easily combined with advanced recognition techniques such as feature-space minimum phone error (fMPE) [47] and semitied transforms [48].

A potential drawback of the front-end-based approaches is that enhanced feature vectors inevitably contain estimation errors, which may degrade the recognition performance of conventional decoders. To mitigate the impact of the estimation errors, uncertainty decoding techniques can be combined with the front-end-based approaches.

A MAJOR ADVANTAGE OF THE FRONT-END-BASED APPROACHES IS THAT THEY DO NOT REQUIRE CHANGES TO THE BACK-END PROCESSING STEPS.

BACK-END-BASED APPROACHES

The back-end-based approaches are less prone to estimation errors because they directly model and recognize observed feature vector sequences. Another potential advantage is that both reverberation compensation and speech recognition are accomplished using the same features and the same acoustic model, and therefore they can be performed coherently. Thus, the mismatch between reverberant observations and the acoustic model can be reduced as effectively as possible.

A common drawback of the back-end-based approaches is that the number of unknown variables to be estimated is proportional to the number of the acoustic model parameters. Therefore, the computational cost scales with the complexity of the recognition task and usually significantly outnumbers that of the front-end-based methods for large tasks. The complexity problem is more severe for the acoustic context-dependent methods compared to the HMM adaptation techniques since the former modify the HMM parameters at each time frame as shown in Figure 7 while the latter perform adaptation only once before recognition. In addition, since most of the back-end-based approaches assume conventional features, such as MFCCs, it is not straightforward to integrate them with the advanced techniques mentioned above [47, 48].

CONCLUSIONS

In this article, we reviewed a variety of methods for recognizing reverberant speech and classified them in a common

framework. As we discussed, if reverberation is modelled as additive interference, then the main difference from common noise and interference is its extreme nonstationarity, which represents the fundamental problem in reverberant speech recognition. The strong relationship between long-term consecutive reverberant frames is an essential clue to compensate for such reverberation interference. Every method reviewed in the article takes this long-term acoustic context into account with some reverberation model. Some models are commonly employed by both front-end- and back-end-based approaches.

The problem of reverberant speech recognition leaves ample room for further research and development. For example, advanced speech recognition systems use discriminative or posterior-based features such as fMPE [47]. These types of features are derived by squeezing long-term consecutive feature vectors and therefore contain some information on the long-term acoustic context. It would be important to examine the effect of reverberation on these advanced features and to develop efficient ways for combining these features and reverberation robustness techniques. Other promising research directions include the improvement of speech and reverberation models and the combination of different approaches to reverberant speech recognition.

Reverberant speech recognition is a subproblem of transcribing distant-talking speech, which will be one of the key features of future speech recognition systems and decisive for the usability of natural human/machine interfaces. However, except for a limited number of methods, many existing techniques for realizing robustness against additive noise, reverberation, and speaker variations were developed independently of each other. Future research should also be directed at combining these different robustness techniques. Thus, we hope that the framework presented in this article will provide a basis for future research on reverberant and distant-talking speech recognition.

ACKNOWLEDGMENTS

The work of the authors from the University of Erlangen-Nuremberg was supported by the German Research Foundation (DFG) under contract number KE 890/4-1.

AUTHORS

Takuya Yoshioka (yoshioka.takuya@lab.ntt.co.jp) is a researcher at NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. He received the B.Eng., M.Inf., and Ph.D. degrees in informatics from Kyoto University, Japan, in 2004, 2006, and 2010, respectively. Since he joined NTT in 2005, he has been working on the development of algorithms for speech dereverberation, source separation, and noise robust speech recognition. He has received research awards, including the 2010 Itakura Prize Innovative Young Researcher Award from the Acoustical Society of Japan.

Armin Sehr (sehr@LNT.de) received the Dipl.-Ing. (FH) degree from the University of Applied Sciences Regensburg, Germany, in 1998 and the Dr.-Ing. as well as the M.Sc. degrees from the University of Erlangen-Nuremberg, Germany, in 2009 and 2010, where he is currently working as chair of multimedia communications and signal processing on robust distant-talking speech recognition. In 2010–2011, he was a visiting researcher at NTT Communications Science Laboratories in Kyoto, Japan, and from 1998 to 2003, he worked as a senior algorithm designer with Ericsson Eurolab in Nuremberg, Germany, on various projects of speech processing and mobile communications.

Marc Delcroix (marc.delcroix@lab.ntt.co.jp) is a research associate at NTT Communication Science Laboratories, Kyoto, Japan. His research interests include speech enhancement, dereverberation, and robust speech recognition. He received the M.Eng. degree from the Free University of Brussels, Belgium, and the Ecole Centrale Paris, France, in 2003 and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University, Japan, in 2007. From 2004 to 2008, he was a researcher at NTT Communication Science Laboratories. From 2008 to 2010, he worked as a software developer at Pixela Corporation, Osaka, Japan. He has received three research awards.

Keisuke Kinoshita (kinoshita.k@lab.ntt.co.jp) received the M.Eng. and Ph.D. degrees from Sophia University in Tokyo in 2003 and 2010, respectively. He is currently a researcher at NTT Communication Science Laboratories, where he is engaged in research on speech and audio signal processing. He received several research awards, including the 2009 Acoustical Society of Japan Outstanding Technical Development Prize.

Roland Maas (maas@LNT.de) received the M.Sc. degree in applied mathematics from the University of Erlangen-Nuremberg, Erlangen, Germany, and the University of Rennes, France, in 2009 within the scope of a joint master program. He is currently working toward the Dr.-Ing. degree in the Audio Research Laboratory, Chair of Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg. His research interests include robust distant-talking speech recognition in reverberant environments.

Tomohiro Nakatani (nakatani.tomohiro@lab.ntt.co.jp) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Japan, in 1989, 1991, and 2002, respectively. He joined Basic Research Laboratory, NTT Corporation, Japan, in 1991. Since then, he has been investigating sound capturing techniques directed at both humans and computers, including speech enhancement and automatic speech recognition. He is now a senior research scientist at NTT Communication Science Laboratories, NTT Corporation. He was an associate editor for *IEEE Transactions on Audio, Speech, and Language Processing* from 2008 to 2010. He is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee.

Walter Kellermann (wk@LNT.de) is a professor for communications at the University of Erlangen-Nuremberg, Germany, since 1999. He received the Dipl.-Ing. (univ.) degree in electrical engineering in 1983, and the Dr.-Ing. degree in 1988. From 1989 to 1990, he was a postdoctoral researcher at AT&T Bell Laboratories, Murray Hill, New Jersey. He has authored or

coauthored 16 book chapters and more than 180 refereed papers. He was an associate editor and a guest editor of various journals and was a general chair for several international conferences. He was also a Distinguished Lecturer for the IEEE Signal Processing Society in 2007 and 2008 and served as chair of the Technical Committee for Audio and Acoustic Signal Processing from 2008 to 2010. He is a Fellow of the IEEE.

REFERENCES

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Hoboken, NJ: Wiley, 2009.

[2] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

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

[4] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, 1996.

[5] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. Int. Conf. Spoken Language Process.*, 2000, pp. 869–872.

[6] J. Droppo and A. Acero, "Environmental robustness," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin: Springer-Verlag, 2008, pp. 653–679.

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

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

[9] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," in *Proc. Int. Conf. Spoken Language Process.*, 2002, pp. 2185–2188.

[10] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 489–498, 2004.

[11] H. Kuttruff, *Room Acoustics*, 5th ed., Abingdon, Oxon: Spon Press, 2009.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[13] T. Nishiura, Y. Hirano, Y. Denda, and M. Nakayama, "Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria," in *Proc. Interspeech*, 2007, pp. 1082–1085.

[14] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliffs, NJ: Prentice Hall, 2001.

[15] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[16] M. J. F. Gales and J.-Q. Wang, "Model-based approaches to handling additive noise in reverberant environments," in *Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays*, 2011, pp. 121–126.

[17] P. A. Naylor and N. G. Gaubitch, Eds., *Speech Dereverberation*. Berlin: Springer-Verlag, 2010.

[18] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 90–100, 2010.

[19] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature normalization for speaker verification in room reverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 4836–4839.

[20] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[21] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear predictor," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[22] A. Krueger and R. Haeb-Umbach, "A model-based approach to joint compensation of noise and reverberation for speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds. Berlin: Springer-Verlag, 2011, pp. 257–290.

[23] S. Haykin, *Adaptive Filter Theory*, 4th ed. Englewood Cliffs, NJ: Prentice Hall, 2001.

[24] B. W. Gillespie and L. E. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2003, pp. 676–679.

[25] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 476–488, 2003.

[26] M. Triki and D. T. M. Slock, "Delay and predict equalization for blind speech dereverberation," in *Proc. Int. Conf. Acoust, Speech, Signal Process.*, 2006, pp. V–97–V–100.

[27] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Berlin: Springer-Verlag, 2010, pp. 311–385.

[28] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 534–545, 2009.

[29] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 499–513, 2012.

[30] M. Delcroix, S. Watanabe, and T. Nakatani, "Variance compensation for recognition of reverberant speech with dereverberation preprocessing," in *Robust Speech Recognition of Uncertain or Missing Data*, R. Haeb-Umbach and D. Kolossa, Eds. Berlin: Springer-Verlag, 2011, pp. 225–256.

[31] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 69–84, 2011.

[32] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Eindhoven Univ. Technology, Eindhoven, The Netherlands, 2006.

[33] J. S. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1746–1765, 2010.

[34] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 45–48.

[35] K. Kumar, B. Raj, R. Singh, and R. Stern, "An iterative least-squares technique for dereverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5488–5491.

[36] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[37] M. Wölfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 2, pp. 312–323, 2009.

[38] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 412–421, 2005.

[39] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Mag.*, vol. 22, no. 5, pp. 101–116, 2005.

[40] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Commun.*, vol. 43, no. 1–2, pp. 123–142, 2004.

[41] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. I–1133–I–1136.

[42] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Commun.*, vol. 50, no. 3, pp. 244–263, 2008.

[43] A. Sehr, M. Gardill, and W. Kellermann, "Adapting HMMs of distant-talking ASR systems using feature-domain reverberation models," in *Proc. European Signal Process. Conf.*, 2009, pp. 540–543.

[44] Y.-Q. Wang and M. J. F. Gales, "Improving reverberant VTS for hands-free robust speech recognition," in *Proc. Workshop. Automat. Speech Recognition, Understanding*, 2011, pp. 113–118.

[45] T. Takiguchi, M. Nishimura, and Y. Ariki, "Acoustic model adaptation using first-order linear prediction for reverberant speech," *IEICE Trans. Inform. Syst.*, vol. E89-D, no. 3, pp. 908–914, 2006.

[46] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1676–1691, 2010.

[47] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: discriminatively trained features for speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 961–964.

[48] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 272–281, 1999.

[SP]