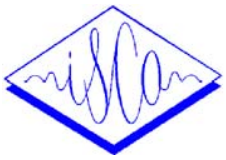
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Audio-Visual Speaker Localization for Car Navigation Systems

Xianxian Zhang[1], *Kazuya Takeda*[2], *John H. L. Hansen*[1], *and Toshiki Maeno*[2]

[1]Robust Speech Processing Group, Center for Spoken Language Research,
University of Colorado at Boulder, Colorado, U.S.A.
[2]Graduate School of Information Science, Nagoya University, Nagoya, Japan
{zhang,jhlh@cslr.colorado.edu} {takeda,maeno@is.nagoya-u.ac.jp}

10.21437/Interspeech.2004-427

Abstract

Human-computer interaction for in-vehicle information and navigation systems is a challenging problem because of the diverse and changing acoustic environments. 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. In this paper, 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 speech accuracy by up to 40.75% compared with audio data alone.

1. Introduction

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 [1]. 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 inteference [2, 3, 4].

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 is 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 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 your wanted to eat?"). For such a case, it would not be appropriate to submit such a request to the dialog system.

In this paper, we focus on integrating basic face tracking and sound localization techniques, which will achieve complex speaker localization tasks for noisy car environments. We also applied the new proposed audio-visual speaker localization technique into a previously developed array processing algorithm [4] for improving speaker tracking accuracy and enhancing desired speech for a car navigation system. Evaluations are based on data collected from the automobile collection platform of the Center for Integrated Acoustic Information Research (CIAIR) [8], Nagoya University, Japan.

2. The Audio-Visual Integration System

Fig. 1 illustrates the proposed audio-visual integration system. The following four stages are considered: audio-visual data synchronization, speaker localization using audio data, face tracking using visual data, and speech enhancement and noise/inteferencing speech suppression using a constrained switched adaptive beamformer [4].

2.1. Audio-Visual Data Synchronization

The goal of this stage is to synchronize audio-visual data for later processing. 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. For example, the sampling rate of audio data collected from the Nagoya University CIAIR automobile collection platform is 16kHz, while the visual data is collected at a rate of 30 frames/sec. Therefore, we need to synchronize the collected audio-visual data before we apply any source and face tracking algorithms. Let x denote the index of the sampled audio signal, and y denote the index of the sampled visual signal, where $x = 1, 2, ..., M$, $y = 1, 2, ..., N$, and $M > N$, the frame ratio is therefore,

$$\hat{y} = \frac{N}{M}x \tag{1}$$

$$y = F(\hat{y}) \tag{2}$$

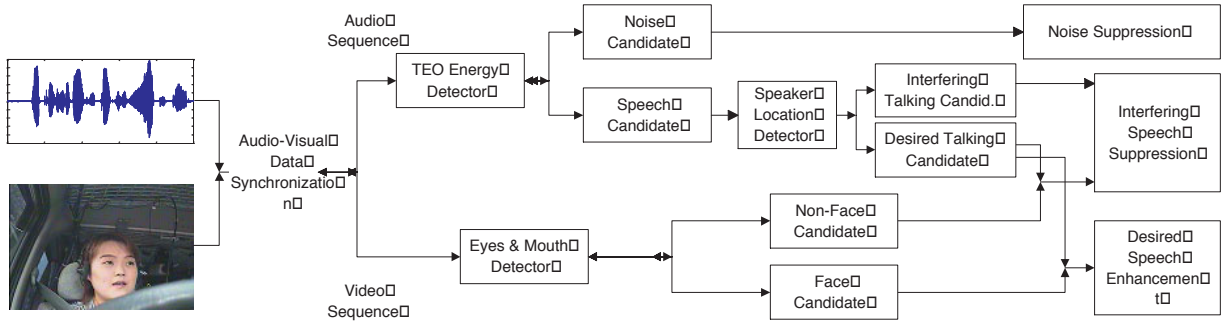


Figure 1: Schematic Diagram of the Proposed Audio-visual Integration System

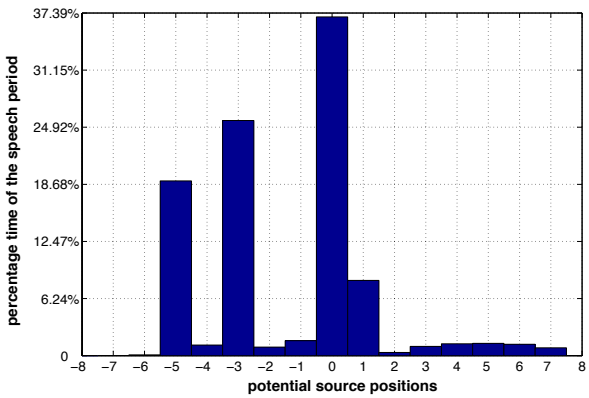


Figure 2: Histogram of Estimated Source Movement inside a Moving Car (Source position is from -45° to $+45^\circ$ in increments of 2.5° as we move from position -8, ..., +8.)

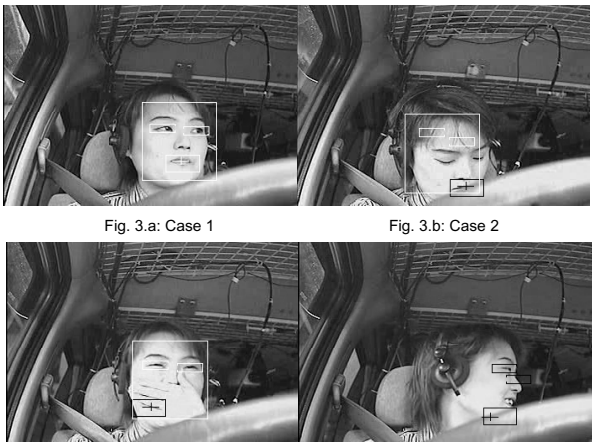


Fig. 3.a: Case 1 Fig. 3.b: Case 2

Fig. 3.c: Case 3 Fig. 3.d: Case 4

Figure 3: Visual Description of Possible Interfering Speech

where

$$F(\hat{y}) = \begin{cases} \lfloor \hat{y} \rfloor & \text{if } \hat{y} - \lfloor \hat{y} \rfloor \leq 0.5 \\ \lceil \hat{y} \rceil & \text{if } \lceil \hat{y} \rceil - \hat{y} \leq 0.5 \end{cases} \quad (3)$$

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 of the speech recognition engine (i.e., lip structure is not presently employed in the ASR task).

2.2. Speaker Localization Using Audio Data

The function of this processing stage is to locate the position of the desired source (i.e., driver's speech) in the car during voice interaction. In our implementation, we first use the Teager Energy Operator (TEO) criterion described in [9, 4] to decide the speech activity for the audio data, then apply the adaptive LMS filter technique [10] to locate the current position of the speech source. In order to analyze the talking activities of the driver, a speaker is randomly selected from the CIAIR-HCC database [8]. Fig. 2 shows the histogram of estimated angular speaker source locations during speech segments for the selected speaker. To describe these angular source locations, we follow the definition described in [4] and use digits to define the potential source angle positions. This speaker face, forward 37% of the time while talking, but turns to the left $7.5^\circ(3 * 2.5^\circ)$ and $12.5^\circ(5 * 2.5^\circ)$ quite often.

2.3. 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 lip 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 .

Fig. 3 shows examples of interfering speech. In the proposed 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.

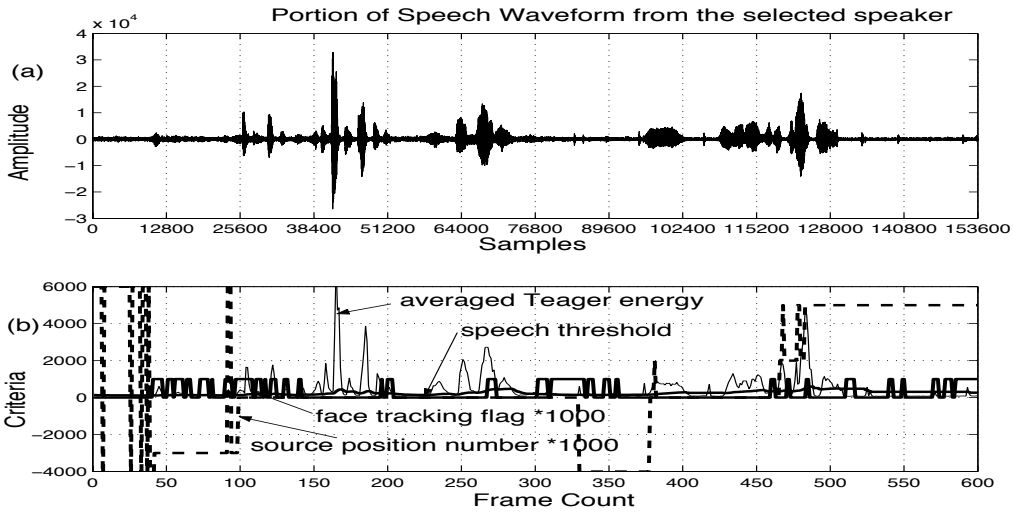


Figure 4: Description of Audio Visual Tracking Results for a Selected Speaker

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 decribed in case 2 &3.

Fig. 4 shows how visual information helps to detect the interfering speech. It is quite straightforward using Fig. 4a to determine when speech activity occurs, but more challenging to say when the speech is directed towords 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 indentified as undesired by audio data only.

2.4. Speech Enhancement and Interferencing Speech / Noise Suppression

Once we detect the nature of the current signal, we propose to use the constrained switched adaptive beamforming algorithm (CSA-BF)[4] to enhance the desired speech and suppress background noise and interfering speech.

3. Performance Evaluation

3.1. CIAIR -HCC Corpus

The CIAIR-HCC audio-visual [8] Corpus contains 300 subjects (male and female) speaking Japanese. The recording conditions includes 4 sessions: Reading out during idle running, conversation during driving, 25 balanced sentences while driving, and 50 balanced sentences during idle driving. Each audio set comprises 16 channels that include 6 distributed microphones (installed at 6 locations within the cabin), 2 head-worn close talking microphones (one each for driver and navigator), and 4 linear arranged microphones in front of the driver. Three cameras are used to record video data from directly in front of the driver, from directly in back looking out the front window, and from the left front (in front of the driver's seat). In our experiments, we use the audio data collected by the 4-channel linear microphone array, and the visual sequence collected by the camera from directly in front of the driver.

3.2. Desired Speech Tracking Verification

In order to evaluate the proposed audio-visual integration system, we first evaluate the accuracy of the desired speech tracking as follows:

- Label the selected speech data manually and compute the entire recording time;
- According to the labeled results, compute both speech activity period and desired speech activity period. As mentioned above, speech includes both desired and undesired speech;
- Detect speech activity and the desired speech activity periods using only audio data with the TEO criterion and LMS adaptive filter technique;
- Detect the speech activity and the desired speech activity periods using both audio and visual data with the TEO criterion, LMS adaptive filter technique, and face features detection and tracking technique.

Table 1 shows the overall 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

Entire Recording Time (in secs)	Manually Computed Periods (in secs)		Detected Periods (in secs) (using audio information)		Detected Periods (in secs) (using audio-visual information)	
Speech/Non-speech	Speech	Desired Speech	Speech	Desired Speech	Speech	Desired Speech
321.5	107.335	47.173	128.392	96.978	128.392	57.456

Table 1: Speech and Desired Speech Periods under Different Experimental Situations

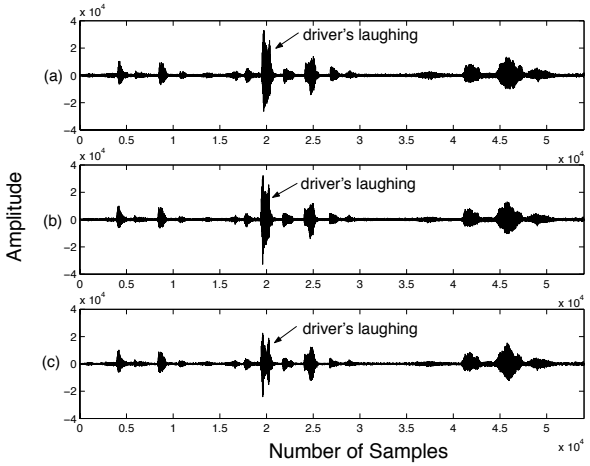


Figure 5: Time Waveforms for a selected speaker: (a) Original unprocessed single microphone, (b) Processed speech waveform by CSA-BF without face tracking, (c) Processed speech waveform by CSA-BF with face tracking.

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.

3.3. Interfering Speech Cancellation Results

Fig. 5 shows the interfering speech cancellation results 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%;

(ii) The proposed system with better source tracking using A-V has also improved interfering speech cancellation.

4. Conclusions and Future Work

In this paper, we have proposed an audio-visual integration system for hands-free voice interaction for in-vehicle route navigation. Based on the proposed system, we have also investigated and compared the difference of the desired speech activities for with and without the assistance of visual information. We demonstrated that the proposed system integration of video analysis can improve the detection of the driver's speech and reduce the impact of the passenger's interfering speech; it also can improve interfering speech cancellation. We believe this contribution will have a more pronounced impact on automatic speech recognition (ASR) versus speech quality as measured by SegSNR.

In the future, we will consider the following issues:

- Use a larger portion of the CIAIR Corpus to evaluate the performance of the proposed Audio-visual integration system, (i.e., consistency across more drivers);
- Improve the performance of the face detection algorithm, especially tracking the movement of the face/head;
- Apply the advances in [11] for driver behavior research to audio-visual integration system, and improve the accuracy of identifying the driver's speech.
- Integrate the proposed audio-visual integration system with an ASR system.

5. References

[1] Komarow, S., "Germany targets dialing and driving", in *USA Today*, Tuesday, September 12, 2000.

[2] Compernolle, D. V., "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings", *Proc. IEEE ICASSP'1990*, Alburquerque, NM, USA, April, 1990.

[3] Visser, E., Otsuka, M., and Lee, T. W., "A spatio-temporal speech enhancement scheme for robust speech recognition", *Proc. IEEE ICSLP'2002*, Denver, CO, USA, September, 2002.

[4] Zhang, X.X., and Hansen, J.H.L., "CSA-BF: A Constrained Switched Adaptive Beamforming for Speech Enhancement & Recognition in Real Car Environments", *IEEE. Trans. on Speech and Audio Processing*, Vol 11(6), pages 733-745, Nov. 2003.

[5] Potamianos, G., and Neti, C., "Audio-Visual Speech Recognition in Challenging Environments", *Proc. Eurospeech'2003*, Geneva, Switzerland, Spet, 2003.

[6] Gatica-Perez, D., Lathoud, G., McCowan, I. Odobez, J. and Moore, D., "Audio-Visual Speaker Tracking with Importance Particle Filter", *Proc. IEEE ICASSP'2003*, HongKong, China, April, 2003.

[7] Beal, M. J., Jojic, N., and Attias, H., "A Self-calibrating Algorithm for Speaker Tracking based on Audio-visual Statistical Models", *Proc. IEEE ICASSP'2002*, Orlando, FL, USA, May, 2002.

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

[9] Kasier, J, "On a Simple Algorithm to Calculate the 'Energy' of a Signa", *ICASSP'90*, Alburquerque, NM, USA, April, 1990.

[10] Reed, F.A., Feintuch, P.L., and Bershad, N.J., "Time delay Estimation Using the LMS Adaptive Filter-Static Behavior", *IEEE Trans. On Acoustics, Speech and Signal Processing*, pages 561-571, vol. ASSP-29, no.3, June, 1981.

[11] Kuge, N., and Yamamura, T., Shimoyama, O., and Liu, A.,"A Driver Behavior Recognition Method Based on a Driver Model Framework", *Proc. of the SAE World Congress*, Detroit, MI, March 6-9, 2000.