ROLE OF HEAD POSE ESTIMATION IN SPEECH ACQUISITION FROM DISTANT MICROPHONES

Shankar T. Shivappa, Bhaskar D. Rao and Mohan M. Trivedi

University of California, San Diego
Department of Electrical and Computer Engineering
9500 Gilman Drive, La Jolla, CA 92093, USA

ABSTRACT

Reverberant environments pose a challenge to speech acquisition from distant microphones. Approaches using microphone arrays have met with limited success. Recent research using audio-visual sensors for tasks such as speaker localization has shown improvement over traditional audio-only approaches. Using computer vision techniques we can estimate the orientation of the speaker's head in addition to the location of the speaker. In this paper we study the utility of using the head pose information for effective beamforming and clean speech acquisition from distant microphones. The improvements in speech recognition accuracy relative to that of a close talking microphone are presented and the results provide sufficient motivation for incorporating head pose information in beamforming techniques.

Index Terms— Speech enhancement, audio-visual fusion, speech recognition, head-pose estimation, intelligent spaces, human-computer interface

1. INTRODUCTION

Speech acquisition from distant microphones in a reverberant environment is a challenging task [1][2]. The signal at the distant microphone is distorted due to echoes and techniques based on SNR measurements cannot be employed to select the best set of microphones. Existing approaches to acquiring clean speech from distant microphones include microphone array based beam-forming techniques. In such systems, recent research has focussed on augmenting the microphone array based system with information from video cameras which are used to track the speakers and provide accurate location information. The sensitivity of the speech acquisition systems to location errors has been studied in [3]. In this paper we explore the sensitivity of distant speech acquisition systems to the orientation of the speaker's head in addition to speaker location. Speech recognition accuracy can be significantly improved by using the correct beamforming parameters for the particular location and the orientation of the speaker's head. The orientation of the speaker's head can be estimated using both the audio and video modalities. An audio-visual head pose estimation system is presented in [4]. A detailed survey of video-only head pose estimation can be found in [5]. In our system we use the head orientation estimates and location estimates from the video modality, to improve the quality of speech enhancement by the microphone array. We adopt a delay, filter and sum strategy and report the improvement in the speech recognition accuracy on a large vocabulary speaker dependent speech recognition task.

This material is based upon work supported by the National Science Foundation under Grant No. (0331690).

2. ROOM ACOUSTIC TRANSFER FUNCTION

Let microphone i be at location (x_i, y_i, z_i) and the head of the speaker be centered at (x_s, y_s, z_s) and oriented in the direction (ϕ_s, θ_s) in the polar co-ordinates relative to the original co-ordinates. The location and directivity of the microphones is assumed to fixed and we do not model changes in those parameters. Let us assume that the source signal $s(t)$, measured using a close talking microphone, encounters a channel whose impulse response is h_i. The transfer function corresponding to this channel will be referred to as the room acoustic transfer function. If we represent the signal received at microphone i by $y_i(t)$, then, $y_i(t) = s(t) * h_i$. In Figure 1, we see an example of $h_i(t)$ for two different source locations.

We claim, h_i depends on $x_i, y_i, z_i, x_s, y_s, z_s, \phi_s$ and θ_s. Since the microphone is assumed to be fixed, we could reduce the dependence to x_s, y_s, z_s, ϕ_s and θ_s. It is easy to see why this is indeed the case. The location of the speaker relative to the microphone and the room determines the relative delay and amplitude of the reflections from the walls and other surfaces in the room, contributing to the tail of the impulse function. The human vocal tract acts as a directed source. This is especially true for frequencies greater than 4kHz. In [4], the head radiation pattern is discussed in detail. From the directional nature of head radiation pattern one can deduce the dependence of the room acoustic transfer function on ϕ_s. In Figure 2, the dependence of h_i on the orientation of the speaker's head confirms this deduction. In the next Section

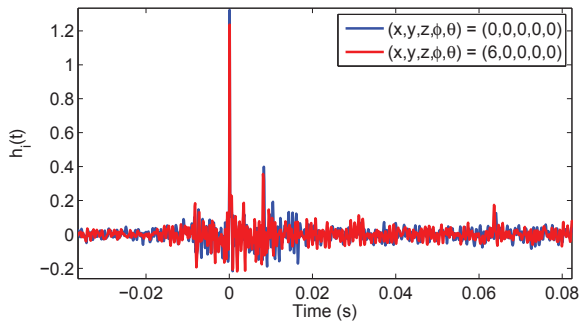

Fig. 1. Room acoustic impulse response, for two source locations 6" apart. The impulse response is estimated by assuming that $s(t)$, measured using a close talking microphone, is the input and $h_i(t)$, the signal received at microphone i, is the output of the channel. The same measurements are then repeated for another location of the speaker, 6" away from the first.

we present a framework to utilize the head pose information for effective speech acquisition from distant microphones.

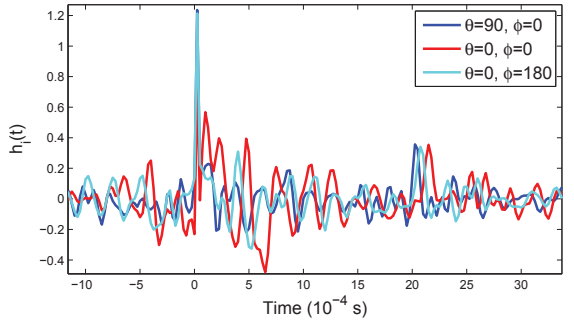

Fig. 2. Room acoustic impulse response, for same location but three different speaker head orientations, estimated as in Figure 1. Note that the impulses responses are very different, indicating the sensitivity to head pose.

3. SPEECH ACQUISITION FROM DISTANT MICROPHONES

Current research on speech acquisition using distant microphones implicitly or explicitly model speaker location. Speaker location is implicitly used in aligning the signals from different microphones with one another [6]. In more advanced schemes, in addition to the proper alignment of the signals using the appropriate delay, location specific parameters are used in beamforming [3]. However, to date, no research has included the orientation of the speaker's head in the beamforming techniques. This is mainly due to the difficulty of estimating the orientation of the head. Using video, however, we can estimate the head pose of the speaker[5] and use this information in acquiring clean speech from distant microphones. This can be done in one of the following ways,

- Use specific microphone array beamformer coefficients for the current location and orientation of the speaker.
- Use a subset of the microphones for acquiring the speech by selecting those microphones that have a strong direct path from the speaker.
- Use the best microphone for the present speaker location and orientation.

Note that the later options are specific instances of the earlier ones. However, they are also easier to implement in practise. Thus there is a trade-off between generality and convenience. In Section 4, we present results that provide practical insights to this trade-off. The other issue that is addressed in Section 4 is that of the sensitivity of automatic speech recognition to the orientation of the speaker's head in each of the three situations considered above. This allows us to implement a practical system by training beamformers for particular orientations of the speaker's head. In more specific instances, such as meeting rooms, the participants tend to face each other while speaking and this would allow the training of beamformers for these particular cases. These cases are also explored in Section 4. Also note that energy/SNR based selection of the 'best' microphones does not convey the same information as a microphone that has a dominant direct path and register clearer signals from the speaker.

4. COMPUTATIONAL FRAMEWORK AND ALGORITHMS

In Figure 3, we present the framework of our proposed scheme. The configuration of the sensors and the layout of the room are shown in 4

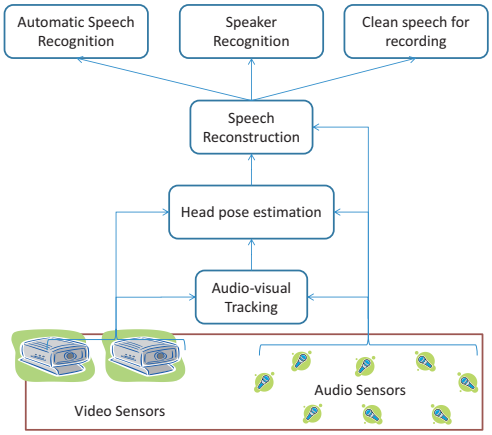

Fig. 3. The overall system flowchart.

4.1. Audio-visual person tracking

The localization of speaker is based on our earlier work. We refer the reader to [7] for details. The audio localization in-

cludes the time difference of arrival (TDOA)estimation as a first step and these TDOA estimates are used in the beamformer for aligning the signals from different microphones.

4.2. Audio-visual head pose estimation

In section 1, we discussed some aspects of audio-visual head pose estimation. Our video head pose estimation algorithm using calibrated video cameras is based on the algorithm discussed in [8]. The audio head pose estimation is not incorporated in the present system, but could be a future addition.

4.3. Filter and sum beamformer

In our experiments we use a filter and sum beamformer to reconstruct the speech signal from the distant microphones . The signal $s_i(t)$ from the ith microphone is delayed by the appropriate delay T_i to align all the microphones with one another. During the training phase, they are aligned with a reference microphone $s_r(t)$ that is placed close to the speaker and the filter taps are trained by a stochastic gradient descent algorithm. Note that by explicitly constraining a subset of the filters to have all zero taps, we can select a subset of the microphones. And in the extreme case, select only one of the microphones. These cases correspond to the three options mentioned in Section 3

4.4. Automatic Speech Recognition

A commercially available speech recognition software, the dragon naturally speaking system is used for recognizing the acquired speech signal. The recognition system is adapted for each speaker separately, using a close talking microphone. This is the same microphone used as the reference microphone in training the beamformer taps (Section 4.3). The results correspond to a person dependent large vocabulary continuous speech recognition task based on the standard dictation mode of the speech recognizer.

5. EXPERIMENTAL EVALUATION

In this Section we describe the experimental setup in the Smartspace lab at UCSD. We present the details of the system used to evaluate the theory presented above. The results presented in Section 5.1 are from this setup. Figure 4 shows the layout of the room in which the audio-visual system is deployed. There are 2 rectilinear cameras and 8 omnidirectional microphones deployed in the room as shown in Figure 4. The cameras and microphones are calibrated with respect to the room co-ordinates. The setup is close to a typical meeting with 4 participants and a presenter. Thus each participant has 4 foci of attention, corresponding to the 4 other participants in the meeting. For each orientation of the speaker, corresponding to the speaker facing one of these foci, we present the following results.

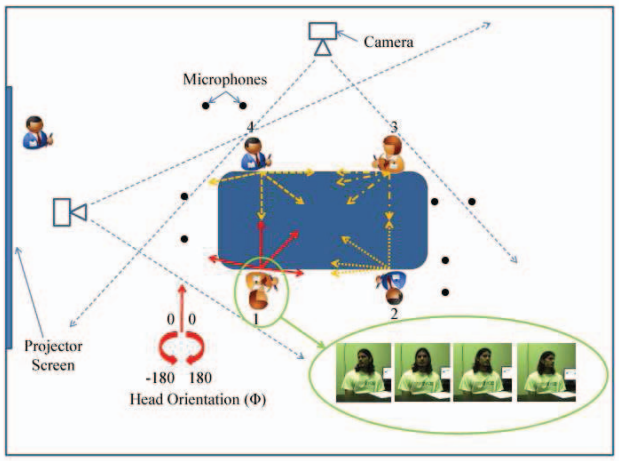

Fig. 4. Layout of the audio-visual testbed at the Smartspace lab at UCSD.

5.1. Results

- Case A: A filter and sum beamformer is trained using all 8 microphones for that particular orientation and speaker location.
- Case B: A filter and sum beamformer is trained using a subset of microphones "in front of" the speaker, for that orientation.
- Case C: The single best microphone, based on speech recognition accuracy, is selected and the signal is directly used for speech recognition.

The baseline results to compare the performance of our scheme are as follows.

- Case D: Close talking microphone.
- Case E: A filter and sum beamformer is trained using all 8 microphones, with the training data including all possible orientations at the given speaker location (orientation agnostic).
- Case F: A filter and sum beamformer is trained using all 8 microphones for a "forward" orientation at the given speaker location.

The results are presented in Table 1. From these results, it is clear that by training the beamformer for particular head orientations in any of the three cases A, B, C, one can achieve an improvement over cases E and F.

In Figure 5, we present the results of head orientation mismatch on the speech recognition accuracy. The baseline for comparison is the accuracy of the close-talking microphone. Then there is the beamformer trained for the correct orientation of the speaker along with the beamformer trained for the nominal orientation (angle zero) and used for other orientations of the head. From this we can conclude that using the right head orientation in selecting the beamformer improves the speech recognition accuracy by 10% in some cases.

Table 1. Comparisons of speech recognition accuracies for the beamformers described above. Note that the first three cases require the estimation of the head pose of the speaker, the last two cases represent the best one can do in the absence of such information.

Location	Case A	Case B	Case C	**Case D**	Case E	Case F
	With	head	pose	**Baseline**	No head	pose
1	85%	87%	85%	**90%**	77%	78%
2	84%	85%	84%	**91%**	78%	77%
3	81%	82%	81%	**85%**	72%	71%
4	85%	87%	85%	**90%**	77%	78%

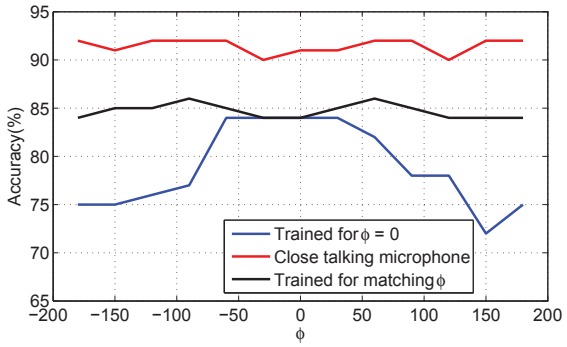

Fig. 5. Sensitivity of the speech recognition task to head orientation mismatch.

6. CONCLUDING REMARKS

We have presented an audio-visual system to effectively acquire speech signals from far-field microphones in a meeting room scenario and demonstrated the improvement in speech recognition accuracy obtained by training beamformers for particular head pose of the speaker. In the more general problems, where the speakers are not constrained to occupy certain locations and face particular directions as in a meeting room, there are open issues that have to be addressed regarding the practicality of storing and using different beamformers for different positions and speaker head orientations. We are working towards reducing the constraints in our system and demonstrate the improvement in speech quality, speech recognition and speaker recognition tasks in a general setting. We would like to acknowledge the detailed comments of the reviewers which helped us to improve the presentation of the paper in the revised version.

7. REFERENCES

[1] T. Gustafsson, B. D. Rao, and M. M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, Nov. 2003.

[2] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Las Vegas, Nevada, USA, 2008.

[3] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 11, pp. 2257–2269, November 2007.

[4] C. Canton-Ferrer, C. Segura, J. R. Casas, M. Pardàs, and J. Hernando, "Audiovisual head orientation estimation with particle filtering in multisensor scenarios," *EURASIP J. Adv. Signal Process*, vol. 2008, no. 1, pp. 1–15, 2008.

[5] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *Accepted, IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[6] M Brandstein and D Ward, *Microphone Arrays*, Springer, 2001.

[7] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Person tracking with audio-visual cues using the iterative decoding framework," in *5th IEEE International Conference On Advanced Video and Signal Based Surveillance*, Santa Fe, New Mexico, USA, 2008.

[8] E. Murphy-Chutorian and M. M. Trivedi, "Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking," *IEEE Intelligent Vehicles Symposium*, pp. 512–517, June 2008.