Single channel dereverberation using example-based speech enhancement with uncertainty decoding technique

Keisuke Kinoshita, Mehrez Souden, Marc Delcroix and Tomohiro Nakatani

NTT Communication Science Labs. NTT Corporation
{kinoshita.k, mehrez.souden, marc.delcroix, nakatani.tomohiro}@lab.ntt.co.jp

Abstract

A speech signal captured by a distant microphone is generally contaminated by reverberation, which severely degrades the audible quality and intelligibility of the observed speech. In this paper, we investigate the single channel dereverberation which has been considered as one of the most challenging tasks. We propose an example-based speech enhancement approach used in combination with non-example-based (conventional) blind dereverberation algorithm, that would complement each other. The term, example-based, refers to the method which has *exact* (not brief and statistical) information about the clean speech as its model. It is important to note that the combination of two algorithms is formulated utilizing the uncertainty decoding technique, thereby achieving the smooth and theoretical interconnection. Experimental results show that the proposed method achieves better dereverberation in severe reverberant environments than the conventional methods in terms of objective quality measures.

Index Terms: single channel speech dereverberation, example-based approach, uncertainty decoding,

1. Introduction

The quality and intelligibility of a speech signal captured by a distant microphone is generally degraded by acoustic interferences such as reverberation and environmental noise. To cope with these interferences in the aim of improving the speech quality, a considerable amount of speech enhancement research has been done from various perspectives [1].

The studies of single channel speech enhancement that have been carried out in the past can be categorized primarily into two types. The first focuses more on the blind estimation of characteristics of interferences such as noise and reverberations [2–9], while the second tries to directly estimate the underlying clean speech component given the noisy observation [10]. These two major trends have different advantages and disadvantages described as follows.

The history of the studies that focus on the estimation of interference component dates back to the 1970s. In late 1970s, the first noise estimation scheme, which may be the simplest and widely used one, had been proposed by [2]. It assumes the stationarity of the target noise and tries to estimate the noise amplitude spectrum during the non-speech period. After solving the estimation problem of stationary noise, the majority of studies turned their attention to the estimation of non-stationary noise, and has solved the problem partially by this day [3, 4]. The estimation of reverberation characteristics has been also viewed as very difficult task [5–9] for a decade. Although many studies have been done [5–9], to the best of our knowledge, the accurate estimation of reverberation characteristics has not been achieved yet, especially in single channel scenario. Some of existing methods suffer from inaccurate modeling of room impulse response [5–7], and some others suffer from theoretical limitation in the estimation accuracy in the single channel scenario [8, 9] which is caused fundamentally by the inversion problem of single channel non-minimum phase impulse response. Although the aforementioned approaches [2–9] sometimes provide inaccurate estimate of the interference, it is still a great advantage that they can reduce acoustic interference in blind manner with reasonable computational complexity.

The other trend of speech enhancement studies put a strong focus, not on the estimation of interference component, but on the estimation of the underlying clean speech itself. For instance, an example-based speech enhancement method proposed in [10] has demonstrated significant advantages in highly non-stationary noisy environments over conventional denoising methods by directly estimating the underlying clean speech given the observed noisy signal. The term *example-based* in this paper refers to the method which has *exact* and *fine* information about the clean speech. The example-based algorithm requires the training preceding the testing stage. In [10], at the training stage, the collection of speech signal with various possible types of noise is prepared as examples, and the exact spectral-temporal patterns of these example sequences are captured with the, so called, example model. At the testing stage, using the example model, the algorithm searches the prototype examples in the training dataset, that most likely match the input noisy signal. And finally, the clean speech examples associated with the matching results are extracted from the training dataset to reconstruct the estimation of the underlying clean speech. By preparing various types of noisy examples in the training stage, the method can omit the noise estimation procedure and focus solely on the robust estimation of the underlying clean speech component. However, it has an apparent disadvantage that, by preparing the vast amount of noisy examples, the cost for searching matched segment becomes huge.

In this paper, we propose an example-based speech enhancement approach which is optimally interconnected with the speech enhancement algorithm based on the estimation of interference component. Hereafter, the method based on the estimation of interference component is called *non-example-based* approach, which is used as opposed to the *example-based* one [1]. It should be noted that the combination of the example- and non-example-based algorithms is formulated in a similar manner as the uncertainty decoding [11] [12]. Thus the smooth and theoretical interconnection is realized. By combining two different types of approaches appropriately, we expect them to work complementary to each other. That is, by putting a strong focus on the estimation of the underlying clean speech sequence, we expect that it can compensate the inaccurate estimation of interference component. Besides, by estimating the interference component based on non-example-based algorithm, we can omit the preparation of various types of noisy/reverberant speech patterns as examples, thereby reducing the computational complexity required in the search of the example-based approach. We believe that this concept is favorable especially in single channel dereverberation scenario since the accuracy in reverberation estimation theoretically deteriorates and has to be compensated in some way. Hence, the evaluation of the proposed framework is done in dereverberation context. The contribution of this paper is twofold: The first is the formulation of the example-based speech enhancement jointly used with non-example-based speech enhancement. The second is the inves-

[1] The non-example-based approach has the information about clean speech up to only the brief and statistical characteristics.

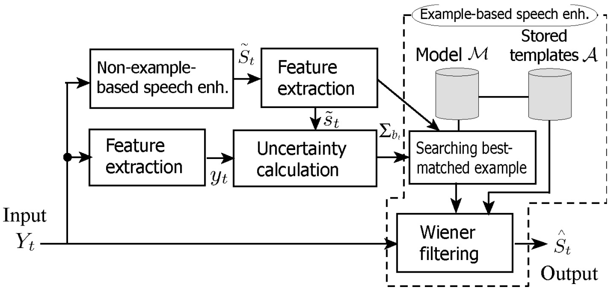

Figure 1: Proposed enhancement framework

tigation of the example-based approach in reverberant environment.

2. Proposed enhancement framework

2.1. Signal Model

In this paper, we consider the following model of the observed signal, in the power spectral domain

$$Y_t = S_t + B_t, \tag{1}$$

and similarly in the feature (e.g., Mel-Frequency Cepstral Coefficient: MFCC) domain

$$y_t = s_t + b_t, \tag{2}$$

where we denote the observed signal at time frame t as Y_t and y_t, the clean signal as S_t and s_t, and the late reverberation as B_t and b_t. As commonly used in the literature [5] [8], the effect of late reverberation is modeled by the additive component b_t in the observation process.

2.2. Overview of the framework

In this section, we briefly review the processing flow of the proposed speech enhancement framework depicted in Fig. 1. First, the power spectrum of the input reverberant speech Y_t is processed with non-example-based speech enhancement algorithm. Secondly, based on the extracted features of observed signal y_t and enhanced signal $\tilde{s}_t$, the uncertainty/unreliability of the results of non-example-based speech enhancement, Σ_{b_t}, is calculated at each time frame t as in uncertainty decoding [12]. The calculated uncertainty is then reflected to the criterion to search the best matched example of the underlying clean speech among the vast amount of the clean speech examples in the training dataset. This search is done based on the example model $\mathcal{M}$. After obtaining the best-matched sequence and extracting the sets of corresponding clean speech examples (i.e. amplitude spectra) from the, so called "stored templates $\mathcal{A}$", we form the clean speech estimate used in the construction of the preceding Wiener filter.

From the next section, we describe each of main components in Fig. 1 in detail, as well as the way of calculating and incorporating the uncertainty decoding technique.

2.3. Example-based speech enhancement

2.3.1. Model capturing the exact and fine information of the training dataset

Here, we explain the model $\mathcal{M}$ which will be used in searching speech examples and show how it models the exact and fine information of the training data. This model and most of the searching criteria in next sections are based on [10].

The model is constructed mainly with two steps. Let $\mathbf{s} = \{s_i : i = 1, 2, \ldots, I_\mathbf{s}\}$ be a complete set of training frame features, s_i be the feature at time frame i, $I_\mathbf{s}$ be the total number of frames in the training dataset. First, we train a Gaussian mixture model (GMM) $\mathcal{G}$ for the set of the features extracted from the training dataset, $\mathbf{s}$

$$\mathcal{G} = \sum_{m=1}^{M} w(m) \underbrace{N(s; \mu_m, \Sigma_m)}_{g(s|m)}, \tag{3}$$

where $g(s|m)$ is the m-th Gaussian component with the mean μ_m and the covariance Σ_m, and $w(m)$ is the corresponding weight. M is the number of mixture component.

Then, based on $\mathcal{G}$, we can build a model that represents the exact pattern of temporal dynamics contained in the training dataset. That is, for each time frame i, we calculate the Gaussian component m in $\mathcal{G}$ that maximizes the likelihood of the frame, and obtain a time sequence of maximum-likelihood Gaussian components. The model can be expressed using the corresponding time sequence of Gaussian indices m_i and $\mathcal{G}$ as:

$$\mathcal{M} = \{\mathcal{G}, m_i : i = 1, 2, \ldots, I_\mathbf{s}\} \tag{4}$$

where m_i is an index of a Gaussian component $g(s|m_i)$ in $\mathcal{G}$, that produces the maximum likelihood for the i-th frame feature, s_i, of the training dataset. We will use $\mathcal{M}$ as a exact and fine spectral-temporal model for the training dataset $\mathbf{s}$. Along with this model, the raw data of clean amplitude spectra is stored as $\mathcal{A}$ in this training stage.

2.3.2. Finding the best-matched example to the input sequence among training sentences

Here, we describe the method to find the segments of the training dataset which most likely represent the underlying clean speech. In this subsection, we temporarily ignore the difference between the observed signal y_t and the clean speech s_t, that is, we assume $b_t = 0$ for simplicity. This issue will be revisited in Section 2.4, which is the main focus of this paper.

Now, how can we find the best-matched example to the input sequence among the training dataset based on $\mathcal{M}$? Let $\mathbf{y} = \{y_t : t = 1, 2, \ldots, T\}$ be a collection of observed signals of T frames and $\mathbf{y}_{t:t+\tau}$ be a test segment taken from time frame t to $t + \tau$ of the sentence $\mathbf{y}$. In addition, let $\mathcal{M}_{u:u+\tau} = \{\mathcal{G}, m_i : i = u, u + 1, \ldots, u + \tau\}$ represent the sequence of Gaussian component modeling consecutive frames from u to $u + \tau$ in the training dataset $\mathbf{s}$. Then, the best-matched segment for t can be identified as the longest segment of the training dataset that matches the observed segment. More specifically, the best-matched segment can be found by maximizing the posterior probability as

$$\mathcal{M}^t_{u:u+\tau} = \arg\max_{\tau} \max_{\mathcal{M}_{u:u+\tau}} p(\mathcal{M}_{u:u+\tau}|\mathbf{y}_{t:t+\tau}), \tag{5}$$

$$p(\mathcal{M}_{u:u+\tau}|\mathbf{y}_{t:t+\tau}) = \frac{p(\mathbf{y}_{t:t+\tau}|\mathcal{M}_{u:u+\tau})p(\mathcal{M}_{u:u+\tau})}{p(\mathbf{y}_{t:t+\tau})}, \tag{6}$$

where $p(\mathcal{M}_{u:u+\tau}|\mathbf{y}_{t:t+\tau})$ is the posterior probability which has an important characteristic: It favors longer matching, i.e. larger τ, between $\mathbf{y}_{t:t+\tau}$ and $\mathcal{M}_{u:u+\tau}$ [10]. So the longer the matched segment length is, the higher the posterior probability become. The concept of longest matching criterion is similar to the idea behind the unit-selection based speech synthesis [13] which aims to resynthesize a natural and clear speech signal to a maximum extent. Hence, we can expect that, seeking longer matched segments with reasonable cost function will result in the naturalness of the processed speech. In (6), we assume an equal prior probability of $p(\mathcal{M}_{u:u+\tau})$ for all the training speech segments, which means that all the possible sequence patterns seen in the training dataset will occur in testing condition with an equal probability. The term in numerator of (6), $p(\mathbf{y}_{t:t+\tau}|\mathcal{M}_{u:u+\tau})$, is the likelihood of the test segment $\mathbf{y}_{t:t+\tau}$

associated with the segment of training dataset modeled with $\mathcal{M}_{u:u+\tau}$. This likelihood can be calculated as:

$$p(\mathbf{y}_{t:t+\tau}|\mathcal{M}_{u:u+\tau}) = \prod_{\epsilon=0}^{\tau} g(y_{t+\epsilon}|m_{u+\epsilon}) \qquad (7)$$

where we assume the conditional independence between adjacent frames. The denominator of (6) can be calculated as the summation of $p(\mathbf{y}_{t:t+\tau}|\mathcal{M}_{u:u+\tau})$ over all the possible pattern of $\mathcal{M}_{u:u+\tau}$ stored in the model $\mathcal{M}$.

To implement (5), we can first set $\tau = 0$ and estimate the most probable segment of length 1 among training dataset $\mathbf{s}$, then increase the value τ to 1 and follow the same procedure. After obtaining the most probable training segment for each τ, we find the maximum matched-segment-length (i.e., $\tau_{\max}$) that should result in the maximum posterior probability.

2.3.3. Forming the clean amplitude spectrum estimate

After finding the longest matched segments in the training dataset for $\mathbf{y}_{t:t+\tau_{\max}}$ at all t, we form an estimate of the underlying clean speech spectra by utilizing the corresponding matched training segments, $\mathcal{M}^t_{u:u+\tau_{\max}}$, and the prototype amplitude spectra stored in connection with $\mathcal{M}^t_{u:u+\tau_{\max}}$. Let ε ($\varepsilon = 1, 2, \ldots, T$) be the frame index of interest where we would like to recast the clean amplitude spectrum, then the resultant amplitude spectrum $\hat{S}'_\varepsilon$ can be constructed as:

$$\hat{S}'_\varepsilon = \frac{\sum_t A(\mathbf{u}^t_\varepsilon) p(\mathcal{M}^t_{u:u+\tau_{\max}}|\mathbf{y}_{t:t+\tau_{\max}})}{\sum_t p(\mathcal{M}^t_{u:u+\tau_{\max}}|\mathbf{y}_{t:t+\tau_{\max}})} \qquad (8)$$

where $A(\mathbf{u}^t_\varepsilon)$ represents a prototype amplitude spectrum associated with the frame of training dataset corresponding to $\mathcal{M}^t_{u:u+\tau_{\max}}$, and $\mathbf{u}^t_\varepsilon$ shows the time frame index of the training dataset corresponding to ε in the obtained most-likely time path $\mathbf{u} = \{u, u+1, \ldots, u+\tau_{\max}\}$ at time t, and $\mathcal{A} = \{A(i) : i = 1, 2, \ldots, I_{\mathbf{s}}\}$. From (8), we see that an estimate of the amplitude spectrum for time frame ε is basically obtained by taking into account all the adjacent matched segments that contain ε and averaging $A(\mathbf{u}^t_\varepsilon)$ over t. In the averaging, we use the posterior probability obtained in (6) as a sort of confidence score.

2.4. Interconnection of the statistical signal based and example-based algorithm

In this section, we revisit the issue about the mismatch between s_t and y_t which can be considered as the interference component b_t contained in observed signal y_t in (2). If s_t and y_t share sufficiently similar static and dynamic acoustic characteristics, it may be straightforward to find the segments in the training dataset which can well represent the underlying clean speech s_t. However, in general, y_t is deviated significantly from s_t due to the interference b_t, which makes it difficult to find the segments sufficiently similar to the underlying clean speech.

It may be possible to some extent to use the enhanced signal $\tilde{s}_t$ instead of directly using y_t to obtain better results. However, besides the fact that we never know how close $\tilde{s}_t$ can get to s_t, especially in the single channel dereverberation case, it is fairly hard to obtain $\tilde{s}_t$ similar to s_t. Therefore, special care has to be taken to use an example-based approach in conjunction with non-example-based approach.

To deal with the deviation of $\tilde{s}_t$ from s_t, we take into account the reliability of the speech enhancement results in a similar way to uncertainty decoding or dynamic variance compensation technique [11] [12]. Now, we would like to formulate the model of the observed signal y_t in a probabilistic way. First let us assume that b_t can be modeled as a Gaussian with

$$p(b_t) = N(b_t; \hat{b}_t, \Sigma_{b_t}) \qquad (9)$$

where $\hat{b}_t$ is an estimate of the mismatch, i.e., $\hat{b}_t = y_t - \tilde{s}_t$, and Σ_{b_t} represents a time-varying covariance matrix of b_t [12]. Then, the likelihood of the observed signal y_t can be obtained by marginalizing the joint probability over clean speech s_t as:

$$\begin{aligned} p(y_t) &= \int_{-\infty}^{+\infty} p(y_t, s_t) ds_t \\ &= \int_{-\infty}^{+\infty} N(y_t - s_t; y_t - \tilde{s}_t, \Sigma_{b_t}) p(s_t) ds_t \\ &= \sum_{m=1}^{M} w(m) N(\tilde{s}_t; \mu_m, \Sigma_m + \Sigma_{b_t}). \end{aligned} \qquad (10)$$

In the development, we used the probability multiplication rule. From (10), we can interpret the time-varying covariance Σ_{b_t} as a measure of uncertainty of speech enhancement. For uncertain features, the corresponding Σ_{b_t} is large, and therefore the influence of these features on the results is reduced. In this paper, we estimated Σ_{b_t} as a diagonal covariance matrix with the k-th diagonal component $\sigma_k = (y_{t,k} - \tilde{s}_{t,k})^2$ where k is the index in feature vector, which worked quite well in a previous study [12].

We insert this dynamic variance compensation rule into the evaluation of the posterior probability in (6) to achieve better interconnection between the non-example- and example-based speech enhancement algorithms.

3. Dereverberation experiment

In this section, we evaluate the effectiveness of the proposed framework in reverberant environments.

3.1. Experimental setup

3.1.1. Training condition

The training data for the GMM $\mathcal{G}$ is the TIMIT core training-set, which consists of 1088 sentences, 136 speakers. The sampling frequency was 8 kHz. The feature vector for the GMM $\mathcal{G}$ is 40th order MFCC with log energy term. The number of mixture component M is 4096. The frame length used for FFT is 20ms, and frame shift is 10ms. These settings are similar to [10].

3.1.2. Testing condition

The impulse response of a reverberant chamber (5m × 5m × 5m) was simulated with the image method [14] with distance of 2.5 m between the microphone and the speaker. The reflection coefficients of the walls are [0.9 0.9 0.85 0.85 0.2 0.2]. The RT$_{60}$ reverberation time of the simulated acoustic environment is about 0.5 sec. The reverberant test speech signals are created by convoluting this impulse response with 64 sentences taken from the TIMIT core test-set which is not used in training. Note that the number of speakers in the test-set is 4, and they are not included in the training dataset.

As a dereverberation method which produces $\tilde{S}_t$, we use the method proposed in [8] with the estimator of late reverberation proposed in [9]. This method may be a current state-of-the-art dereverberation and hereafter referred to as the conventional dereverberation.

The performance of the proposed framework ("proposed proc.") is compared with (a) observed signal ("no proc."), (b) the signal processed with conventional dereverberation ("conv. proc."), (c) the signal processed with the naive combination of conventional dereverberation and the example-based approach, i.e. $\Sigma_{b_t} = \mathbf{0}$ ("naive comb.").

3.2. Experimental results

3.2.1. Improvement in spectrogram

Fig. 2 shows the spectrograms of each signal. As we expected, the conventional dereverberation suppresses reverber-

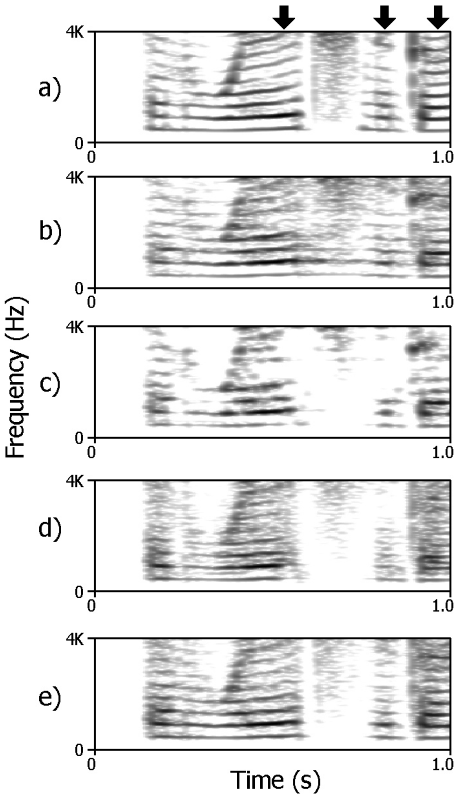

Figure 2: The spectrograms of a) clean speech, b) observed speech (no proc.), c) conv. proc., d) naive comb. and e) proposed proc.

ation but at the same time distorts the underlying target signal component. Conversely, naive combination produces a cleaner speech. The proposed framework contributed to further quality improvement for example in the voiced regions indicated with arrows in the figure. Naive combination provided somewhat blurry harmonicity, while the proposed framework recovers it much more clearly. From this figure, we can confirm the great advantage of the proposed framework in severe reverberant environment.

3.2.2. Objective measurement of speech quality

We measured the effect of the proposed method in terms of segmental SNR and log-spectral distance. The results obtained from the test utterances are averaged and shown in Fig. 3. As can be seen, the proposed framework successfully improves both segmental SNR and log-spectral distance measure. Although the improvement obtained by incorporating uncertainty decoding technique is not very drastic in terms of these measures, we can confirm that it improves the quality consistently. The improvement in audible quality can be confirmed in [15].

4. Summary

In this paper, we investigated the single channel dereverberation, which is based on the example-based speech enhancement approach used in combination with the non-example-based blind dereverberation algorithm. The combination of example- and non-example-based algorithm was formulated using the uncertainty decoding technique, thereby realizing the smooth and theoretical interconnection. In the experiment, we compared the proposed framework with the conventional method, and naive combination of example- and non-example-based algorithm, in severe reverberant environment. The results indicated the consistent improvements in terms of the spectrogram as well as the objective quality measures by the proposed framework. Future work includes thorough performance evaluation and the application of the proposed method to automatic speech recognition task.

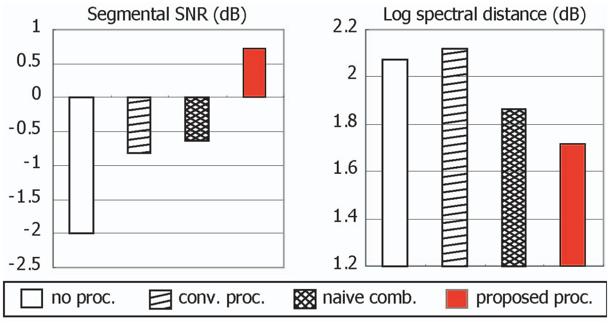

Figure 3: The result of objective quality measurement

5. References

[1] H.-W. H. X. Huang, A. Acero, *Spoken language processing*. Upper Saddle River, NJ: Prentice Hall, 2001.

[2] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE TSAP*, vol. 27(2), pp. 113–120, 1979.

[3] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE TASLP*, vol. 9, no. 5, pp. 504 – 512, 2001.

[4] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE TASLP*, vol. 11, no. 5, pp. 466 – 475, 2003.

[5] K. Lebart and J. Boucher, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[6] E. A. P. Habets, “Multi-channel speech dereverberation based on a statistical model of late reverberation,” in *ICASSP*, vol. 5, 2005, pp. 173–176.

[7] H. W. Löllmann and P. Vary, “Low delay noise reduction and dereverberation for hearing aids,” *EURASIP Journal on advances in signal processing*, 2009, article ID: 437807.

[8] 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 TASLP*, vol. 17, no. 4, pp. 534–545, 2009.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech dereverberation based on variance-normalized delayed linear prediction,” *IEEE TASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[10] M. Ji, R. Srinivasan, and D. Crookes, “A corpus-based approach to speech enhancement from nonstationary noise,” *IEEE TASLP*, vol. 19, no. 4, pp. 822–836, 2011.

[11] D. Li, J. Droppo, and A. Acero, “Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion,” *IEEE TASLP*, vol. 13, no. 3, pp. 412 – 421, 2005.

[12] M. Delcroix, T. Nakatani, and S. Watanabe, “Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing,” *IEEE TASLP*, vol. 17, no. 2, pp. 324 – 334, 2009.

[13] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *ICASSP*, vol. 1, 1996, pp. 373–376.

[14] J. B. Allen and D. A. Berkeley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, vol. 65(4), pp. 943–950, 1979.

[15] http://www.kecl.ntt.co.jp/icl/signal/kinoshita/publications/IS2011/.