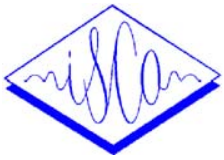

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Harmonicity based monaural speech dereverberation with time warping and F_0 adaptive window

Tomohiro Nakatani, Keisuke Kinoshita, Masato Miyoshi, Parham S. Zolfaghari

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
nak@cslab.kecl.ntt.co.jp

10.21437/Interspeech.2004-318

Abstract

Although a number of dereverberation methods have been reported, dereverberation is still a challenging problem especially when a single microphone is used. To overcome this problem, we proposed a harmonicity based dereverberation method (HERB). HERB can blindly estimate the inverse filter of a room transfer function based on the harmonicity of speech signals and dereverberate the signals. However, HERB uses an imprecise assumption that hinders the dereverberation performance, that is, the fundamental frequency (F_0) of a speech signal is assumed to be constant within a short time frame when extracting the features of harmonic components. In this paper, we combine HERB with time warping analysis and an F_0 adaptive window to remove this bottleneck. This extension makes it possible to estimate harmonic components precisely even when their frequencies change rapidly. Experiments show that time warping analysis with an F_0 adaptive window can effectively improve the dereverberation effect of HERB.

1. Introduction

Speech dereverberation is desirable in applications such as robust automatic speech recognition (ASR). Although several adaptation techniques have been proposed for recognizing reverberant speech signals [1], they can only deal with short reverberation. The recognition performance cannot be improved sufficiently when the reverberation time is longer than 0.5 sec, even for the case where acoustic models that have been trained with a matched reverberation condition are used.

Several blind dereverberation techniques have been proposed that employ microphone array systems. A typical technique involves estimating the directions of arrival (DOAs) of a direct speech signal, and enhancing signal components coming from that direction. The delay-and-sum beamformer is often used for this purpose. However, it requires a large number of microphones to achieve a large dereverberation gain. By contrast, another technique based on inverse filtering can suppress reverberation using a small number of microphones. Several blind techniques for estimating the inverse filter have been proposed based on the assumption that a source signal is a statistically independent and identically distributed (i.i.d.) sequence [2]. However, they cannot appropriately deal with speech because speech has inherent properties, such as harmonicity and formant structure, making its sequences statistically dependent.

To overcome these problems, we have recently proposed a new dereverberation method, known as *Harmonicity based dEReverBeration* (*HERB*) [3]. HERB can estimate the inverse filter of a room transfer function based on the harmonicity of speech signals. The experiments showed that HERB can effectively dereverberate speech signals when sufficiently long observed signals are given. However, HERB does not have as great a dereverberation effect for male speech signals as for female ones. In addition, HERB cannot deal appropriately with higher frequency components whose frequencies change rapidly with time.

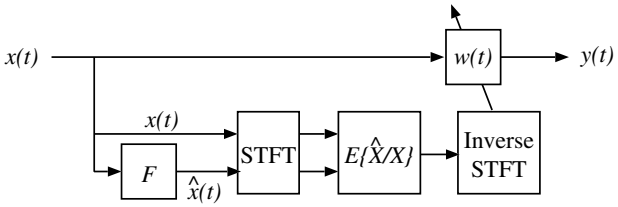

Figure 1: Diagram of HERB: each observed signal, $x(t)$, is first filtered by an adaptive harmonic filter, F, to obtain $\hat{x}(t)$. The average ratio of $\hat{x}(t)$ to $x(t)$ in the frequency domain over different observations is calculated to determine the dereverberation filter, $w(t)$. Finally, $x(t)$ is dereverberated by being convolved with $w(t)$.

In this paper, we present an extended version of HERB, referred to as *HERB with TimE wArping* (*HERB-TEA*), to improve the preciseness of the dereverberation. We believe the above problems with HERB to be caused by its imprecise treatment of harmonic components, that is, F_0 is assumed to be a constant within a short time frame. To overcome this problem, we introduce time warping analysis with an F_0 adaptive window, which expands and contracts the time axis of the signals to make their F_0s approximately constant. Time warping analysis allows us to extract the features of harmonic components precisely [4]. Our experiments show that this extension successfully improves the performance of HERB, especially for male speech signals.

2. Dereverberation method

In this section, we describe the dereverberation filter estimated by HERB, discuss its problems, and provide details of its extension.

2.1. Dereverberation filter estimated by HERB

Figure 1 shows a diagram of HERB. HERB estimates the dereverberation filter as an average filter that transforms observed reverberant signals into the output of an adaptive harmonic filter, which roughly estimates the direct harmonic components of the observed signals. The dereverberation filter, $W(\omega)$, is calculated as follows[1]:

$$W(\omega) = E\left\{\frac{\hat{X}(\omega)}{X(\omega)}\right\}, \quad (1)$$

[1] In this paper, time and frequency domain signals are represented by lower and upper case symbols, respectively. Arguments "(ω)" that represent the center frequencies of the short time Fourier transformation (STFT) bins are often omitted from frequency domain signals.

where X and $\hat{X}$ are the observed reverberant signal and the output of an adaptive harmonic filter, respectively. $E\{\cdot\}$ is an average function that calculates the average value of $\hat{X}/X$ for different observed signals from a sound source.

This filter can be proven to approximate the inverse filter of a room transfer function.

2.1.1. Interpretation of dereverberation filter

A speech signal, $S(\omega)$, can be modeled by the sum of the harmonic components, $S_h(\omega)$, and non-harmonic components, $S_n(\omega)$. A room transfer function, $H(\omega)$, can be divided into a direct component $D(\omega)$ and a reverberation part $R(\omega)$ where $H = D + R$. Then, the observed signal, X, can be represented as follows:

$$X(\omega) = H(\omega)S_h(\omega) + H(\omega)S_n(\omega), \quad (2)$$

$$= D(\omega)S_h(\omega) + (R(\omega)S_h(\omega) + H(\omega)S_n(\omega)), \quad (3)$$

where DS_h is a direct signal of S_h, RS_h is the reverberation of S_h, and HS_n is an observed signal of S_n. Of these components, DS_h can be approximated from X by an adaptive harmonic filter. This approximated direct signal $\hat{X}(\omega)$ can be modeled as follows:

$$\hat{X}(\omega) = D(\omega)S_h(\omega) + (\hat{R}_h(\omega) + \hat{H}_n(\omega)), \quad (4)$$

where $\hat{R}_h(\omega)$ and $\hat{H}_n(\omega)$, respectively, are part of the reverberation of S_h and part of the observed signal of S_n, which remain in $\hat{X}$ after the harmonic filtering. We assume that all the estimation errors in $\hat{X}$ are caused by $\hat{R}_h$ and $\hat{H}_n$ in eq. (4).

By substituting X and $\hat{X}$ in eq.(1) with eqs. (2) and (4), we can derive the following equation [5]:

$$W(\omega) \simeq \frac{D(\omega) + \hat{R}(\omega)}{H(\omega)} P\{|S_h(\omega)| > |S_n(\omega)|\}, \quad (5)$$

where

$$\hat{R}(\omega) = E\left\{ \frac{\hat{R}_h(\omega)}{S_h(\omega)} \right\}_{|S_h(\omega)|>|S_n(\omega)|}, \quad (6)$$

where $P\{\cdot\}$ is a probability function, and $E\{\cdot\}_A$ represents an average function under a condition where A holds.

Equation (5) means that W approximately coincides with the product of $(D + \hat{R})/H$ and $P\{|S_h| > |S_n|\}$. The former, $(D + \hat{R})/H$, strictly equals the inverse filter, D/H, when an adaptive harmonic filter can completely eliminate $\hat{R}_h$ in eq. (4) without any errors. Although it is very difficult to eliminate $\hat{R}_h$ completely, a major part of RS_h can be eliminated with an adaptive harmonic filter. In addition, $\hat{R}$ is defined as an average filter that transforms S_h to $\hat{R}_h$. Therefore, $\hat{R}$ is expected to become a transformation that produces reduced reverberation. As a consequence, the signal obtained by multiplying the observed signal X by $(D + \hat{R})/H$ is expected to be the sum of the direct signal and the reduced reverberation, that is, $((D + \hat{R})/H)X = DS + \hat{R}S$. By contrast, $P\{|S_h| > |S_n|\}$ in eq. (5) is the probability that the harmonic component has a larger energy than the non-harmonic component, and has a real value between 0.0 and 1.0. This term changes the gain of eq. (1) but does not affect its dereverberation function.

2.2. Problems

As discussed above, the precise extraction of direct harmonic components with an adaptive harmonic filter is very important for dereverberation. There are, however, the following problems involved with the adaptive harmonic filter used in HERB.

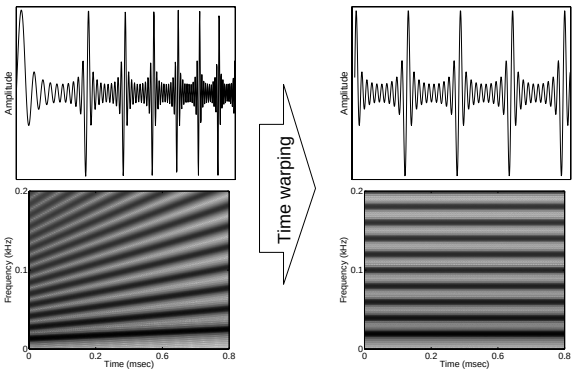

Figure 2: Waveforms (upper panels) and spectrograms (lower panels) of a signal before and after time warping. In this example, the fundamental frequency of the signal increases with time in the original time domain while it is constant in the warped time domain.

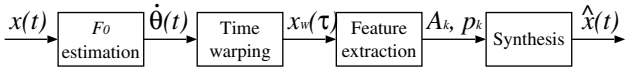

Figure 3: Processing flow of adaptive harmonic filtering with time warping

1. Features of harmonic components are extracted by assuming that the F_0 of speech signals is constant within a short time frame, although F_0 generally changes even in a local time region. This causes estimation errors in a direct signal, $\hat{X}(\omega)$, and thus degrades the dereverberation filter estimation.
2. When the F_0 of a speech signal is low, it is difficult to distinguish its direct signal from its reverberation part using an adaptive harmonic filter. This is because the differences between adjacent harmonic frequencies at a frame are small and relatively large parts of the reverberation overlap the direct signal.

As a consequence, HERB even increased the energies of the reverberations in certain experiments using male speech signals, compared with those of room impulse responses in time regions long after the direct signals had arrived (see section 3.1).

2.3. Extension of HERB

To improve the dereverberation performance of HERB, we extended it by combining it with time warping analysis and an F_0 adaptive window. This extended method is referred to as *HERB with TimE wArping* (*HERB-TEA*) in this paper.

2.3.1. Adaptive harmonic filter with time warping

Figure 2 illustrates the idea of time warping and Fig. 3 shows the flow of adaptive harmonic filtering when time warping analysis is employed. Time warping analysis first uses a time-warping function that expands and contracts the time axis of a signal in the original time domain to obtain a signal with an approximately constant F_0 in the warped time domain. The amplitudes and phases of the sinusoidal components are extracted from the signals in the warped time domain. A harmonicity enhanced signal is then synthesized in the original time domain using the extracted features and the time warping function.

Let $\tau = \mathcal{W}_l(t)$ be the time-warping function that transforms $x(t)$ within a short time frame, whose center time is t_l in the original time domain, into $x_{\mathcal{W}_l}(\tau)$ in the warped time domain, then the relation between $x(t)$ and $x_{\mathcal{W}_l}(\tau)$ is represented as

$$x_{\mathcal{W}_l}(\mathcal{W}_l(t)) = x(t) \ \text{ for } \ |t - t_l| < \frac{T}{2}. \tag{7}$$

where T is the length of the frame. In particular, let $\theta(t)$ be the phase of the F_0 component of $x(t)$, and $\phi(\tau)$ be that of $x_{\mathcal{W}_l}(\tau)$, then the relation between $\theta(t)$ and $\phi(\tau)$ is represented as

$$\phi(\mathcal{W}_l(t)) = \theta(t) \ \text{ for } \ |t - t_l| < \frac{T}{2}. \tag{8}$$

With time warping analysis, we determine $\mathcal{W}_l(t)$ so that it makes $\dot{\phi}(\tau)$ constant within a time frame as

$$\frac{d\phi(\tau)}{d\tau} = \dot{\phi}_{\tau_l} \ \text{ for } \ |{\mathcal{W}_l}^{-1}(\tau) - t_l| < \frac{T}{2}, \tag{9}$$

where $\tau_l = \mathcal{W}_l(t_l)$. τ_l and $\dot{\phi}_{\tau_l}$ are parameters that can be set to an arbitrary number[2]. In addition, to simplify the calculation, we assume that the time derivative of F_0 is constant within a short time frame in the original time domain, that is:

$$\frac{d^2\theta(t)}{dt^2} = \ddot{\theta}_{t_l} \ \text{ for } \ |t - t_l| < \frac{T}{2}, \tag{10}$$

where $\ddot{\theta}_{t_l}$ is the derivative of F_0 at time t_l. Then, the time warping function, $\mathcal{W}_l(t)$, that satisfies eqs. (8), (9) and (10) is derived as

$$\mathcal{W}_l(t) = (t - t_l)^2 \frac{\ddot{\theta}_{t_l}}{2\dot{\phi}_{\tau_l}} + (t - t_l)\frac{\dot{\theta}_{t_l}}{\dot{\phi}_{\tau_l}} + \tau_l, \tag{11}$$

$$\mathcal{W}_l{}^{-1}(\tau) = \begin{cases} \dfrac{(\dot{\theta}_{t_l}^2 + 2(\tau - \tau_l)\dot{\phi}_{\tau_l}\ddot{\theta}_{t_l})^{\frac{1}{2}} - \dot{\theta}_{t_l}}{\ddot{\theta}_{t_l}} + t_l, \\ \qquad\qquad\qquad\qquad \text{for } \ddot{\theta}_{t_l} \neq 0, \\ (\tau - \tau_l)\dfrac{\dot{\phi}_{\tau_l}}{\dot{\theta}_{t_l}} + t_l, \quad \text{for } \ddot{\theta}_{t_l} = 0. \end{cases} \tag{12}$$

The signal, $x_{\mathcal{W}_l}(\tau)$, in the warped time domain can then be obtained from $x(t)$ as

$$x_{\mathcal{W}_l}(\tau) = x({\mathcal{W}_l}^{-1}(\tau)). \tag{13}$$

The F_0 of this signal is expected to be constant because of the assumption of eq. (9), and thus, it is appropriate to model the signal with a sinusoidal representation. Let $X_{\mathcal{W}_l}(\omega)$ be the STFT of $x_{\mathcal{W}_l}(\tau)$, then the amplitude $A_{k,l}$ and phase $p_{k,l}$ of the k-th harmonic component in the warped time domain are extracted as

$$X_{\mathcal{W}_l}(\omega) = \sum_n g(\tau_n - \tau_l) x_{\mathcal{W}_l}(\tau_n) e^{-j\,\omega\,(\tau_n - \tau_l)}, \tag{14}$$

$$A_{k,l} = |X_{\mathcal{W}_l}(r\{k\dot{\phi}_{\tau_l}\})|, \tag{15}$$

$$p_{k,l} = \angle X_{\mathcal{W}_l}(r\{k\dot{\phi}_{\tau_l}\}). \tag{16}$$

where $g(t)$ is a window function and $r\{\cdot\}$ is a function that quantizes a continuous frequency into the discrete center frequency of the nearest STFT bin. Then, the output of the harmonic filter in the original time domain at this frame can be synthesized as

$$\hat{x}_l(t_n) = \sum_k A_{k,l} \cos(r\{k\dot{\phi}_{\tau_l}\}(\mathcal{W}_l(t_n) - \tau_l) + p_{k,l}), \tag{17}$$

Finally, overlap-add synthesis is used in order to combine signals over succeeding frames.

[2]For example, $\tau_l = 0$ and $\dot{\phi}_{\tau_l} = \dot{\theta}_{t_l}$ are reasonable parameter settings.

2.3.2. F_0 adaptive window

When the F_0 of a signal is constant, the frequency resolution of an adaptive harmonic filter can be increased simply by extending the window length. Therefore, we expect time warping analysis with a longer window to improve the accuracy with which harmonic components are estimated. However, a tradeoff still remains between frequency and time resolution because time warping cannot make F_0 completely constant. Therefore, we introduced an adaptive window length control, in which a long analysis window is applied only to signals with low F_0s in the low frequency region.

Let L be the default window length, and f_0 be the estimated F_0 at a frame, then we determine the adaptive window length L_a as

$$L_a = \begin{cases} (f_L/f_0)L & \text{for } \ f_0 < f_L, \\ L & \text{otherwise,} \end{cases} \tag{18}$$

where f_L is a constant.

2.3.3. Processing flow of HERB-TEA

HERB-TEA is implemented using the same processing flow as HERB, except that it uses time warping analysis with an F_0 adaptive window. The implementation is described in detail in [6].

3. Experiments

We evaluated the performance of HERB-TEA in terms of the energy decay curves of the impulse responses and ASR. The task used in our experiments was the dereverberation of reverberant word utterances. We used 5240 Japanese word utterances provided by a male and a female speaker (MAU and FKM) included in the ATR database as source signals, $s(t)$. We used four impulse responses measured in a reverberant room whose reverberation times were about 0.1, 0.2, 0.5, and 1.0 sec. Reverberant signals, $x(t)$, were obtained by convolving $s(t)$ with the impulse responses. Each dereverberation filter was estimated using all male word utterances or all female word utterances.

The length of the dereverberation filter was 131,072 taps, that is, we used a 10.9 sec rectangular window for the X and $\hat{X}$ calculations. We used a much shorter time frame, that is, a 42 msec hanning window and 1 msec window shift for the F_0 estimation. As regards the F_0 adaptive window, we adopted $L = 42$ msec, and $f_L = 200$ Hz when calculating eq. (14) for $\omega < 2000$ Hz. We used signals sampled at 12 kHz.

3.1. Energy decay curves of impulse responses

Figures 4 and 5 show energy decay curves of room impulse responses and dereverberated impulse responses obtained by HERB and HERB-TEA while controlling the reverberation time. Each dereverberated impulse response was obtained by convolving a room impulse response with its dereverberation filter, and each decay curve was calculated using Schroeder's method [7].

These figures show that HERB-TEA could effectively reduce the reverberation energy when the reverberation time was longer than 0.1 sec. HERB-TEA reduced the energy more successfully than HERB in all cases. This improvement was especially clear with male speech signals, that is, the energies of the dereverberated impulse responses in higher time regions were, in certain cases, increased by HERB compared with the energies of the room impulse responses, while they were effectively reduced by HERB-TEA. In addition, HERB-TEA also reduced the energy just after the direct signal more successfully than HERB in most cases. Because this part of the reverberation energy has the largest effect on speech intelligibility [8], HERB-TEA is expected to improve the intelligibility.

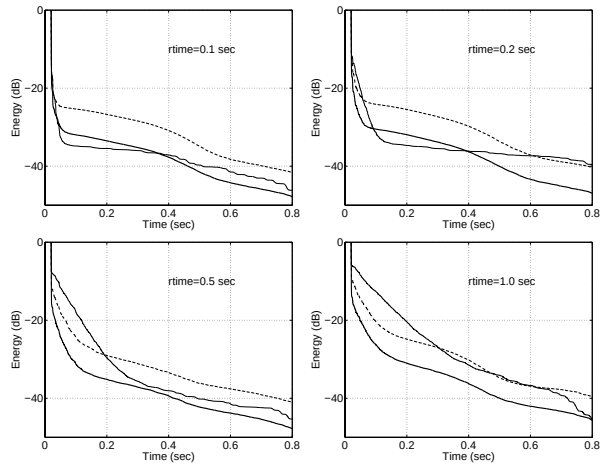

Figure 4: Energy decay curves of the room impulse responses (thin solid line) and dereverberated impulse responses (HERB: thin dashed line, HERB-TEA: thick line) for different reverberation times (rtime) when using male speech signals as training data.

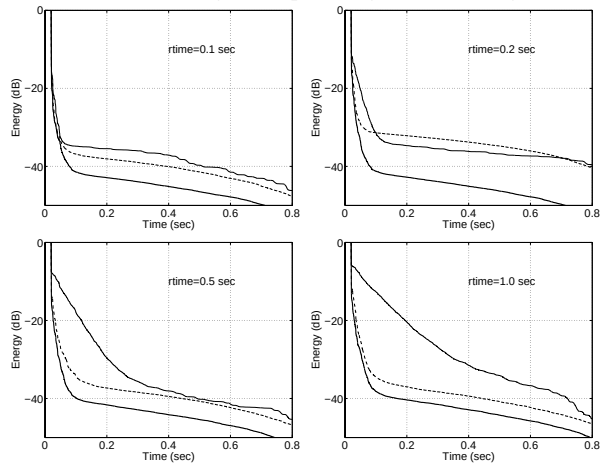

Figure 5: Energy decay curves of the room impulse responses (thin solid line) and dereverberated impulse responses (HERB: thin dashed line, HERB-TEA: thick line) for different reverberation times (rtime) when using female speech signals as training data.

3.2. Speaker dependent word recognition rate

We evaluated the speaker dependent word recognition rate (WRR) of reverberant and dereverberated speech signals. Reverberant signals were recognized using an acoustic monophone model trained on source signals. Signals dereverberated by HERB and by HERB-TEA were recognized using models trained on signals obtained by applying HERB and HERB-TEA to the source signals, respectively. We used 4740 words randomly selected from 5240 words as training data, and used the remaining 500 words as test data. We adopted 12-th order MFCCs, 12-th order delta MFCCs, three state HMMs, five mixture Gaussian distributions, a 25 msec frame length, and a 5 msec frame shift as the analysis conditions.

Figure 6 shows the resulting WRRs. The WRRs of the signals dereverberated by HERB-TEA were much improved compared with the others, and were more than 90% under all reverberation conditions. This means HERB-TEA can successfully reduce the spectral variations in speech signals produced by reverberation without losing the speech features essential for ASR.

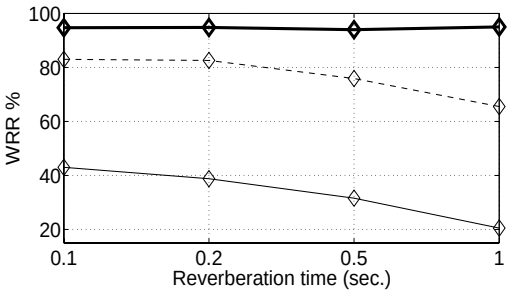

Figure 6: Average word recognition rates (WRRs) for reverberant signals (thin solid line) and dereverberated signals (HERB: thin dashed line, HERB-TEA: solid line) over male and female utterances under different reverberation time conditions.

4. Conclusion

This paper proposed a method for improving the dereverberation effect of the harmonicity based dereverberation method (HERB) by introducing time warping analysis with an F_0 adaptive window into its adaptive harmonic filtering. This extension allows us to extract features of harmonic components precisely even when their frequencies change within a short time frame. Experimental results showed that HERB with time warping (HERB-TEA) provided better dereverberation performance than HERB in terms of the energy decay curves of the impulse responses and ASR under various reverberation conditions. In particular, speaker dependent word recognition rates could be increased to more than 90% even with a 1.0 sec reverberation time. Future work will include an investigation of how such high quality speech dereverberation can be achieved with fewer speech data.

5. References

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

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

[3] Nakatani, T., and Miyoshi, M., "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP-2003*, vol. 1, pp. 92–95, Apr., 2003.

[4] Abe, T., and Honda, M., "Sinusoidal modeling based on instantaneous frequency attractors," *Proc. ICASSP-2003*, 2003.

[5] Nakatani, T., Miyoshi, M., and Kinoshita, K., "One microphone blind dereverberation based on quasi-periodicity of speech signals,", *NIPS-2003*, Dec., 2003.

[6] Nakatani, T., Miyoshi, M., and Kinoshita, K., "Implementation and effects of single channel dereverberation based on the harmonic structure of speech," *Proc. IWAENC-2003*, Sep., 2003.

[7] Schroeder, M.R., "New method of measuring reverberation time," *JASA*, 37, pp. 409–412, 1965.

[8] Yegnanarayana, B., and Ramakrishna, B.S., "Intelligibility of speech under nonexponential decay conditions," *JASA*, vol. 58, pp. 853–857, Oct. 1975.