SPEECH UNDERSTANDING IN NOISE PROVIDED BY A SIMULATED COCHLEAR IMPLANT PROCESSOR BASED ON MATCHING PURSUIT

Abigail A. Kressner and Christopher J. Rozell

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, Georgia 30332 USA, {abbiekre,crozell}@gatech.edu

ABSTRACT

Speech reception is poor for cochlear implant recipients in listening environments with interfering noise. This study investigates the speech understanding provided in interfering noise by a coding strategy based on the sparse approximation algorithm matching pursuit (MP) and additionally proposes two modifications to the strategy. The levels of spectral information provided by the MP strategy and the modified MP strategy are compared to that of continuous interleaved sampling (CIS) and a strategy based on the ideal binary mask (IBM) using vocoded speech and the normalized covariance metric (NCM). We demonstrate objective intelligibility improvements in quiet, and total and partial objective intelligibility restoration in steady-state and fluctuating noise, respectively.

Index Terms— Cochlear implant, speech coding, sparse coding, matching pursuit (MP), continuous interleaved sampling (CIS), normalized covariance metric (NCM)

1. INTRODUCTION

For cochlear implant (CI) recipients, speech perception is poor in listening environments with interfering noise [1]. The poor outcomes are due in part to the fact that CIs provide limited frequency resolution, weak temporal pitch cues, a small dynamic range, and severely degraded temporal fine structure [2]. With these limitations in place, it is especially important to encode information in speech with a strategy that is both intelligent and efficient.

Two strategies that remain in widespread use in current CI systems are the Continuous Interleaved Sampling (CIS) and the Advanced Combination Encoder (ACE) [3]. Studies have shown that CI recipients generally prefer ACE over CIS [4]. In noise however, the peak amplitude criterion that is at the root of ACE's n-of-m channel selection algorithm can be problematic since it will prioritize encoding channels with high envelope amplitudes even if the channels contain only interfering noise [5]. To address this issue, Hu and Loizou [5] proposed a selection criterion based on the ideal binary mask (IBM): channels with a local signal-to-noise ratio (SNR) greater than or equal to 0dB are selected. They demonstrated restoration of intelligibility for noisy speech when compared to CIS in quiet. However, this selection criteria requires oracle knowledge of the target and interfering signal.

By stimulating a limited number of electrodes, n-of-m coding strategies in general create T-F representations that are more sparse (i.e., has fewer non-zeros) than the original acoustic signal. Many recent advances in signal processing for both signal modeling and signal acquisition have revolved around this notion of sparsity. Alongside these advances in signal modeling, there is mounting evidence from computational neuroscience that neural systems may also use sparse coding to represent sensory information (e.g., [6]). Furthermore, the signal processing community is actively developing methods to compute sparse approximations in real-time and using low-power (e.g., [7]). The fact that CI coding strategies create T-F representations that are sparse leads to the possibility that the recent advances in sparse approximation algorithms may be utilized for intelligent channel selection in CI processing. In line with this effort, Taal et al. [8] proposed a coding strategy based on a sparse approximation algorithm called matching pursuit (MP). In their preliminary evaluation using tone-vocoded speech to simulate a CI processor, they demonstrated that their strategy improves intelligibility relative to the ACE strategy (as measured by objective measures) for speech in quiet.

In this paper, we use tone-vocoded speech and the normalized covariance metric (NCM) (which correlates well with the intelligibility of vocoded speech [9]) to predict how the MP coder performs in noise. We identify two areas of weakness in the MP strategy and propose modifications to address them. Finally, we compare MP and MPm (without and with modification, respectively) to the CIS and IBM coding strategies. We demonstrate objective intelligibility improvements in quiet, and total and partial objective intelligibility restoration in steady-state and fluctuating noise, respectively.

2. CODING STRATEGIES

2.1. Basic strategies

CIS and ACE remain in widespread use within CI processors and are often the basis for comparison. In the CIS strategy, signals are decomposed into a small number of bands and the envelopes are extracted from each band. Variations in the envelope amplitudes are represented at corresponding electrodes in the CI through proportional modulation of trains of biphasic electrical pulses. The pulse trains for each of the channels and electrodes are interleaved in time so that the pulses across channels are non-simultaneous [3].

The ACE strategy belongs to the more general class of n-of-m algorithms. In n-of-m coding strategies, input signals are passed through m bandpass filters, and from these m bandpass filters, n channels are selected for stimulation. In ACE, the channels with the n largest envelope amplitudes are selected [3]. Studies with the Nucleus CI 24M system have shown that most users prefer the ACE over the CIS strategy [4]. However, while the peak amplitude criterion works well in quiet, it can introduce confusion in the presence of dominating noise because it prioritizes encoding channels with high envelope amplitudes even if the channels contain only interfering noise [5].

This research was made with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

978-1-4799-0972-8/13/$31.00 ©2013IEEE

2.2. IBM strategy

To address the shortcomings of the ACE strategy, Hu and Loizou proposed using channel-specific SNR for selecting which n channels to stimulate [5]. This coding strategy is strongly based on the ideal binary mask (IBM); it selects channels with local SNRs that are larger than or equal to 0dB and discards channels with local SNRs less than 0dB. Unlike the n-of-m strategy which selects n channels to stimulate in every cycle, the IBM strategy chooses as few as 0 channels and as many as m channels each cycle. When in quiet, the IBM strategy is equivalent to CIS. With six postlingually deafened CI users, Hu and Loizou demonstrated nearly full restoration of speech intelligibility in noise when compared to CIS in quiet. However, this coding strategy requires oracle knowledge of the target and interferer signals.

2.3. MP strategy

In contrast to IBM, the MP strategy [8] does not use oracle knowledge. Instead, it uses a sparse coding model to efficiently encode signals. Sparse coding models treat a signal as a linear combination of elements from a dictionary, and sparse approximation uses these models to find approximations to signals with as few of the dictionary elements as possible. A signal, Y, is represented by a linear superposition of a basic set of dictionary elements (D_j). The sparse coding model is given as $Y \approx \sum_{j=1}^{J} a_j D_j$. To code signals efficiently, one generally needs to find the optimal set of coefficients $\{a_j\}$ (*encoding* or *inference*), as well as find the optimal set of D_j (*learning*). Note that inference is done on a signal-by-signal basis, whereas learning is done once to capture signal statistics. Many algorithms exist for inferring the optimal set of coefficients, and matching pursuit (MP) is one example [10]. MP first chooses the basis that has the largest inner product with the signal, then subtracts the contribution due to that basis, and repeats the process iteratively until the signal is satisfactorily approximated.

Taal et al. [8] reformulate CI channel selection as an inference problem in the time-frequency domain. They use a dictionary of complex exponentials characterized by frequencies that align with the CI electrodes. However, since inference is done on envelope amplitudes in the time-frequency domain, the dictionary is actually made up of the short-time filterbank magnitude responses to the set of windowed complex exponentials. Taal et al. [8] also propose a weighting scheme that is derived from the short time objective intelligibility (STOI) measure, an objective intelligibility measure that correlates well with time-frequency weighted noisy speech [11]. The weighting scheme is incorporated into MP to emphasize and de-emphasize channels in order to encourage or discourage selection of the corresponding dictionary elements. At each stimulation cycle, the MP strategy chooses the n channels that correspond to the first n dictionary elements that the weighted-MP program identifies as optimal.

2.4. Implementation details

We first introduce some general notation, and then provide detail on how we implemented the CIS, IBM, and MP strategies. Let $x(t)$ denote the acoustic waveform, let $x_m(t)$ denote the m^{th} windowed segment of $x(t)$, and let $X_m(i)$ denote the short-time gammatone filterbank magnitude for the i^{th} band of $x_m(t)$ (X is the "gammatonegram") [12]. For CIS and IBM, the analysis rate and filterbank size match the stimulation rate and number of CI channels. Therefore, the stimulation pattern at the c^{th} cycle and the j^{th} channel ($E_c(j)$) is given as a function of the gammatonegram for the corresponding m^{th} frame and i^{th} frequency band as

$$E_c^{\text{CIS}}(j) = \sqrt{\hat{X}_m(i)} \tag{1}$$

$$E_c^{\text{IBM}}(j) = \sqrt{\hat{X}_m(i)} \quad \text{if} \quad \text{SNR}_m(i) \geq 0\text{dB} \tag{2}$$

$$E_c^{\text{IBM}}(j) = 0 \qquad\qquad \text{if} \quad \text{SNR}_m(i) < 0\text{dB}, \tag{3}$$

where $\text{SNR}_m(i)$ is the local SNR. For these algorithms, we use a window length of 8ms (Hann window with 50% overlap) and record a stimulation pattern every 4ms.

For MP, let B_m be a block of $\hat{X}_m(i)$ over all I bands for the last M frames,

$$B_m = \text{vec}\left(\begin{bmatrix} \hat{X}_{m-M+1}(1) & \cdots & \hat{X}_m(1) \\ \vdots & \ddots & \vdots \\ \hat{X}_{m-M+1}(I) & \cdots & \hat{X}_m(I) \end{bmatrix}\right) \tag{4}$$

and let A_m be a set of channel-specific weights that are inversely proportional to the variation within each of the channels.

$$A_m = \text{diag}\left(\text{vec}\left(\begin{bmatrix} \frac{\sqrt{M}}{\sigma_m(1)} & \cdots & \frac{\sqrt{M}}{\sigma_m(1)} \\ \vdots & \ddots & \vdots \\ \frac{\sqrt{M}}{\sigma_m(I)} & \cdots & \frac{\sqrt{M}}{\sigma_m(I)} \end{bmatrix}\right)\right), \tag{5}$$

where $\sigma_m(i)$ is the standard deviation of the vector $\begin{bmatrix}\hat{X}_{m-M+1}(i) & \cdots & \hat{X}_m(i)\end{bmatrix}$. For this strategy, we use an analysis rate that is half as low as the stimulation rate (analysis window is a 16ms Hann window with 50% overlap and a stimulation pattern is generated every 4ms) so we compute two stimulation patterns for each analysis frame. Thus, we have two different sets of dictionaries that are each characterized by how the stimulation cycle aligns with the analysis window (see Fig. 1 of [8]). The first channel to be selected in the stimulation pattern at the c^{th} cycle corresponds to the dictionary element that can represent the largest amount of energy in B_m. Put another way, MP chooses the dictionary element that minimizes what the residual signal will become.

$$\hat{a}_j = \text{argmin}\, ||A_m R_x - a_j A_m D_j||\,, \tag{6}$$

$$a_j = \frac{\langle A_m B_m, A_m D_j \rangle}{||A_m D_j||^2} \tag{7}$$

$$E_c^{\text{MP}}(j) = \sqrt{\hat{a}_j} \tag{8}$$

where R_x is the part of B_m not yet represented by stimulation in previous cycles (i.e., the residual). After selecting $\hat{a}_j$ and recording it to $E_c^{\text{MP}}(j)$, the residual is updated and the process is repeated until MP has identified the optimal set of n channels.

For all of the strategies, we use $J = 20$ channels logarithmically spaced between 150Hz and 5kHz, and for MP, we select $n = 9$ of those channels each cycle. For the analysis filterbank, we use $I = 20$ gammatone filters with center frequencies linearly spaced on an ERB scale between 150Hz and 5kHz for CIS and IBM and $I = 32$ with linear spacing and the same range for MP. Furthermore, we use $M = 48$ to form the time-frequency block B_m, which equates to approximately 400ms. Finally, to create the vocoded stimuli, we initialize sinusoidal carriers with frequencies that match the center frequencies of the stimulation channels, modulate 8ms frames (Hann windowed with 50% overlap) of the carriers with the stimulation pattern, and then sum all of them together.

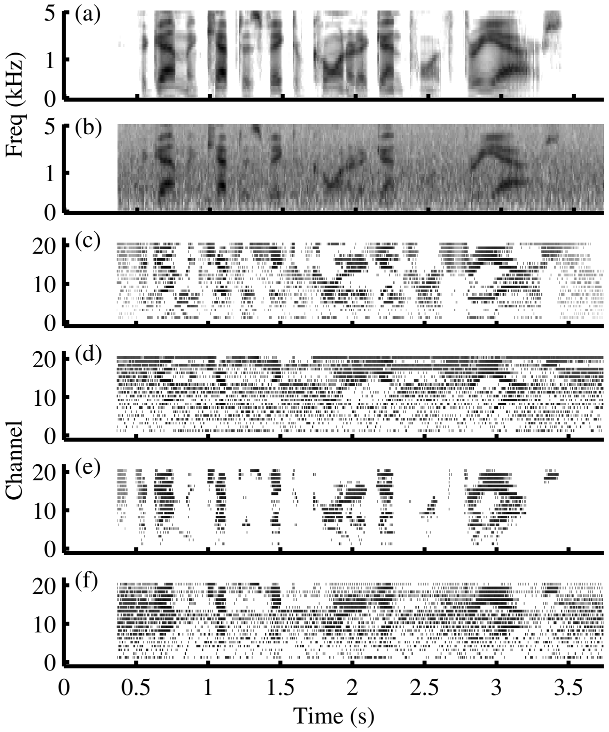

Figure 1: (a) Spectrogram of a TIMIT sentence in quiet (*the gunman kept his victim cornered at gunpoint for three hours*); (b) spectrogram in SSN (0dB SNR); (c) electrodogram in quiet using the MP strategy; (d) electrodogram in SSN using the MP strategy; (e) electrodogram in quiet using the MPm strategy; and (f) electrodogram in SSN using the MPm strategy. The color map is on a logarithmic scale with black indicating large magnitudes, white indicating small magnitudes, and a range of 50dB.

3. MP STRATEGY MODIFICATIONS

For an example sentence, we plot the gammatonegrams in quiet and in speech-shaped noise (SSN) at 0dB SNR in Fig. 1a-b. In Fig. 1c-d, we show the corresponding electrodograms (a visualization of the stimulation pattern over time) using the MP strategy. In quiet, the electrodogram tracks the spectral peaks well, and there are clear boundaries between speech segments and gaps across all channels. In SSN however, channel selection tends to favor the highest channels.

In Fig. 2a, we plot the set of weights $\{ \sqrt{M} \Big/ \sigma_m(i) \}$ as they evolve for the noisy sentence in Fig. 1b. At the onset of a large spectral peak, the standard deviations of associated channels increase significantly, and they remain high for the 400ms that the peak is within B_m. Since the weights are inversely proportional to the standard deviations, these channels are de-emphasized or inhibited. Furthermore, since MP is consequently less likely to choose dictionary elements that encode the inhibited channels, the electrodograms tend to exhibit periods of low stimulation after large spectral peaks. Although Taal et al. [8] designed this weighting scheme because of its correspondence with STOI and not because they specifically wanted to encode spectral peaks and gaps accurately, the weighting scheme tends to produce clear ON/OFF modulations at the boundaries between strong speech segments and the subsequent gaps; even in noise, almost completely white regions follow the dark black regions in the electrodogram in Fig. 1d. Qazi et al. [2] looked in detail at the effect of noise on stimulation strategies and intelligibility and concluded that preserving ON/OFF modulations is the most important factor for preserving intelligibility in noise.

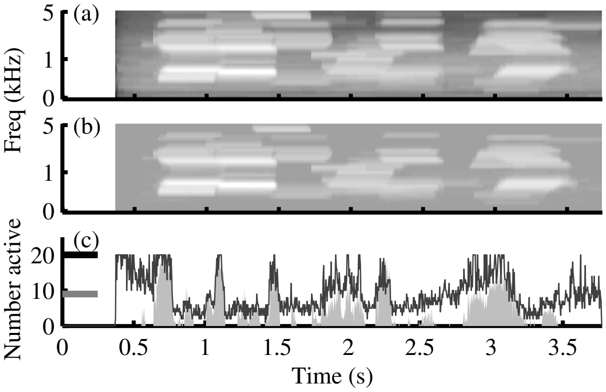

Figure 2: (a) MP weights for the noisy sentence shown in Fig. 1b (color map is on a logarithmic scale with black indicating large magnitudes, white indicating small magnitudes, and a range of about 80dB); (b) MPm weights; and (c) the number of active channels during each stimulation cycle (CIS: black dash, MP: medium gray dash, IBM: light gray shaded area, MPm: dark gray thin line).

Although we have identified some of MP's very promising characteristics, we can also identify two weaknesses. First, MP favors higher channels, and second, MP selects n channels even during speech gaps in quiet. Borrowing inspiration from IBM, MP would likely benefit from being able to recruit more or less channels when appropriate. To address these issues, we propose two modifications: (1) clip large weights so that the weighting scheme only produces inhibition (i.e., eliminate channel emphasis), and (2) let MP recruit as many channels as it wants as long as the contribution to reducing the residual is large enough.

In Fig. 2b, we plot the set of modified weights we obtain when we limit them to 0.004 (heuristically chosen). This modification substantially increased the robustness of the weights to noise, and as a result, reduced over-activity in the highest stimulation channels (Fig. 1f compared to 1d). In Fig. 2c, we show a comparison of the number of active channels during each stimulation cycle. In CIS, the number of active channels is fixed to 20, and in MP, the number of active channels is fixed to 9. In IBM, the number of active channels fluctuates with the number of channels that are target-dominated. When we allow MPm to select all channels within a cycle that reduce the residual by at least 0.01% relative to the previous iteration (again, heuristically chosen), MPm tends to recruit additional channels during the same segments as IBM despite not having oracle knowledge.

4. RESULTS

We use NCM to predict intelligibility outcomes for MP and MPm in quiet and in noise as compared to CIS and IBM. NCM is an intelligibility measure based on the covariance between the envelopes of the test signal and its reference (the un-vocoded sentence in quiet) in each frequency band, and it has been shown to correlate highly with vocoded speech [13, 9]. With 100 randomly selected TIMIT sentences, we create mixtures at 0dB with SSN and 5dB with AURORA babble noise, encode the quiet and noisy versions using CIS, IBM, MP, and MPm, and then synthesize vocoded stimuli. We plot mean NCM scores in Fig. 3 for the vocoded sentences and perform

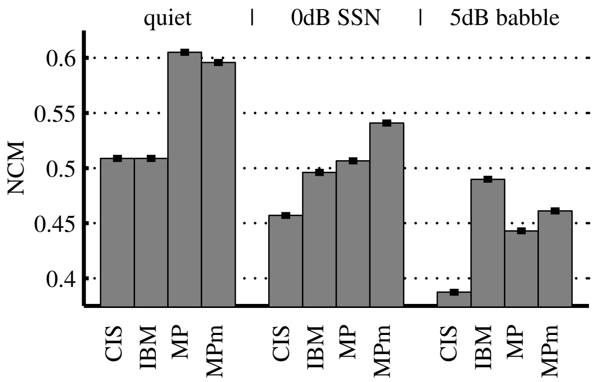

Figure 3: Mean NCM scores of vocoded signals as a function of the coding strategy for TIMIT sentences in quiet, SSN (0dB SNR), and babble (5dB SNR). Standard error bars drawn in black.

a three-way ANOVA (factors are sentence, interferer type, and coding strategy). We find significant main effects as well as two-factor interactions ($p < 0.01$ for all). Furthermore, we perform a multiple comparisons test with a Bonferroni adjustment (to control the familywise error rate) over the interferer and coding strategy factors (i.e., population marginal means are computed for each combination of these two factors while removing the effects of the other factor). We find significant differences for all pairs except CIS in quiet compared to IBM in quiet, both CIS and IBM in quiet compared to MP in SSN, CIS in SSN compared to MPm in babble, and IBM in SSN compared to IBM in babble.

NCM predicts that IBM almost fully restores intelligibility in both steady-state and fluctuating noise to the level predicted in quiet, which aligns well with the results obtained in CI recipients [5]. In quiet, NCM predicts that MP will significantly improve intelligibility (which supports previous results [8]), and although the mean score for MPm is slightly lower, it also yields high scores. In SSN, NCM predicts that MP will fully restore intelligibility, and with MPm, it will actually increase intelligibility as compared to that of CIS in quiet. For both MP and MPm, the means are greater than the mean for IBM—a surprising result given that IBM utilizes oracle knowledge. In babble, NCM predicts that MP will significantly improve intelligibility as compared to CIS in babble, but that it only partially restores intelligibility as compared to CIS in quiet. Again, the modifications to the MP strategy yield significantly higher NCM scores. Unlike in the case of steady-state noise, MP and MPm do not outperform IBM in fluctuating noise.

5. DISCUSSION

We have demonstrated that (1) in quiet, MP and MPm obtain significantly larger NCM scores than both CIS and IBM; (2) in steady-state and fluctuating noise, MP and MPm obtain significantly larger NCM scores than CIS; (3) in steady-state noise specifically, MP and MPm obtain significantly larger NCM scores than IBM despite not having oracle knowledge; and (4) in steady-state and fluctuating noise, MPm obtains significantly larger NCM scores than MP. Although tone-vocoded speech is an imperfect model of CI recipient outcomes, vocoded speech intelligibility at least demonstrates how much of the spectral information is represented in a coding strategy's output. However, since we are using NCM and not listener studies, we will need to conduct further testing to be sure of the speech understanding the MP and MPm strategies can provide.

6. ACKNOWLEDGMENT

A.A.K. thanks Cees H. Taal for his work on the MP strategy.

7. REFERENCES

[1] M. F. Dorman, P. C. Loizou, J. Fitzke, and Z. Tu, "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," *The Journal of the Acoustical Society of America*, vol. 104, pp. 3583–3585, 1998.

[2] O. u. R. Qazi, B. van Dijk, M. Moonen, and J. Wouters, "Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility," *Hearing Research*, 2013.

[3] B. S. Wilson and M. F. Dorman, "Cochlear implants: a remarkable past and a brilliant future," *Hearing Research*, vol. 242, no. 1, pp. 3–21, 2008.

[4] M. W. Skinner, L. K. Holden, L. A. Whitford, K. L. Plant, C. Psarros, and T. A. Holden, "Speech recognition with the nucleus 24 SPEAK, ACE, and CIS speech coding strategies in newly implanted adults," *Ear and Hearing*, vol. 23, no. 3, pp. 207–223, 2002.

[5] Y. Hu and P. C. Loizou, "A new sound coding strategy for suppressing noise in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 124, pp. 498–509, 2008.

[6] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 23, pp. 978–982, 2006.

[7] S. Shapero, A. S. Charles, C. J. Rozell, and P. Hasler, "Low power sparse approximation on reconfigurable analog hardware," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 2, no. 3, pp. 530–541, 2012.

[8] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in cochlear implants based on an intelligibility metric," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 504–508.

[9] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded speech," *Ear and hearing*, vol. 32, no. 3, pp. 331–338, 2011.

[10] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125–2136, 2011.

[12] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 1292–1304, 2005.

[13] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.