FAST ESTIMATION OF A PRECISE DEREVERBERATION FILTER BASED ON SPEECH HARMONICITY

Keisuke Kinoshita *Tomohiro Nakatani* *Masato Miyoshi*

NTT Communication Science Laboratories, NTT Corporation

{kinoshita,nak,miyo}@cslab.kecl.ntt.co.jp

ABSTRACT

A speech signal captured by a distant microphone is generally smeared by reverberation. This severely degrades both the speech intelligibility and Automatic Speech Recognition (ASR) performance. We have already proposed a novel dereverberation technique named "Harmonicity based dEReverBeration (HERB)", which utilizes essential properties of speech, harmonics, and estimates an inverse filter for an unknown impulse response. If a large amount of acoustically stable training data is available, HERB is able to estimate an accurate inverse filter even in severely reverberant environments. In general, however, a dereverberation algorithm has to work with small amounts of training data, because the acoustic property of a real world environment changes according to various factors such as the speaker's position and room temperature. In this paper, we propose a new dereverberation scheme based on HERB, aiming primarily at reducing the amount of training data needed to estimate an inverse filter. We show experimentally that our new dereverberation scheme successfully achieves high quality dereverberation with much smaller amounts of training data, and is very effective at improving both audible quality and ASR performance, even in unknown severely reverberant environments.

1. INTRODUCTION

The quality of a speech signal captured by a distant microphone is generally degraded by surrounding acoustic interference such as reverberation and environmental noise. A recorded speech signal can be modeled as

$$x(n) = s^T(n)h(n) + b(n) \tag{1}$$

where $s(n)$ refers to clean speech, $h(n) = [h(0,n) \dots h(M-1,n)]^T$ to an M-tap room impulse response, and $b(n)$ to environmental additive noise. Among these interference, reverberation is known to severely degrade both Automatic Speech Recognition (ASR) performance and speech intelligibility. In particular, in a reverberant environment with a reverberation time (RT) of more than 0.5 seconds, the ASR performance cannot be improved even with an acoustic model trained with a matched reverberation condition [1].

Considerable research has been undertaken to find a way of dealing with additive noise. For example, Spectral Subtraction (SS) can greatly reduce the effect of additive noise and leads to a sufficient improvement in ASR performance and speech intelligibility [2]. In an ASR system, a Parallel Model Combination (PMC) [3] can also be utilized as a good solution to additive noise.

Although noise reduction techniques have been developed, no effective dereverberation technique has yet been proposed despite considerable effort over many years. The most widely used dereverberation remains microphone array [4]. It first estimates the direction of arrivals (DOAs), and steers the "nulls" of a microphone array (null beam-forming) to best suppress the reflections. Since the number of reflections is much greater than the nulls formed by the microphone array, it only works in moderately reverberant environments. Other major approaches attempting to estimate the inverse filter for an unknown impulse response are based on blind equalization methods, such as Independent Component Analysis (ICA) [5]. This method works effectively if the signals are statistically independent and identically distributed non-Gaussian sequences. However, it cannot appropriately handle speech signals because they have inherent properties, such as periodicity and formants, making the sequence statistically dependent. Another approach that focuses on audible quality has been proposed by Yagnanarayana [6]. Improvement was achieved by attenuating the relative amplitude of LPC residuals where the speech to reverberation ratio is smaller. Even though this method might help to improve speech intelligibility, it does not improve ASR performance because it makes no changes to the spectral features that are essential for ASR.

To achieve an improvement in both ASR performance and speech intelligibility, we have proposed a novel dereverberation methodology named "Harmonicity based dEReverBeration (HERB)" [7, 8]. It utilizes essential properties of speech, harmonics, and estimates an inverse filter for an unknown impulse response. HERB is very effective at improving both speech intelligibility and ASR performance even under unknown severely reverberant environments (e.g. Reverberation Time:RT=1.0s, 0.5s) [9], if a large amount of acoustically stable training data is available (e.g. more than 60 minutes in [7]). In general, however, the real world acoustic environment changes according to various factors such as the speaker's position and room temperature, and never remains stable for a long time. Therefore, in practice, the conventional HERB estimation scheme cannot be applied directly to a real world environment.

In this paper, we would like to propose a new dereverberation scheme named "Fast HERB" that utilizes the the framework of HERB. The primary aim is to reduce the amount of training data and thus enable the fast estimation of the inverse filter. In sections 2 and 3, we describe the problem of HERB in more detail and derive a way of reducing the amount of training data without degrading its dereverberation performance. In section 5, we report an experimental result obtained with Fast HERB and using much smaller amount of training data.

2. PROBLEM OF HARMONICITY BASED DEREVERBERATION

HERB estimates the inverse filter without any knowledge of the input signal, by regarding the harmonic structure within reverberant speech as the voiced portion of a direct sound. The inverse filter derived with this scheme is proven to approximate that of a room impulse response precisely, if there is sufficient training data. We begin by describing the HERB dereverberation scheme in more detail, and then address the problem it poses.

2.1. Harmonicity based dereverberation

In HERB, the dereverberation filter W_H is calculated as

$$W_H(f) = \mathcal{E}\left\{\frac{\mathcal{H}\{X(\tau,f)\}}{X(\tau,f)}\right\}, \tag{2}$$

where $X(\tau, f)$ is a discrete short-time Fourier transformation of an observed reverberant signal $x(n)$ with a center frequency f at a time frame τ. $\mathcal{H}\{\cdot\}$ is a function that extracts harmonic components from X. An adaptive comb filter is used to implement this function in HERB. $\mathcal{E}\{\cdot\}$ is a function that calculates the average value at each f over time frames.

To prove that W_H can precisely approximate the true inverse filter W_T, we first introduce the model of speech signals and reverberation. Let S be the clean speech, S_h be its harmonic component (i.e. harmonics within voiced vowels, voiced consonants) and S_n its non-harmonic component (i.e. noisy components within unvoiced vowels, consonants). Then, the clean speech is modeled as

$$S(\tau, f) = S_h(\tau, f) + S_n(\tau, f). \tag{3}$$

If the speech is reverberated with a transfer function H, which can be decomposed into reverberation component R and direct component D, reverberant speech X is given by

$$\begin{aligned} X(\tau, f) &= H(f)S(\tau, f), \\ &= D(f)S(\tau, f) + R(f)S(\tau, f), \end{aligned} \tag{4}$$

where DS is the direct signal component of S. Using this notation, we define that the true inverse filter W_T can be modeled as

$$W_T(f) = \mathcal{E}\left\{\frac{D(f)S(\tau, f)}{X(\tau, f)}\right\}, \tag{5}$$

$$= \frac{D(f)}{H(f)}. \tag{6}$$

Now, by assuming that the output of harmonic filter $\mathcal{H}\{X\}$ is an approximation of harmonic components of the direct signal as[1],

$$\mathcal{H}\{X(\tau, f)\} \simeq D(f)S_h(\tau, f), \tag{7}$$

the property of W_H can be analyzed as follows. Substituting eq. (7) for eq. (2) and using the Taylor expansion, W_H can be rewritten as

$$W_H(f) \simeq \frac{D(f)}{H(f)}\mathcal{E}\left\{\frac{S_h(\tau, f)}{S_h(\tau, f) + S_n(\tau, f)}\right\}, \tag{8}$$

$$= W_T(f)\{P_h(f) + C(f)\}. \tag{9}$$

Here $P_h(f)$ is a probability that $S_h(\tau, f)$ has a larger energy than $S_n(\tau, f)$ at each f,

$$C(f) = P_h(f)C_1(f) + (1 - P_h(f))C_2(f), \tag{10}$$

$$C_1(f) = \sum_{k=1}^{\infty}(-1)^k\mathcal{E}_{|Q(\tau,f)|<1}\{Q(\tau, f)^k\} \tag{11}$$

$$C_2(f) = -\sum_{k=1}^{\infty}(-1)^k\mathcal{E}_{|Q(\tau,f)|>1}\{Q(\tau, f)^{-k}\}, \tag{12}$$

$$Q(\tau, f) = \frac{S_n(\tau, f)}{S_h(\tau, f)}, \tag{13}$$

and $\mathcal{E}_{f(Q)}\{Q^k\}$ is an average of Q^k for Q that satisfies $f(Q)$. With speech signals, $\angle Q$ is expected to be uniformly distributed within $[-\pi, \pi)$ and independent of $|Q|$. Therefore, the following equation is expected to hold if a sufficiently large number of time frames are available for calculating the average value.

$$\mathcal{E}_{|Q|<1}\{Q^k\} = \mathcal{E}_{|Q|>1}\{Q^{-k}\} = 0 \text{ for } k > 0. \tag{14}$$

Consequently, W_H is shown to have the following property.

$$W_H(f) \simeq P_h(f)W_T(f). \tag{15}$$

Equation (15) indicates that the dereverberation effect of W_H is approximately the same as that of W_T, although it has an additional effect caused by P_h. Note that, P_h is a term that manipulates only the gain of the filter at each f because it takes a real value between 0.0 and 1.0.

[1] A physical interpretation of this assumption is discussed in [7]

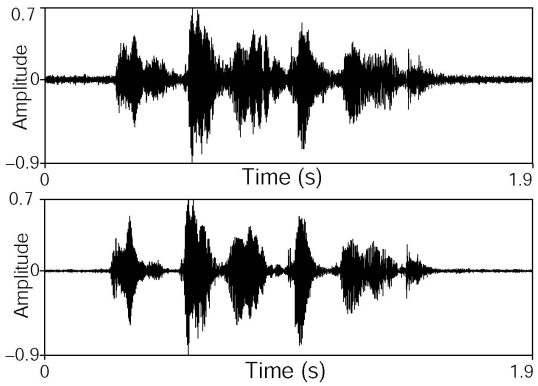

Fig. 1. Waveforms of speech signals dereverberated by W_H (top) and W_{FH} (bottom) when both inverse filters were estimated using one-minute training data. (RT=1s)

2.2. Problem

Equation (14), however, does not hold if a sufficiently large number of time frames are not available. In that case, $C(f)$ in eq. (9) cannot be disregarded, and it generates an additive noise component when applied to an observed signal. Therefore, HERB requires a large number of time frames to achieve high quality dereverberation.

The top of Fig. 1 shows the waveform of a speech signal dereverberated by W_H, which was derived using small amount of observed signal. We can see that the additive noise component generated by W_H appears as random noise in the signal. This can easily be confirmed by multiplying eq. (9) with an observed signal X to obtain the dereverberated signal Y as follows.

$$\begin{aligned} Y(\tau, f) &= W_H(f)X(\tau, f), \\ &= P_h(f)D(f)S(\tau, f) + C(f)D(f)S(\tau, f). \end{aligned} \tag{16}$$

The second term on the right hand side of eq. (16) can be considered the additive noise component that remains after dereverberation. According to eqs. (11) and (12), $C(f)$ is the weighted sum of sub-functions Q and its powers over time frames. The value Q at each time frame can be considered a function that transforms a harmonic component S_h into a non-harmonic component S_n. Since S_h and S_n have no fixed linear mapping from one to the other, Q is expected to be a random function. Therefore, it transforms a speech signal into random noise, as can be seen in Fig. 1. A similar discussion is also given for Q^{-1}. As a consequence, $C(f)$ is expected to be a function that generates random noise.

3. SOLUTION

Here we describe our solution to the problem presented in the previous section. First, we apply a noise reduction algorithm denoted as function $\mathcal{F}\{\cdot\}$ to both sides of eq. (16), with the aim of eliminating only the additive noise term $C(f)D(f)S(\tau, f)$. Care should be taken to implement function $\mathcal{F}\{\cdot\}$. As shown by eq. (16), the noise included in $Y(\tau, f)$ changes its amplitude according to the gain of input signal $S(\tau, f)$. Therefore the noise reduction algorithm should be able to track non-stationary noise. In addition, with blind dereverberation, the Speech to Noise Ratio (SNR) in $Y(\tau, f)$ is totally unpredictable, and potentially very low. Therefore the performance of a traditional speech activity detector cannot be guaranteed. In other words, it is appropriate to use an algorithm that has a robust speech activity detector, or an algorithm that does not require one such as a Kalman filter [10]. For this purpose, we introduce a previously proposed noise reduction method based on minimum statistics [11]. This method is capable of adaptively estimating the background noise level without any distinction between speech activity and speech pause. Using an algorithm that

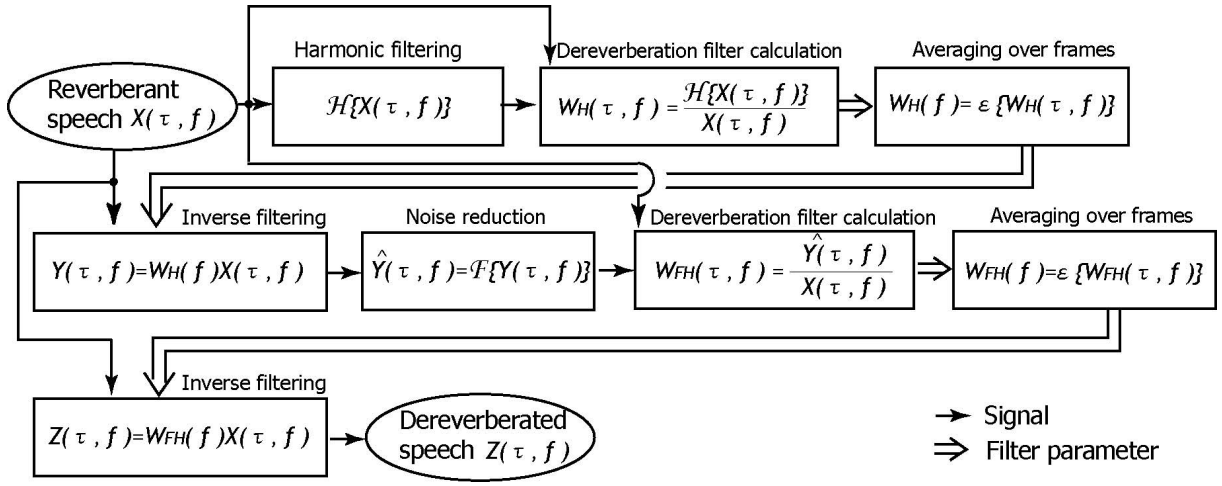

Fig. 2. Block diagram of Fast HERB

meets the requirements, we were able to obtain an approximation of a direct signal from a HERB dereverberated signal as

$$\hat{Y}(\tau,f) = \mathcal{F}\{W_H(f)X(\tau,f)\}, \quad (17)$$
$$= \mathcal{F}\{P_h(f)D(f)S(\tau,f) + C(f)D(f)S(\tau,f)\}, \quad (18)$$
$$\simeq P_h(f)D(f)S(\tau,f). \quad (19)$$

However, another problem arises here. Although it might be possible to consider $\hat{Y}$ as resultant dereverberated speech, $\hat{Y}$ is very likely to contain some deviations from the direct signal such as a residual noise component, or an estimation errors. To solve this problem, we consider $\hat{Y}$ to be a reference signal, rather than a final dereverberated signal, thus allowing us to re-estimate a more accurate inverse filter. Estimation errors in the inverse filter caused by $\hat{Y}$ can be greatly reduced by averaging over several frames. Now a more accurate inverse filter can be estimated as

$$W_{FH}(f) = \mathcal{E}\left\{\frac{\mathcal{F}\{Y(\tau,f)\}}{X(\tau,f)}\right\}, \quad (20)$$
$$\simeq \mathcal{E}\left\{\frac{P_h(f)D(f)S(\tau,f)}{X(\tau,f)}\right\}, \quad (21)$$
$$= P_h(f)W_T(f). \quad (22)$$

Finally, we were able to obtain an accurate estimation of the true inverse filter W_T using an even smaller amount of training data. This entire dereverberation procedure is referred to as "Fast HERB". The bottom of Fig. (1) shows the waveform of a speech signal dereverberated by W_{FH}.

4. IMPLEMENTATION OF NEW METHOD

A block diagram of Fast HERB is shown in Fig. 2. First, input reverberant speech X is divided into 5.5 second frames with rectangular windows each overlapping by 75%. The value τ in Fig. 2 denotes a time frame index. The window length and its overlapping rate were chosen arbitrarily. In each frame, harmonic components within $X(\tau,f)$ are extracted using harmonic filtering $\mathcal{H}\{\cdot\}$ as described in [7, 8]. Based on the extracted harmonic components $\mathcal{H}\{X(\tau,f)\}$ and $X(\tau,f)$, we calculate the initial estimation of the dereverberation filter as $\frac{\mathcal{H}\{X(\tau,f)\}}{X(\tau,f)}$ in each time frame. By averaging the initial estimated value over several frames, we obtain a first dereverberation filter $W_H(f)$. Proceeding to the next step in Fig. 2, $W_H(f)$ is applied to $X(\tau,f)$ to obtain $Y(\tau,f)$. We use an overlap-add technique to synthesize Y. Since the resultant dereverberated speech $Y(\tau,f)$ has an additive noise as shown in the previous section, a noise reduction algorithm $\mathcal{F}\{\cdot\}$ is employed to obtain an estimation of the direct signal $\hat{Y}(\tau,f)$ as in eq. (19). The noise reduction algorithm used here is a spectral subtraction based on minimum statistics proposed in [11], which meets the requirements described in section 3. After the noise reduction, a more accurate dereverberation filter is calculated as $\frac{\hat{Y}(\tau,f)}{X(\tau,f)}$. Then, $W_{FH}(\tau,f)$s are averaged over the frames, and converted to $W_{FH}(f)$. Finally, $W_{FH}(f)$ is applied to $X(\tau,f)$ to obtain the final dereverberated speech $Z(\tau,f)$.

5. EXPERIMENT

In this section, we evaluate the effectiveness of Fast HERB in terms of audible quality and ASR performance, compared with conventional HERB.

5.1. Experimental conditions

Five spoken Japanese sentences were obtained from ATR data set B for each gender (male:MHT, female:FKN), as the training data for Fast HERB. The signals were sampled at 12kHz and quantized with 16-bit resolution. To simulate a reverberant environment, each sentence was convoluted with each of 4 impulse responses (RT: 0.1, 0.2, 0.5, 1.0 sec.) measured in advance. The total duration of the 5 reverberant sentences was about 1 minute. Dereverberation was performed on 4 impulse responses $\times$ 2 genders.

To compare the effectiveness of Fast HERB with that of conventional HERB, we performed HERB dereverberation for two different cases (case 1: 60-minute training data, case 2: 1-minute training data). The HERB dereverberation procedure is summarized in [9]. We also used ATR data set B as the filter training data for HERB; 503 sentences for each gender for case 1, 5 sentences for each gender for case 2. The total duration of the 503 reverberant sentences was about 56 minutes.

5.2. Subjective evaluation of audible quality

We informally evaluated the audible quality of the dereverberated speech processed by Fast HERB. Even though the training data for Fast HERB was only a minute long, all of 4 subjects judged the speech quality and intelligibility to be greatly improved. Speech samples are available at [12].

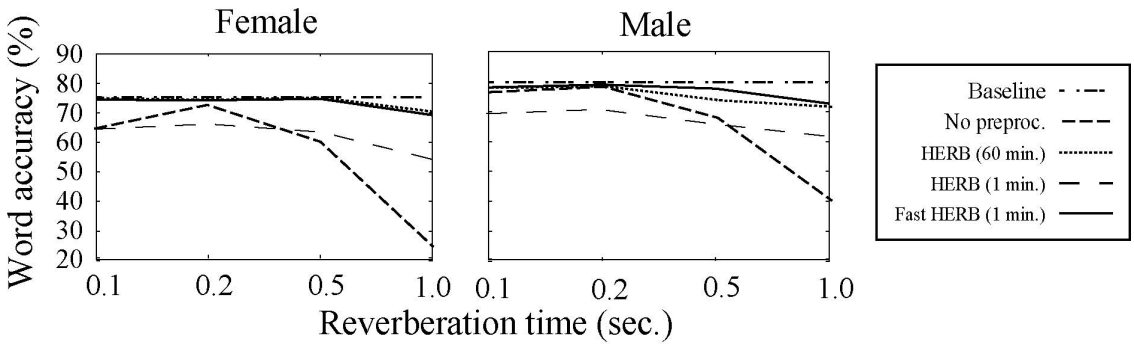

Fig. 3. The result of ASR experiment

5.3. Improvement of ASR performance

We then investigated the effectiveness of Fast HERB as a preprocessing algorithm for ASR. The ASR performance was evaluated in terms of speaker dependent word accuracy. In the acoustic model, we used the following parameters : 12 order MFCCs, 12 order delta MFCCs, 3 state HMMs, and 4 mixture Gaussian distributions. The language model that was used was trained on Japanese newspaper articles written over a ten-year period.

The acoustic model was trained on speech signals observed under various reverberant environments excluding the recognition target environment. We call this "multicondition training". For example, an acoustic model trained on RT 0.1s, 0.2s and 0.5s dereverberated speech was used to recognize RT 1.0s dereverberated speech. The model trained on reverberant (dereverberated) speech was used to recognize reverberant (dereverberated) speech. Cepstral Mean Normalization [13] was used for all the recognition tasks. This model training methodology is summarized in [9]

As a result of the experiments, in all reverberant environments, the recognition rate for the Fast HERB dereverberated speech was comparable to that of clean speech.

Figure 3 shows the word accuracy of the baseline performance, no preprocessing case, HERB using 60 min. training data, HERB using 1 min. training data and Fast HERB using 1 min. training data, under each reverberant environment, for each gender. The baseline in Fig. 3 represents the word accuracy for clean speech recognized with the clean acoustic model. If we do not employ a dereverberation algorithm as a preprocessing, the word accuracy decreases especially in a severely reverberant environment. If a large amount of training data is available, HERB can estimate an accurate inverse filter and bring the word accuracy close to the baseline performance. In contrast, when there is only 1 min. of training data, the HERB dereverberated speech score is lower than without preprocessing, especially in a moderately reverberant environment. It should be noted that, with Fast HERB, the word accuracy recovered to approximately the level obtained with HERB using 60 min. training data.

6. CONCLUSION

A speech signal captured by a distant microphone is generally smeared by reverberation, which severely degrades both speech intelligibility and the Automatic Speech Recognition (ASR) performance. In this paper, we proposed a new dereverberation scheme, named "Fast HERB". When the amount of training data for the estimation of an inverse filter is insufficient, the dereverberated speech processed by conventional HERB was analyzed and found to contain direct sound and additive random noise. To deal with this problem, Fast HERB employed a noise reduction algorithm, and enabled the fast estimation of a precise inverse filter. Our experiments showed that Fast HERB successfully achieved high quality dereverberation with a much smaller amount of training data than conventional HERB, and was very effective at improving ASR performance even under unknown severely reverberant environments. In addition, the audible quality of the dereverberated speech was found to be fairly good.

7. REFERENCES

[1] B. Kingsbury, N. Morgan, "Recognizing reverberant speech With Rasta-Plp," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1259-1262, 1997

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2), pp. 113-120, 1979

[3] F. Martin, K. Shikano and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models," *Proc. of Eurospeech*, pp. 1031-1034, 1993.

[4] J. L. Flanagan, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of Acoustical Society of America*, 78 (11), pp. 1508-1518, Nov., 1985

[5] S. Amari, S. C. Douglas, A. Cichocki and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," *Proc. of IEEE Workshop on Signal Processing Advances in Wireless Communications*, pp. 101-104, April, 1997

[6] B. Yegnanarayana and P .Satyanarayana, "Enhancement of reverberant speech using LP residual," *IEEE Trans. on Speech and Audio Processing*, 8 (3), pp. 267-281, May 2000.

[7] T. Nakatani and M. Myoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 92–95, 2003

[8] T. Nakatani, M. Myoshi and K. Kinoshita, "Implementation and effects of single channel dereverberation based on the harmonic structure of speech," *Proc. of International Workshop on Acoustic Echo and Noise Control*, pp. 91-94, 2003.

[9] K. Kinoshita, T. Nakatani, and M. Myoshi, "Improving automatic speech recognition performance and speech intelligibility with harmonicity based dereverberation," *Proc. of International Conference on Spoken Language Processing*, (to appear), 2004.

[10] J. D. Gibson, B. Koo and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Processing*, 39 (8), pp. 1732-1741, August 1991.

[11] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. of European Association for Signal Processing*, pp. 1182-1185, 1994.

[12] http://www.kecl.ntt.co.jp/icl/signal/kinoshita/publications/icassp05/

[13] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of Acoustical Society of America*, 55(6), pp. 1304-1312, 1974.