STATISTICAL NOISE COMPENSATION FOR COCHLEAR IMPLANT PROCESSING

Hui Jiang[†], *Qian-Jie Fu*[‡]

† Department of Computer Science and Engineering, York University,
4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA
‡ House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057, USA
Email: *hj@cs.yorku.ca* and *qfu@hei.org*

ABSTRACT

A statistical noise compensation algorithm is proposed for cochlear implant processing to improve cochlear implant patients' speech performance in noise. With the well-known environmental model for speech in additive noise, the MMSE (minimum mean square error) estimation of clean speech signals was derived according to the noisy speech observation based on a linear approximation of the original nonlinear environmental model. Words-in-sentences recognition by four cochlear implant subjects was tested under different noisy listening conditions (steady white noise and 6-talker speech babble at +15, +10, +5, and 0 dB SNR) with and without the noise compensation algorithm. For steady white noise, a mean improvement of 36% correct of sentence recognition scores was obtained across the SNR levels when the noise compensation algorithm was applied to cochlear implant processing. However, the amount of improvement was highly dependent on the SNR levels with the speech babble noise. The improvement was gradually increased from 7% to 32% correct when the SNR levels increased from 0 dB to 15 dB. The results suggest that cochlear implant patients may significantly benefit from the proposed noise compensation algorithm in noisy listening.

1. INTRODUCTION

Cochlear implants (CIs) have provided profoundly deaf individuals with hearing sensation. Many post-lingually deafened patients, fitted with the latest multi-channel speech processors, perform quite well in quiet listening situations. However, speech performance deteriorates rapidly with increased levels of background noise, even for the best CI users. Previous studies have shown that, for CI users, the absence of fine spectro-temporal cues may contribute to poorer performance in noise, especially when the noise is dynamic. Fu et al. in [3] also found that CI users' susceptibility to noise may also be partially caused by the high degree of spectral smearing associated with channel interaction. Improving the effective number of spectral channels and/or reducing channel interactions are an intuitive approach to improve CI patients' performance in noise. Unfortunately, the limited number of spectral channels and the channel interactions may be partially caused by the nerve survival in some CI patients, which may not be easily addressed with the settings in the speech processor. For these CI patients, alternative approaches should be considered. One possible approach is to apply the noise cancellation algorithm to noisy speech before transmitting to the speech processor.

Some speech enhancement algorithms originally developed for normal hearing (NH) persons have been applied to CI speech processing to reduce the effects of background noise. These algorithms were able to somewhat improve CI users' performance in noisy listening conditions. Recently, Yang and Fu [10] also evaluated the effect of single-channel noise cancellation algorithms utilizing speech pause detection and nonlinear spectral subtraction on sentence recognition score in seven CI patients. The spectral subtraction algorithm estimates the short-time spectral magnitude of speech by subtracting the estimated noise spectral magnitude from the noisy speech spectral magnitude. The results suggest that the speech enhancement algorithm may be beneficial for CI users in noisy listening. Unfortunately, with single-channel spectral subtraction algorithm, the benefit is mostly limited to the situation when the interfering noise was relatively steady. The benefit is only marginal or not at all in the presence of dynamically fluctuated noise, e.g. speech babble noise or interfering speaker. To achieve better performance in the presence of dynamically fluctuated noise, more efficient algorithms should be developed for cochlear implant processing.

Motivated by the successes in automatic speech recognition [6, 7], we apply statistical noise compensation methods to remove noise for cochlear implant processing. With the well-known environmental model for speech in additive noise, the MMSE (minimum mean square error) estimation of clean speech signals was derived given the noisy speech observation based on a linear approximation of the original nonlinear environmental model. Words-in-sentences recognition by four cochlear implant subjects was tested under different noisy listening conditions (steady white noise and 6-talker speech babble at +15, +10, +5, and 0 dB SNR) with and without the noise compensation algorithm. Sentence recognition experiments show that the proposed noise compensation algorithm significantly improve recognition performance of patients in noisy conditions.

2. STATISTICAL NOISE COMPENSATION ALGORITHM

In the following, we first introduce the well-known environmental model for speech in additive noise. Then we derive the MMSE (minimum mean square error) estimation of clean speech signals given the noisy speech observation based on a linear approximation of the original nonlinear environmental model.

2.1. Environmental Model for Speech in Additive Noise

Assume we have clean speech $x(t)$ in the time domain and $x(t)$ is corrupted by an independent ambient noise $n(t)$ (also in the time domain). The resultant noisy speech can be expressed in the time domain as:

$$y(t) = x(t) + n(t) \qquad (1)$$

Usually we can assume $x(t)$ and $n(t)$ are statistically independent.

If we convert the signals into the log-spectrum domain (either linear or Mel-scale or other scales), the above simple relation becomes a complex nonlinear function (see [1]). For d-th filter bank (or d-th frequency bin), we have

$$\mathbf{y}_d = \mathbf{x}_d + \ln\left(1 + e^{\mathbf{n}_d - \mathbf{x}_d}\right) \qquad (2)$$

If we assume the independence between all different filter bands, then we can drop the subscript d for clarity. We just repeat the same operation for all different filter bands (or feature dimensions).

Apparently, the above environmental model is highly nonlinear. In practice, for the sake of simplicity, the model is usually approximated by a linear function based on a low-order vector Taylor series (VTS) expansion. Here, we can adopt the zero-th order VTS expansion to approximate the log term in environmental model as in eq.(2). In this case, the function can be expanded around any a point $(\mathbf{x}_0, \mathbf{n}_0)$ as follows:

$$\begin{aligned}\mathbf{y} &= \mathbf{x} + \ln\left(1 + e^{\mathbf{n}-\mathbf{x}}\right)\\ &\approx \mathbf{x} + \ln\left(1 + e^{\mathbf{n}_0-\mathbf{x}_0}\right)\end{aligned} \qquad (3)$$

2.2. Statistical Models for Clean Speech and Noise

Based on the above environmental model, for any given noisy speech feature vector $\mathbf{y}$, we will try to estimate a clean speech $\hat{\mathbf{x}}$ in the MMSE (minimum mean square error) sense.

First of all, we assume the clean speech feature vector $\mathbf{x} = \{x_1, x_2, \cdots, x_D\}$ in *the log-spectral domain* follows a multivariate Gaussian mixture model (GMM) as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_{xk}, \sigma^2_{xk}) = \sum_{k=1}^{K} w_k \cdot \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2_{xkd}}} \cdot e^{-\frac{(x_d-\mu_{xkd})^2}{2\sigma^2_{xkd}}} \qquad (4)$$

where $\mu_{xk} = \{\mu_{xk1}, \mu_{xk2}, \cdots, \mu_{xkD}\}$ and $\sigma_{xk} = \{\sigma_{xk1}, \sigma_{xk2}, \cdots, \sigma_{xkD}\}$ are mean and variance vectors of k-th Gaussian mixture, and w_k is the weight of k-th mix, with the constraint $\sum_{k=1}^{K} w_k = 1$.

Secondly, we also assume noise signals in *the log-spectral domain* follows a single Gaussian distribution as:

$$p(\mathbf{n}) = \mathcal{N}(\mathbf{n} \mid \mu_n, \sigma^2_n) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2_{nd}}} \cdot e^{-\frac{(n_d-\mu_{nd})^2}{2\sigma^2_{nd}}} \qquad (5)$$

where $\mu_n = \{\mu_{n1}, \mu_{n2}, \cdots, \mu_{nD}\}$ and $\sigma_n = \{\sigma_{n1}, \sigma_{n2}, \cdots, \sigma_{nD}\}$ are mean and variance vectors of noise signals.

The clean speech model is estimated prior to noise compensation and it is kept unchanged during the entire noise compensation procedure. We first collect a small set of clean speech data and extract feature vectors in the log-spectral domain. Then all these feature vectors are used to estimate the Gaussian mean vectors and variance vectors based on the standard EM (Expectation-Maximization) algorithm.

The noise model is estimated separately for each utterance. Firstly, the first N frames (We typically use $N = 10$.) of each utterance are used to initialize the noise mean μ_n and noise variance σ_n because the beginning part of each utterance usually contains only noise signals. Given the clean speech model $p(\mathbf{x})$, we can refine the above noise model (mainly noise mean) according to the EM algorithm based on the whole utterance as follows.

For simplicity, we use the zero-order VTS linear approximation for the environmental model as in eq.(3). In this case, the p.d.f. of noisy speech $\mathbf{y}$ is also a GMM model as:

$$p(\mathbf{y}) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{y} \mid \mu_{yk}, \sigma^2_{yk}) \qquad (6)$$

with

$$\mu_{yk} = \mu_{xk} + C_k \qquad (7)$$

$$\sigma^2_{yk} = \sigma^2_{xk} \qquad (8)$$

where $C_k = \ln(1 + e^{\mu_n - \mu_{xk}})$ denotes coefficient when expanding zero order vector Taylor series around the means of Gaussian components (μ_{xk}, μ_n).

Given the whole noisy speech utterance $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T\}$, the iterative maximum likelihood (ML) estimation for μ_n can be derived based on the EM algorithm as:

$$\mu_n^{(i+1)} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} \Pr(k|\mathbf{y}_t)(\mathbf{y}_t - \mu_{xk} - C_k)}{\sum_{t=1}^{T}\sum_{k=1}^{K} \Pr(k|\mathbf{y}_t)} \qquad (9)$$

where

$$\Pr(k|\mathbf{y}_t) = \frac{w_k \cdot \mathcal{N}(\mathbf{y}_t \mid \mu_{yk}, \sigma^2_{yk})}{\sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{y}_t \mid \mu_{yk}, \sigma^2_{yk})} \qquad (10)$$

In which μ_{yk} is derived from eq. (7) based on the current noise model parameter $\mu_n^{(i)}$.

The above iterative estimation in eq.(9) continues until some convergency conditions are met.

2.3. MMSE Estimation of Clean Speech Signal

Given the pdf's of clean speech $\mathbf{x}$ and noise $\mathbf{n}$ in eqs.(4) and (5), as well as the environmental model for noisy speech $\mathbf{y}$ as described in section 2.1, we are interested in deriving an MMSE estimation, $\hat{\mathbf{x}}$, of clean speech signal based on any noisy speech observation, $\mathbf{y}_0$. To derive a closed-form solution, we adopt the VTS-based linear approximation of the environmental model in (3) in this work. The detailed derivation of the MMSE estimation of clean speech based on the original nonlinear environmental model can be found in [6].

Given a noisy speech vector, $\mathbf{y}_0$, it is well known that the MMSE estimation of clean speech, $\hat{\mathbf{x}}$, is calculated as $\hat{\mathbf{x}} = \mathrm{E}_{\mathbf{x}}[\,\mathbf{x} \mid \mathbf{y}_0\,]$.

Given the clean speech model as in eq.(4) and noise model in eq.(5), we adopt a piece-wise linear approximation approach to expand the environmental model eq.(2) separately around its mean vectors, $\mu_{xk},\ k = \{1, 2, \cdots, K\}$ and the noise mean μ_n as follows:

$$\mathbf{x} = \mathbf{y} - \ln\left(1 + e^{\mu_n - \mu_{xk}}\right) \quad \text{for } k = 1, 2, \cdots, K \qquad (11)$$

In this case, given a noisy speech observation $\mathbf{y}_0$, the MMSE estimation of the clean speech is derived as:

$$\hat{\mathbf{x}} = \mathrm{E}_{\mathbf{x}}[\,\mathbf{x} \mid \mathbf{y}_0\,] = \mathbf{y}_0 - \sum_{k=1}^{K} \Pr(k \mid \mathbf{y}_0) \cdot \ln\left(1 + e^{\mu_n - \mu_{xk}}\right) \qquad (12)$$

where $\Pr(k \mid \mathbf{y}_0)$ denotes the posterior probability of its clean speech vector belonging to k-th Gaussian component given its noisy speech observation $\mathbf{y}_0$. For k-th Gaussian component, if clean speech $\mathbf{x}$

follows a Gaussian distribution, $\mathcal{N}(\mathbf{x} \mid \mu_{xk}, \sigma^2_{xk})$, and noisy speech $\mathbf{y}$ is a linear function of $\mathbf{x}$ as in eq.(11), then $\mathbf{y}$ is also a Gaussian distribution as $\mathcal{N}(\mathbf{y} \mid \mu_{xk} + \ln\left(1 + e^{\mu_n - \mu_{xk}}\right), \sigma^2_{xk})$. Therefore,

$$\Pr(k \mid \mathbf{y}_0) = \frac{w_k \cdot \mathcal{N}(\mathbf{y}_0 \mid \mu_{xk} + \ln\left(1 + e^{\mu_n - \mu_{xk}}\right), \sigma^2_{xk})}{\sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{y}_0 \mid \mu_{xk} + \ln\left(1 + e^{\mu_n - \mu_{xk}}\right), \sigma^2_{xk})} \quad (13)$$

2.4. Overview of the Whole Noise Compensation Process

First of all, we train a GMM model for clean speech in the log-cepstral domain, i.e. $p(\mathbf{x})$, based on clean speech data in training set. In this work, the clean speech model is trained in the MFCC domain and then transformed into the log-spectral domain by applying the inverse DCT. Then model parameters for $p(\mathbf{x})$ will be fixed during noise compensation procedure.

For each test noisy speech utterance, we compute the feature vectors, denoted as $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T\}$, in the log-spectral domain. Then we do

1. Initialize the mean μ_n and variance σ_n of the noise distribution $p(\mathbf{x})$ using the first N frames of the utterance. We typically use $N = 10$.[1]
2. Given clean speech model $p(\mathbf{x})$, refine the noise mean μ_n according to the EM algorithm based on the whole utterance $\mathbf{Y}$.
3. Based on the refined noise model $p(\mathbf{n})$ and clean model $p(\mathbf{x})$, we compensate $\mathbf{Y}$ a frame by a frame. More specifically, for each vector $\{\mathbf{y}_t \mid 1 \leq t \leq T\}$, we use equation (12) to obtain its MMSE estimation.
4. The compensated feature vectors in the log-spectral domain are converted into the appropriate format and will be sent to the cochlear implant processing device.

3. EXPERIMENTS

3.1. Subjects

Four post-lingually deafened adults using the Nucleus cochlear implant device participated in this study. All were native speakers of American English and had at least nine years experience with the device. All implant subjects had 20 active electrodes available for use and had extensive experience in speech recognition experiments.

3.2. Signal Processing

Custom 6-channel speech processors with Continuous Interleaved Sampling (CIS) strategy [9] were implemented in the present study. The detailed implementation was showed as follows. A speech signal was first windowed with a Hamming window and the magnitude of the short-time Fourier transformed (STFT) data was calculated. Next, six band-pass filters (filter bank) was applied. The filter bank was designed based on the Greenwood formula in [4] and the corner frequencies of these filters were 200 Hz, 427 Hz, 803 Hz, 1426 Hz, 2458 Hz, 4167 Hz, and 7000 Hz. For the speech processor without noise compensation, the output (acoustical amplitude) of the filter bank (60-dB range) was directly transformed into electric amplitude by a power-law function with an exponent of 0.2 ($E = A^{0.2}$) between each subject's threshold (T-level) and upper level of loudness (C-level). This transformed amplitude was then used to modulate the amplitude of a continuous biphasic 250-pulse-per-second pulse train with a 200 μs/phase pulse duration, and delivered to six electrode pairs interleaved in time: (20,22), (16,18), (12,14), (8,10), (5, 7) and (2,4) via a custom research interface [8]. For the speech processor with noise compensation algorithm, a logarithm operation was applied to the outputs of the filter bank to generate the log-spectral feature vectors. If necessary, the discrete Cosine transform (DCT) was used to convert the log-spectral feature into the MFCC features for noise compensation algorithm. After noise compensation, the compensated feature vectors in the log-spectral domain were converted into the appropriate format (acoustical amplitude), which were sent to the cochlear implant processing device as above. The entire speech signal processing procedure is shown in Fig. 1.

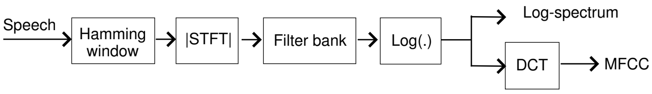

Fig. 1. Basic steps in the speech feature extraction.

3.3. Test Materials and Procedures

Recognition of words in sentences was measured using sentences from the IEEE sentence corpus [5]. The sentences were digitized recordings spoken by 1 male talker (recorded at House Ear Institute). The IEEE sentences were of easy to moderate difficulty. The IEEE sentences (totally 720 utterances) were splitted into two parts. The first part (including 320 utterances) was used to train the clean speech GMM models. The GMM speech models were built with 32 mixtures. The remaining 400 utterances were used to evaluate sentence recognition performance of CI patients under different noise conditions. In the experiments, two kinds of noise were added into clean speech in the time domain at various SNR levels. First, the computer-generated white Gaussian noise was added into the 400 utterances at the following four SNR levels, i.e., 15dB, 10dB, 5dB and 0 dB, to generate the first set of noisy speech. Next, speech babble noise was added at the SNR levels of 15dB, 10dB, 5dB and 0 dB to generate the second set of noisy speech.

During behavioral testing, a list was chosen pseudo-randomly among 40 lists, and sentences were chosen randomly, without replacement, from the 10 sentences within that list. Subjects responded by repeating the sentence as accurately as possible; the experimenter tabulated correctly identified words and sentences. Subjects were presented with at least 2 sets per condition. The percent correct of words-in-sentences was measured as a function of the SNR levels for each subject with and without the noise compensation algorithm. The SNR levels and noise type were randomized and counterbalanced across subjects.

3.4. RESULTS AND DISCUSSION

Fig. 2 shows words-in-sentences recognition as a function of the SNR levels with and without noise compensation algorithm for the

[1] We assume the first 10 frames, i.e. 100 msec in usual frame rate, of each utterance are non-speech segment, which is reasonable in most situations.

individual CI patients. Individual data are shown with (open symbols) and without (filled symbols) the noise compensation algorithm for steady white noise (circles) and speech babble (triangles). The recognition performance of words in sentences generally increased when the SNR levels increased. Fig. 3 shows the amount of improvement in words-in-sentences recognition as a function of the SNR levels. Individual and mean CI data are shown for steady noise (filled symbols/solid line) and speech babble (open symbols/dashed line). For the steady white noise, the improvement was about 36% regardless of the SNR levels. For the speech babble noise, the improvement was only moderate (about 7%) in 0 dB SNR. The improvement gradually increased when the SNR level increased. About 31% percentage correct in improvement was observed when the SNR reached 15 dB.

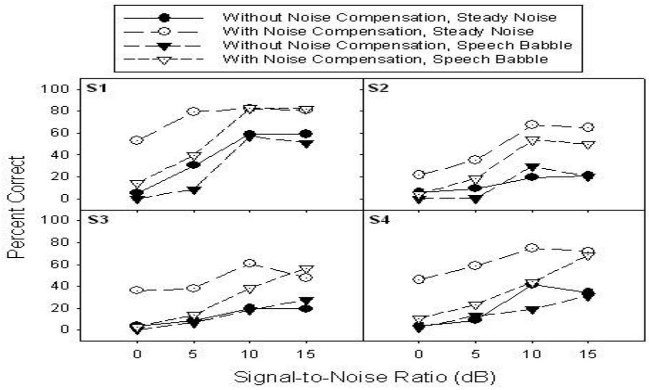

Fig. 2. Words-in-sentences recognition as a function of the SNR levels with and without noise compensation algorithm for the individual CI patients.

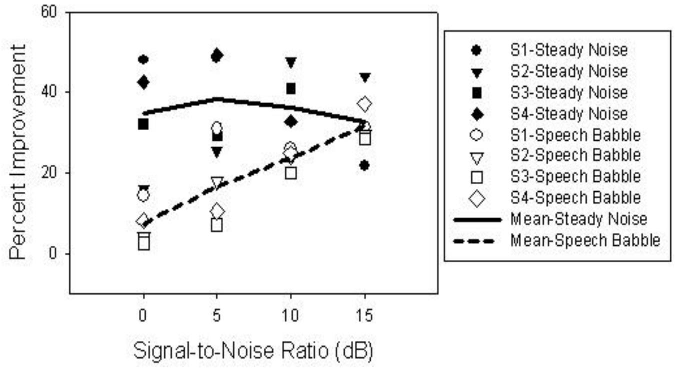

Fig. 3. The amount of improvement in words-in-sentences recognition as a function of the SNR levels.

The data from the present study demonstrate that statistical noise compensation algorithm is an efficient approach to improve CI patients' speech performance in the presence of noise. Similar to the previous single-channel spectral subtraction approach, the algorithm performed differently for the different kind of noise. For steady white noise, the improvement was highly significant at all SNR levels in the present study (from 0 dB to 15 dB). The amount of improvement was also comparable across these SNR levels. However, the amount of improvement was highly dependent on the SNR levels in the presence of speech babble noise. When the SNR level was relatively high (e.g. 15 dB), the amount of improvement was comparable to that observed in the steady white noise. The improvement was gradually reduced when the SNR level gradually decreased. When the SNR level reached to 5 dB or lower, no significant improvement was observed with the current noise compensation.

4. CONCLUSIONS

In general, the noise compensation algorithm proposed in the present study provided significantly better benefit than the previous single-channel spectral subtraction algorithm in [10], especially in the presence of speech babble noise. Taken together, these results suggest that the statistical noise compensation proposed in the present study may be a useful alternative approach to improve CI patients' speech performance in noisy environment.

ACKNOWLEDGEMENT

We appreciate the efforts of all subjects. We also would like to thank John J. Galvin III for assistance in editing the manuscript. The research was partially supported by grants R01-DC004993 from NIDCD and Minor Research Grant from York University.

5. REFERENCES

[1] A. Acero, *Acoustic and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic, 1993.

[2] Q.-J. Fu and R. V. Shannon, "Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners," *Journal of the Acoustical Society of America*, Vol.104, No. 5, pp.2570-2577, 1998.

[3] Q.-J. Fu and G. Nogaki, "Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing," *JARO in press*, 2005.

[4] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *Journal of the Acoustical Society of America* Vol. 87, pp.2592-2605, 1990.

[5] IEEE, "IEEE recommended practice for speech quality measurements," *Institute of Electrical and Electronic Engineers*, New York, 1969.

[6] H. Jiang, Q. Wang, "Nonlinear Noise Compensation in Feature Domain for Speech Recognition with Numerical Methods," *Proc. of 2004 International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2004)*, Montreal, Canada, May 2004.

[7] P.J. Moreno, B. Raj and R.M. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. of ICASSP'96*, pp.733-736, Atlanta, GA, May 1996.

[8] R.V. Shannon, D. D. Adams, R. L. Ferrel, R. L. Palumbo and M. Grantgenett,"A computer interface for psychophysical and speech research with the Nucleus cochlear implant," *Journal of the Acoustical Society of America*, Vol. 87, pp.905-907, 1990.

[9] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington and W. M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature*, Vol. 352, 236-238, 1991.

[10] L.-P. Yang and Q.-J. Fu, "Spectral subtraction based speech enhancement for cochlear implant patients in background noise," *Journal of the Acoustical Society of America*, 117(3), 1001-1005, 2005.