SPARSITY LEVEL IN A NON-NEGATIVE MATRIX FACTORIZATION BASED SPEECH STRATEGY IN COCHLEAR IMPLANTS

Hongmei Hu[1,2] , Nasser Mohammadih[3], Jalil Tahgia[3], Arne Leijon[3], Mark E Lutman[1], Shouyan Wang[1]

1. Institute of Sound and Vibration Research, University of Southampton, SO17 1BJ, Southampton, UK
2. Department of Testing and Control, Jiangsu University, 212013, Zhenjiang, China
3. School of Electrical Engineering, Royal Institute of Technology, Stockholm, Sweden

ABSTRACT

Non-negative matrix factorization (NMF) has increasingly been used as a tool in signal processing in the last years, but it has not been used in the cochlear implants (CIs). To improve the performance of CIs in noisy environments, a novel sparse strategy is proposed by applying NMF on envelops of 22 channels. In the new algorithm, the noisy speech is first transferred to the time-frequency domain via a 22- channel filter bank and the envelope in each frequency channel is extracted; secondly, NMF is applied to the envelope matrix (envelopegram); finally, the sparsity condition is applied to the coefficient matrix to get more sparse representation. Speech reception threshold (SRT) subjective experiment was performed in combination with five objective measurements in order to choose the proper parameters for the sparse NMF model.

Index Terms— Non-negative matrix factorization, cochlear implants, sparse coding, objective measurements, speech perception threshold

1. INTRODUCTION

Cochlear implants (CIs) are electrical devices that help to restore hearing to the profoundly deaf. The main principle of CIs is to stimulate auditory nerves via electrodes surgically inserted in the inner ear. With the development of new speech processors and algorithms, the majority of implanted users benefit from this device, some of them to some degree allow users to communicate via telephone without much difficulty. However, average performance of most CIs users still falls below normal hearing (NH) listeners, and speech quality and intelligibility generally deteriorate in the presence of background noise. Specifically, users often complain that their CIs do not work well in background noise. It is well known that one of the most relevant differences between NH and CIs users in terms of speech perception is the dynamic range: the dynamic range of the impaired ear is much smaller than that of the normal ear. Thus the electrical stimulation provides a severe bottleneck of the information transfer, which only allows limited acoustic information to be transmitted to the auditory neurons [1]. Our recently developed sparse speech processing strategies [2] [3]significantly improve the speech intelligibility in patients with cochlear implants by reducing the level of noise and increasing dynamic range simultaneously to overcome the bottleneck of the information transmission.

Non-negative matrix factorization (NMF) is a method to factorize a non-negative matrix into two non-negative matrices. After being improved by Lee [4], NMF has increasingly been used as a tool in signal processing in the last years, such as image processing, speech processing, and pattern classification[5],[6],[7],[8],[9],[10]. Instead of learning holistic presentations, NMF usually results to parts-based decomposition[4] and reconstruction of the signal by using non-negativity constraints.

In this paper, a NMF based sparse coding strategy is proposed to improve the performance for CIs users in noisy environments. The basic motivation to use NMF is that the envelope in each channel is non-negative and the firing rates of neurons are never negative. Assuming that speech and noise signals are independent and that the observed noisy signal is obtained by adding the speech and noise signals, NMF is used to factorize the envelopegram, the matrix of 22 channels envelopes, into NMF basis and coefficient matrices. The application of sparse NMF can now be interpreted as a noise reduction by assuming that the smaller NMF coefficients correspond either to the noise basis vectors, or they do not contribute significantly in explaining the speech signal. Hence, by applying sparseness constraint to the factorization, the NMF coefficients which are small will be removed (set to zero) and a more sparse signal will be obtained by performing noise reduction. That is to say, the proposed algorithm can enhance the noisy speech by increasing the sparsity level of the reconstructed signal.

Here, considering computation complexity and the real-time implementation in the future, a basic NMF with sparsity constrain is used aiming to improve the performance of CIs users in noisy environment. In order to select a proper sparsity constraint parameter, five objective evaluation algorithms combined with speech perception threshold (SRT) subjective experiments were carried out for choosing the proper sparse parameter to obtain proper

tradeoff between the sparsity and the approximation of the signal.

2. NON-NEGATIVE MATRIX FACTORIZATION

Given a non-negative matrix $\mathbf{Z}$, NMF is a method to factorize $\mathbf{Z}$ into two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ so that $\mathbf{Z} \approx \mathbf{WH}$. To do the factorization, a cost function $D(\mathbf{Z} \| \mathbf{WH})$ is usually defined and minimized. Since the basic NMF allows a large degree of freedom, different types of cost functions and regularities have been used in the literature to derive meaningful factorizations for a specific application [7],[8], [9].

In this paper the square Euclidean distance $D(\mathbf{Z} \| \mathbf{WH}) = \frac{1}{2}\|\mathbf{Z} - \mathbf{WH}\|_2^2$ is used as the cost function, which is equivalent to Maximum Likelihood (ML) estimation of $\mathbf{W}$ and $\mathbf{H}$ in additive independent and identically distributed (i.i.d.) Gaussian noise. In order to impose additional sparseness, the standard NMF is combined with a sparseness penalty function based on L_1-norm through a least absolute shrinkage and selection operator (LASSO) framework, i.e., the sparsity is measured by L_1 norm. The sparseness weight (λ in the following sections) can be optimized to get a good trade-off between the sparseness and approximation of the signal which is convenient to tune according to individual preference for CIs users in the future.

In our application, $\mathbf{Z}$ denotes an $N \times M$ envelope matrix of one analysis block where N and M indicate the number of channels and the number of frames, respectively. NMF is applied to factorize the non-negative envelope matrix into basis matrix $\mathbf{W}$ and coefficient matrix $\mathbf{H}$ respectively, the additional sparseness constraint is to explicitly control the sparsity of the NMF coefficients matrix $\mathbf{H}$ that represents the activity of each basis vector over time such that

$$D(\mathbf{Z} \| \mathbf{WH}) = \frac{1}{2}\|\mathbf{Z} - \mathbf{WH}\|_2^2 + \lambda g(\mathbf{H}) \quad (1)$$

is minimized, under the constraints $\forall_{ij} : W_{ij} \geq 0, H_{ij} \geq 0$, $\lambda \geq 0$, where

$\mathbf{W} = \begin{bmatrix} w_{11} & \cdots & w_{1K} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NK} \end{bmatrix}_{N \times K}$, $\mathbf{H} = \begin{bmatrix} h_{11} & \cdots & h_{1M} \\ \vdots & \ddots & \vdots \\ h_{K1} & \cdots & h_{KM} \end{bmatrix}_{K \times M}$, $\mathbf{w}_i$ denotes the i^{th} column of $\mathbf{W}$, $g(\mathbf{H}) = \sum_{j=1}^{M}\sum_{i=1}^{K} h_{ij}$.

An iterative algorithm is implemented as proposed in [8] to minimize equation (1), in which basis matrix $\mathbf{W}$ and coefficient matrix $\mathbf{H}$ are updated by gradient descent and multiplicative update rules, respectively.

The parameter λ in equation (1) is an important factor, it is a compromise between the regulation and the NMF cost function. One novelty of this work is the two-step optimization approach, which is proposed to find a proper λ to heuristically optimize the performance of the subjective and various objective measures. This approach is described in more detail in section 4.

3. NMF SPARSE STRATEGY

The dynamic range for electrical stimulation for CIs users is much smaller than acoustic dynamic range in the normal ear. Thus the electrical stimulation has a severe bottleneck to overcome, which only allows limited acoustic information to be transmitted to auditory neurons. However, many experiments have showed that speech has a high degree of redundancy and only few components are needed to allow people to understand speech [11, 12]. Most existing CIs strategies, such as continuous interleaved sampling (CIS), spectral peak (SPEAK) and advanced combination encoder (ACE) indeed try to reduce the redundancy property of speech by selecting only few channels or only using envelope information to stimulate auditory neurons. In order to further solve the information bottleneck problem by stimulating auditory neurons sparsely and efficiently, a serials PCA and ICA based sparse algorithms working on the spectral envelope for CIs was proposed, evaluated and improved in our group[2], [3].

Since the envelope in each channel is non-negative and the firing rates of neurons are never negative, the following part will introduce how NMF can be used in the sparse strategy for CIs. Suppose $z(t)$ is the measured noisy signal, $Z_{i,j}(f)$ is the envelope bin in the i^{th} channel of the j^{th} frame, which is calculated by weighting and summing the short time Fourier transform (STFT) spectrum according to the ACE strategy. $\mathbf{Z}$ is an $N \times M$ envelope matrix, where each column consists of $N = 22$ channel envelop bins, and each row consists of $M = 10$ frames in each analysis block, which is the same as the one used in [2],[3] in order to guarantee the same input signal is used in each analysis block.

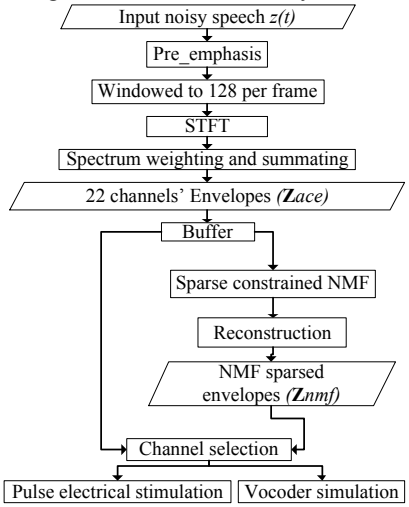


Figure 1 – NMF SPARSE strategy

Figure 1 shows ACE and the proposed NMF sparse strategy for CIs stimulation. The pre-emphasis filter in

Figure 1 is to compensate for the −6dB/octave natural slope in the long term speech spectrum, starting at 500 Hz. After transforming the input speech signal into spectrogram by Fourier analysis, the envelope is extracted in 22 frequency bands by summing the power within each band. These three steps are similar as those in the standard ACE strategy, hence we define it as ACE envelope (although ACE has additional steps such as channel selection). Then NMF sparse are applied to the spectrum envelope on a block by block basis by buffering certain numbers of continuous frames in each channel. In order to produce stimuli for CIs, the envelopes are reconstructed from the NMF components respectively. Finally, appropriate channels are selected and used to stimulate the auditory neurons or to obtain the vocoder simulation signals. In the stimulation stage, the electrical pulse trains driving the stimulation channels are modulated by the envelopes of the signals in the corresponding band pass filters. In addition, the pulse trains are separated in time and interleaved in order to avoid interaction among the electrodes. While the vocoder [13] simulated signals are produced by modulate white noise with the obtained envelope after channel selection.

4. OBJECTIVE EXPERIMENTS AND RESULTS

In this section, a two-step parameter selection procedure is introduced to find the λ in equation (1) : first, various objective measures are introduced to select a range of sparsity levels; then a subjective experiment was performed to set the final value of λ to get better speech intelligibility performance. In detail, since the subjective optimization is time consuming and expensive, five objective evaluation measurements are selected and evaluated for a wide range of $\lambda = [0.01 : 0.01 : 0.2]$ as a pre-selection procedure. A fine range of λ is obtained in this stage and is used in the subjective evaluation experiments to determine the final value.

4.1. Objective evaluation methods and test materials

Because of the space limitation, the introduction of each evaluation method is omitted. Table 1 lists the five objective evaluation methods chosen in this paper and with short descriptions to them.

As shown in Table 1 most of the objective evaluation methods (except kurtosis) require time domain input, while the reconstruction of the NMF is an envelope matrix. In order to evaluate the performance of the sparse NMF algorithms for CIs, the test data are resynthesized vocoder [13] acoustical signal based on the spectrum envelope to simulate the perception of a CIs user, which have been used widely as an extremely valuable tool in the CIs field to simulate the perception of a CIs user [14]. Although the simulations cannot absolutely predict individual user's performance, vocoder simulations have been shown to predict well the pattern or trend in performance observed in CIs users[14]. In this paper, the vocoder simulated signals are produced by modulate white noise with the ACE and NMF sparse strategies processed envelope after channel selection.

The same Bamford-Kowal-Bench (BKB) sentences as in [2] [3] are used as the clean speech in both the objective and subjective experiments. Babble noises at three different long-term signal to noise ratios (SNR) (0, 5, 10 dB) are added to the speech material.

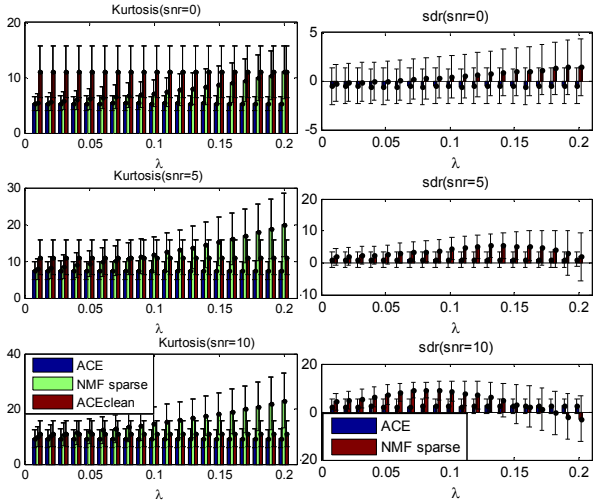
Table 1 Five objective measurements chosen in this research

Objective measurement	Short descriptions
Kurtosis	Since one of the most important goals of these algorithms is to transform the stimuli to be in a more sparse distribution than noisy speech in order to resemble the natural code of auditory neurons better. The kurtosis of the signal is selected to measure the sparseness as used in [2].
Signal-to-distortion ratio (SDR)	The signal-to-distortion ratio (SDR) is shown to be valid as a global performance measure [15].
Normalized covariance metric (NCM)	NCM measure is based on the covariance between the input and output envelope signals. The NCM measure is expected to highly correlate with the intelligibility of vocoded speech due to the similarities in the NCM calculation and CIs processing strategies[16].
Short-time objective intelligibility (STOI)	STOI measure is based on a correlation coefficient between the temporal envelopes of the clean and degraded speech, in short-time overlapping segments. The basic structure of STOI is described in the reference [17].
SNR /Segment SNR	The SNR, frame-based signal-to-noise ratio (SNR) and the corresponding segmental SNR are used as objective measure of speech quality [18] in this paper.

4.2. Results

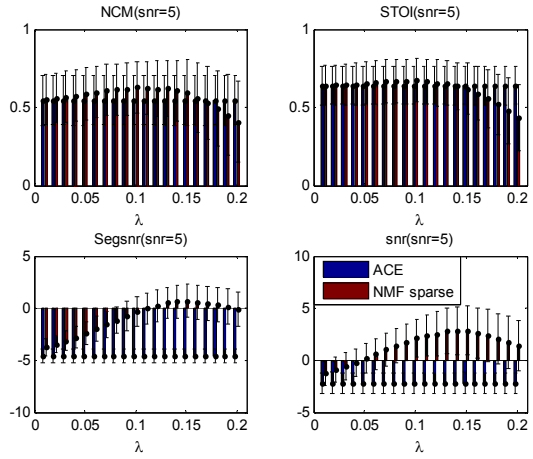
Figure 2 (a) shows the kurtosis of the vocoder sounds of the clean speech's ACE (ACEclean) envelop, the corresponding noisy speech's ACE envelop and sparse NMF envelope at three SNR levels (0, 5 and 10dB) respectively.

To evaluate the sparseness of the processed signal, the vocoder simulated output waveforms is used to calculate the kurtosis of the entire time series. These results are consistent with the results of [2] in that the outputs of the ICA sparse algorithms are more sparse than the output of ACE algorithm. Figure 2 (b) shows the SDR of the vocoder sounds of the noisy speech's ACE envelop and NMF envelope respectively. Figure 3 only shows the NCM, STOI, Segment SNR (Segsnr) and SNR of speech processed by different strategies at two SNR levels (5 and 10 dB) as examples.

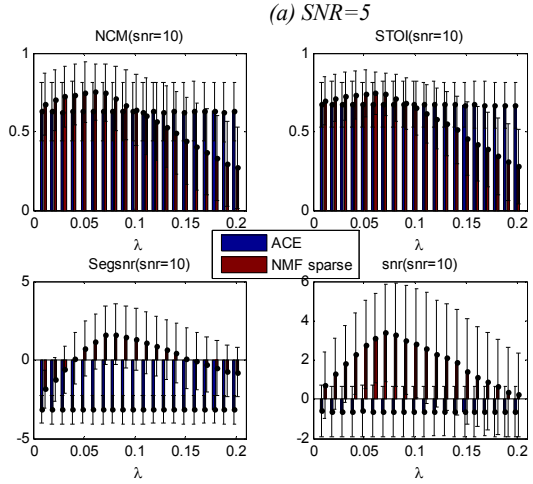


(a) Kurtosis *(b) SDR*

Figure 2 – Kurtosis and SDR of speech processed by different strategies at three SNR levels of 0, 5 and 10 dB



(a) SNR=5



(b) SNR=10

Figure 3 – NCM, STOI, SNR and Segment SNR (Segsnr) of speech processed by different strategies at three SNR levels of 0, 5 and 10 dB

Figure 2 and figure 3 show that for different scenario and measurements, different value of λ should be set to get the corresponding optimized value. Here comes how to choose one λ from this range of optimal values to get better global better performance. In this study, a pilot experiment is designed aimed at finding one optimal λ among this range to obtain better speech intelligibility.

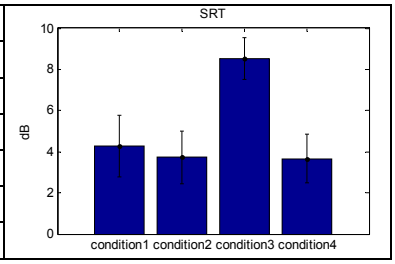
5. SUBJECTIVE SPEECH INTELLIGIBILITY EXPERIMENTS AND RESULTS

Speech reception threshold (SRT) has been proven to faithfully represent speech perception reliability in [19]. To enable comparison with subjective results, speech recognition was assessed using a method and system that described in [20] to provide a speech-in-noise threshold in dB. In this paper, the SNR is changed adaptively with 1 dB step size. All experiments are performed in a sound-isolated room with the sounds presented through a SENNHEISER HDA 200 headphone with the Creek OBH- 21SE headphone amplifier. The BKB sentence lists are presented in a version spoken by a female talker. The sample ratio of the stimulus was 16 kHz. 5 NH (3 males, 2 females, and aged 18-26) paid native English speaking volunteers with no previous experience of the BKB sentence lists participated in these experiments.

Table 2 shows the test materials in different conditions. In condition1, 2 and 3, the vocoder sound was reconstructed from NMF envelope with the sparsity constraint parameter λ = 0.08, 0.13 and 0.18 for all the SNR (from -1dB to 10 dB in the SRT adaptive procedure) respectively. While in condition 4, different λ applied within different SNR range, e.g., λ =0.08 when SNR between 7dB to 10 dB, λ =0.13 when SNR between 3dB to 6 dB and λ =0.18 when SNR between -1dB to 10 dB according to the SNR dependent optimization value of λ showed in Figure 2 and Figure 3.

Table 2. The subjective experiment conditions and results.

Cond.	λ	*SNR(dB)*	
1	*0.08*	*-1 :1 :10*	
2	*0.13*	*-1 :1 :10*	
3	*0.18*	*-1 :1 :10*	
4	*0.08*	*7, 8,9,10*	
	0.13	*3,4, 5,6*	
	0.18	*-1, 0,1,2*	



The bar chart in table 2 shows that condition 2 and condition 4 have significant better SRT than the other two conditions. It is reasonable that condition 2 and 4 have very similar SRT when we notice that their SRT values are around 4 dB, in this situation, both condition have the same λ =0.13, which in another way prove the reliability of the SRT test used in this paper. So the optimized λ according to SRT should between 0.08 and 0.13. Figure 4 shows the bar chart of five objective evaluation measurement values

when λ was set to 0.13 according to the SRT experiments which is chosen heuristically to maximize the performance of the whole algorithm by subjective informal listening. It indicates that λ =0.13 can improve most of the objective measurements for all three SNR although it is not always the golden value for different measures and SNR conditions.

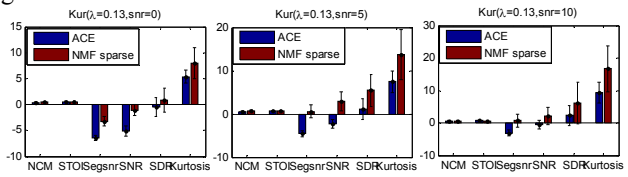


Figure 4 – Five objective measurement values of the NMF sparse processed vocoder sound at three SNR levels of 0, 5 and 10 dB

6. DISCUSSIONS AND CONCLUSIONS

Normal hearing listeners understand speech well in a noisy environment, but this is a very challenging situation for CIs users. Sparse strategies proposed in our previous work showed prospect for CIs users in both noise reduction and sparsity enhancement in order to deliver key information to CIs users via limited frequency channels. The characteristics of the non-negativity of both the envelope in each channel and that of the firing rates of neurons draw our attention to the NMF which has increasingly been used as a tool in various applications, while it has not been used in the CIs yet. In this paper, a basic NMF was applied to the envelope matrix with sparsity constraint on the coefficient matrix to get more sparse representation. Since the choice of sparsity parameter is important, five objective evaluations and a pilot subjective experiment were used together in this study aimed to choose the parameters of sparse NMF properly to trade-off between the objective measurements and speech intelligibility. Finally the objective results for the parameter chosen in the pilot experiment were applied and five objective evaluations were calculated for three different SNR, most of the objective evaluation measurements showed improvement compared to the noisy ACE strategy. In the future more participants of NH and CIs will be recruited to further evaluate the proposed CIs strategy.

7. ACKNOWLEDGEMENTS

This work was supported by the European Commission within the ITN AUDIS (grant agreement number PITN-GA-2008-214699). The authors appreciate Cochlear Europe Ltd. providing the NIC software and participants' hard work in subjective experiments.

8. REFERENCES

[1] S. Greenberg, W. A. Ainsworth, A. N. Popper *et al.*, "Speech Processing in the Auditory System: An Overview," *Speech Processing in the Auditory System*, Springer Handbook of Auditory Research, pp. 1-62, New York: Springer, 2004.

[2] G. Li, "Speech perception in a sparse domain," PhD thesis, Institute of Sound and Vibration, University of Southampton, Southampton, 2008.

[3] H. Hu, G. Li, L. Chen *et al.*, "Enhanced sparse speech processing strategy for cochlear implants," in 19th European Signal Processing Conference (EUSIPCO 2011) Barcelona, Spain, 2011, pp. 491-495.

[4] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, no. 6755, pp. 788-791, 1999.

[5] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), , Mohonk Mountain House, New Paltz, NY, 2011, pp. 418-423.

[6] Z. Yang, G. Zhou, S. Xie *et al.*, "Blind Spectral Unmixing Based on Sparse Nonnegative Matrix Factorization," *Image Processing, IEEE Transactions on,* vol. 20, no. 4, pp. 1112-1125, 2011.

[7] A. Cichocki, R. Zdunek, and S. Amari, "New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation," in Acoustics, Speech and Signal Processing, 2006 IEEE International Conference on, 2006, pp. V-V.

[8] P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *The Journal of Machine Learning Research,* vol. 5, pp. 1457-1469, 2004.

[9] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 3, pp. 1066-1074, 2007.

[10] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 2008, pp. 1833-1836.

[11] K. Kasturi, P. C. Loizou, M. Dorman *et al.*, "The intelligibility of speech with ``holes" in the spectrum," *The Journal of the Acoustical Society of America,* vol. 112, no. 3, pp. 1102-1111, 2002.

[12] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.,* vol. 119, pp. 1562-1573, 2006.

[13] R. V. Shannon, F.-G. Zeng, V. Kamath *et al.*, "Speech Recognition with Primarily Temporal Cues," *Science* vol. 270, no. 5234, pp. 303-304 1995.

[14] P. C. Loizou, "Speech processing in vocoder-centric cochlear implants," *Cochlear and Brainstem Implants*, A. R. Møller, ed., pp. 109-43, Basel, New York: Karger, 2006.

[15] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, no. 4, pp. 1462-1469, 2006.

[16] F. Chen, and P. C. Loizou, "Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech," *The Journal of the Acoustical Society of America,* vol. 128, no. 6, pp. 3715-3723, 2010.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens *et al.*, "An Algorithm for Intelligibility Prediction of Time and Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 7, pp. 2125-2136, 2011.

[18] P. C. Loizou, *Speech Enhancement: Theory and Practive*: CRC Press, 2007.

[19] R. Plomp, and A. M. Mimpen, "Improving the Reliability of Testing the Speech Reception Threshold for Sentences," *International Journal of Audiology,* vol. 18, no. 1, pp. 43-52, 1979.

[20] M. Dahlquist, M. E. Lutman, S. Wood *et al.*, "Methodology for quantifying perceptual effects from noise suppression systems," *International Journal of Audiology,* vol. 44, no. 12, pp. 721-32, Dec, 2005.