SPEAKER SOURCE LOCALIZATION USING AUDIO-VISUAL DATA AND ARRAY PROCESSING BASED SPEECH ENHANCEMENT FOR IN-VEHICLE ENVIRONMENTS[⊥]

Xianxian Zhang[2], *John H. L. Hansen*[1,2,4],
Kazuya Takeda[3], *Toshiki Maeno*[3], *Kathryn Arehart*[4]

[1]Center for Robust Speech Systems, Dept. of Electrical Engineering, University of Texas at Dallas, Richardson TX USA
[2]Robust Speech Processing Group, CSLR, University of Colorado at Boulder, Boulder CO USA
[3]Nagoya University, Nagoya, Japan
[4]Dept. of Speech, Language and Hearing Sciences, University of Colorado at Boulder, Boulder CO USA
{xianxian.zhang@ti.com john.hansen@utdallas.edu} http://crss.utdallas.edu

ABSTRACT

Human-Computer interaction for in-vehicle systems requires effective audio capture, tracking of who is speaking, environmental noise suppression, and robust processing for applications such as route navigation, hands-free mobile communications, and human-to-human communications for hearing impaired subjects. In this paper, we consider two interactive speech processing frameworks for in-vehicle systems. First, we consider integrating audio-visual processing for localization the primary speech for a driver using a route navigation system. Integrating both visual and audio content allows us to reject unintended speech to be submitted for speech recognition within the route dialog system. Second, we consider a combined multi-channel array processing scheme based on a combined fixed and adaptive array processing scheme (CFA-BF) 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 evaluate the audio-visual localization scheme for route navigation dialogs and show improved speech accuracy by up to 40% using the CIAIR in-vehicle data corpus from Nagoya, Japan. For the combine array processing and speech enhancement methods, we demonstrate consistent levels of noise suppression and voice communication quality improvement using a subset of the TIMIT corpus with four real noise sources, with an overall average 26dB increase in SegSNR from the original degraded audio corpus.

1. INTRODUCTION

Human-computer interaction for in-vehicle information access, human communications, and navigation systems are challenging problems because of the diverse and changing acoustic environments inside cars. There are many situations where it is important to be able to identify and track which subject inside the vehicle is talking, as well as perform speech processing using multi-microphone arrays and enhancement algorithms to improve the perceived quality of speech. 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] - Chap. 8).

In this paper, we consider two aspects of signal processing for in-vehicle systems: (i) audio-visual processing for localization of the primary talker for in-vehicle route dialog systems, and (ii) combine array processing and auditory based speech enhancement to improve quality for hearing impaired subjects. In the first area, it is proposed that the integration of video and audio information can significantly improve dialog system performance, since the visual modality is not impacted by acoustic noise. Here, we propose a robust audio-visual integration system for source tracking and speech enhancement for an in-vehicle speech dialog system. The proposed system integrates both audio and visual information to locate the desired speaker source. Using real data collected in car environments, the proposed system can improve desired speech accuracy by up to 40.75% compared with audio data alone.

In the second area, speech enhancement for hearing impaired subjects inside car environments requires FM technology where speech from non-hearing impaired speakers are captured and transmitted via a wireless link directly to a hearing assist device worn by the hearing impaired subject. One way to discuss trade-offs in speech enhancement algorithms in this area 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 this area, we incorporate array processing with single channel speech enhancement methods to suppress noise for in-vehicle applications. In the next section, we consider localization in the car environment.

[⊥] This research was supported in part by U.S. Air Force Research Laboratory, and the University of Texas at Dallas.

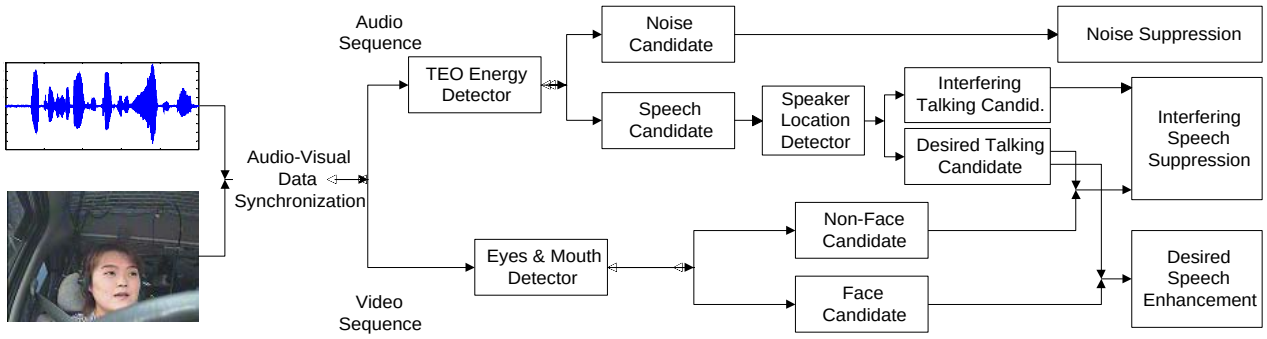

Figure 1: Schematic diagram of the proposed audio-visual integration system

2. LOCALIZATION VIA AUDIO-VISUAL

The increased use of mobile telephones and voiced controlled features for human-machine dialog system in cars has created a greater demand for hands-free, in-car installations. Many countries now restrict handheld cellular technology while operating a vehicle. As such, there is a greater need to have reliable voice capture within automobile environments.

However, the distance between a hands-free car microphone and the speaker will cause a severe loss in speech quality due to changing acoustic environments. Therefore, the topic of capturing clean and distortion-free speech under distant talker conditions in noisy car environments has attracted much attention. Microphone array processing and beamforming is one promising area which can yield effective performance. Currently, most beamforming algorithms must integrate speaker/source localization techniques in order to enhance the desired speech and suppress interference [9,10,8].

Here, speaker localization is the ability to estimate the position of a speaker in the car, and involes the following:

(i) Complex in-vehicle noise situations will severely degrade performance of speaker localization techniques.

(ii) Speaker localization techniques cannot distinguish between desired and undesired speech if both speech sources are from the same direction.

In car environments, the desired speech for a navigation system is assumed to be the driver's, while the undesired speech includes both the passengers and a portion of driver's (e.g., the driver murmurs while looking up or down, the driver laughs and chats with other people inside car, etc.). One way to address this problem is to integrate visual based object localization techniques.

Audio-visual (A-V) speaker localization has recently received significant interest [5,6,7] mainly because the visual modality is not affected by varying acoustic noise and sound localization is unaffected by rapidly varying room lighting. However, there are situations where the integration of video information can significantly improve in-vehicle human-machine dialog system performance. For example, determining the movement of the driver's mouth, body, and head position can impact how a dialog system should respond. If the driver's mouth does not move while speech is detected from the driver's position, then most likely the passenger who sits behind the driver is talking. If the driver asks a question while facing forward, then we can expect the request is being directed towards the in-vehicle dialog system. If the driver is turned towards individuals sitting in the backseat, then the question is most likely directed at someone in the car (e.g., "Where did you say you wanted to eat?"). For such a case, it would not be appropriate to submit such a request to the dialog system. In this area, we discuss the development of an audio-visual system for in-vehicle localization which was originally developed in [2]. Evaluations are based on data collected from the automobile collection platform of the Center for Integrated Acoustic Information Research (CIAIR) [4], Nagoya University, Japan.

Fig. 1 illustrates the proposed audio-visual integration system which includes the following four stages: audio-visual data synchronization, speaker localization using audio data, face tracking using visual data, and speech enhancement and noise/interfering speech suppression using a constrained switched adaptive beamformer [8].

A-V Data Synchronization: Since sampling rates of the audio and visual signals are different, the proportion of the number of the sampled audio data to that of the visual data in general is a fractional frame number. In our case, the CIAIR database from Nagoya Univ.[4] uses a 16kHz speech sample rate with a visual data rate of 30 frames/sec. After synchronization, we keep the temporal mismatch error between audio and visual data at less than 0.033 sec. This mismatch level is acceptable since visual data is only used for speaker localization and activation.

Source Tracking Using Audio: We first use the Teager Energy Operator (TEO) criterion to decide the speech activity for the audio data, then apply the adaptive LMS filter technique to locate the current position of the speech source. Further details are discussed in [2].

Face Tracking Using Visual Data: The function of this processing stage is to detect interfering speech which cannot be identified by sound localization techniques alone. We apply basic eye and mouth detection and tracking techniques in this processing stage.

From our observation and experiments using the CIAIR in-vehicle corpus, we found that most of the interfering speech versus that from the driver occurs in the following situations:

Case 1: The passenger talks and the driver listens. Under this situation, the driver's lips will not move often;

Case 2: The driver murmurs while looking up or down, which causes part of his/her face to be obscured by the steering wheel;

Case 3: The driver laughs or coughs while covering his/her mouth with their hands;

Case 4: The driver chats with the interfering person while he/she is driving. Under this situation, the driver will likely shift his/her head or body slightly towards the interfering person, which makes a portion of the face features disappear.

Manually Computed Periods (in secs)		Detected Periods (in secs) Using Audio Data		Detected Periods (in secs) Using Audio-Visual Data	
Speech	**Desired speech**	**Speech**	**Desired speech**	**Speech**	**Desired speech**
107.335	**47.173**	**128.392**	**95.978**	**128.392**	**57.456**

Table 1: Performance of Detected Route Dialog Directed Speech using (i) manual labeling, (ii) audio processing alone, and (iii) audio and visual processing (note: this particular audio-visual stream consists of 321.5 sec of data, of which there is 107.335sec of speech activity).

Fig. 2 shows examples where there is speech interference (i.e., speech not always intended for the route navigation system). In the formulated audio-visual integration system, we use template based eyes and mouth detection software to detect face features. We also track the distance between the eyes and changing mouth shape across frames. For example, if a driver's mouth shape does not change within a certain period, the current speech most likely comes from the passenger (i.e., Case 1); if the distance between driver's eyes is smaller than a certain value for a time period, then he/she likely has shifted their head backwards (i.e., Case 4); if part of face features, such as the mouth, cannot be detected, then most likely the driver is under situations described in Case 2 and 3.

Enhancement and Interfering Speech / Noise Suppression: Once we detect the nature of the current signal, we propose to use the constrained switched adaptive beamforming algorithm (CSA-BF) [8] to enhance the desired speech and suppress background noise and interfering speech.

Fig. 3 shows how visual information helps to detect the interfering speech or non-dialog directed speech. Here, it is straightforward to determine when speech activity occurs, but larger challenge is to say when the speech is directed towards the microphone array based in-vehicle navigation system. For example, the signal during the period from frame count 150 to 200 corresponds to when the driver is laughing. Here, the averaged Teager energy (TEO) is high enough to pass the speech threshold, and sound localization results also confirm that the speech comes from position number 0, (i.e., the driver is talking and facing forward). Therefore, if only audio information is used, the speech during this period will be identified as desired speech. However, from the results of face feature detection, we find that the driver's mouth cannot be detected since it is covered by the driver's hand, and therefore this segment is correctly labeled as interfering speech. Similarly, when the driver is talking with the passenger during frame count 400 to 500, our face tracking algorithm is also able to classify this speech as interfering speech, since the driver shifts her head backward frequently while chatting with the passenger, and the detected distance between the eyes is shorter than that while facing forward. This speech also cannot be identified as undesired by audio data only.

Table 1 shows the accumulated speech and desired speech activity periods under different experimental situations. From this table, we can see that by using visual data processing in addition to the TEO criterion with the LMS filter, the accuracy of the desired speech detection is improved 40.75%, (i.e., a reduction in the desired speech duration from 96 sec to 57 sec, approaching the goal of 47 sec). While this improvement is important (i.e., reduction of 39.412 sec), there is still approximately 10 seconds of interfering speech that is still included.

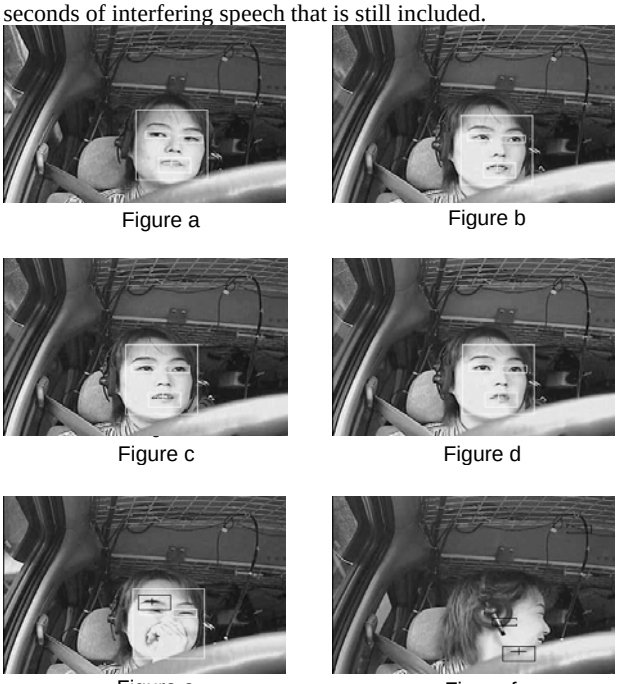

Figure a Figure b Figure c Figure d Figure e Figure f

Fig. 2: (a-d): Face is detected as "existing" and source comes from quantized angular locations 1, 0, -3, and -5 respectfully. (e-f): face is detected as "not existing" and source is from direction 0 (driver laughing) and -5 (interference from passenger talking with the driver).

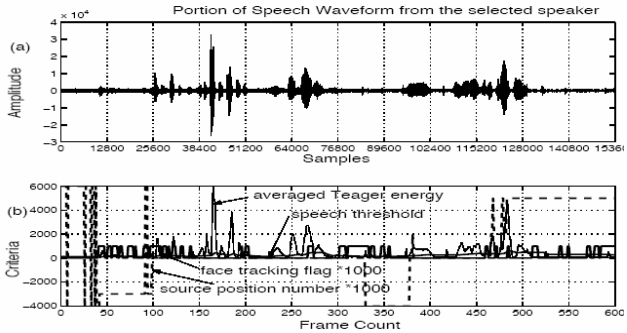

Fig. 3: Audio-Visual Tracking results for a selected speaker using actual in-vehicle audio-visual data from the CIAIR corpus.

Interfering Speech Cancellation Results: For the system from Fig. 1, the interfering speech cancellation is possible with/without face tracking results as one of the constraints for the constrained switched adaptive beamforming (CSA-BF). From these results, we can make the following observations:

(i) Employing the proposed audio-visual integration system can improve the accuracy of the desired source tracking by up to 40.75% (i.e., we can remove non-desired speaker speech prior to ASR for the dialog system);

(ii) The proposed system with better source tracking using Audio-Visual also improves interfering speech cancellation.

3. SPEECH ENHANCEMENT: ARRAY + SINGLE CHANNEL PROCESSING SCHEMES

Next, we consider speech/array processing for hearing impaired subjects for in-vehicle environments. Background car noise and competing speakers interference represents challenges for hearing-impaired subjects in car environments. For the application for hearing impaired subjects in vehicle, we first present a data collection experiment for a proposed FM wireless transmission scenario using a 5-channel microphone array in the car, and followed by several alternative speech enhancement algorithms. After formulating 6 different processing methods, we evaluate the performance using SegSNR improvement with data recorded in a moving car environment. Among the 6 processing configurations, the combined fixed/adaptive beamforming (CFA-BF) obtains the highest level of SegSNR improvement by up to 2.65 dB. An earlier version of this work was presented in [3], and here we discuss the framework and further results in detail.

To motivate the proposed method, we consider a previous proposed combined fixed/adaptive beamforming algorithm (CFA-BF) [7] 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 [11] 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. 4 shows the spectrogram of the original degraded speech, and enhanced speech by CFA, GAE, and Auto-LSP respectively [11]. 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. 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.

3.1 Overall Algorithm Description

In our 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. 5 summarizes an overall description of the proposed algorithm.

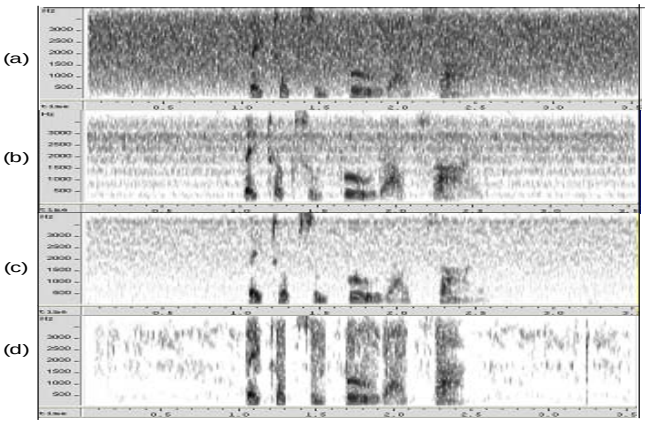

Fig. 4: Spectrogram of Speech Data with: (a). Original FLN degraded noisy speech; (b). CFA Enhanced speech; (c). GMMSE-AMT-ERB Enhanced speech; (d). Auto-LSP Enhanced speech.

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.

Fig. 5: Formal description of the proposed algorithm.

3.2 Detailed Algorithm Design

3.2.1. Front-end processing

The block diagram of the structure of the proposed algorithm is shown in Fig. 6. 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) [8], 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.

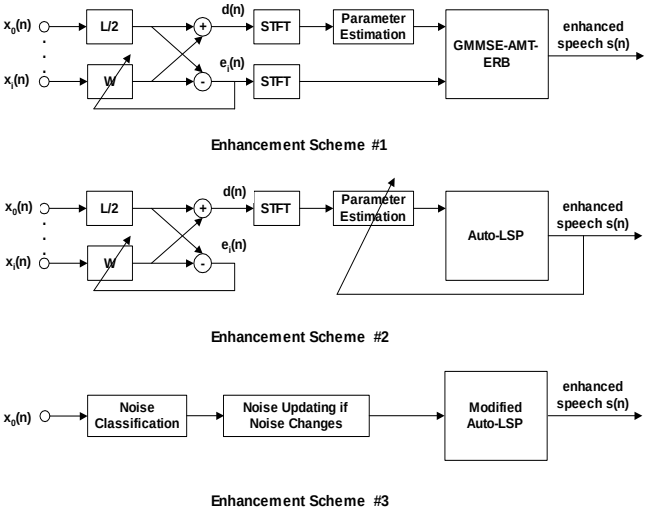

Fig. 6: Block Diagram of the Proposed Algorithm

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.

3.3 Performance Evaluation

3.3.1. Experimental Database & Setup

In order to evaluate the performance of the proposed algorithm, we select 10 sentences from the TIMIT database, and degrade these sentences with 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.

3.3.2. Experiment Results

Fig. 7 illustrates average SegSNR improvement using sentences degraded with FLN noise. Table 2 show the Segmental SNR measure for the degraded speech with 4 different noises and enhanced speech by 5 different schemes.

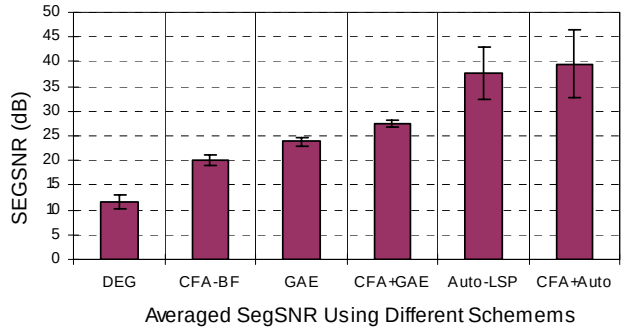

Fig. 7: SegSNR Results for Degraded and Enhanced Speech

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 2: Averaged Segmental SNR (dB) for Different Schemes

4. SUMMARY & CONCLUSIONS

In this paper, we have consider two interactive speech processing frameworks for in-vehicle systems. First, we considered integrating audio-visual processing for localization the primary speech for a driver using a route navigation system. Integrating both visual and audio content allows us to reject unintended speech to be submitted for speech recognition within the route dialog system (i.e., a 40.75% improvement). Next, we considered 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. These contributions suggest that improvements for interactive systems for in-vehicle systems such as multi-sensor based schemes and assist frameworks for hearing-impaired users can expand the use of in-vehicle route navigation systems as well as hands-free and human communication devices for cars.

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] X.X. Zhang, K. Takeda, J.H.L. Hansen, T. Maeno, "Audio-Visual Integration for Hands-Free Voice Interaction in Automobile Route Navigation," ICA-2004: International Congress on Acoustic, vol. 4, pp. 2821-2824, Kyoto, Japan, April 2004.

[3] X.X. Zhang, J.H.L. Hansen, K. Arehart, "Speech Enhancement based on a Combined Multi-Channel Array with Constrained Iterative and Auditory Masked Processing," IEEE ICASSP-2004, vol. 1, pp. 229-232, Montreal, Canada, May 2004

[4] http://www.ciair.coe.nagoya-u.ac.jp/

[5] C. Neti, G. Potamianos, et. al, "Audio-Visual Speech Recognition," Final Workshop Report, Johns Hopkins Univ., 2000.

[6] M. Beal, N. Jojic, H. Attias, "A self-calibrating algorithm for speaker tracking based on audio-visual statistical models, IEEE ICASSP-02.

[7] X.X. Zhang, J.H.L. Hansen, "CFA-BF: A Novel Combined Fixed/Adaptive Beamformer for Robust Speech Recognition in Real Car Environments," Eurospeech-2003, pp. 1289-92, Sept. 2003.

[8] X. Zhang, J.H.L. Hansen, "CSA-BF: A Constrained Switched Adaptive Beamformer for Speech Enhancement and Recognition in Real Car Environments," IEEE Trans. Speech & Audio Proc., vol. 11, no. 6, pp. 733-745, Nov. 2003.

[9] D. von Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," IEEE ICASSP-90, April 1990.

[10] E. Visser, M. Otsuka, T.W. Lee, "A spatio-temporal speech enhancement scheme for robust speech recognition," ICSLP-2002, Denver, CO, Sept. 2002.

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

[12] 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.

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

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

[15] 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.

[16] A. Natarajan, J.H.L. Hansen, K.H. Arehart, J. Rossi-Katz, "An Auditory Masked Threshold based Noise Suppression Algorithm GMMSE-AMT[ERB] for Listeners with Sensorineural Hearing Loss," EURASIP Journal of Applied DSP: Special Issue on Signal Processing Hearing Aids and Cochlear Implants, vol. xx, no. x, pp. xxx-xxx, 2005.

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

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

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

[20] 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.