AUGMENTING INFORMATION CHANNELS IN HEARING AIDS AND COCHLEAR IMPLANTS UNDER ADVERSE CONDITIONS

Yasir Suhail [1], *Karim G. Oweiss* [2]

[1] Biomedical Engineering Department, Johns Hopkins University, USA
[2] Electrical & Computer Engineering Department, Michigan State University, USA

ABSTRACT

We conceptualize a new signal processing strategy to better represent the temporal and spectral cues in speech signals for Hearing Aid (HA) and Cochlear Implant (CI) applications under severe adverse conditions. The proposed approach rests on two well studied methods for signal separation and noise suppression, namely, the denoising and function approximation capabilities of the wavelet transform, blended with signal subspace decomposition through low rank approximation. The technique targets suppression of "competing voice" type noises. A cost function is defined to obtain a "best basis" representation of the desired speech signal for which an inherent invariance property of the signal subspace is observed. This allows better separation of the *speech-like* noise in contrast to classical bandpass filtering currently employed in CI and HA devices. We demonstrate the efficiency of the proposed method in capturing the rapid dynamics of speech signals, while minimizing the masking effects of noise, in addition to improved recognition rates in normal hearing listeners. The technique remains to be tested on actual patients.

1. INTRODUCTION

Cochlear implant (CI) and Hearing Aid (HA) devices have undergone significant development over the past three decades due to continuous advances in micro- and nano-circuit fabrication, the concomitant development of multi-electrode arrays, and the accompanying development of signal processing technologies. Quality gains in patient performance can be assessed in Hearing Aid (HA) applications because the level of uncertainty in many individual factors is negligible. In the case of CIs, it is widely believed that multichannel CI devices provide the users with substantially better speech recognition capabilities than single channel implants [1]. Nevertheless, measuring quality gains in the CI case is a challenging task because of the difficulty in assessing the significance of individual factors like neuron survival rate, electrode insertion depth and alignment, pre-surgical hearing and language skills, etc.

It is clearly established, however, that performance gains in both cases quickly corrode in the presence of competing speakers, and of transient and persistent sources of environmental noise. This problem remains troublesome, even in the face of remarkable technical advances that allow CI devices to deliver increasing amounts of information to the auditory nerve. As current and future research produces continuous technical improvement in CI devices, it appears that the ultimate success of CI technology will be persistently mitigated by the performance of the associated signal processing technologies under adverse conditions.

Current signal processing techniques for HA and CIs are merely based on classical Band Pass filtering (BPF) [2]. Filter-bank techniques provide poor temporal resolution of transient sound events which are often critical to proper speech recognition, as well as poor frequency resolution if low order filters are used. Higher temporal and spectral resolution are needed to better encode the necessary temporal, spectral and acoustic cues by providing a compact, nonlinear, multiresolution mapping of sound to the auditory system.

Some recent efforts focused independently on exploiting two well known methods for noise reduction, namely the wavelet transform as a *denoising* tool [3][4], and signal subspace decomposition as an optimal orthogonal noise suppression tool [5]0. However, the link between both methods has not been fully exploited. The objective of this paper is to introduce a novel signal processing strategy that relies on exploiting signal subspace decomposition in the multiresolution domain that provides a number of advantages over existing techniques, and can be easily tailored to improve the patient performance under severe, *speech*-like, noisy conditions. Therefore, we adequately formulate a blend of both methods, and demonstrate that superior noise suppression can be obtained, besides improved temporal resolution, frequency specificity, and electrode channel selectivity.

2. THEORY

2.1. Signal Model

Suppose that $\boldsymbol{s}_1 = [s_1[0] \ldots s_1[N-1]]$ denotes the desired clean speech signal to be transduced through the device over a time frame of length N. In the presence of an unknown number P of independent speakers, the mixture model takes the form

$$\boldsymbol{x}_m = \sum_{p=1}^{P} a_{mp} \boldsymbol{s}_p \qquad (1)$$

where a_{mp} denotes the weight of speaker p in the m^{th} speech frame. An additive noise model assumes the m^{th} frame to be expressed as

$$\boldsymbol{y}_m = \boldsymbol{x}_m + \boldsymbol{z}_m \qquad (2)$$

where $\boldsymbol{z}_m$ denotes a zero-mean additive white noise comprising the thermal and electrical noise from the electronics of the associated HA (CI) device circuitry. Over M consecutive speech

frames, the observations can be conveniently expressed in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z} = \boldsymbol{AS} + \boldsymbol{Z} \tag{3}$$

Classical signal subspace decomposition using Singular Value Decomposition (SVD) can be used to spectrally factor $\boldsymbol{Y}$ as

$$\boldsymbol{Y} = \boldsymbol{U}_Y \boldsymbol{D}_Y \boldsymbol{V}_Y^T = \sum_{i=1}^{M} \delta_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{4}$$

where δ_i denotes the i^{th} singular value corresponding to the i^{th} diagonal entry in $\boldsymbol{D}_Y$. The eigenvectors $\boldsymbol{u}_i$, $i = 1,...,P$ span the subspace spanned by the columns of $\boldsymbol{A}$. The remaining $M - P$ eigenvectors span the noise subspace. Since the objective is to isolate the principal speech signal in the observed mixture, it is sufficient to consider the first eigenvalue/eigenvector pair $\delta_1 / \boldsymbol{v}_1$ since $\boldsymbol{v}_1$ spans the row space of ***Y.*** This analysis is guaranteed to separate the principal speech signal provided that the desired signal has the largest energy. In such case, the observations are projected onto the signal subspace (now spanned by $\boldsymbol{u}_1$) and used to activate the device interface.

2.2. Subband Decomposition and Best Basis Selection

The approach outlined above has some serious drawbacks. *First*, it assumes that the mixing among multiple, consecutive speech frames remains stationary within the analysis window. That is, the desired speech signal always has higher energy than the competing speakers. This may not always be the case, especially in speech-like background noise, where the variance of the desired signal can be lower than that of the competing ones. In this case, artifacts can occur and can significantly degrade the outcome of the algorithm by causing spurious activation of the device. *Second*, the fixed frame length does not capture the rapid dynamics of unvoiced sounds such as fricatives, which are crucial to proper speech intelligibility and recognition [1].

To separate the desired signal, the m^{th} frame undergoes a Discrete Wavelet Packet Decomposition (DWPT) up to L levels ($J = 2^{L+1} - 1$ subbands) such that (2) can be expressed as

$$\boldsymbol{y}_m^j = \boldsymbol{x}_m^j + \boldsymbol{z}_m^j \tag{5}$$

Where $\boldsymbol{y}_m^j = (y_m^j[0], y_m^j[1]..., y_m^j[N_j - 1])$ denotes the DWPT of $\boldsymbol{y}_m$ in the j^{th} subband, and $N_j = N / Q^j$, where Q denotes the order of the DWPT. In doing so, one obtains an overcomplete representation of the observations in the form of a dictionary of basis $\Delta\{J\}$ to choose from. Similar to (3), the transformed observations can be expressed in matrix form as

$$\boldsymbol{Y}^j = \boldsymbol{X}^j + \boldsymbol{Z}^j \tag{6}$$

It is reasonable to assume that speech signals from independent speakers are mutually independent, as well as independent of the speech to be recognized. A singular value decomposition applied to (6) yields

$$\boldsymbol{Y}^j = \boldsymbol{U}_Y^j \boldsymbol{D}_Y^j \boldsymbol{V}_Y^{j^T} = \sum_{i=1}^{M} \delta_i^j \boldsymbol{u}_i^j \boldsymbol{v}_i^{j^T} \tag{7}$$

The objective is thus to identify the signal and noise subspace in each subband. This can be efficiently carried out with a "best basis" approach [7]. To adequately prune the DWPT binary tree obtained, a cost function has to be defined. In [7], the cost is based on entropy minimization. Other criteria were proposed in the context of mean square error (MSE) minimization [8]. If the cost of the children is less than that of the parent, the parent node is further split and the process repeats. The algorithm stops when all the dictionary of basis $\Delta\{J\}$ is exhausted, which occurs at the last decomposition level. The outcome is a characteristic best basis tree

In our case, the cost function needs to be expressed in terms of the "features" of the parent subband that are preserved in the children subbands. It is reasonable to assume that the principal speech signal in those subbands would dominate over those of the competing speakers if the cost function is defined in terms of second order statistics. This implies that the best basis search would rely on identifying the dominant eigenvector that remains *invariant* in those subbands that best represent the desired signal temporal and spectral features. However, in the presence of competing voices, the desired signal may not always correspond to the principal eigenvector. This is particularly true if the desired signal undergoes temporal decay in its energy content over the time interval spanned by *M* frames, and can be easily superseded by the competing speakers signal energy. This can be expected near the start and/or the end of a given word where hearing impaired patients experience the foremost difficulty that greatly impacts their recognition capabilities. In such case, the best basis selection would rely on identifying the principal eigenvector of a given parent node and search in the eigenvectors of the children nodes for a *similar* eigenvector in a MSE error sense. If one is located -which may not necessarily be the dominant one- then the child node is marked as a good candidate for further splitting. Mathematically, this can be expressed as

$$Cost(j, p) = \min_{j \in \Im_p} \left\| \boldsymbol{u}_p^{Parent} - \boldsymbol{u}_p^{Child} \right\|^2 \tag{8}$$

where p denotes the index of the eigenvector in a given node for which the desired source signal p was preserved from its parent and $\Im_p \subset \Delta\{J\}$ denotes the set of best basis indices representing the desired signal.

It is worth noting that our assumption of independence among competing speakers guarantees that the desired signal eigenvector $\boldsymbol{u}_p^{Child}$ can be found among the columns of $\boldsymbol{U}_Y^j$ if and only if the child's wavelet basis best approximates the desired signal temporal and spectral content. In that context, existing techniques such as independent component analysis (ICA) for blind separation of independent speakers from mixtures are worth discussing. In the context of ICA, the separation is based on minimizing mutual information and thus is based on higher order statistics. The limitation in such case is the inability of ICA to separate more than one Gaussian source in the mixture, because the Gaussian distribution has maximum entropy. In our case, it is clear that the technique can be applied to observations containing more than one Gaussian source. Indeed, it is guaranteed to yield its best performance if the observations are a mixture of Gaussian sources since it exploits second order statistics in the multiresolution domain, thus can be shown to yield superior performance over ICA in this problem. We omit the details of the derivations for the lack of space.

2.3. Noise Suppression

Besides the best basis selection approach we outlined above, an added advantage of the proposed technique is the existence of a large body of literature (see [9], [10] for example), that demonstrate the denoising capabilities of wavelets in many different contexts. In this section, we outline the multistage noise suppression capabilities of the proposed technique. Noise components orthogonal to the signal subspace (i.e., uncorrelated with the speech of interest) can be directly suppressed using the low-rank approximation by truncating the summation in (7) as

$$\widetilde{\boldsymbol{Y}}_{\boldsymbol{j}} = \sum_{i=1}^{P_j} \delta_i^j \boldsymbol{u}_i^j \boldsymbol{v}_{\boldsymbol{i}}^{j\,T} \tag{9}$$

where P_j denotes the dimension of the signal subspace in subband *j*. Nevertheless, some competing voice components may be spectrally correlated with the signal of interest and therefore will not be orthogonal to the signal subspace. It is obvious that such a realistic assumption, while inadvertently ignored in exiting filter bank techniques, is essential to provide a pragmatic solution to the problem. This assumption is tantamount to say that a *special competing* voice spectrum overlaps significantly with the signal of interest and can therefore involuntarily activate the stimulating electrode array causing degradable performance. To suppress this noise component, the matrix $\widetilde{\boldsymbol{Y}}^j$ is first whitened using the eigenvector matrix $\boldsymbol{U}^j$ obtained from (7) to yield

$$\hat{\boldsymbol{Y}}^j = \boldsymbol{U}^{j\,T} \widetilde{\boldsymbol{Y}}^j \tag{10}$$

This step segregates the signal and the correlated noise components in each subband. The matrix $\hat{\boldsymbol{Y}}^j$ is then thresholded by setting to zero (*hard* thresholding) or shrinking (*soft* thresholding) all the coefficients that are below a certain threshold. The resulting above-threshold coefficients are used for signal reconstruction in HA devices or energy extraction to modulate the carrier pulse train activating the electrodes of CI devices. One possibility for setting the denoising threshold is by using *universal thresholding* rule [9], defined as $T_j = \sigma_j \sqrt{2 \log N}$, in which σ_j^2 is the noise variance in the j^{th} subband. For the m^{th} frame, σ_j^2 can be estimated according to $\hat{\sigma}_j^2 = MAD\{\boldsymbol{y}_m^j\} / 0.6475$, where MAD{.} is the median absolute deviation.

As the number of speech frames *M* becomes large, it is anticipated that the technique captures more *discriminant* temporal and spectral features between the desired and the undesired signals. This can be seen by noting that increasing *M* augments the size of the eigenvector $\boldsymbol{u}_p$. It follows directly from equation (8) that this contributes largely to improved tracking capability of the algorithm when searching for the invariant signal subspace of the children subbands. In fact, the ℓ_2 norm criterion is known to provide better performance as the dimension of the vectors increase.

In terms of stimulation parameters in CI applications, it is clearly seen that the matrices $\hat{\boldsymbol{Y}}^j$, $j \in \Im_p$ comprise the combined temporal and spectral features needed to activate the appropriate electrode locations in the cochlea in accordance with the "place theory" [11], thus can contribute to improved electrode channel specificity and minimize channel interaction [12]. This can be inferred by realizing that the wavelet coefficients constituting the features in the matrices $\hat{\boldsymbol{Y}}^j$ are sparse, and therefore the envelope detection scheme currently used in CI device technology can be well tailored to minimize the time in which the simulation pulse train is turned on. The magnitude of the eigenvalue δ_p^j directly indicates the relative energy of the signal in the j^{th} subband and is used as a voice activity detector in our algorithm. This measure can be effectively used to pre-determine how much energy is needed at the input of the stimulation pulse train stage. Modulation of pulse trains with temporal characteristics of the coefficient envelopes can be directly utilized in existing stimulation technology.

3. RESULTS

The technique described above was fully implemented in MATLAB® and tested with data obtained from the IEEE sentence database available in [13]. The noisy database contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. We selected in our tests the noisy files that contain babble and exhibition hall noises. The samples were tested with normal hearing listeners consisting of two males and two females. Figure (1) illustrates one sample sentence for which the signal subspace voice activity detector indicated the presence of 'significant' signal energy. The detector is based on the magnitude of the principal eigenvalues that exceed a predetermined threshold. These were eventually used to identify the candidate subbands in which high energy coefficients reside. Next, the best basis selection mechanism was applied to identify the characteristic subbands that efficiently track the desired speech signal across the decomposition. A small interval at the start of a word is magnified to illustrate the negative signal to noise ratio in certain instants demonstrating the ability of the algorithm to separate the desired signal from the background babble noise. The quasi periodic nature of residual noise (indicated in blue) and the phase difference between the clean signal and the residual noise indicate that the noise is mostly composed of speech-like components. This can be clearly seen in the reconstructed spectrogram in the bottom of Figure 1. Another single utterance example is illustrated in Figure 2. The performance for variable SNR for different choices of *M* is illustrated in Figure 3. The results perfectly agree with our expectation in terms of improved recognition accuracy.

4. CONCLUSION

We presented a novel algorithm that combines the advantages of signal subspace decomposition with those of best basis selection to provide adaptive and dynamic allocation of frequency contents of a desired speech signal corrupted by competing-voice type noise. More compact temporal resolution is desired in the context of CI applications to achieve sparse activation of neuronal populations thus minimizing channel interaction with high-density electrode configurations (250μm pitch). The selectivity and specificity of the algorithm can be inferred by considering the variable number of best basis obtained in each batch of consecutive frames

simultaneously processed. Moreover, the scalability to an arbitrary number of physical electrode channels depending on the available technology [13] can be readily seen by varying the order *Q*. The technique remains to be tested on actual HA and CI patients.

5. REFERENCES

[1] P. Loizou, "Signal processing techniques for cochlear implants," *IEEE EMBS Magazine*, 18(3), pp. 34-46, 1999.

[2] D. Lawson, B. Wilson and M. Zerbi "Speech processors for auditory prostheses," NIH Project N01-DC-2-2401, 2nd QPR, 1993

[3] J. Yao, and Y-T. Zhang, "The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations," *IEEE TBME*, 49,1299, 2002

[4] Y. Hu, and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. On Speech and Audio Proc.*, vol. 12, pp. 59-67, 2004.

[5] J.-P. Maj, *et al*. "SVD-Based optimal filtering for Noise reduction in dual microphone hearing aids: A real-time implementation and perceptual evaluation," *IEEE Trans. BME*, 32-9, Sep 2005.

[6] Y. Hu, and P. Loizou, "A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 334-341, 2003

[7] R.Coifman and M.Wickerhauser, "Entropy based algorithms for best basis selection," *IEEE Trans. on IT*, 38: 713-718, 1992.

[8] H. Krim *et.al.*, "Best basis algorithm for signal enhancement," *Proc. ICASSP*, pp. 1561-1564, May 1995

[9] D.L. Donoho, "De-Noising by Soft Thresholding," *IEEE Trans. on IT*, vol. 41, pp.613-627, 1995.

[10] K.G. Oweiss and D. J. Anderson, "A New Approach to Array Denoising," *Proc. of the IEEE 34th ASSC*, 1403-7, Nov 2000

[11] H. Helmhotz, *On the Sensations of Tone*, New York: Dover, 1954. (Originally published in German, 1877)

[12] J. Middlebrooks, "Effects of Cochlear-implant pulse rate and inter-channel timing on channel interactions and thresholds", *JASA* 116 (1): 452:468, July 2004.

[13] http://www.utdallas.edu/~loizou/speech/noizeus

[14] K. Wise *et al*., "Wireless implantable Microsystems: High density electronic interface to the nervous system," *Proc. IEEE* vol. 92 , pp. 76 - 97. 2004

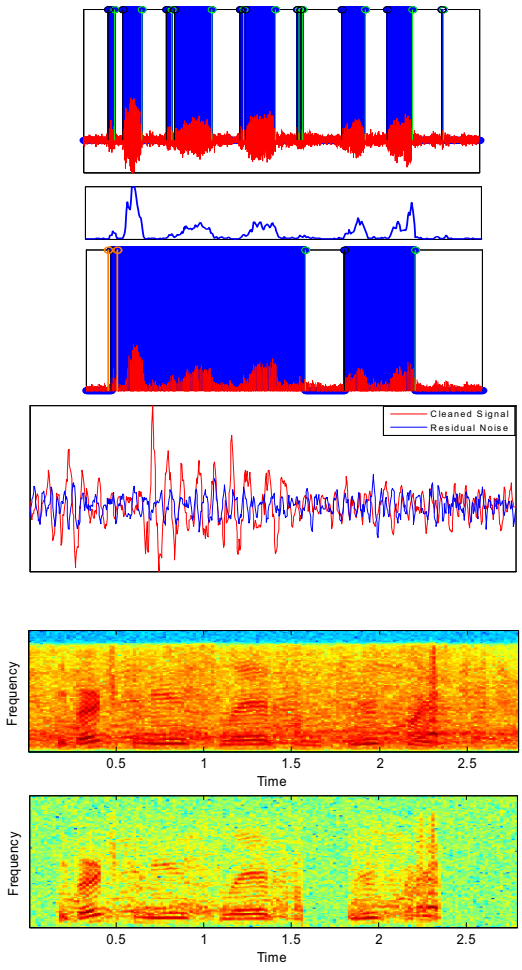

Figure 1: Sample sentence from IEEE database of speech corrupted by babble noise. Envelope of the eigen-based voice activity detector. Zoom-in on the start of the first word. Spectrogram of the overall sentence preserving all the spectral content of the principal speech

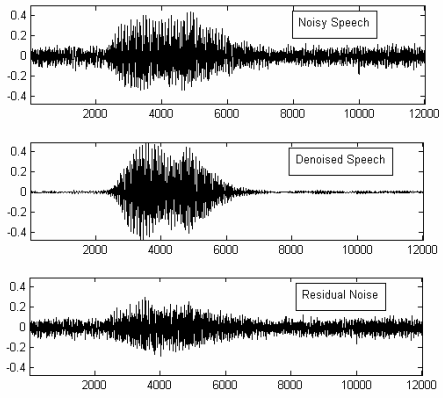

Figure 2: Female utterance of /two/ in the presence of coherent babble and vacuum cleaner noise. Notice the smooth transient start and end of the word in the reconstructed speech segment (middle panel) where the SNR is negative.

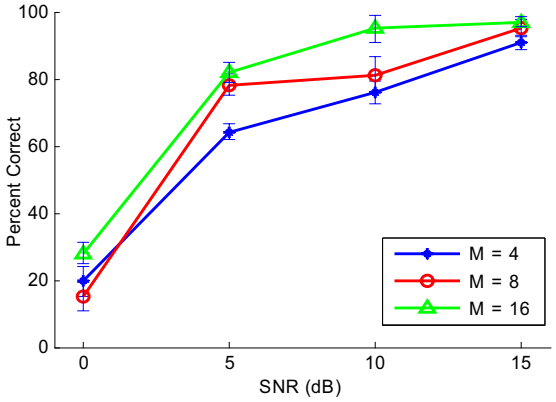

Figure 3: Recognition test scores versus SNR for normal–hearing listeners for multiple numbers of simultaneously processed frames.