Noise reduction in speech signals using a cochlear model

Mladen Russo, Maja Stella and Nikola Rožić

Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture,
University of Split, Split, HR-21000, Croatia
mrusso@fesb.hr

Abstract

Smart systems and artificial intelligence technology are becoming increasingly popular and are continuously finding more applications in real-life situations. Many systems require human-computer interaction and the natural language interface. One of the major issues in speech recognition systems is their performance in real world (noisy) environment. Over the past decades many techniques for noise reduction were developed. Motivated by human auditory processing, and it is well known that humans are remarkably good at detecting speech in the background noise, we propose a noise reduction technique based on a biophysical cochlear model. Using a model of signal reconstruction from the cochlear output, we observed an improvement in the quality of noisy speech and a significant increase in speech recognition performance.

1. Introduction

In a natural environment speech signals are almost always immersed in ambient noise and it is essential for speech processing systems to apply some noise reduction techniques to extract the desired speech signal. Noise reduction is a very challenging problem since the characteristics of a noise signal vary significantly in different environments and moreover in time. Over the past decades many approaches 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].

Smart systems and artificial intelligence technology are becoming increasingly popular and are continuously finding more applications in real-life situations. Many systems require human-computer interaction and the natural language interface. One of the major issues in speech recognition systems is their performance in real world (noisy) environment. Comparisons using many speech corpora demonstrate that word error rates of machines are often more than an order of magnitude greater than those of humans for quiet, wideband, read speech. Machine performance degrades further below that of humans in noise, with channel variability, and for spontaneous speech [11]. Motivated by human auditory processing, we propose a noise reduction technique based on a biophysical cochlear model. Biophysical cochlear models are generally developed using many simplified assumptions about the cochlear fluid dynamics and mechanics of cochlear microstructures. We use the model of Mammano and Nobili [12],[13], since it models the cochlea at a level adequate to the complexity of realistic cochlear structures. Thus it is logical to assume that, because of the model realism, it could closely resemble a response of the real human cochlea.

Since this work is still in progress, we are here presenting only preliminary results and findings, but from what can be observed so far, applying this biophysical cochlear model to noisy speech and reconstructing the speech signal results with improvement in the quality of noisy speech and with significant increase in speech recognition performance.

2. Biophysical Cochlear Model

Many cochlear models have been developed in the past decades. These models can be roughly divided into biophysical models and signal-processing models [14]. Signal-processing models typically use filtering operations to produce cochlea-like output (e.g. gammatone filterbank by Patterson et al. [15]). The filters are tuned to different frequencies with response shape and spacing inspired by psychophysical and psychological data. Biophysical cochlear models, such as the macromechanical and micromechanical cochlear models, seek to explain how the cochlea works through simplified assumptions about the cochlear fluid dynamics and mechanics of cochlear microstructures. For example, in the model of Netten and Duifhuis [16], mass of the cochlea, cross-sectional area of cochlear channels and the width of the partition are all assumed to be constant, and the stiffness exponentially decreases from the base to the apex of the cochlea. In micromechanical models, the cochlear partition is modelled more realistically.

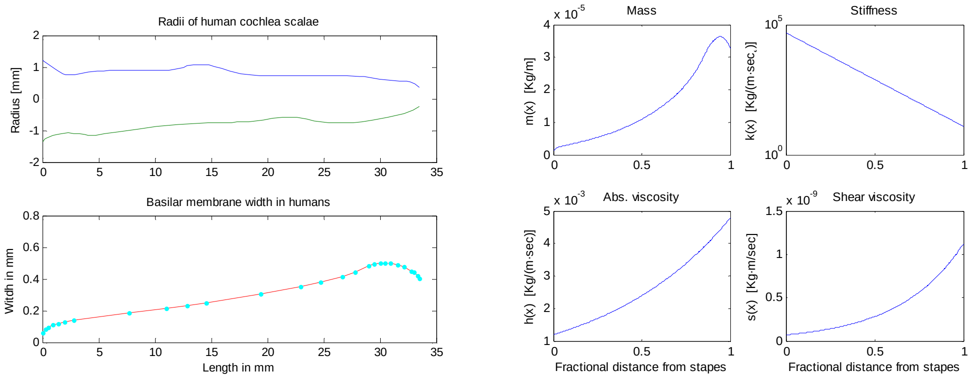


Figure 1. Geometrical and physical parameters of the biophysical cochlear model

In the model of Mammano and Nobili [13],[14], both mechanical and hydrodynamical aspects are treated at a level adequate to the complexity of realistic cochlear structures. 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) [17]. Geometrical and physical properties of this model are shown in Figure 1.

3. Noise reduction experiments and results

For the passive cochlear model and pure tone input, discretized output of the model (shown in Figure 2) will consist of a series of sinusoids distributed along the basilar membrane (with peek amplitude corresponding to characteristic frequency location) and delayed in phase towards the apex of the cochlea. *X*-axis represents time, *y*-axis represents space/frequencies along the basilar membrane and *z*-axis represents normalized intensity.

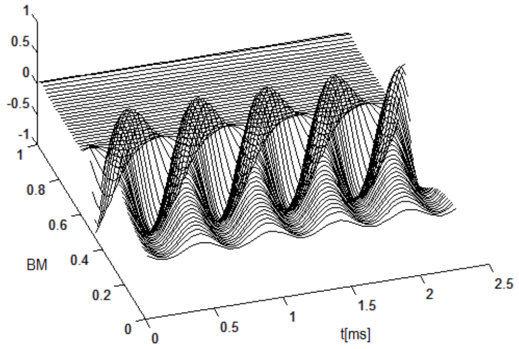


Figure 2. Basilar membrane response for a 2 kHz tone (passive model).

In order to reconstruct the original sinusoid from the model output, a few simple steps can be applied: 1) correcting the model output by the delay in phase towards the apex; 2) integrating over BM space in order to obtain a single sinusoid; 3) dividing by an area of the travelling wave profile in order to obtain the correct amplitude. The method is described in more detail in [18]. Since any audio signal can be considered as a sum of sinusoid components, and we are here talking about linear cochlear system, superposition is valid and the signal can be reconstructed if the phase delay and the areas of travelling wave profiles are known for all frequencies/sinusoids.

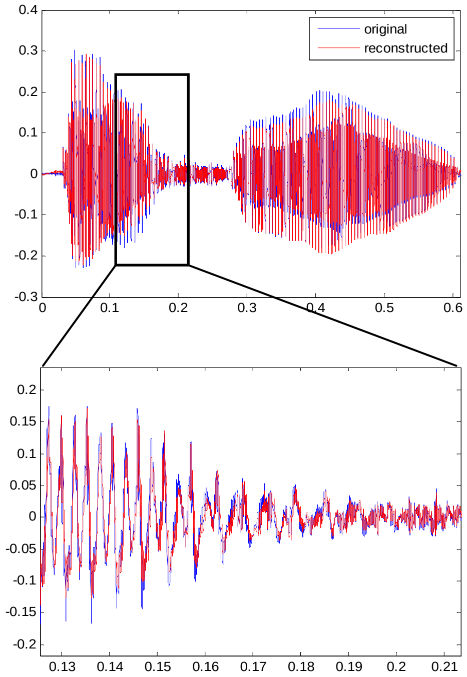
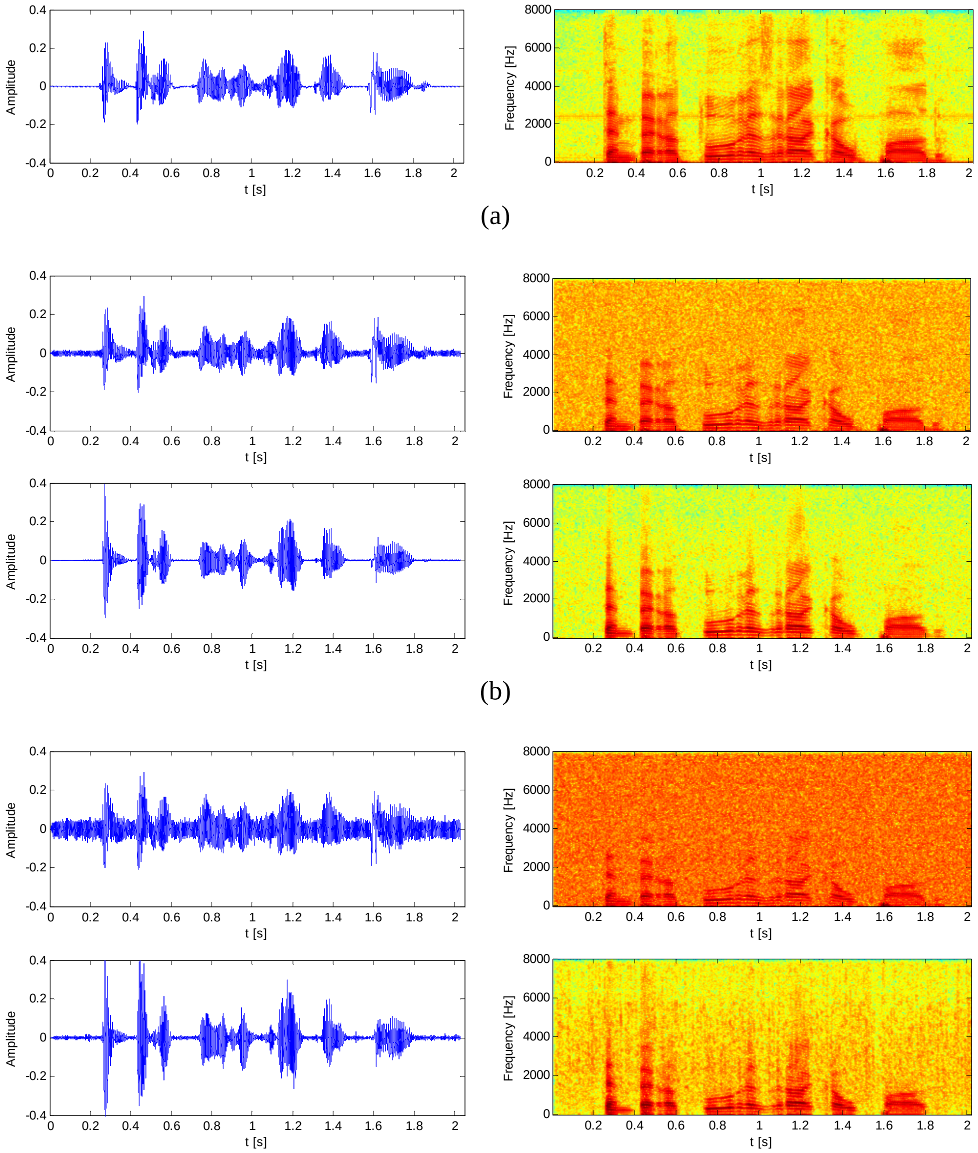


Figure 3. Example of speech signal reconstruction

Figure 3 shows an example of speech signal reconstruction. It is clearly visible that very faithful reconstruction is possible - if the PESQ MOS (Perceptual Evaluation of Speech Quality Mean Opinion Score) [19] is used as a measure of quality, reconstructed signal will have a 4.49 PESQ score (maximum value for PESQ score is 4.5).

Applying this method on time-frequency response of the active model would not be mathematically correct, since the model is nonlinear and superposition is not valid and phenomena like frequency masking are present. But it was observed that, when applied to active cochlear model response of a noisy speech signal, this reconstruction method can result with improved signal quality (with some spectral enhancements and noise suppression). Reconstructed signal is obviously more similar to the one which human ear actually hears, and it is well known that humans are remarkably good at detecting speech in the background noise.



(a)

(b)

(c)

Figure 4. 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

Figure 4 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. Level of improvement can be expressed with PESQ score – for the case when SNR=10dB, PESQ score was improved from 2.78 to 3.31, and for the case when SNR=0dB, PESQ score was improved from 2.02 to 2.31, or expressed in percentages 19% and 14% respectively.

In order to evaluate the performance of the proposed technique more thoroughly, we have developed a speech recognition system. Our speech recognition system is based on continuous density HMM models, and is developed using the HTK toolkit. The speech material consists of 673 sentences in Croatian (5731 words) recorded by 12 male speakers from the texts of short weather forecasts for the Adriatic coast. It was recorded in quiet office and sampled at 16 kHz with 16 bits. Vocabulary size is 362 words. The data and the speakers were divided in two sets: one for training and one for testing. Acoustical modelling was made at phone level by continuous Gaussian density HMMs with 3 states (left-right topology) per phone and 6 mixture components per state with diagonal covariance matrices. Standard configuration of the HTK setup was used. Models were trained with MFCC (Mel-frequency cepstral coefficients) feature vectors of 39 elements (13 static + 13 velocity + 13 acceleration coefficients) representing 25 ms segments of speech, every 10 ms. Bigram language model was used.

SNR (dB)	ASR	ASR_NR
clean	93,92	93,71
25	92,87	93,08
20	91,4	92,45
15	81,13	90,57
10	42,77	84,28
5	4,82	65,41
0	1,26	21,59

Table 1. Recognition rates (%)

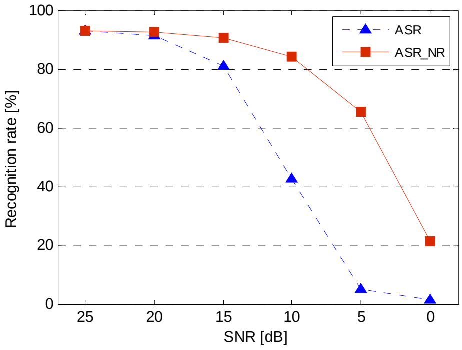


Figure 5. Recognition rates (%)

Table 1 and Figure 5 show the recognition results of the speech recognition systems with and without noise reduction, denoted as ASR_NR and ASR respectively.

It can be observed that as the noise level increases, the recognition rates of the NR based system become significantly higher than the standard ASR system. The largest improvement in performance occurs for SNR=5dB, where recognition rate goes from 4,82% to 65,41% when noise reduction technique was used.

4. Conclusions

As the technology advances, many modern intelligent systems will require natural language interface and thus speech recognition capabilities. One of the major issues in speech recognition systems is their performance in real world (noisy) environment. Motivated by human auditory processing, and it is well known that humans are remarkably good at detecting speech in the background noise, in this work we have proposed a noise reduction technique based on cochlear processing of speech signals. Using a model of signal reconstruction from the cochlear output, we observed an improvement in the quality of noisy speech and a significant increase in speech recognition performance. This work is still in progress and it yet remains to be tested how the method performs in other types of noise, but results obtained so far are quite promising.

5. References

1. Boll, S. F. 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. Vary, P. Noise suppression by spectral magnitude estimation-mechanism and theoretical limits, Signal Processing, vol. 8, pp. 387–400, Jul. (1985)
3. Ephraim, Y. and Van Trees, H. L. A signal subspace approach for speech enhancement, IEEE Trans. Speech Audio Process., vol. 3, no. 4, pp. 251–266, Jul. (1995)
4. Lev-Ari, H. and Ephraim, Y. 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. Chen, J 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. Widrow, B. and Stearns, S. D. Adaptive Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, (1985)
7. Paliwal, K. K. and Basu, A. A speech enhancement method based on Kalman filtering, in Proc. IEEE ICASSP, 1987, pp. 177–180.
8. Gannot, S. 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. Ephraim, Y. 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. Sameti, H. 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. Lippmann, R. P. Speech recognition by machines and humans, Speech Commun., vol. 22, no. 1, pp. 1–15, July (1997)
12. Mammano, F. and Nobili, R. Biophysics of the cochlea: linear approximation, J. Acoust. Soc. Amer., vol. 93, no. 6, pp. 3320–3332 (1993)
13. Mammano, F. and Nobili, R. Biophysics of the cochlea. ii: Stationary nonlinear phenomenology, J. Acoust. Soc. Amer., vol. 99, no. 4, pp. 2244–2255 (1996)
14. Munkong, R. and Juang, B.-H. Auditory perception and cognition, IEEE Signal Proc. Mag., vol. 25, no. 3, pp. 98–117, May (2008)
15. Patterson, R. et al. An efficient auditory filterbank based on the gammatone function, APU report, vol. 2341, (1988)
16. van Netten, S. M. and Duifhuis, Modelling an active, nonlinear cochlea, in Mechanics of Hearing, Nijhoff/Delft Univ. Press, pp. 143–151 (1983)
17. Nobili, R. 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)
18. Russo, M. et al. Biophysical cochlear model: Time-frequency analysis and signal reconstruction, Acta Acustica united with Acustica, vol. 97, no. 4, pp. 632–640 (2011)
19. 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)