Interference Robust DOA Estimation of Human Speech by Exploiting Historical Information and Temporal Correlation

Wei Xue, Shan Liang, Wenju Liu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

{wxue,sliang,lwj}@nlpr.ia.ac.cn

Abstract

Although various DOA estimation methods for human speech have been presented, most of them assume noises received by different microphones are undirected. However, strong directional interferences often also exist in practical scenarios and the performances of existing methods degrade seriously in such case. In this paper, we present a new interference robust DOA estimation method for human speech. Historical information and temporal correlation are taken advantage to deal with the problem. Firstly, utilizing the historical DOA estimates, we perform "post-beamforming" in the last frame to suppress the directional interferences. Then exploiting temporal correlation of speech spectra, frequency weights which highlight the effects of speech frequency bins are calculated based on the estimated *a priori* SNR of enhanced signal. Finally, we propose a new DOA cost function using frequency-weighted spatial correlation matrix to estimate the DOA of speech source. Experimental results show that the proposed method outperforms existing algorithms in reverberant environments with additive white Gaussian noises in the presence of different kinds of interferences.

Index Terms: direction of arrival estimation, microphone array signal processing, directional noise

1. Introduction

Direction of arrival (DOA) estimation mainly refers to determining the direction of sound sources in an acoustic environment using microphone arrays. It has drawn much attention due to its wide applications, such as in hand free devices, audio-visual conferences, automotive systems and surveillance. In most scenarios the sources of interest are human speech sources.

Generally, conventional DOA estimation methods can be classified into three categories: high-resolution spectral (HOS) estimation[1,2,3], steered beamformer response power (SRP)[4,5], and time difference of arrival (TDOA) estimation[6,7]. Although many approaches have been proposed, these classical methods assume that noises in different microphones are undirected. However, in practical scenarios, directional interferences often also exist. If the interference signals are weak, conventional methods may stay robust by finding the maximum peak in the DOA cost function. However, as the speech is non-stationary, the peak of cost function corresponding to DOA of speech source is not always global maximum, especially when strong interferences exist. One may take interferences as regular sources, and solve the problem in the framework of multi-source DOA estimation. However, multi-source estimation methods always require the knowledge of the source number which is often unknown in advance in the real environment. Furthermore, additional efforts must be paid to distinguish which direction corresponds to the speech source, which is also a difficult task.

In this paper, we aim at estimating the DOA of the target speech source by finding only the maximum peak in the DOA cost function. A novel DOA estimation method for human speech which is robust to the non-speech interferences is proposed. We utilize the historical information and temporal correlation of consecutive speech spectra in the proposed algorithm. First, depending on the historical DOA estimates, we perform post-beamforming to obtain a single-channel signal in which the interference signals are suppressed. Then by exploiting the correlation of frequency spectra in consecutive speech frames, *a priori* SNR based frequency weights are computed. The SNR-based frequency weights have large values in the speech frequency bins and zeros in the non-speech frequency bins. Finally, a new DOA cost function is defined based on the frequency-weighted spatial correlation matrix (FWSCM) utilizing the computed frequency weights. The new cost function enables us to enlarge the impact of speech frequency bins by frequency weighting, but avoids estimating DOAs separately in each band. Experimental results demonstrate that the proposed method outperforms existing algorithms in reverberant and noisy environments in the presence of different kinds of interferences.

The rest of the paper is organized as follows. Section 2 formulates the problem and provides assumptions. In section 3 we introduce the proposed DOA estimation method. Experimental results are given in section 4, and finally section 5 concludes the paper.

2. Problem statement

Suppose that there exists a reverberant environment with an N-element uniform linear array (ULA), a speech source and several non-speech interferences. The sound source and interferences are all in the far field [8] and uncorrelated with each other.

As we are only interested in the speech source and the aim is to estimate the DOA of the interested source, for the sake of simplicity, we consider the received signal at the microphone array as a combination of the direct path component and the combined noise. The combined noise consists of the echo of the speech signal, the received interference signals, and additive white Gaussian noises. Using short-time Fourier transform (STFT), the received signal in time frame t and frequency bin f can be expressed in a vector form as:

$$Y(t,f) = A(f)S(t,f) + N_c(t,f), \tag{1}$$

where $Y(t,f)$ and $N_c(t,f)$ are both $N \times 1$ vectors denoting the STFT of the received signal and the combined noise, respec-

Copyright © 2013 ISCA

tively. $S(t,f)$ is the STFT of speech source, and $A(f)S(t,f)$ stands for the direct path part of received speech signal. $A(f)$ is called the steering vector of microphone array.

For ULA, according to the geometry, the steering vector is expressed as follows:

$$A(f) = [1, e^{-i2\pi f \frac{\sin(\hat{\theta}) f_s d}{c}}, \ldots, e^{-i2\pi f(N-1) \frac{\sin(\hat{\theta}) f_s d}{c}}]^T, \quad (2)$$

where c is the propagation speed of sound in the air, which is usually set to 343m/s, f_s is the sampling rate, d is the spacing between two adjacent microphones, and $\hat{\theta}$ is the true DOA to be estimated.

3. Proposed method

As described above, in the proposed method, the frequency weights are calculated firstly, and then used by the DOA cost function. In this section, we choose to begin with introducing the DOA cost function which utilizes the frequency weights. From the derivation of the cost function, we can make clear the needed characteristics of frequency weights for robust DOA estimation. Then we will show how to calculate the frequency weights with these characteristics.

3.1. Cost function for DOA

As we all know, speech signals and non-speech signals always have different frequency distributions, so better DOA estimation results can be expected by accentuating the effect of "speech" frequency bins in the overall DOA cost function. However, it is time-consuming if we estimate DOA in each frequency bin and weighted-sum-up the narrowband results to get the wideband estimation. Here a new cost function for DOA estimation is proposed based on the FWSCM, which enables frequency weighting and avoids the separated narrowband DOA estimation.

It can be seen from Eq.(2) that for the direct path speech signal, if the signal impinges on the microphone array from non-perpendicular directions, phase shifts are generated on different microphones. In other words, if we estimate the DOA correctly, the phase shifts of direct path signal can be totally compensated.

Define a phase compensation vector for a hypothesized DOA θ in frequency bin f as:

$$C(\theta, f) = [1, e^{i2\pi f \frac{\sin(\theta) f_s d}{c}}, \ldots, e^{i2\pi f(N-1) \frac{\sin(\theta) f_s d}{c}}]^T. \quad (3)$$

Then the phase compensation of the received signal is derived as follows:

$$\begin{aligned} Y^c(\theta,t,f) &= C(\theta,f) \circ Y(t,f) \\ &= C(\theta,f) \circ (A(f)S(t,f)) + C(\theta,f) \circ N_c(t,f), \end{aligned} \quad (4)$$

where $Y^c(\theta,t,f)$ is the phase-compensated signal vector, the symbol "$\circ$" stands for the Hadamard product. Obviously, once the hypothesized DOA θ is equal to $\hat{\theta}$, then

$$Y^c(\theta,t,f) = \Gamma S(t,f) + C(\theta,f) \circ N_c(t,f), \quad (5)$$

where $\Gamma = [1,1,\ldots,1]^T$, which indicates that the speech signals received by different microphones are phase aligned.

As the frequency distributions of speech and interference signals are different, we assume that in speech frequency bins, the interference component is absent. If we further neglect the effect of white Gaussian noise and reverberant part of speech signal in these frequency bins, Eq.(5) can be simplified as:

$$Y^c(\theta,t,f) = \Gamma S(t,f), f \in \Omega_s, \quad (6)$$

where Ω_s is the set of speech frequency bins.

One common way of selecting the speech frequency bins from the entire frequency bins is to use a set of weights and set zero weights to the non-speech bins. Furthermore, even for the speech frequency bins, different bins are not equally affected by non-speech components, as a result, larger weights should be given to the more "pure" speech frequency bins. Given the weight $w(t,f)$ for the fth frequency bin in time frame t ("how to compute the weight" will be described later), we define the FWSCM as:

$$\begin{aligned} \mathbf{R}(\theta,t) &= \sum_{f\in\Omega} [w(t,f)Y^c(\theta,t,f)][w(t,f)Y^c(\theta,t,f)]^H \\ &= \sum_{f\in\Omega} w^2(t,f)Y^c(\theta,t,f)[Y^c(\theta,t,f)]^H, \end{aligned} \quad (7)$$

where Ω is the set of all frequency bins, "$[\cdot]^H$" denotes Hermitian transpose.

If $w(f)$ has non-zero values only in speech frequency bins and the hypothesized DOA θ is equal to $\hat{\theta}$, according to Eq.(6),

$$\begin{aligned} \mathbf{R}(\theta,t) &= \sum_{f\in\Omega_s} w^2(t,f)|S(t,f)|^2 \Gamma\Gamma^H \\ &= \eta(t) \cdot \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}, \end{aligned} \quad (8)$$

where $\eta(t) = \sum_{f\in\Omega_s} w^2(f)|S(t,f)|^2$ which is a constant in a certain frame given $w(f)$. In this case, $\mathbf{R}(\theta,t)$ is matrix of rank 1. If $\theta \neq \hat{\theta}$, the $\mathbf{R}(\theta,t)$ will be semi-definite, and its rank will be greater than 1.

Let us perform the eigenvalue decomposition of $\mathbf{R}(\theta,t)$, and let $\lambda_1(\theta,t) \geq \lambda_2(\theta,t) \geq \ldots \geq \lambda_N(\theta,t)$ denote the N eigenvalues of $\mathbf{R}(\theta,t)$. Obviously, if $\mathbf{R}(\theta,t)$ is of rank 1, $\lambda_2(\theta,t) = \ldots = \lambda_N(\theta,t) = 0$. Therefore, if we form the following cost function

$$\mathbf{J}(\theta,t) = \frac{1}{\sum_{i=2}^{N} |\lambda_i(\theta,t)|}, \quad (9)$$

the cost function reaches the maximum if $\theta = \hat{\theta}$, then the estimated DOA $\tilde{\theta}(t)$is calculated as:

$$\tilde{\theta}(t) = arg\max_{\theta} \mathbf{J}(\theta,t). \quad (10)$$

In practice, in order to overcome the fluctuation of FWSCM estimation, we modify FWSCM by recursively smoothing it over time:

$$\hat{\mathbf{R}}(\theta,t) = \alpha\hat{\mathbf{R}}(\theta,t-1) + (1-\alpha)\mathbf{R}(\theta,t), \quad (11)$$

where $\hat{\mathbf{R}}(\theta,t)$ is the modified FWSCM of time frame t, and α is the smoothing factor which is usually chosen close to 1. $\mathbf{R}(\theta,t)$ is substituted by $\hat{\mathbf{R}}(\theta,t)$ for eigenvalue decomposition. By exploiting the recursively smoothing procedure, more reliable FWSCM estimation is achieved, which improves performance of DOA estimation.

3.2. Frequency weight calculation

As mentioned before, the frequency weights should have large values in speech frequency bins and zero values in non-speech frequency bins. In this subsection, by exploiting the historical DOA estimates and the correlation of frequency spectra in consecutive speech frames, the frequency weights are calculated based on post-beamforming and *a priori* SNR estimation.

3.2.1. Post-beamforming

As DOAs are estimated frame by frame continuously, the estimation results of past frames may provide valuable information to improve the accuracy of the estimation in the current frame. Supposing we have estimated DOA of speech source correctly in the last frame, if we perform beamforming on the multichannel signals of the last frame towards the estimated DOA, a signal in which the speech signal is enhanced and the interference signals are suppressed will be obtained.

It should be noted that the beamforming is performed "after" the DOA estimation of the last frame (so we call it post-beamforming), and only the multichannel signals of the last frame are utilized. Therefore, even though the historical estimates are used, we don't make any assumption of the DOA of the current frame. The role of post-beamforming is to provide us some "priori" knowledge of what the speech spectrum may look like in the current frame.

We adopt the simple delay-and-sum(DS) beamformer [9] to enhance the signal of one frame. The reason for choosing DS beamformer is that it compromises between the computational complexity and the capacity of suppressing the interferences. In the frequency domain, the beamformed signal towards direction θ is calculated as:

$$Y^B(t,f,\theta) = \frac{1}{N}\sum_{i=1}^{N} y_i(t,f) e^{i2\pi f(i-1)\frac{\sin(\theta) f_s d}{c}}, \qquad (12)$$

where $Y^B(t,f,\theta)$ is the beamformed signal, $y_i(t,f)$ denotes the STFT of the ith microphone signal in time frame t and frequency bin f.

Unfortunately, the DOA estimate of the last frame is not always correct. To reduce the negative effect of beamforming towards wrong directions, the enhanced signal of last frame is actually computed as the weighted sum of beamformed signals towards the last K DOA estimates:

$$Y^E(t-1,f) = \frac{1}{\sum_{i=1}^{K} e^{-\beta(i-1)}} \sum_{i=1}^{K} Y^B(t-1,f,\theta_i) e^{-\beta(i-1)}, \qquad (13)$$

where $Y^E(t-1,f)$ is the enhanced signal, θ_i is the ith DOA estimate ahead from the current frame, and $e^{-\beta(i-1)}$ is a fading weight with β controlling the fading rate. K and β is set to be 15 and 0.3 respectively here. With this formula, the enhanced signal does not depend only on the latest DOA estimate, although the latest estimate has larger impact than other estimates.

In the first K frames with limited historical DOA estimates, the enhanced signal is replaced by the signal received by the first microphone in the current frame. The enhanced signal will be used for the *a priori* SNR estimation in the following step.

3.2.2. A priori SNR based frequency weight

The interference signals are suppressed after post-beamforming. With the enhanced signal, we estimate the *a priori* SNR, which can be viewed as an indicator of speech presence in a certain frequency bin. Then the frequency weights are calculated by computing the correlation between the estimated SNR and the spectrum of current frame.

Here the *a priori* SNR is estimated based on selective cepstro-temporal smoothing, which has been presented in our group's previous work [10]. We measure the estimated *a priori* SNR by dB. The frequency weight $w(t,f)$ is further defined by a recursively smoothing procedure:

$$\begin{aligned} w(t,f) =& \rho w(t-1,f) \\ &+ (1-\rho)\{(\mathcal{L}[y_1(t,f)] + \zeta) \cdot max[\xi(t-1,f),0]\}^p \end{aligned} \qquad (14)$$

where $\xi(t-1,f)$ is the *a priori* SNR of $Y^E(t-1,f)$, ρ is the smoothing factor, p is a factor controlling the impact of high SNR frequency bins on the DOA cost function, $\mathcal{L}[y_1(t,f)]$ is the log-spectral amplitude of first microphone signal. A variable ζ is added to $\mathcal{L}[y_1(t,f)]$ to make the log-spectral amplitudes non-negative. A frequency bin is considered as non-speech bin if $SNR < 0dB$, and the weight for this bin is set to be 0, which is expressed by $max[\xi(t-1,f),0]$. In this paper, we set $\rho = 0.3$, $p = 2$, and $\zeta = 15$.

Fig.1 shows an example of the calculated frequency weight and its corresponding spectrograms of the clean speech and noisy speech. It can be observed that under the noisy condition, the calculated frequency weight can roughly reflect the spectrogram of the speech signal, and the effect of the interference signal has been almost totally removed.

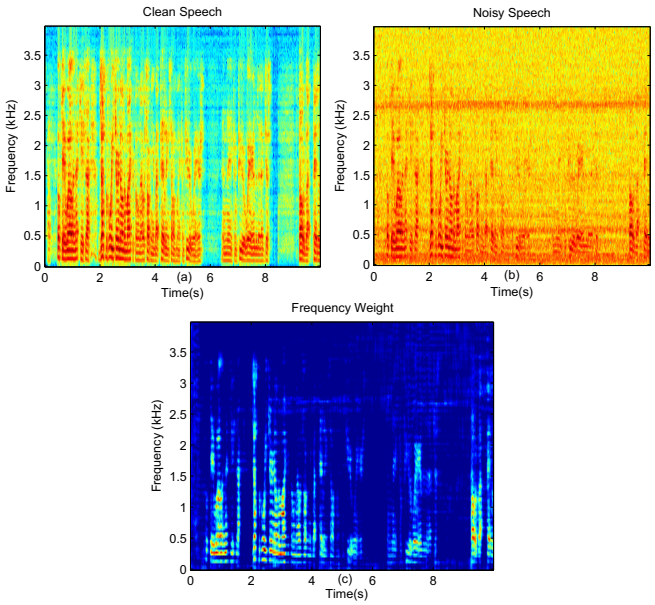

Figure 1: *Spectrograms of clean speech (a) and noisy speech (b), calculated frequency weight (c) in time-frequency domain. Noisy environment: 8dB SIR F16 cockpit noise + 15dB SNR white Gaussian noise + T60 250ms reverberation*

4. Experiment

In order to evaluate the performance of the proposed algorithm, we conduct experiments on the synthetic data. Two well-known conventional algorithms, the SRP-PHAT [5] and broadband MUSIC algorithm[3], are used for comparison.

4.1. Experimental setup

A rectangular room with size $6 \times 4 \times 3$ meters is modeled in the experiment. We employ a ULA which consists of eight omni-directional microphones, the spacing between adjacent microphones is 10 cm. The microphones at two ends of ULA are at (2.5,2.0,1.5), (3.2,2.0,1.5) respectively. Although we don't limit the number of interferences in the proposed algorithm, in order to facilitate the test, we assume that only one interference exists in the experiment. The speech source and interference are both located on a horizontal plane (x,y,1.5) with a distance of 2m to the center of the microphone array. We consider that the DOAs of speech source ranges from -90° to 90° with a step size of 20°. For the interference, we consider three possible DOAs, which are 20°, 40° and 60° respectively. As the ULA is symmetric to the normal line, there is no need to consider the case that the interference appears on the opposite side.

The speech source is with 16-bit resolution, sampling rate of 8KHz and durations of 10 seconds. We utilize three different types of noises (white Gaussian noise, car interior noise and F16 cockpit noise) taken from Noisex92 [11] as the interference signals. The room impulse responses from source to microphones are modeled by the image-source method [12], and a Matlab code implementation [13] is used to generate the signals received by different microphones. In the experiment, we set the reverberant time T60 of the room to be 250ms. The received speech signal and interference signal are separately generated, and mixed together after being scaled to control the signal-to-interference ratio (SIR).

Algorithms are evaluated under different SIR conditions. The SIR changes from -10dB to 20dB, with a step size of 5dB. We also add 15dB uncorrelated white Gaussian noise to each microphone, to simulate a more adverse environment. For all evaluated algorithms, the analysis frame size is set to be 256 samples with 50% overlap. 50 Monte Carlo simulations are conducted for each scenario (DOAs of the speech and interference, SIR).

4.2. Experimental results

We use two frame level metrics, denoted as Accuracy and Root Mean Square Error (RMSE), to evaluate the performance of different algorithms. The estimation is considered to be correct if $|\tilde{\theta}(t) - \hat{\theta}| < Th$, where Th is a threshold which is commonly set to be 5°. Then Accuracy and RMSE are defined as:

$$Accuracy = \frac{N_c}{N}, \tag{15}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tilde{\theta}(i) - \hat{\theta})^2}, \tag{16}$$

where N_c is the number of speech frames which have the correct estimation, N is the number of total speech frames. We only consider speech frames for evaluation, which are labeled manually in advance using clean speech. We should point out that these labels are never used by any of the three algorithms.

The results of all the evaluated algorithms under different interferences are compared in Fig.2. The performance under white Gaussian noise is illustrated in Fig.2(a) and Fig.2(b). It can be seen clearly that proposed algorithm yields the highest estimation accuracy in all SIRs considered, and gets the lowest RMSE in almost all scenarios. Similar results can be observed in the F16 cockpit noise and car interior noise cases, as shown in Fig.2(c) $\sim$ Fig.2(f). By comparing between the

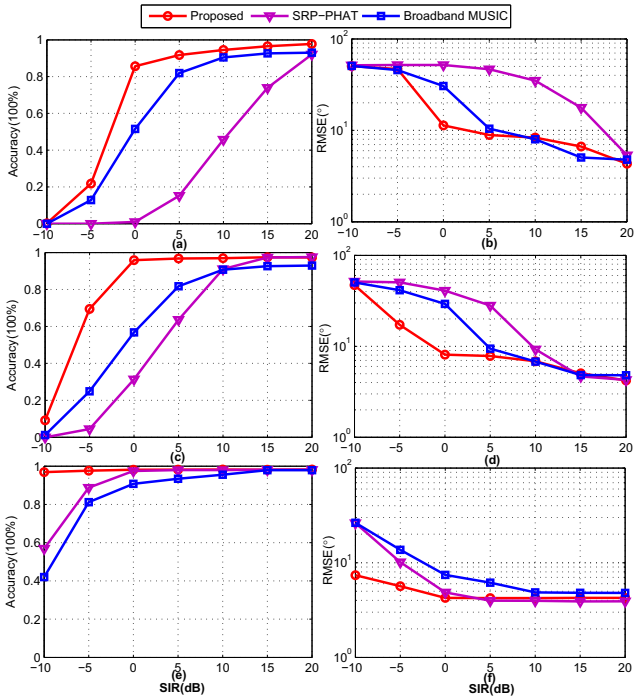

Figure 2: *Estimation performance of different algorithms under different SIRs and interference signals. Accuracy and RMSE with interference as: white Gaussian noise (a)(b), F16 cockpit noise (c)(d), car interior noise (e)(f). The error tolerance for Accuracy is* 5°

performances under different interferences, we can see that all algorithms achieve the performance improvement as the frequency distribution of the interference signal gets narrow from white Gaussian noise to car interior noise. It's probably because the strong interference signal has less probability to effect the speech frequency bins. Nevertheless, the proposed algorithm gets the most significant improvement in low SIR conditions. From Fig.2(e)(f), it can be seen even under the strong car interior noise interference, the proposed algorithm performs extremely robust.

5. Conclusion

In this paper, we present a new DOA estimation method for human speech in noisy and reverberant environments when the interferences exist. By exploiting the historical DOA estimates and the temporal correlation of speech frequency spectra, we perform post-beamforming to suppress the interference signals and calculate the frequency weight based on the estimated *a priori* SNR and the spectrum of the current frame. Then we introduce a new DOA cost function based on FWSCM which enables efficient frequency weighting. Experimental results on different kinds of interferences demonstrate that the proposed algorithm can achieve higher estimation accuracy and lower RMSE for human speech than conventional algorithms.

6. Acknowledgements

This research was supported in part by the China National Nature Science Foundation (No.91120303, No.61273267 and No.90820011).

7. References

[1] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas and Propagation AP-34, vol. 3, pp. 276-280, 1986.

[2] M. McCloud and L. Scharf, "A new subspace identification algorithm for high resolution DOA estimation," IEEE Trans. Antennas Propag., vol. 50, pp.1382-1390, 2002.

[3] J. P. Dmochowski , J. Benesty and S. Affes, "Broadband MUSIC: opportunities and challenges for multiple source localization," in Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 18-21, 2007.

[4] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2510-2526, 2007.

[5] M. Brandstein, D. Ward, "Microphone Arrays: Signal Processing Techniques and Applications," Springer, 2001.

[6] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," Signal Processing, vol. 85, no. 1, pp. 177-204, 2005

[7] M. Brandstein, J. E. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," Computer Speech and Language, vol. 9, no. 2, pp. 153-169, 1995

[8] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, "Springer Handbook of Speech Processing," Springer, 2008.

[9] Johnson D H, Dudgeon D E, "Array signal processing: concepts and techniques[M]," Simon & Schuster, 1992.

[10] Chao Li, Wenju Liu, "Improved a posteriori Speech Presence Probability Estimation Based on Cepstro-Temporal Smoothing and Time-Frequency Correlation," Twelfth Annual Conference of the International Speech Communication Association, 2011.

[11] A. Varga, H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to studythe ePect of additive noise on speech recognition systems," Speech Communication, 12(3): 247-251, 1993

[12] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," Journal of the Acoustical Society of Amesrica, vol. 124, no. 1, pp. 269-277, 2008.

[13] Online: "http://www.eric-lehmann.com/ism_code.html," accessed on 12 Mar 2013.