# Survey on Approaches to Speech Recognition in Reverberant Environments

Takuya Yoshioka*, Armin Sehr†, Marc Delcroix*, Keisuke Kinoshita*, Roland Maas†,
Tomohiro Nakatani*, and Walter Kellermann†,
* NTT Communication Science Laboratories, Kyoto, Japan † University of Erlangen-Nuremberg, Germany

***Abstract*—This paper overviews the state of the art in reverberant speech processing from the speech recognition viewpoint. First, it points out that the key to successful reverberant speech recognition is to account for long-term dependencies between reverberant observations obtained from consecutive time frames. Then, a diversity of approaches that exploit the long-term dependencies in various ways is described, ranging from signal and feature dereverberation to acoustic model compensation tailored to reverberation. A framework for classifying those approaches is presented to highlight similarities and differences between them.**

## I. Introduction

Automatic speech recognition technology has matured substantially, enabling a wide range of innovative voice-driven applications and radically changing our way of accessing digital services and information. Although most of today's applications still require microphones located near the talker for reliable operation, the rapid spread of speech recognition technology has fueled the need for speech recognizers that work in distant-talking situations, where talkers are able to speak at some distance from the microphones. Distant-talking speech recognition technology has potential to significantly extend the availability of speech recognizers, playing an essential role in realizing diverse applications including automatic meeting transcription, automatic annotation of user-generated audio and video, speech-to-speech translation in teleconferencing, and hands-free interfaces for controlling consumer products.

It is paramount for distant-talking speech recognizers to deal appropriately with the background noise and reverberation that result from the large speaker-to-microphone distance. Various successful techniques have been developed to combat the additive noise and the distortion caused by short impulse responses [1]. On the other hand, reverberation is characterized by long impulse responses, whose effect span a number of consecutive frames. Although several pioneering efforts were made [2]–[4], compensating for such long-term distortion is very challenging and has not gained wide attention until recently, precluding the wider use of distant-talking speech recognizers.

Recently, substantial progress has been achieved in reverberant speech processing and recognition. A diversity of solutions has been developed, ranging from dereverberation techniques in both signal and feature domains to model compensation approaches tailored to reverberation. At the heart of these emerging techniques lie multidisciplinary approaches that combine ideas from room acoustics, optimal filtering, machine learning, speech modeling, enhancement, and recognition. Some of these techniques are now ready to be evaluated for real-world speech recognition applications.

To help the reader have a quick overview of the current state of this important research topic, this paper briefly describes promising approaches to handling reverberation in speech recognition. Section II illustrates that reverberant speech recognizers must account for long-term dependencies between reverberant observations obtained from consecutive time frames. Section III overviews different approaches for exploiting these long-term dependencies, and Section IV concludes the paper. The present work is a summary of the tutorial review entitled "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition", which will be published in the IEEE Signal Processing Magazine in November 2012 [5]. The tutorial review in [5] provides detailed descriptions of various methods, some experimental results, discussion on the relative merits of different approaches, and a comprehensive reference list.

## II. Problem in Reverberant Speech Recognition

Let us imagine capturing speech with one or more distant microphones in a room. The speech emanated from the speaker's mouth is repeatedly reflected at the walls and other objects in the room and then picked up by the microphones. The sequence of these reflections is perceived as reverberation.

Recognizing such reverberant speech is very difficult with the current technology because the corruption of feature vector sequences by reverberation cannot be satisfactorily compensated for by conventional robustness techniques. While conventional techniques focus mainly on intra-frame distortion, it is essential for reverberant speech recognizers to deal appropriately with dependencies between consecutive reverberant feature vectors. The rest of this section explains this problem in more detail.

### A. Characteristics of reverberation

Reverberation can be described as a linear convolution of a speech signal and a room impulse response [6]. When the clean speech signal, the reverberant speech signal, the room impulse response, and the additive noise are denoted by $x(t)$, $y(t)$, $h(t)$, and $d(t)$, respectively, $y(t)$ is written as

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

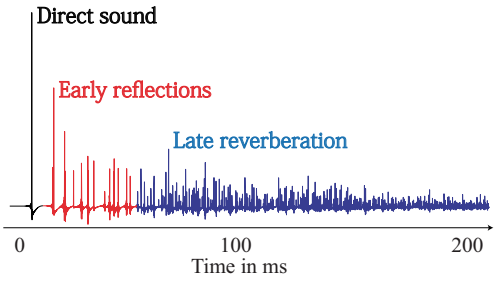

Fig. 1. Example room impulse response.

where $\circledast$ represents convolution and $T_h$ is the length of the room impulse response. To focus on reverberation, this paper neglects the additive noise $d(t)$. Discussion on dealing jointly with additive noise and reverberation can be found in [5].

The room impulse response can be divided into three portions as shown in Fig. 1 [6]. After the arrival of the direct sound, several strong reflections, called early reflections, occur at first sporadically, and later comes a series of numerous indistinguishable reflections, called late reverberation. The magnitude of the late reverberation decays approximately exponentially, and the decay rate is largely independent of the speaker and microphone positions. In contrast, the characteristics of the early reflections depend strongly on the positions. The time required for the late reverberation to decay by 60 dB relative to the direct sound level is called the reverberation time. For typical office and home environments, the reverberation time ranges from 0.2 to 1 s, which is much longer than an analysis frame used for speech recognition. 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 $h_l(t)$ as follows:

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

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

### B. Nonstationarity of interference caused by reverberation

The problem that makes reverberant speech recognition so challenging is the extreme nonstationarity of the interference caused by reverberation. This can be illustrated by recasting 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 (2), the observed reverberant signal $y(t)$ is written as

$$y(t) = h_i(t) \circledast x(t) + h_l(t) \circledast x(t-\Delta) = h_i(t) \circledast x(t) + r(t), \tag{3}$$

where $r(t)$ is the late reverberation component of $y(t)$. The late reverberation component $r(t)$ is often assumed uncorrelated to $x(t)$, allowing us to consider $r(t)$ as additive noise [7]. Since $r(t)$ results from filtering a delayed clean speech signal, it has substantial time variations and thus can be seen as very nonstationary additive noise. This nonstationarity renders most conventional noise robustness techniques ineffective for combatting the late reverberation because they assume stationary or slowly varying noise for reliable noise parameter estimation.

Fortunately, the late reverberation $r(t)$ at time $t$ can be predicted from past observations $(y(\tau))_{\tau \leq t}$ of the reverberant signal, since both $r(t)$ and $y(t)$ are filtered copies of the clean signal $x(t)$. This implies that a reverberant speech recognizer must account for the long-term acoustic context to predict and compensate for the late reverberation. The long-term acoustic context can be leveraged in various ways as described in the next section.

## III. Overview of Approaches

The approaches for reverberant speech recognition can be classified according to which element of speech recognition is modified. As shown in Table I, the approaches can be broadly categorized as either front-end-based or back-end-based. The front-end-based approaches modify the feature extraction process, i.e., the front-end, to enhance reverberant feature vectors. By contrast, the back-end-based approaches change elements of the back-end, specifically the acoustic model or the decoding algorithm, to directly process the reverberant feature vectors.

The front-end-based approaches are further categorized into three classes according to where enhancement takes place in the chain of processing steps for feature extraction. The linear filtering approach dereverberates time-domain signals or short-time Fourier transform (STFT) coefficients. The spectrum enhancement approach dereverberates corrupted power spectra. Finally, the feature enhancement approach directly enhances the corrupted feature vectors.

There are two subcategories in the class of back-end-based approaches. The hidden Markov model (HMM) adaptation approach adjust the parameters of the HMMs of the acoustic model once before recognition. The adapted acoustic model is used to transform the reverberant feature vectors into words with the conventional Viterbi algorithm. On the other hand, with the acoustic context-dependent likelihood evaluation approach, reverberation compensation is interwoven with the decoding process in order to account for the acoustic context effectively when evaluating the likelihood of an HMM state.

The following summarizes the basic concepts that the above approaches employ to deal with reverberation.

### A. Front-end-based approaches: Linear filtering

The linear filtering approach seeks a linear filter that cancels the effect of reverberation in the time or STFT domain. The filter must be long enough to cover the relevant part of the reverberation. Therefore, a dereverberated signal at the current time frame is computed from observations at the current and previous frames. After transforming the dereverberated signal into feature vectors it is fed into the back-end.

In one successful method, called long-term linear prediction [8], the linear dereverberation filter is estimated by temporally decorrelating the observed signals with linear prediction. This method includes two modifications to the conventional

TABLE I
CLASSIFICATION OF APPROACHES ACCORDING TO WHICH ELEMENT OF SPEECH RECOGNITION IS MODIFIED.

| | Front-end-based approaches | | | Back-end-based approaches | |
|---|---|---|---|---|---|
| | Linear filtering | Spectrum enhancement | Feature enhancement | HMM adaptation | Acoustic context-dependent likelihood evaluation |
| Element to be modified | Signal or STFT | Power spectrum | Feature vector | Acoustic model | Decoding algorithm |

linear prediction algorithm. First, the cost function for filter optimization is derived based on a time-varying speech model to account for the non-stationarity of the speech signals while the conventional linear prediction minimizes the sum of the squares of the prediction errors. Secondly, multi-step prediction is employed to predict the observation at frame $n$ from those at frames $n-T_\delta, \cdots, n-T_\tau$, where $T_\delta \geq 2$ and $T_\tau$ is the filter order. This is contrary to the conventional algorithm, which uses frames $n-1, \cdots, n-T_\tau$ to predict the observation at frame $n$. The $T_\delta$-frame moratorium period in prediction prevents from removing the temporal correlation inherent in the speech signals. Both of the two concepts were shown to play a critical role in achieving dereverberation and have been implemented in various ways in different dereverberation methods [9], [10].

The linear filtering approach can leverage both the amplitudes and phases of observed signals while the other approaches ignore the phases. The benefit from using the signal phases is pronounced especially in multi-microphone situations because both the amplitude and phase differences between multiple microphones can be exploited to improve the dereverberation performance [11]. Furthermore, the use of the phases allows close coupling with beamforming techniques to jointly achieve noise reduction and dereverberation [12].

### *B. Front-end-based approaches: Spectrum enhancement*

The spectrum enhancement approach attempts to restore a sequence of clean power spectra in various ways, given the corresponding sequence of reverberant power spectra. The disregard of the signal phases endows this approach with a higher degree of robustness against speaker movement than the linear filtering methods. The high degree of robustness results 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 [13].

For successful dereverberation, it is paramount to accurately estimate the degree of power spectral distortion caused by reverberation. Many spectrum enhancement methods exploit the knowledge of reverberation time to estimate the power spectral distortion [7], [13], where the reverberation time is automatically estimated, for example, by using the maximum likelihood method [14]. A few methods bypass the reverberation time estimation and directly estimate the clean power spectra by exploiting speech characteristics such as the frame-level temporal uncorrelatedness [15].

### *C. Front-end-based approaches: Feature enhancement*

The feature enhancement approach estimates a clean feature vector sequence, given the corresponding sequence of reverberant feature vectors, by capitalizing on a pre-trained model of clean feature vectors and a feature-domain reverberation model. By combining these two models, the mismatch between reverberant feature vectors and the acoustic model used for speech recognition can be effectively reduced. A Gaussian mixture model (GMM) or similar models can be used as the clean feature model.

Design of the feature-domain reverberation model is the most difficult element in developing feature enhancement algorithms. Since the relationship between clean and reverberant feature vectors is nonlinear due to the logarithmic compression in feature vector extraction, appropriate approximation of this nonlinear relationship is necessary. Existing methods use techniques such as a particle filter [16] and an extended Kalman filter [17].

While most front-end-based methods are computationally efficient, feature vectors estimated with these methods inevitably include a time-varying degree of estimation errors. Uncertainty decoding provides a way for taking these estimation errors into account during decoding, therefore deriving benefits from both the front-end and the back-end. While the concept of uncertainty decoding was proposed for noisy speech recognition [18], it can be applied to reverberant data [17], [19].

### *D. Back-end-based approaches: HMM adaptation*

Unlike the above three approaches, the HMM adaptation approach directly transcribes the observed reverberant feature vectors after adjusting the parameters of the HMMs of the clean acoustic model to the observed feature vectors. While the HMMs can be adjusted with general adaptation methods, such as maximum likelihood linear regression [20], there are several adaptation methods that capitalize on reverberation models. These methods, dedicated to reverberation, require a much smaller quantity of reverberant feature vectors to adjust the HMM parameters than the generic methods.

There are two strategies for modeling the effect of reverberation on the HMMs. One is to use a state-level convolution model, which describes the average energy dispersion from each state to the succeeding states [21], [22]. The effect of each state on its succeeding states is described by a state-level reverberation model, which is determined based on the knowledge of the reverberation time or by using pilot reverberant data. While the methods using the state-level reverberation model achieve noticeable improvement in speech recognition performance compared to the clean acoustic model, they are not optimal because the state-level convolution model cannot accurately describe the process in which each feature vector is corrupted by the preceding feature vectors. The other strategy allows the use of a frame-level convolution model by utilizing

an extended feature vector [23]. The extended feature vector is obtained by concatenating consecutive feature vectors within a sufficiently long window to broaden the temporal coverage of each feature vector. With this strategy, pre-trained HMMs of clean extended feature vectors and the frame-level reverberation model is combined to create HMMs of reverberant feature vectors. The parameters of the reverberation model are estimated based only on the observed feature vectors with the maximum likelihood method.

### *E. Back-end-based approaches: Acoustic context-dependent likelihood evaluation*

The acoustic context-dependent likelihood evaluation approach uses the current and past observations to evaluate the likelihood of each HMM state at the current frame, i.e., it takes the acoustic context into account. This is contrary to the HMM adaptation approach, which uses modified HMMs to evaluate the state likelihood based only on the reverberant feature vector at the current frame. The context-dependent approach allows for a more accurate description of the long-term dependencies between reverberant feature vectors at the expense of an increased computational complexity.

To utilize the current and past observations for state likelihood evaluation, it is necessary to modify the decoding algorithm. One method to do this is to modify the HMM parameters on a frame-by-frame basis so that the HMMs used for the current frame account for interference from the past reverberant feature vectors [24]. Another method is to decompose each reverberant feature vector into contributions from the clean acoustic model and a reverberation model for each frame and each state [25]. This method, called REMOS (REverberation-Modeling for Speech recognition), exploits the clean acoustic model, the reverberation model, and an estimate of the interference from past observations to identify the contribution from the clean acoustic model to each observed feature vector.

## IV. Final Remark

In this paper, we described the fundamental problem in reverberant speech recognition and overviewed a variety of approaches for overcoming this problem while categorizing them into five classes. Although significant progress has been achieved, the problem of reverberant speech recognition is far from being solved, calling for further research and development. The reader can consult [5] for more details.

## References


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

[2] 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.

[3] 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.

[4] 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.

[5] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, 2012, to appear.

[6] H. Kuttruff, *Room acoustics*, 5th ed. Spon Press, 2009.

[7] 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, pp. 359–366, 2001.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Eindhoven University of Technology, 2006.

[14] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, Jr., C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Amer.*, vol. 114, no. 5, pp. 2877–2892, 2003.

[15] 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 Processing*, vol. 18, no. 7, pp. 1746–1765, 2010.

[16] 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.

[17] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1692–1707, 2010.

[18] 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.

[19] 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. Springer, 2011, pp. 225–256.

[20] 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.

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

[22] 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.

[23] 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.

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

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