Target-to-Non-target Directional Ratio Estimation Based on Dual-Microphone Phase Differences for Target-Directional Speech Enhancement

Seon Man Kim and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{kobem30002, hongkook}@gist.ac.kr

Abstract

In this paper, we propose a dual-microphone target-directional speech enhancement system utilizing target-to-non-target directional signal ratio (TNR) based on dual-microphone phase differences in adverse noise environments. One of the important issues associated with multi-microphone speech enhancement methods is the effective utilization of spatial cues such as phase differences for target-speech estimation within noisy speech. To this end, a TNR estimation method is presented based on phase differences between dual-microphone signals. Then, the estimated TNR is incorporated into a Wiener filter to obtain a masking filter for speech enhancement. Consequently, it is shown from a perceptual evaluation of speech quality that the performance of the proposed speech enhancement system outperforms those of conventional single- or dual-microphone speech enhancement systems based on a Wiener filter, beamformer, and phase-error-based filter under noise conditions with a signal-to-noise ratio ranging from 0 to 20 dB.

Index Terms: dual-microphone speech enhancement, target-to-non-target directional signal ratio, direction-of-arrival, phase difference, signal-to-noise ratio, beamforming

1. Introduction

The goal of speech enhancement is to suppress additive background noise components while maintaining the quality and intelligibility of speech [1]. This task is usually accomplished by preserving the characteristics of speech using the short-term spectral amplitude (STSA), which results in better performance for a multichannel speech enhancement system than for single-channel ones [2]. Multichannel speech enhancement utilizes the direction-of-arrival (DOA) information of a target speech, assuming that speech and noise are coming from different directions. The DOA is strongly linked to the phase difference between multichannel signals [2][3]. Thus, an important issue is the effective utilization of spatial cues such as phase differences for target speech estimation within noisy speech.

Among the available methods, a beamformer was utilized to construct spatial directional sensitivity and attenuate spatially unwanted noises arising from non-target directions [4][5]. However, the performance obtained via a dual-microphone system might be relatively unsatisfactory when compared with mask-based methods [6]. In addition, a post-filtering technique was also utilized to further improve the noise reduction performance of the beamformer [2][6]. Moreover, a soft-mask filtering method for dual-microphone speech enhancement, referred to as a phase-error-based filter (PEF), was proposed in [6]. The PEF method employed the spectral gain function of phase differences, which was motivated by the fact that phase difference errors between dual-microphone signals were related with the signal-to-noise ratio (SNR) of the observed noisy speech signal. Consequently, it was shown that the PEF method provided higher digit recognition accuracy than the beamformer [7] and the beamformer with a postfilter [8]. This implies that the phase difference plays a crucial role for the spatial cue in terms of reliability when estimating the target speech within noisy mixtures. Therefore, how to utilize the spatial cues such as phase differences to enhance speech becomes an important problem.

In this paper, to provide an effective solution to this problem, a method for estimating target-directional speech by exploring spatial cues is proposed. To this end, the proposed method attempts to reliably estimate the target-to-non-target directional signal ratio (TNR) by exploiting beamformers, which is derived as the function of phase difference between dual-microphone signals. Using the estimated TNR, a masking filter is constructed to obtain target speech from noisy signals.

Following this introduction, Section 2 reviews a conventional dual-microphone speech enhancement system based on the PEF method. Then, Section 3 describes an overview of the proposed speech enhancement system that incorporates the TNR obtained by phase differences. Next, the performance of the proposed system is evaluated by measuring the perceptual evaluation of speech quality (PESQ) in Section 4. Section 5 shortly discusses the performance of the proposed method in reverberant environments. Finally, concluding remarks are presented in Section 6.

2. PEF-based dual-microphone speech enhancement

Let us assume that input signals can be classified into target-directional speech and non-target-directional noise, and that the target speech is located far enough from the microphone array, which is referred to as the acoustic far-field condition. Let $X_{m,k}(\ell)$ be the k-th spectral component of the m-th microphone signal at the ℓ-th frame. Then, $X_{m,k}(\ell)$ can be represented as [2][6]

$$X_{1,k}(\ell) = T_k(\ell) + N_{1,k}(\ell) \tag{1}$$

$$X_{2,k}(\ell) = T_k(\ell)e^{-j\omega_k\tau_{12}} + N_{2,k}(\ell) \tag{2}$$

where $k = 0,1,\cdots,K-1$, $T_k(\ell)$ are the k-th spectral component representing the target-directional speech, and $N_{m,k}(\ell)$ (m = 1,2) are the m-th microphone signal corresponding to the non-target-directional noise $N_k(\ell)$ [2][6]. In addition, τ_{12} in (2) is the time difference of arrival (TDOA) between two microphones, and ω_k is the angular frequency in radians at the k-th frequency bin. Here, an estimate of τ_{12}, $\hat{\tau}_{12}$, can be obtained through localization algorithms such as the generalized cross

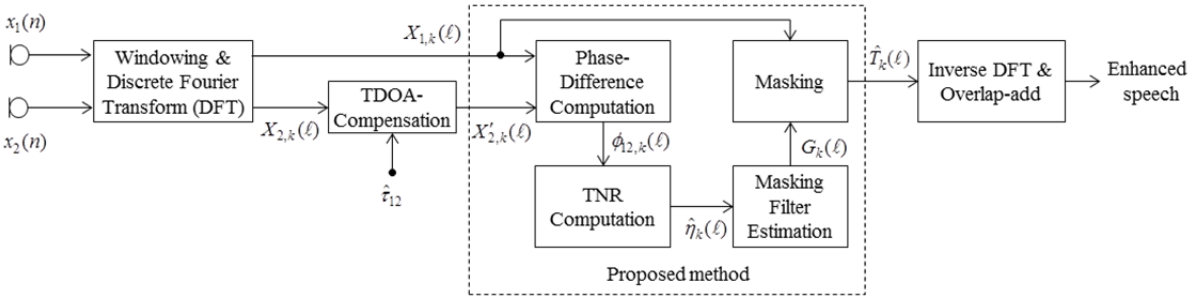

Figure 1: *Block diagram of the proposed dual-channel speech enhancement system.*

correlation (GCC) with phase transform (PHAT) weighting [2][6]. Then, by multiplying $\exp(j\omega_k\hat{\tau}_{12})$ into $X_{2,k}(\ell)$ in (1), we have

$$X'_{2,k}(\ell) = T_k(\ell) + N'_{2,k}(\ell) \tag{3}$$

where $N'_{2,k}(\ell) = N_{2,k}(\ell)\cdot\exp(j\omega_k\hat{\tau}_{12})$ also becomes a non-target-directional noise. In particular, the PEF method is motivated by the fact that a phase difference between $X_{1,k}(\ell)$ and $X'_{2,k}(\ell)$ is related with the TNR of the observed noisy speech signal [6]. In other words, PEF estimates the TNR based on the phase difference, which subsequently estimates a masking filter to enhance a target-directional speech by incorporating the estimated TNR to a Wiener filter.

First, let us define a TNR as $\eta_k(\ell) \equiv |T_k(\ell)|^2 / |N_k(\ell)|^2$. Then, the masking filter, $G_k^{TNR}(\ell)$, can be derived by employing the Wiener filter formulation as

$$G_k^{TNR}(\ell) = \frac{\eta_k(\ell)}{\eta_k(\ell)+\mu} \tag{4}$$

where μ is a constant used to control the degree of noise attenuation. It has been reported in PEF [6] that the TNR, $\eta_k(\ell)$, in (4) was related with the phase difference $\phi_{12,k}(\ell)$ between $X_{1,k}(\ell)$ and $X'_{2,k}(\ell)$, and the inverse of phase difference square, $1/|\phi_{12,k}(\ell)|^2$, could be used to approximate $\eta_k(\ell)$. Thus, the PEF method uses $1/|\phi_{12,k}(\ell)|^2$, instead of $\eta_k(\ell)$ to obtain a masking filter, which is denoted as $G_k^{PEF}(\ell)$ [6]. That is,

$$G_k^{PEF}(\ell) = \frac{1}{1+\mu\cdot|\phi_{12,k}(\ell)|^2}. \tag{5}$$

Even though $\eta_k(\ell)$ is estimated by approximating as $1/|\phi_{12,k}(\ell)|^2$, it was shown that this approximation was more effective in non-target-directional noise suppression than the beamformer-based ones [6].

3. Proposed dual-microphone speech enhancement based on TNR estimation

In this section, a dual-microphone speech enhancement system is proposed, where a masking filter is designed to estimate target speech from a noisy speech signal by utilizing phase differences. Fig. 1 shows a block diagram of the proposed speech enhancement system. First, the phase difference, $\phi_{12,k}(\ell)$, between $X_{1,k}(\ell)$ and $X'_{2,k}(\ell)$ in (3) is computed, which is then utilized to obtain an estimate of $\eta_k(\ell)$, $\hat{\eta}_k(\ell)$. Next, a masking filter, $G_k(\ell)$, is obtained based on $\hat{\eta}_k(\ell)$. which is subsequently applied to the first microphone signal in this paper.

3.1. TNR estimation based on phase differences

In this subsection, we explain a TNR estimation method based on only phase differences, similar to the PEF described in (5). Since a beamformer can be generally represented as a function of the phase difference [9], a beamformer is applied to a pair of inputs denoted as (1) and (3), resulting in the beamforming output, $BF_k(\ell)$. Thus, the transfer function of the beamformer for the target-directional signal, $G_k^{BF}(\ell)$, is represented as

$$G_k^{BF}(\ell) = \frac{BF_k(\ell)}{X_{1,k}(\ell)} = W_{1,k}^*(\ell) + W_{2,k}^*(\ell)\cdot\frac{|X'_{2,k}(\ell)|}{|X_{1,k}(\ell)|}\exp(j\phi_{12,k}(\ell)) \tag{6}$$

where $W_m\ (m=1,2)$ denotes a beamformer weight of the m-th microphone, and $*$ is a complex conjugate operator [2].

Compared with a conventional beamformer structure [2][4], which multiplies a complex-valued weight by each channel and then sums all the channels together, $G_k^{BF}(\ell)$ in (6) is represented as phase differences. In addition, because $|X'_{2,k}(\ell)|/|X_{1,k}(\ell)|$ in (6) can be approximated to unity according to the acoustic far-field assumption, $G_k^{BF}(\ell)$ becomes insensitive to a microphone gain mismatch [4]. Furthermore, by normalizing $\phi_{12,k}(\ell)$ as $\phi_{12,k}^{norm}(\ell) = \phi_{12,k}(\ell)\cdot c/(\omega_k\cdot d)$ with the microphone space, d, and the speed of sound, c, the performance of the beamformer can be improved further [10][11].

If the beamformer in (6) is a delay-and-sum beamformer (DSB) [2], then the weights are denoted as $W_{1,k}(\ell)=0.5$ and $W_{2,k}(\ell) = 0.5\exp(j\omega_k\tau_{12})$. Note that $\tau_{12}=0$ under the TDOA matched condition. Thus, the target-directional signal, $G_k^{DSB}(\ell)$, is estimated as

$$G_k^{DSB}(\ell) \approx 0.5(1+\exp(j\phi_{12,k}^{norm}(\ell))). \tag{7}$$

On the other hand, the transfer function for rejecting the target-directional signal, which is called a blocking matrix (BM)

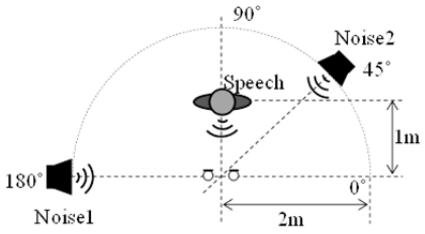

Figure 2: *Experimental setup for the performance evaluation of dual-microphone speech enhancement.*

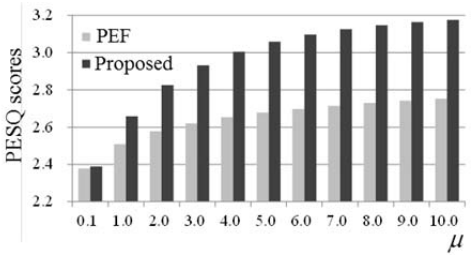

Figure 3: *Comparison of PESQ scores between PEF and the proposed method according to different values of μ.*

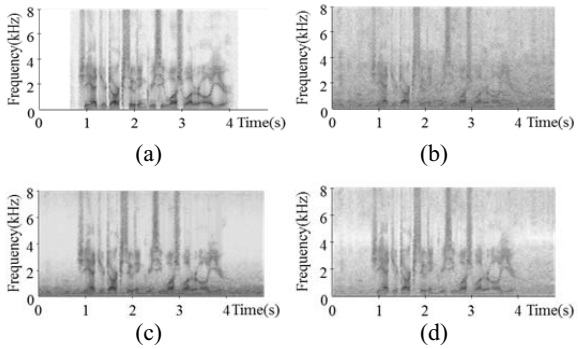

Figure 4: *Spectrograms of (a) clean male speech, (b) babble noisy speech at 5 dB SNR, (c) the output signal processed by PEF, and (d) the output signal processed by the proposed method.*

[4], is obtained using $W_{1,k}(\ell)=1$ and $W_{2,k}(\ell)=-\exp(j\omega_k\tau_{12})$. That is, the transfer function of BM, $G_k^{BM}(\ell)$, can be obtained as

$$G_k^{BM}(\ell)\approx 1-\exp(j\phi_{12,k}^{norm}(\ell))\,. \tag{8}$$

Finally, the TNR estimate is obtained by

$$\hat{\eta}_k(\ell)=\frac{|\hat{T}_k(\ell)|^2}{|\hat{N}_k(\ell)|^2} \tag{9}$$

where $\hat{T}_k(\ell)$ and $\hat{N}_k(\ell)$ are the estimates of $T_k(\ell)$ and $N_k(\ell)$, respectively, and they are represented as $\hat{T}_k(\ell)=G_k^{DSB}(\ell)\cdot X_{1,k}(\ell)$ and $\hat{N}_k(\ell)=G_k^{BM}(\ell)\cdot X_{1,k}(\ell)$. Then, combining (7) and (8), $\hat{\eta}_k(\ell)$ in (9) can be rewritten as

$$\hat{\eta}_k(\ell)\approx\left|\frac{1+\exp(j\phi_{12,k}^{norm}(\ell))}{2(1-\exp(j\phi_{12,k}^{norm}(\ell)))}\right|^2=\frac{1+\cos\phi_{12,k}^{norm}(\ell)}{4(1-\cos\phi_{12,k}^{norm}(\ell))}\,. \tag{10}$$

3.2. Masking filter and speech reconstruction

Similarly to (4), the masking filter for attenuating noise, $G_k(\ell)$, can be represented as

$$G_k(\ell)=\frac{\hat{\eta}_k(\ell)}{\hat{\eta}_k(\ell)+\mu} \tag{11}$$

where $\hat{\eta}_k(\ell)$ is the TNR estimate obtained from (10). In the above equation, μ is also a constant parameter to control the degree of noise attenuation. Finally, the target-directional speech spectral estimate, $\hat{T}_k(\ell)$, can be obtained by multiplying $G_k(\ell)$ and $X_{1,k}(\ell)$, which is then transformed into an enhanced speech signal.

4. Speech enhancement experiments

In this section, we evaluated the performance of the proposed target-directional speech enhancement method under simulated noisy conditions. In particular, it was assumed that τ_{12} in (2) was known *a priori* for all dual-microphone algorithms in the experiment.

4.1. Experimental setup

Fig. 2 illustrates an experimental setup for obtaining dual-microphone noisy speech signals in a non-reverberant room. The speech source was located at 1 m apart from the center of the dual-microphone array, which had a microphone space of 4 cm. For the experiment, two speech utterances (one male and one female) and babble noise were excerpted from the TIMIT database [13] and the NOISEX-92 database [14], respectively.

To simulate the target and non-target environments, two directional babble noise sources were panned so that they were positioned along the azimuth directions of 180° or 45° on a circle with a radius of 2 m, as shown in Fig. 2. Then, the panned babble noise was mixed with the target speech. In this paper, the noisy target speeches located at 180° and 45° were referred to as *S*1 and *S*2, respectively. In addition, noise 1 and 2 were added together and then mixed with speech, which was referred to as *S*3. Consequently, 50 test noisy speech signals (10 signals at SNRs of 0, 5, 10, 15, and 20 dB) were prepared for each simulation condition (*S*1, *S*2, and *S*3). Each test noisy signal was segmented using a 32 ms long hamming window, which corresponded to 512 samples at a sampling rate of 16 kHz. Furthermore, each segment was overlapped with the previous segment by half, thus an overlap-and-add method was used for generating an enhanced speech signal.

4.2. Experimental results

We first compared the perceptual evaluation of speech quality (PESQ) [12] performance of the processed signals by the PEF method in (5) and the proposed method in (10), where μ varied from 0.1 to 10, as shown in Fig. 3. The PESQ scores in the figure were averaged over all the sets and SNRs. It was shown from the figure that the proposed method provided higher PESQ scores than the PEF method for all the values of μ. Thus, we also set μ = 5 for the experiments in this paper, which was the same in the PEF method [6].

Table 1. *Comparison of average PESQ scores* over *S*1, *S*2, and *S*3 *simulation conditions for different speech enhancement methods under various SNR conditions.*

SNR (dB)	None	Wiener [1]	SDB [7]	GSC [5]	SDB+PW [2]	PEF [6]	Proposed
20	2.98	3.16	3.04	3.16	3.26	3.31	3.58
15	2.62	2.81	2.69	2.81	2.90	2.95	3.24
10	2.24	2.46	2.31	2.48	2.55	2.60	2.88
5	1.86	2.08	1.93	2.15	2.18	2.24	2.51
0	1.46	1.70	1.55	1.79	1.83	1.87	2.12
Avg.	2.23	2.44	2.30	2.48	2.55	2.59	2.87

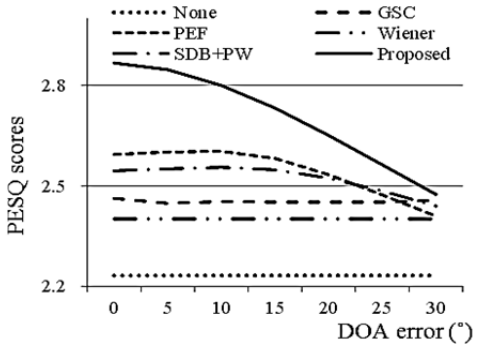

Figure 5: *Comparison of average PESQ of different speech enhancement methods depending on DOA errors from 0° to 30°.*

Next, we compared the spectrograms of the processed signals by the PEF and proposed methods. Fig. 4 shows the spectrograms of clean male speech, babble noisy speech at 5 dB SNR, the output signal processed by the PEF method, and the output signal processed by the proposed method, where the simulation condition was *S*1. As shown in the figure, the proposed method achieved better suppression performance for the non-target-directional noise than the PEF method. This result implies that proposed TNR estimate method could provide a more robust masking filter for speech enhancement than the PEF approach.

Third, we compared the quality of the processed speech signals by measuring PESQ scores. Table 1 compares the average PESQ scores of noisy speech signals processed by the different speech enhancement methods under different SNR conditions. Here, the PESQ scores were averaged over the three simulation conditions: *S*1, *S*2, and *S*3. As shown in the table, compared to conventional methods, the proposed method provided the highest PESQ scores for all the SNR conditions.

Next, we investigated the performance of the proposed method against the target DOA error. Fig. 5 compares the PESQ scores between the different speech enhancement methods when the target DOA was assumed to be estimated incorrectly from 0° to 30°. As shown in the figure, the Wiener filter and GSC provided nearly constant PESQ scores regardless of the DOA error. On the other hand, the super directive beamformer (SDB) with a Wiener filter [2], the PEF method, and the proposed method were shown to be sensitive to the DOA error. In particular, the proposed method showed the best performance near a DOA error of 0°, denoting an accurate target DOA estimate. Note that the proposed method degraded PESQ scores, when the DOA error increased. Nevertheless, the proposed method was deemed to be more effective for speech enhancement than the conventional methods.

Table 2. *Comparison of average PESQ scores over all simulation conditions (S1, S2, and S3) and SNRs (0, 5, 10, 15, and 20dB) of the processed speech signals using different speech enhancement methods under different noise conditions.*

Noise Type	None	Wiener [1]	SDB [7]	GSC [5]	SDB+PW [2]	PEF [6]	Proposed
Babble	2.23	2.44	2.30	2.48	2.55	2.59	2.87
Factory	2.19	2.48	2.30	2.48	2.62	2.66	2.86
Vacuum Cleaner	2.02	2.50	2.14	2.31	2.58	2.63	2.74
White	2.05	2.63	2.25	2.37	2.84	2.90	2.98
Avg.	2.12	2.51	2.25	2.41	2.65	2.70	2.86

Finally, we applied the proposed method to noisy speech signals under different types of noise. Table 2 compares the average PESQ scores over all simulation conditions (*S*1, *S*2, and *S*3) and SNRs (0, 5, 10, 15, and 20dB) of the different speech enhancement methods under four different noises. As can be seen from the table, the proposed method also yielded the highest PESQ scores among all the methods under all the different noise types.

5. Performance comparison of the proposed method in reverberant environments

In this paper, the proposed method was developed without considering any reverberant environments. Nevertheless, we evaluated the performance of the proposed method in reverberation environments. Similar to the previous discussion, we repeated the experiment after filtering noisy speech signals with a reverberant filter with different RT60's of 100, 200, and 300 ms. After that, the PESQ scores were measured for the different speech enhancement methods. The proposed method yielded the highest scores at 100 ms RT60 and comparable scores to the PEF method at 200 and 300 ms RT60. This implies that compared with the PEF method, the proposed method could effectively utilize phase differences even under reverberation conditions.

6. Conclusion

In this paper, we proposed a dual-microphone target-directional speech enhancement method. To achieve this, with the target speech DOA information, we could estimate a masking filter using a TNR estimate based on dual-microphone phase differences between dual-microphone signals. By performing the performance comparison with conventional signal and dual-microphone speech enhancement methods, it was shown that the proposed method outperformed conventional methods under different types of noise signals and different SNRs ranging from 0 to 20 dB.

7. Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No.2012-010636).

8. References

[1] Benesty, J., Makino, S., and Chen, J., Speech Enhancement, New York: Springer-Verlag, 2005.

[2] Brandstein, P. M. and Ward, D., Microphone Arrays, New York: Springer-Verlag, 2001.

[3] Yilmaz, O. and Rickard, S., "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. on Signal Proc., vol. 52, no. 7, 1830-1847, July 2004.

[4] Dolco, S. and Moonen, M., "Superdirective beamforming robust against microphone mismatch," IEEE Trans. on Speech Audio Proc., vol. 15, no. 2, pp. 617-631, Feb. 2007.

[5] Han, S., Hong, J., Jeong, S., and Hahn, M., "Probabilistic adaption mode control algorithm for GSC-based noise reduction," IEICE Trans. on Fundamentals of Electr., Communications and Computer Sciences, vol. E93-A, no. 3, pp.627-630, Mar. 2010.

[6] Aarabi, P. and Shi, G., "Phase-based dual-microphone robust speech enhancement," IEEE Trans. Systems, Man, and Cybernetics-Part B: Cyber., vol. 34, no. 4, pp. 1763-1773, Aug. 2004.

[7] Bitzer, J., Simmmer, K. U., and Kammeyer, K. D., "Multi-microphone noise reduction techniques for hands-free speech recognition-a comparative study," in Proc. of Robust Methods for Speech Recognition in Adverse Conditions (ROBUST 99), Tampere, FI, pp. 171-174, May 1999.

[8] Marro, C., Mahieux, Y., and Simmer, K. U., "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," IEEE Trans. Speech Audio Proc., vol. 6, pp. 240–259, May 1998.

[9] Kim, S. M. and Kim, H. K., "Hybrid probabilistic adaptation mode controller for generalized sidelobe canceller-based target-directional speech enhancement", in Proc. of ICASSP, Prague, CZ, May 2011.

[10] Araki, S., Sawada, H., Mukai, R., and Makino, S., "Underdetermined blind sparse source separation of arbitrarily arranged multiple sensors", Signal Proc., vol. 87, no. 8, pp.1833-1847, Mar. 2007.

[11] Mukai, R., Sawada, H., Araki, S., and Makino, S., "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," IEICE Trans. Fundamentals, vol. E87-A, no. 8, pp. 1941–1948, Aug. 2004.

[12] ITU-T Recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ), and Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Coders, Feb. 2001.

[13] Garofolo, J., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V., TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia: Linguistic Data Consortium, 1993.

[14] Varga, A., Steenneken, H. J. M., Tomilson, M., and Jones, D., The NOISEX–92 Study on the Effect of Additive Noise on Automatic Speech Recognition, Documentation on the NOISEX–92 CD–ROMs, June 1992.