Cochlear Model based Enhancement of Noisy Speech Signals

Mladen Russo, Maja Stella, Maroje Kurajica

***Abstract*—**Many noise reduction techniques were developed over the past decades and noise reduction is still a major problem in many applications. Although reducing noise, these algorithms typically introduce some distortion to speech signal. Humans are quite capable of detecting speech in background noise, so we propose a method for noise reduction based on the model of cochlear processing of speech signals. Using our model of signal reconstruction from the cochlear output we have achieved improvement in speech quality. Our experiments show that the proposed approach performs better than several other noise reduction methods.

***Keywords*—**Cochlear model, noise reduction, speech signal, PESQ.

I. Introduction

SPEECH communication is an essential part of our lives. It takes place practically everywhere and is always affected by random noises. Whether the characteristics of these noises are known or not, they all corrupt the quality of speech signals. The problem of noise reduction has been widely studied over the past decades and in still an active research field.

Many techniques for noise reduction have been developed, including spectral magnitude estimation [1][2], signal subspace [3],[4], Wiener filtering [5],[6], Kalman filtering [7],[8] and hidden Markov models [9],[10].

Generally, their noise reduction performance was evaluated by assessing the improvement of signal-to-noise ratio (SNR), subjective speech quality or automatic speech recognition (ASR) performance. Noise reduction algorithms typically achieve noise reduction by introducing some distortion to speech signal, and some, like the subspace method, are even explicitly formulated based on the trade-off between noise reduction and speech distortion [5]. Many noise reduction techniques have also been used in image and video signals [11,12].

We propose a method for noise reduction based on the model of cochlear processing of speech signals. Humans are quite capable of detecting speech in background noise, so the idea is to mimic the behavior of human cochlea in order to achieve better noise reduction. However, the output of a cochlear model is not a speech signal so we proposed a method for speech signal reconstruction from the cochlear output. This method of signal reconstruction has been presented in [13] and this paper is an extension of our previous works [13],[14]. We observed that, when applied to noisy speech signal, this method results with improved signal quality. In order to evaluate our approach, we compare it to other state-of-the-art noise reduction techniques. Speech quality is measured with Perceptual Evaluation of Speech Quality (PESQ) score [15]. It is standardized as ITU-T recommendation P.862 and is used for objective assessment of speech quality.

This paper is organized as follows. In Section II. we present our cochlear model based method for noise reduction in speech signals. In Section III. several other methods for noise reduction are briefly described and compared to our approach in Section IV. Section V. concludes the paper.

II. Cochlear Model based Noise Reduction

A. *Biophysical cochlear model*

The human cochlea transforms incoming sound pressure into vibrations of the basilar membrane which result in generation of neural impulses. Besides amplifying incoming sound waves and converting them into neural signals, it also acts as a mechanical frequency analyzer and can be seen as a system designed to analyze frequency components present in complex sounds. Each position along the basilar membrane (BM) corresponds to a particular frequency.

The cochlea is a small tube-like coiled structure about 1 cm long and 3.5 cm wide. From its base to apex it is internally bisected with a flexible basilar membrane containing the organ organ of Corti - a very sophisticated structure which responds to basilar membrane vibrations and generates neural impulses, Fig. 1. It consists of one row of inner hair cells (IHC) and mostly three rows of outer hair cells (OHC). The IHCs are the actual sensory receptors; based on their hairs' movements they generate neural impulses. The hair cell is an evolutionary triumph that solves the problem of transforming vibrational energy into an electrical signal. At the limits of human hearing, hair cells can faithfully detect movements of atomic dimensions and respond in the tens of microseconds.

This work has been fully supported by Croatian Science Foundation under the project UIP-2014-09-3875.

Authors are with the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Croatia (e-mails: {mrusso, mstella, mkurajic}@fesb.hr).

Furthermore, hair cells can adapt rapidly to constant stimuli, thus allowing the listener to extract signals from a noisy background [16].

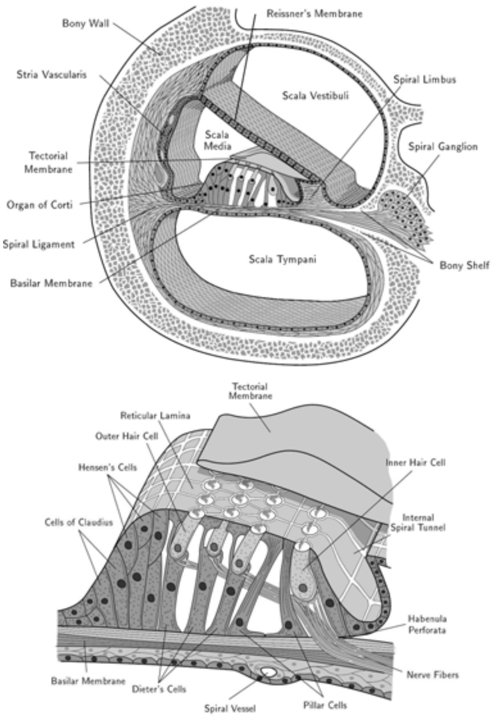


Fig. 1 Cross-section of the cochlea with enlarged organ of Corti [17]

In our work, we use a biophysical cochlear model developed by Mammano and Nobili [18],[19]. It is a micromechanical cochlear model and models the cochlea at a level adequate to the complexity of realistic cochlear structures (varying cross-sectional area of cochlear channels, width of the partition, basilar membrane mass, stiffness, absolute and shear viscosity, OHC action). It fits very nicely with experimental data and can explain some auditory system phenomena like two-tone suppression, two-tone distortion, otoacoustic emissions including spontaneous (SOAE), transient-evoked (TEOAE) and stimulus frequency otoacoustic emissions (SFOAE) [20].

B. Signal reconstruction and noise reduction

Based on a time-frequency response of the biophysical cochlear model, we propose a method for reconstructing the original signal.

For the passive cochlear model and pure tone input, in order to reconstruct the original sinusoid from the model output (Fig. 2), we perform several steps:

1) correct the model output by the delay in phase towards the apex;

2) integrate over BM space in order to obtain a single sinusoid;

3) divide by an area of the travelling wave profile in order to obtain the correct amplitude.

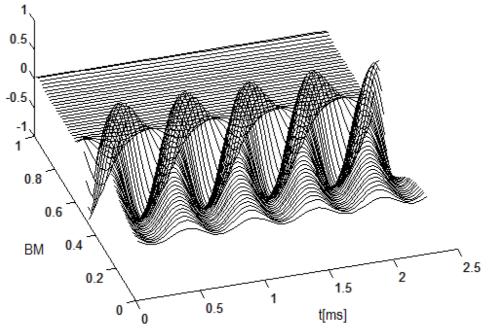


Fig. 2 Basilar membrane response for a 2 kHz tone (passive model)

When phase delay and the areas of travelling wave profiles are known for all frequencies/sinusoids, in a linear cochlear system (where superposition is valid) any audio signal can be easily reconstructed as it can be considered as a sum of its sinusoid components.

On the other hand, the active cochlear model is nonlinear and superposition is not valid, but nevertheless, when we applied the same method on active model output of a noisy speech signal, we observed improved signal quality in the reconstructed speech signal.

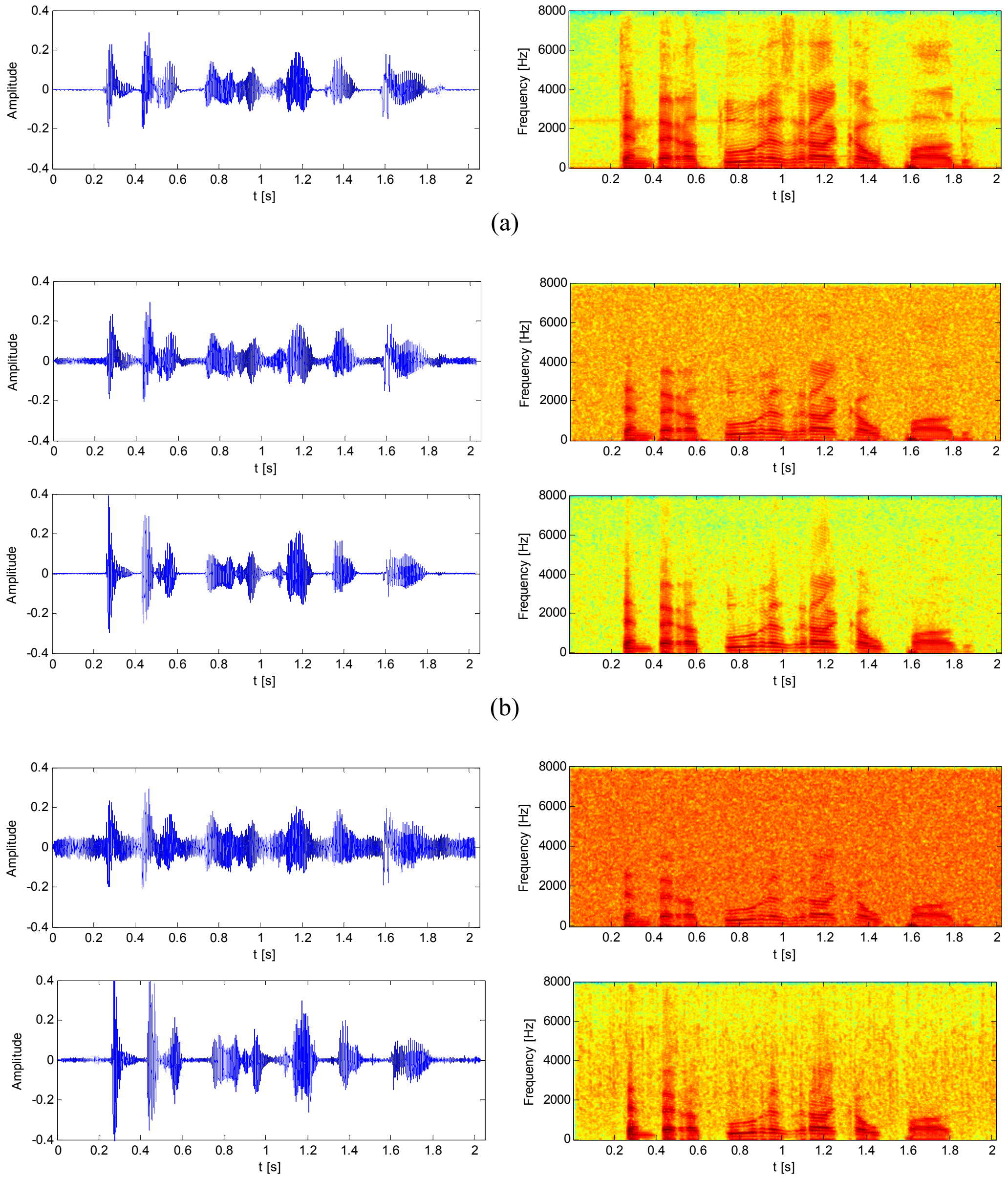
Fig. 3 shows an example of white noise reduction for a test sentence from male speaker. Speech signal is first applied to the active cochlear model input and then reconstructed from the basilar membrane response. Two examples of noise reduction for SNR=10dB and SNR=0dB are given. For each example, both signal waveform and its power spectrogram are shown. It is clearly visible that the proposed technique results with reduction in noise levels.

In order to evaluate the performance of the proposed method, quality of the reconstructed speech has to be quantitatively measured and compared to other noise reduction techniques.

III. Other Noise Reduction Techniques used in Comparison

A. Spectral subtraction with noise spectrum measured during non-speech activity

This method was proposed in a well-known and extensively cited paper by S. Boll published in IEEE Trans. on Acoust., Speech, Signal Process. in 1979 [1]. The objectives of the paper was to develop a noise suppression technique, implement a computationally efficient algorithm, and test its performance in actual noise environments. The approach used was to estimate the magnitude frequency spectrum of the underlying clean speech by subtracting the noise magnitude spectrum from the noisy speech spectrum. This estimator required an estimate of the current noise spectrum.



(a)

(b)

(c)

Fig. 3 Noise reduction example: (a) clean speech and its spectrogram; (b) noise reduction for SNR=10dB – top panels noisy speech and its spectrogram, bottom panels noise reduced speech and its spectrogram; (c) noise reduction for SNR=0dB – top panels noisy speech and its spectrogram, bottom panels noise reduced speech and its spectrogram

Rather than obtain this noise estimate from a second microphone source, it was approximated using the average noise magnitude measured during nonspeech activity. Using this approach, the spectral approximation error was then defined, and secondary methods for reducing it were also described.

This approach is implemented in a Matlab function which uses amplitude spectral subtraction and includes magnitude averaging and residual noise reduction. Noise spectrum is estimated during the initial silence period.

B. Spectral subtraction and Wiener filtering with noise power power spectral density estimation based on optimal smoothing and minimum statistics

This highly cited method of estimating the power spectral density of nonstationary noise when a noisy speech signal is given was presented by R. Martin in IEEE Trans. on Speech and Audio Proc. in 2001 [21]. In contrast to other methods, this approach does not use a voice activity detector. Instead it tracks spectral minima in each frequency band without any distinction between speech activity and speech pause. By minimizing a conditional mean square estimation error criterion in each time step, authors derived the optimal smoothing parameter for recursive smoothing of the power spectral density of the noisy speech signal. Based on the optimally smoothed power spectral density estimate and the analysis of the statistics of spectral minima an unbiased noise estimator was developed.

This method can be combined with any speech enhancement algorithm which requires a noise power spectral density estimate and we are using it in a Matlab function which implements both spectral subtraction and Wiener filtering.

C. Wiener filter with Two-Step Noise Reduction (TSNR) and Harmonic Regeneration Noise Reduction (HRNR)

This is the state-of-the-art approach based on the work of C. Plapous, C. Marro and P. Scalart published in IEEE Trans. on Audio, Speech, and Lang. Proc. in 2006 [22]. They proposed a method, called two-step noise reduction (TSNR), to refine the estimation of the a priori SNR which removes the drawbacks of the decision-directed (DD) approach (proposed by Ephraïm and Malah in [23]) while maintaining its advantage, i.e., highly reduced musical noise level. The major advantage of this approach is the suppression of the frame delay bias leading to the cancellation of the annoying reverberation effect characteristic of the DD approach. However, classic short-time noise reduction techniques, including TSNR, introduce harmonic distortion in enhanced speech because of the unreliability of estimators for small signal-to-noise ratios. This is mainly due to the difficult task of noise power spectrum density (PSD) estimation in single-microphone schemes. To overcome this problem, the authors proposed a method called harmonic regeneration noise reduction (HRNR) that takes into account the harmonic characteristic of speech. In this approach, the output signal of any classic noise reduction technique (with missing or degraded harmonics) is further processed to create an artificial signal where the missing harmonics have been automatically regenerated. This artificial signal helps to refine the a priori SNR used to compute a spectral gain able to preserve the harmonics of the speech signal.

This approach is implemented in a Matlab function which uses Wiener filter based on tracking a priori SNR using decision-directed method with TSNR and HRNR algorithms.

D. The wavelet approach

Another popular approach in noise reduction are wavelet-based techniques. Wavelet thresholding methods for signal denoising were firrst introduced by D. L. Donoho in IEEE Trans. on Inf. Theory in 1995 [24].

A wavelet is a mathematical function used to divide a given function or continuous-time signal into different scale components. Usually one can assign a frequency range to each scale component. Each scale component can then be studied with a resolution that matches its scale. A wavelet transform is the representation of a function by wavelets. The wavelets are scaled and translated copies (known as "daughter wavelets") of a finite-length or fast-decaying oscillating waveform (known as the "mother wavelet"). Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks, and for accurately deconstructing and reconstructing finite, non-periodic and/or non-stationary signals.

Wavelets are natively supported in Matlab using Wavelets Toolbox and in our experiments we used these commands with the same parameters as in audio signal denoising example from [25].

IV. Comparison Results

The proposed approach is compared against several popular noise reduction techniques briefly described in previous section. In order to quantize the reconstructed speech quality, we used an objective voice quality measurement based on ITU-T recommendation P.862 – PESQ [15]. It analyzes the speech signal sample-by-sample after a temporal alignment of corresponding excerpts of reference and test signal. PESQ results principally model mean opinion scores (MOS). We find the PESQ score well suited for this measurement since all that finally matters in speech enhancement is human perception of denoised signal.

An example of white noise reduction based on our cochlear model based speech enhancement is given in Fig. 3 for two levels of additive white noise (SNR=10dB and SNR=0dB). It is clearly visible that the proposed technique results with reduction in noise levels. Noisy speech at SNR=10dB has PESQ score 2.78, and noisy speech at SNR=0dB has PESQ score 2.02. We also calculated the PESQ score after signal reconstruction: for SNR=10dB, PESQ score was improved from 2.78 to 3.31, and for SNR=0dB, PESQ score was improved from 2.02 to 2.31, or expressed in percentages 19% and 14%, respectively.

We conducted the same experiment with other noise reduction techniques and results are given in tables and figures below. They are denoted as methods A, B, C, D as described in previous section. Method B1 represents spectral subtraction with noise power spectrum estimation, method B2 represents Wiener filter with noise power spectrum estimation, and methods C1 and C2 represent Wiener filter with TSNR and HRNR, respectively. The proposed method is denoted as "our".

TABLE I. WHITE NOISE PESQ RESULTS

White noise	Noisy PESQ	Denoised speech PESQ						
		A	B1	B2	C1	C2	D	our
SNR=10dB	2,78	3,18	3,45	3,10	2,81	2,84	2,95	3,31
SNR=0dB	2,02	1,48	2,46	2,26	2,37	2,43	2,31	2,31

Table I and Fig. 4 show the performance of the proposed method compared to other noise reduction techniques in terms of PESQ score for white noise. Lines in Fig. 4 denote noisy speech PESQ level. Results clearly indicate that our method performs better or comparably to other state-of-the-art methods.

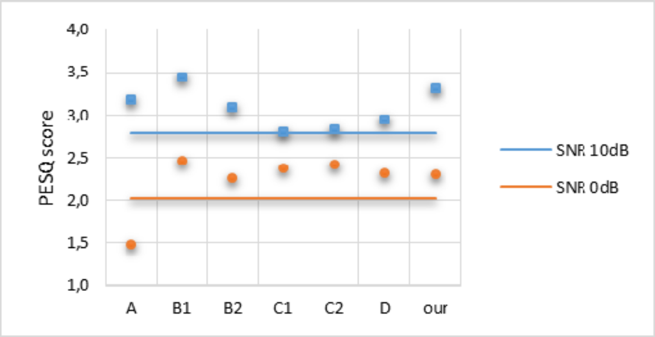


Fig. 4 Performance comparison in white noise

Practically all the methods result with improved speech quality (except for method A in low SNR scenario). It should also be noted here that white noise is stationary and is easily estimated from non-speech intervals. Non-stationary noise is quite another problem and much harder to estimate. In order to evaluate how our method performs in non-stationary noise conditions, we have corrupted the original speech with babble noise (speech-like noise), Table II and Fig. 5.

TABLE II. BABBLE NOISE PESQ RESULTS

Babble noise	Noisy PESQ	Denoised speech PESQ						
		A	B1	B2	C1	C2	D	our
SNR=10dB	2,80	2,88	2,95	2,86	2,53	2,54	2,81	3,06
SNR=0dB	1,87	1,78	2,11	1,99	1,51	1,54	1,90	2,16

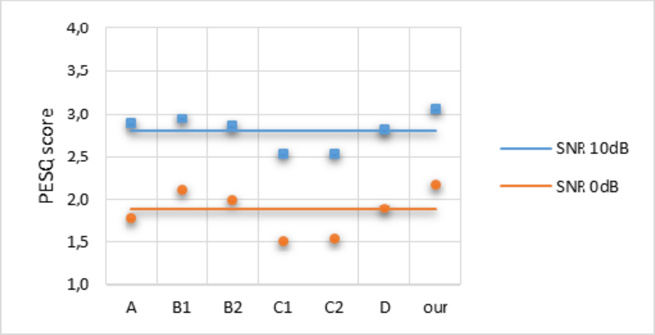


Fig. 5 Performance comparison in babble noise

Results clearly indicate that our method outperforms all other state-of-the-art methods. Some methods even degrade the signal, and methods B1 and B2 are better than the others since they do not estimate noise from non-speech intervals, but continuously minimize a conditional mean square estimation error criterion in each time step.

V. Conclusion

In this paper we have presented our approach for noise reduction based on the model of cochlear processing of speech signals. Humans are quite capable of detecting speech in background noise, so the idea is to mimic the behavior of human cochlea in order to achieve better noise reduction.

In order to evaluate the proposed method, we have conducted experiments in stationary white noise conditions and non-stationary babble noise conditions. Speech quality was measured with PESQ score. Our method was compared with several state-of-the-art noise reduction techniques and results showed that the proposed approach performs similarly or outperforms the other methods.

References

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Trans. Acoust., Speech, Signal Process*., vol. 27, no. 2, pp. 113–120, Apr. 1979.
[2] P.Vary, “Noise suppression by spectral magnitude estimation-mechanism and theoretical limits”, Signal Processing, vol. 8, pp. 387–400, Jul. 1985.
[3] Y.Ephraim, and H. L. Van Trees, “A signal subspace approach for speech enhancement”, *IEEE Trans. Speech Audio Process*., vol. 3, no. 4, pp. 251–266, Jul. 1995.
[4] H. Lev-Ari and Y. Ephraim, “Extension of the signal subspace speech enhancement approach to colored noise”, *IEEE Signal Process. Lett*., vol. 10, no. 4, pp. 104–106, Apr. 2003.
[5] J. Chen, et al., “New insights into the noise reduction Wiener filter”, *Trans. Acoust., Speech, Signal Process., IEEE*, vol. 14, no. 4, pp. 1218-1233. 2006.
[6] B. Widrow and S. D. Stearns, “Adaptive Signal Processing”, Englewood Cliffs, NJ: Prentice-Hall, 1985.
[7] K. K. Paliwal and A. A Basu, “Speech enhancement method based on Kalman filtering”, in Proc. IEEE ICASSP, 1987, pp. 177–180.
[8] S. Gannot et al., “Iterative and sequential Kalman filter-based speech enhancement algorithms”, *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, Jul. 1998.
[9] Y.Ephraim, et al., “On the application of hidden Markov models for enhancing noisy speech”, *IEEE Trans. Acoust., Speech, Signal Process*., vol. 37, no. 12, pp. 1846–1856, Dec. 1989.
[10] H. Sameti, et al., “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise”, *IEEE Trans. Speech Audio Process*., vol. 6, no. 5, pp. 445–455, Sep. 1998.
[11] S. N. Sulaiman, et al., "Denoising of noisy MRI brain image by using switching-based clustering algorithm", in Proc. of INASE conference on Energy, Environment, Biology and Biomedicine, Prague, pp. 33-39, 2014.
[12] V. Ponomaryov, “Fuzzy method for suppressing of different noises in color videos”, in Proc. of INASE Conference on Circuits, Systems, Signal Processing, Communications and Computers (CSSCC 2015), Vienna, pp.65-74., 2015.
[13] M. Russo, N. Rožić, M. Stella, “Biophysical cochlear model: Time-frequency analysis and signal reconstruction”, *Acta Acustica united with Acustica*, vol. 97, no. 4, pp. 632–640, 2011.
[14] M. Russo, M. Stella, N. Rožić, “Noise reduction in speech signals using a cochlear model”, *Advances in Smart Systems Research*. vol. 2, 1; pp.7-12, 2012.
[15] ITU-T, Rec. P.862. “Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs”, 2001.
[16] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, J. O. McNamara, S. M. Williams, Neuroscience, 2004.
[17] R. G. Kessel, R. H. Kardon, Tissues and organs: a text-atlas of scanning electron microscopy, WH Freeman San Francisco, 1979.

[18] F. Mammano and R. Nobili, “Biophysics of the cochlea: linear approximation”, *J. Acoust. Soc. Amer*., vol. 93, no. 6, pp. 3320–3332, 1993.

[19] F. Mammano, and R. Nobili, “Biophysics of the cochlea ii: Stationary nonlinear phenomenology”, *J. Acoust. Soc. Amer*., vol. 99, no. 4, pp. 2244–2255, 1996.

[20] R. Nobili et al., “Otoacoustic emissions from residual oscillations of the cochlear basilar membrane in a human ear model”, *Journal of the Association for Research in Otolaryngology*, vol. 4, no. 4, pp. 478–494, 2003.

[21] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol.9, no.5, pp.504-512, Jul 2001.

[22] C. Plapous, C. Marro, P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, issue 6, pp. 2098 - 2108, Nov. 2006.

[23] Y. Ephraïm and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[24] D. L. Donoho, “Denoising by soft-thresholding,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, Mar. 1995.

[25] A. E. Villanueva-Luna et al., “De-Noising Audio Signals Using MATLAB Wavelets Toolbox”, in *Engineering Education and Research Using MATLAB*, InTech, 2011.