SPEECH ENHANCEMENT BASED ON A COMBINED MULTI-CHANNEL ARRAY WITH CONSTRAINED INTERATIVE AND AUDITORY MASKED PROCESSING ⊥

Xianxian Zhang[1], *John H. L. Hansen*[1,2], *and Kathryn Arehart*[2]

[1]Robust Speech Processing Group, Center for Spoken Language Research
[2]Department of Speech, Language and Hearing Sciences
University of Colorado at Boulder, USA
{zhang, jhlh}@cslr.colorado.edu http://cslr.colorado.edu

ABSTRACT

While a number of studies have investigated various speech enhancement and noise suppression schemes, most consider either a single channel or array processing framework. Clearly there are potential advantages in leveraging the strengths of array processing solutions in suppressing noise from a direction other than the speaker, with that seen in single channel methods that include speech spectral constraints or psychoacoustically motivated processing. In this paper, we propose to integrate a combined fixed/adaptive beamforming algorithm (CFA-BF) for speech enhancement with two single channel methods based on speech spectral constrained iterative processing (Auto-LSP), and an auditory masked threshold based method using equivalent rectangular bandwidth filtering (GMMSE-AMT-ERB). After formulating the method, we evaluate performance on a subset of the TIMIT corpus with four real noise sources. We demonstrate a consistent level of noise suppression and voice communication quality improvement using the proposed method as reflected by an overall average 26dB increase in SegSNR from the original degraded audio corpus.

1. INTRODUCTION

There are numerous areas where it is necessary to enhance the quality of speech degraded by background noise. Some example environments include: in-vehicle hands-free voice communications, mobile phone use in public noisy environments, hearing impaired persons in large classrooms or meeting halls, and others. A number of speech enhancement algorithms have been proposed in the past, and a survey can be found in [1, 2].

One way to discuss trade-offs in speech enhancement algorithms is to separate those that are single-channel, dual channel, or multi-channel array based approaches. For single-channel applications, only a single microphone is available. Characterization of noise statistics must be performed during periods of silence between utterances, requiring (i) a stationary or short-time varying assumption of the background noise, and (ii) that the speech and noise are uncorrelated. In one study, Lim and Oppenheim[3] developed a sequential MAP estimation method. In a later study, Hansen and Clements [4] compared the performance of Boll's spectral subtraction method[1] with that of a traditional Wiener filtering and proposed an alternative formulation based on iterative Wiener filtering augmented with speech-specific constraints in the spectral domain (Auto-LSP). Auditory Based constraints using loudness perception, lateral neural inhibition, and critical band analysis are used in conjunction with constraints, applied to speech feature sequence in [5]. Later studies considered the introduction of an auditory masked threshold, bandwidth spreading, and evaluated thresholds for hearing impaired subjects [6]. However, stationary noise conditions do not always exist in real world environments, and noise updating during silent sections can result in distortion since voice activity detection techniques do not always work well. Dual channel methods are more successful in noise suppression when the reference microphone can track changing noise conditions assuming limited speech cross-talk from the primary microphone. In multi-channel array algorithms, the acoustic sound waves arrive at each sensor at slightly different times (one is normally a delayed version of the other). Currently, most multi-channel enhancement techniques employ a beamforming solution, which use the spatial differences between speech and noise to remove the noise. A survey on array processing methods can be found in [7]. Theoretically, multi-channel techniques offer more information about the acoustic environment, and therefore should indeed offer the prospect of improved noise suppression especially in the case of reverberant environments. This is due to the multi-path effects and severe noise conditions known to affect the performance of state-of-the-art single channel techniques. However, compared with some successful single-channel enhancement algorithms, a beamforming solution in general can only provide limited noise suppression because of the correlation between noise and speech. Therefore, some researchers have considered methods that extend successful mono noise reduction techniques to multiple channels. An example is the study by Rosca, *et. al*. [8], where a single channel psychoacoustic masking filter is extended to a multi channel speech enhancement solution.

In this paper, we propose an algorithm that combines the multi-channel beamforming algorithm and single-channel spectral constrained based iterative and auditory masked

⊥ This research was supported in part by U.S. Navy (SPAWAR Systems) Grant No. N66001-03-1-8905, and a grant from the Whitaker Foundation.

processing method, which seeks to leverage advantages from both. We first use the noisy speech data collected by multi-sensors to do front-end processing and noise classification, which offers a first stage of enhanced speech by removing high-frequency noise and providing more information concerning the acoustic environment; then we use a successful spectral based single-channel enhancement algorithm to do post processing, which can suppress the environmental noise thoroughly and improve speech quality by employing the known estimated acoustic information from the first stage.

2. PROPOSED ALGORITHM MOTIVATION

To motivate the proposed method, we consider a previous proposed combined fixed/adaptive beamforming algorithm (CFA-BF) [9] for a TIMIT sentence degraded by Flat Channel Communication Noise (FLN). We use the same microphone array set up, and found that this method can improve SegSNR (Signal-to-Noise Ratio) by up to 11.75dB. Next, we also applied a recently proposed GMMSE-AMT-ERB algorithm (GAE) [6]that uses an auditory masked threshold with equal rectangular bandwidth filters, and an earlier spectral constrained iterative speech enhancement algorithm Auto-LSP [4] on the same noisy data, and found that the SegSNR improvements are 16dB and 20.5dB respectively. However, these algorithms cannot entirely suppress the FLN noise. Fig. 1 shows the spectrogram of the original degraded speech, and enhanced speech by CFA, GAE, and Auto-LSP respectively. Our original objective of choosing FLN noise was to focus on the design of an algorithm that can obtain the best performance under this stationary noise condition, and then to extend it to more complex noise environments.

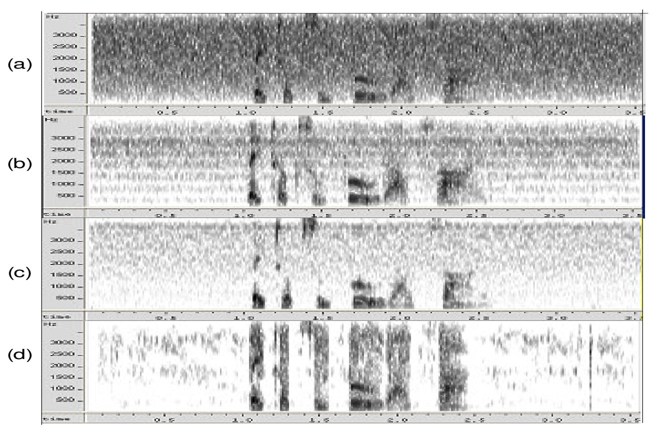

Figure 1: Spectrogram of Speech Data with: (a). Original FLN degraded noisy speech; (b). CFA Enhanced speech; (c). GMMSE-AMT-ERB Enhanced speech; (d). Auto-LSP Enhanced speech.

From the above experimental results, we see that CFA is able to suppress high frequency noise, GAE suppresses noise uniformly, and Auto-LSP suppresses noise efficiently across the entire frequency band, but there is still some residual noise in the high frequency region. These results are quite common in speech enhancement algorithms using multi-channel or single-channel configurations. The results here clearly support our idea that speech enhancement algorithms can be successful in given noise conditions, but that perhaps combinations could offer both improved as well as more consistent solutions in diverse changing noisy environments.

3. ALGORITHM DESIGN

3.1 Overall Algorithm Description

In our proposed algorithm, we first apply combined fixed/adaptive beamforming (CFA-BF) for front-end processing to obtain a first stage enhanced speech signal by suppressing high frequency noise as well as generating a corresponding residual noise. Secondly, according to the nature of the noise and the angle between the direction of speech and interference, we select a back-end processing method from 3 possible spectral based speech enhancement algorithms to suppress residual noises (i.e. enhancement scheme #1, #2 or #3). Fig. 2 summarizes an overall description of the proposed algorithm.

Let: ϕ be the angle between the speech source and the axis of the microphone array, ψ be the angle between the interference and the axis of the microphone array, θ_1 be the lower bound of the angle threshold, θ_2 be the upper bound of the angle threshold; then,

1. if $|\phi-\psi| \geq \theta_1$, then go to Step 4;
2. if $|\phi-\psi| \leq \theta_2$, then select scheme #2;
3. if $\theta_1 < |\phi-\psi| < \theta_2$, then we are between performance bounds for the methods, so we can randomly select one of the schemes to use, or employ other criteria to select the proper scheme to use;
4. if the current noise has strong low frequency content, then select scheme #2; else select scheme #1.

Here, both the angle and threshold are decided by the geometry of the microphone array, the distance from the sources to the array, and the nature of the interference.

Figure 2: Formal description of the proposed algorithm.

3.2 Detailed Algorithm Design

3.2.1. Front-end processing

Fig. 3 is the block diagram of the structure of the proposed algorithm. We know that most of adaptive beamforming algorithms will select one of the microphones as the primary microphone, and build an adaptive filter between it and each of the other microphones. These filters compensate for the different transfer functions between the speaker and the microphone array. Therefore, there are two kinds of outputs from the adaptive beamforming algorithm: namely the enhanced speech $d(n)$ and noise signal $e_i(n)$. Here, when we use the combined fixed/adaptive beamforming algorithm (CFA-BF) [9], we choose microphone 0 as the primary microphone, therefore, the enhanced speech $d(n)$ and noise signal $e_i(n)$ are given as in Eqn. (1) and (2).

$$d(n) = \frac{1}{N}\sum_{i=0}^{N-1} w_i^T(n) x_i(n) \quad (1)$$

$$e_i(n) = w_0^T(n) x_0(n) - w_i^T x_i(n) \quad (2)$$

where, N is the total number of microphones, x_i is the i^{th} microphone input signal with $i = 0, 1, ..., N-1$. Compared with the original noisy speech, the enhanced speech $d(n)$ suppresses noise mainly in the high-frequency band, and the corresponding noise outputs $e_i(n)$ are the residual noises that are synchronous with $d(n)$ in time, but asynchronous with $d(n)$ in phase.

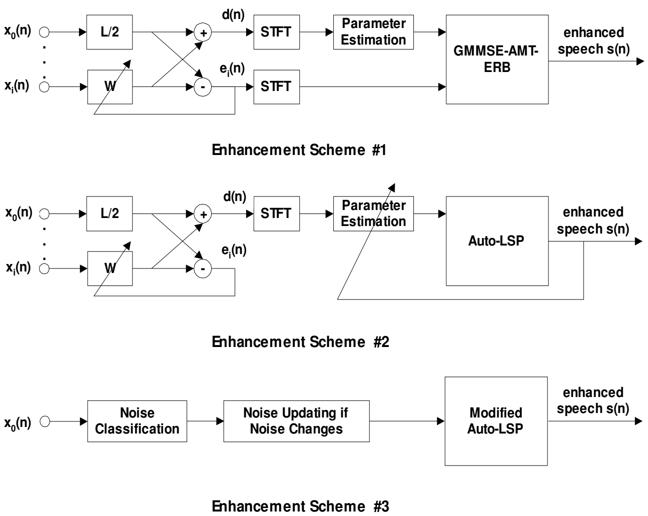

Figure 3: Block Diagram of the Proposed Algorithm

3.2.2. Back-end processing

For the back-end processing, we propose 3 possible enhancement schemes, which are classified into 2 categories:

- Category 1: includes scheme #1 and #2. Both enhancement schemes use the outputs of front-end processing as the input for back-end processing;
- Category 2: includes scheme #3 only. This scheme uses the microphone array as a tool to classify the current noise. If the current noise changes, noise updating will be performed to provide current noise estimation for back-end processing. The input of the back-end processing here will be the original input signal of the primary microphone.

In scheme #1, we adapt a modified GMMSE-AMT-ERB (mGAE), which builds on the original MMSE method[11]. The original GAE is proposed in [2] and assumes that the speech is degraded with additive noise and the speech and noise segments are uncorrelated as in Eqn (3):

$$y(n) = x(n) + n(n) \quad (3)$$

The short term power spectrum is calculated by applying a Hamming window to a frame of speech. Under this assumed model, one can obtain a family of MMSE speech spectral estimators as,

$$\hat{X}_p = (E\{X_p^{\alpha} \mid Y_p\})^{1/\alpha} \quad (4)$$

Here, let P_{nk} be the noise power spectrum for the k^{th} subband, and P_{yk} be the noisy speech power spectrum for the k^{th} subband. The values of P_{nk} and P_{yk} are calculated as follows,

$$P_{nk}[n] = \gamma P_{nk}[n-1] + \frac{1-\gamma}{1-\beta}(P_{yk}[n] - \beta P_{yk}[n-1]) \quad (5)$$

$$P_{yk}[n] = \alpha P_{yk}[n-1] + (1-\alpha)(Y_k \,|[n]|^2) \quad (6)$$

In our implementation, the first ten frames of noisy speech, which consists of only noise, is taken as the estimation of the noise for the entire noisy speech sentence. This assumption is valid if the noise does not change. However, once the noise spectrum changes, enhancement performance will decrease, resulting in either under or over noise suppression. Therefore, in the modified GAE (mGAE) algorithm, we use the residual noise $e_i(n)$ that is generated by beamform front-end processing instead of the noise spectrum estimation of GAE in scheme #1. Under the proposed model, Eqn (5) now becomes,

$$P_{nk}[n] = \sum_{i=1}^{N-1} \lambda_i P_{e_i k}[n] \quad (7)$$

$$P_{e_i k}[n] = |e_i[n]|^2 \quad (8)$$

where λ_i is a scaling factor, and we use $\lambda_i = 1/N$ for all $i = 1, ..., N-1$.

In scheme #2, we use the enhance speech $d(n)$ as an input of the Auto-LSP algorithm to remove the residue noise. This algorithm is discussed in more detail in [1] and [5].

Scheme #3 is selected only when the speech source and interference are very close to each other. Since beamforming algorithms (delay-and-sum beamforming or adaptive beamforming) obtain the enhanced signal by selecting the appropriate delays (fixed or adaptive) between each microphone and summing the delayed signals in phase for direction angle θ, we will have destructive interference for signals arriving from other angles. Fortunately, we can obtain a good noise estimate using single channel processing under this situation. Once a noise change is detected, noise spectrum updating is performed. We do not update the noise spectrum frame by frame, since we believe this will increase speech distortion. With the aid of a noise classification stage, a modified Auto-LSP algorithm (mAutoLSP) is used here as the back-end processing solution. The difference between mAuto-LSP and Auto-LSP is the presence (e.g. with/without) of the noise classification stage.

4. PERFORMANCE EVALUATION

4.1. Experimental Database & Setup

In order to evaluate the performance of the proposed algorithm, we select a corpus of 10 sentences from the TIMIT database, and degrade these sentences with the following four different noise sources: (i) White Gaussian Noise (AWG), (ii) Flat Channel Communication Noise (FLN), (iii) Large Crowd Room Noise (LCR), and (iv) Automobile Highway Noise (HWY). The sample frequency of both the sentences and noises is 8kHz. The noise level is adjusted to be an overall average 5dB SNR. For evaluations, we use the Segmental Signal-to-Noise Ratio

(SegSNR) measure [10], which represents a noise reduction criterion for voice communications.

4.2. Experiment Results

Fig. 4 shows plots of the (a) clean, (b) degraded, and enhanced speech by (c) enhancement scheme #1 & (d) #2 for the sentence, "They took some food outside". Fig. 5 illustrates average SegSNR improvement using sentences degraded with FLN noise. Table 1 show the Segmental SNR measure for the degraded speech with 4 different noises and enhanced speech by 5 different schemes.

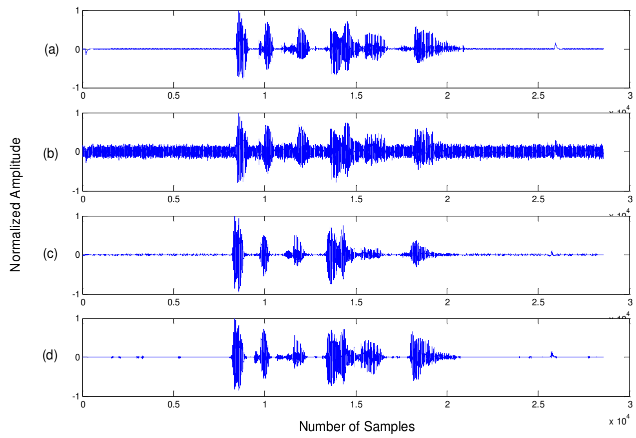

Figure 4: Time Waveforms for a single TIMIT speech file: (a) Clean speech waveform, (b) Degraded flat channel communication noise (FLN), (c) Enhanced speech waveform using scheme #1 (CSA-BF & mGMMSE-AMT-ERB), (d) Enhanced speech waveform using scheme #2 (CSA-BF & Auto-LSP)

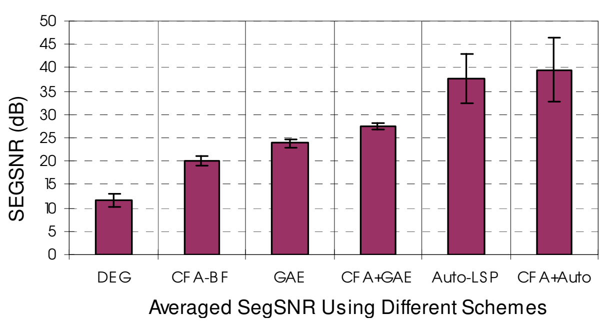

Figure 5: SegSNR Results for Degraded and Enhanced Speech

From these results, we can see that employing the proposed algorithm (array processing combined with either the psychoacoustically motivated GMMSE-AMT-ERB or speech based spectral constrained Auto-LSP), increases SegSNR significantly compared with any one individually. The SegSNR improvement is up to 26dB over the original degraded corpus set. Finally, an informal listener test evaluation confirmed the level of noise suppression and quality improvement for the proposed method.

NOISE	DEG	CFA-BF	GAE	CFA-BF + GAE	Auto-LSP	CFA + Auto-LSP
FLN (5dB)	11.55	20.1	23.775	27.575	37.55	39.525
LCR (5dB)	13.775	21.35	23.875	29.825	27.125	37.525
HWY (5dB)	12.1	13.35	18.975	16.225	36.925	39.4
AWN (5dB)	8.15	14.175	18.275	19.975	32.525	32.5
Avg. across noises	11.39	17.24	21.23	23.4	33.53	37.24

Table 1: Averaged Segmental SNR (dB) for Different Schemes

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a combined multi-channel array processing scheme based on CFA with a spectral constrained iterative Auto-LSP and auditory masked GMMSE-AMT-ERB processing for speech enhancement. The combined scheme takes advantage of the strengths offered by array processing methods in noisy environments, as well as speed and efficiency for single channel methods. We evaluated the enhancement methods on a section of the TIMIT corpus using four different actual noise conditions. We demonstrated a consistent level of noise suppression and voice communication quality improvement using the proposed method as reflected by an overall average 26dB increase in SegSNR from the original degraded audio corpus. In the future, we plan to study algorithm sensitivity to more time varying noise sources as well as reverberant environments.

REFERENCES

[1] J. R., Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, Ch. 8, Speech Enhancement, (2nd Edition), IEEE Press, New York, NY, 2000.

[2] Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proceeding of the IEEE*, **80**(10):1526-1555, 1992.

[3] J.S., Lim and A.V., Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech,Sig. Proc.*, vol. 26, June 1978.

[4] J.H.L. Hansen, M. Clements, "Constrained iterative speech enhancement with application to speech recognition", *IEEE Trans. On Signal Processing*, vol. 39, no. 4, April, 1991.

[5] J.H.L. Hansen, S. Nandkumar, "Robust Estimation of Speech in Noisy Backgrounds Based on Aspects of the Auditory Process," *Journal of the Acoustical Society of America*, vol. 97, no. 6, June 1995.

[6] A. Natarajan, J.H.L. Hansen, K. Arehart, and J. Rossi-Katz, "Perceptual Based Speech Enhancement for Normal-Hearing & Hearing-Impaired Individuals", *Interspeech/Eurospeech-'2003*, pp.1425-1428 Geneva, Switzerland.

[7] M. Brandstein and D. Ward (Eds.), *Microphone Arrays*, Springer-Verlag, New York, NY, 2001.

[8] J. Rosca, R. Balan, C. Beaugeant, "Multi-channel Psychoacoustically Motivated Speech Enhancement", *ICASSP'2003*, HongKong, China.

[9] X.X. Zhang and J. H. L. Hansen, "CFA-BF: A novel combines Fixed/Adaptive Beamforming for Robust Speech Recognition in Real Car Environments", *Interspeech/Eurospeech-2003*, pp.1289-1292, Geneva, Switzerland.

[10] http://www.nist/gov/

[11] Y. Ephriam and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimatior", *IEEE. Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, 1984.