Improving Audio-Visual Speech Recognition with an Infrared Headset

Jing Huang, Gerasimos Potamianos, Chalapathy Neti

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{jghg,gpotam,cneti}@us.ibm.com

Abstract

Visual speech is known to improve accuracy and noise robustness of automatic speech recognizers. However, almost all audio-visual ASR systems require tracking frontal facial features for visual information extraction, a computationally intensive and error-prone process. In this paper, we consider a specially designed infrared headset to capture audio-visual data, that consistently focuses on the speaker's mouth region, thus eliminating the need for face tracking. We conduct small-vocabulary recognition experiments on such data, benchmarking their ASR performance against traditional frontal, full-face videos, collected both at an ideal studio-like environment and at a more challenging office domain. By using the infrared headset, we report a dramatic improvement in visual-only ASR that amounts to a relative 30% and 54% word error rate reduction, compared to the studio and office data, respectively. Furthermore, when combining the visual modality with the acoustic signal, the resulting relative ASR gain with respect to audio-only performance is significantly higher for the infrared headset data.

1. Introduction

Visual speech information from the speaker's mouth region has been shown to improve the accuracy and noise robustness of *automatic speech recognition* (ASR) systems for both small- and large-vocabulary tasks [1]-[7]. However, almost all research on *audio-visual ASR* (AV-ASR) has concentrated on databases recorded under ideal visual conditions. Such sets contain high-resolution video of the subjects' frontal face, with very limited variation in head pose and subject-camera distance, rather uniform lighting, and, in most cases, constant background. This kind of recording restricts subjects from free movement, and also requires tracking of facial features in order to extract visual speech information. Such tracking becomes less robust in realistic, non-ideal environments, where users move, for example, drivers in cars and agents on trading floors. As a result, in such scenarios, performance of AV-ASR degrades significantly [6].

This motivated us to design a special audio-visual headset that captures the video of the speaker's mouth region, independently of the speaker's movement and head pose, and simultaneously records the corresponding audio signal. Since the headset consistently focuses on the desired mouth region, face tracking is no longer required. Eliminating this step improves the visual front end robustness and reduces its computational requirements.

Figure 1 shows the frontal and side views of the infrared headset and how it is worn. The headset contains both audio and video components in the boom. A microphone is placed in the side of the boom, offset from the mouth, in order to reduce audio distortion caused by breath noise. Located in the tip of the boom are the camera and infrared components. On either side of the camera are two diodes that illuminate the speaker's face with infrared light. Both the camera and infrared diodes sit in a plastic housing, covered by a rectangular infrared filter, which blocks visible light and allows only infrared light to pass through, where it is picked up by the camera lens. The audio and video signals are carried wirelessly by means of an RF transmitter housed in the earpiece of the headset. The designed headset is significantly smaller and less intrusive than the one reported in [8].

This paper represents our first attempt to investigate the benefit of using such an infrared headset for AV-ASR. Since the distance between the infrared camera and the speaker's mouth can be slightly adjusted, we recorded two sets of headset data: one with a close-up (CU) view of the speaker's mouth (see the first row of images in Figure 2), the other with a wider view (WV) of the mouth region (depicted in the second row of Figure 2). Our aim is to study whether including some portion of the cheeks is helpful to visual speech recognition.

We utilize our state-of-the-art AV-ASR system [7] to benchmark AV-ASR on the headset data against traditional frontal, full-face videos, collected both at an ideal studio-like environment and at a more challenging office domain. Of particular interest is the performance of the system visual front end, as well as its audio-visual fusion module. The first is measured by visual-only ASR accuracy, whereas the latter by the achieved relative word error rate reduction, compared to audio-only performance. In all cases, the recognition task considered is connected-digit ASR.

The paper is structured as follows: Section 2 reviews the main components of the AV-ASR system, whereas Section 3 briefly describes the audio-visual corpora considered in this work. Section 4 is devoted to the experimental study and comparison of AV-ASR performance across these datasets. Section 5 concludes the paper.

Figure 1: *First row: frontal and side view of the infrared headset; Second row: frontal and side view of a subject wearing the headset.*
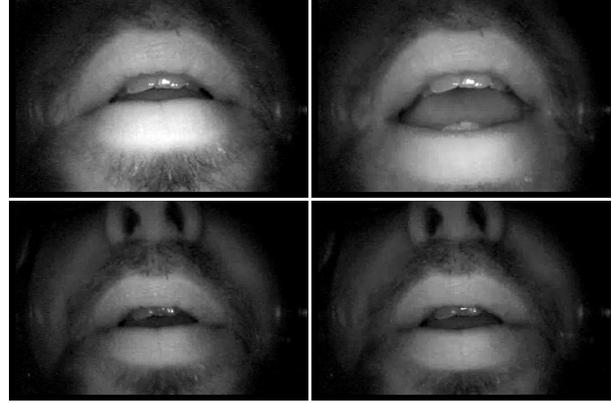

Figure 2: *First row: examples of close-up (CU) view video images; Second row: examples of wide view (WV) video images recorded using the infrared headset.*

2. The audio-visual ASR system

There are three main areas that differentiate AV-ASR systems [2]: The visual front end design, the audio-visual integration strategy, and the speech recognition method used. With respect to the first area, given video data, there exist three possibilities for visual speech representation [2, 7]: Appearance-based features that typically seek a suitable transform of the pixel values within a visual *region of interest* (ROI) [1, 7], shape-based features that consist of a geometric or statistical representation of the lip contours [3]-[5], and combination of the two strategies [3]. Concerning audio-visual integration, most methods fall within the feature or decision fusion framework. The former approach combines the speech information at the feature level and utilizes a single classifier for recognition [4, 7], whereas the latter combines the two single-modality classification decisions typically at the likelihood level [3]-[7]. Finally, a *hidden Markov model* (HMM) with Gaussian mixture emission probabilities [3, 7], or alternatively, an artificial neural network classifier [1, 5] can be used for AV-ASR. The system considered in this work employs appearance-based features based on the ROI *discrete cosine transform* (DCT), HMMs for ASR, and feature fusion (see also Figure 3).

In more detail, since the headset directly captures the mouth region, we extract a 64×64 pixel ROI by a simple truncation and subsampling of the original 720×480 pixel frame. Histogram equalization is then applied to the ROI to enhance its details (see Figures 2 and 4). In the CU setting, the ROI contains only the speaker's mouth region. The WV setting includes also some part of the nose and cheeks (see Figure 4).

Following ROI extraction, a two-dimensional, separable DCT is applied to the ROI, and the 100 highest-energy DCT coefficients are retained. To reduce dimensionality and improve discrimination among the speech classes, an *intra-frame linear discriminant analysis* (LDA) projection is applied, resulting in a 30-dimensional feature vector. This is followed by a *maximum likelihood linear transformation* (MLLT) that improves maximum likelihood based statistical data modeling [7]. To facilitate audio-visual fusion, linear interpolation is employed that synchronizes the features to the 100 Hz rate of their audio counterpart, whereas feature mean normalization is used to further compensate for lighting variations, providing the visual-only *static* features. Fifteen consecutive such features are then concatenated, and subsequently projected/rotated by means of an *inter-frame* LDA/MLLT combination, thus giving rise to *dynamic* visual features $\mathbf{o}_{v,t}$ of dimension 41 (see also Figure 3).

In addition to visual features, time-synchronous audio features are extracted at 100 Hz. First, 24 mel-frequency cepstral coefficients of the speech signal are computed over a sliding window of 25 msec, and are mean normalized to provide static features. Then, nine consecutive such frames are concatenated and projected by means of LDA and MLLT onto a 60-dimensional space, producing dynamic audio features $\mathbf{o}_{a,t}$ (see also Figure 3).

Following feature extraction, we consider a simple, well-known *feature fusion* strategy for AV-ASR, based on projecting the 101-dimensional concatenated audio-visual vectors $\mathbf{o}_{av,t} = [\mathbf{o}_{a,t}, \mathbf{o}_{v,t}]$ onto a 60-dimensional space by an LDA/MLLT [7]. Such features are then fed to a suitable HMM-based decoder for speech recognition.

3. The audio-visual databases

As mentioned in the introduction, most AV-ASR research has concentrated on databases collected in ideal visual conditions. One such set is the IBM ViaVoice$^{\text{TM}}$ audio-visual corpus, recorded in a quiet studio-like environment, with uniform lighting and background. The subjects' head pose remains frontal with little variation in the database, due to the use of a teleprompter that displays the dictation text. High-quality video is captured, and is MPEG2 encoded at 30 fps and a 704×480 pixel frame size. In addition to video, high quality wideband audio is synchronously collected at a rate of 16 kHz. A
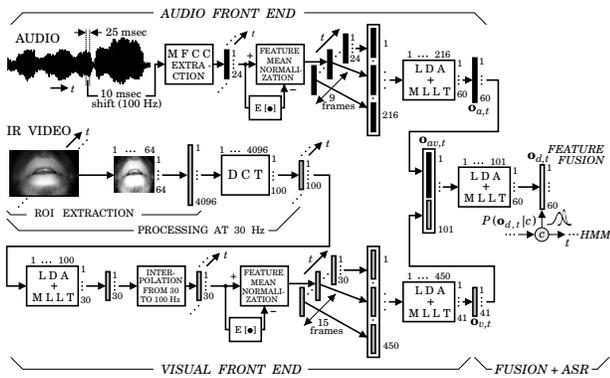
Figure 3: *Block diagram of our AV-ASR system [7]. Synchronous, 60-dimensional audio features, $\mathbf{o}_{a,t}$, and 41-dimensional visual observations, $\mathbf{o}_{v,t}$, are extracted, both at 100 Hz. Subsequently, a feature fusion strategy is used for speech recognition by hidden Markov models.*

Figure 4: *First row: examples of close-up view ROI images; Second row: examples of wide view ROI images.*

50-subject subset of this dataset ("STUDIO"), containing connected-digit utterances, is used here as a reference for audio-visual ASR under ideal conditions.

A second full-face database is captured using a laptop-based audio-visual data collection system. The system records 22 kHz audio using the built-in laptop microphone and uncompressed video by means of an inexpensive web-cam, utilizing the USB 2.0 interface. Compared to the STUDIO corpus, the video quality is now poorer, with automatic gain control present, and 30 fps available at only a 320×280 pixel size. In addition, the database subjects are recorded in their own offices without the use of a teleprompter, and thus, lighting, background, and head-pose vary greatly. A total of 109 subjects uttering connected digit strings are available in this set, which is referred to as "OFFICE".

The infrared headset data are also captured using the laptop-based audio-visual collection system. The audio and S-video signals are carried wirelessly by means of an RF transmitter housed in the earpiece of the headset. The transmitted wireless signal is picked up by a 2.4 GHz, 4-channel receiver, and fed into the laptop after DV-conversion via the Firewire interface. The system records audio at 22 kHz and video at 30 fps and a 720×480 pixel resolution. A total of 55 subjects uttering connected digit strings under the close-up view (CU) and wide view (WV) settings are available in two sets, which are referred to in Tables 1 and 2 as "IR(CU)" and "IR(WV)".

4. Recognition experiments

We now proceed to report a number of ASR experiments on the previously discussed databases. Each corpus is split into two sets, one for training HMMs and the remaining for testing ASR performance, as depicted in Table 1. Our training/testing paradigm is of a *multi-speaker* style, where separate data from *all* subjects are used for both training and testing.
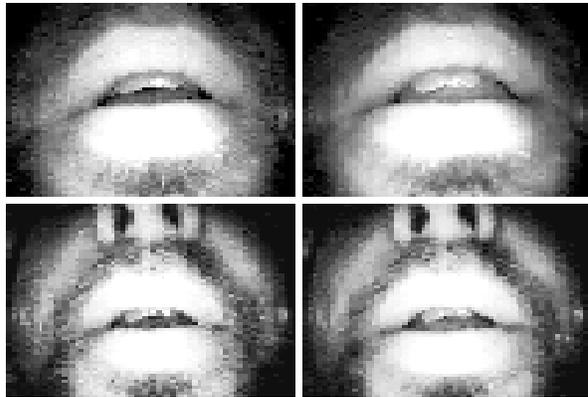
For connected digit ASR (11-word vocabulary), a two-stage stack decoding algorithm is employed, with unknown digit-string length. HMMs with 101 context dependent states, 1.7k Gaussian mixture components are used for all data sets. Compared to the STUDIO and OFFICE data sets in [6], we randomly chose much smaller training data sets in order to have fair comparison to the IR data sets. For all sets, in addition to the original database acoustic signal, audio-only and AV-ASR are also considered on artificially corrupted audio by additive "speech babble" at abound 7 dB SNR, using HMMs trained on the original data.

It is only natural to expect that using the infrared headset we could easily obtain a better and more accurate ROI than by face tracking, and thus better visually-only ASR performance, compared to the full-face video datasets STUDIO and OFFICE. Indeed (see Table 2), the visual-only *word error rate* (WER) improves by 30% relative, when moving from the STUDIO to the IR(CU) data (31.68%→22.08%) and by 54% as compared to the OFFICE data (47.94%→22.08%), due to the visually challenging nature of the latter [6]. The IR(WV) data give a somewhat smaller improvement (WER is 25.44%), thus suggesting that including more of the facial area is not

Database	Subj.	Set	Utter.	Dur.
STUDIO	50	Train	1540	2:15
		Test	395	0:35
OFFICE	109	Train	1522	2:02
		Test	439	0:35
IR(CU)	55	Train	1518	2:18
		Test	392	0:34
IR(WV)	55	Train	1497	2:07
		Test	381	0:31

Table 1: *Four audio-visual databases used in this paper for connected-digit ASR (studio, office, infrared close-up view, and infrared wide view). Their partitioning into training and test sets is depicted (number of utterances and duration (in hours) are shown for each set).*

Database	VI	Clean		Noisy	
		AU	AV	AU	AV
STUDIO	31.68	1.19	1.33	26.85	17.69
OFFICE	47.94	2.14	2.87	25.58	15.92
IR(CU)	22.08	3.36	2.50	23.92	11.06
IR(WV)	25.44	2.02	1.95	24.28	10.11

Table 2: *Single-modality (visual-only (VI), audio-only (AU)) and audio-visual (AV) ASR performance by means of feature fusion on the test sets of the four audio-visual databases of Table 1 (top-to-bottom: studio, office, infrared close-up view and infrared wide view). Results (in word error rate, %) in two acoustic conditions are considered: The original database audio (clean), and a degraded version (noisy) by additive babble noise. All HMMs are trained on the clean acoustic condition.*

beneficial for such data.

Notice that audio-only ASR on the IR(CU) and IR(WV) sets is worse than that on the STUDIO data. Indeed, the WER degrades from 1.19% to 3.36% for the IR(CU) data, and to 2.02% for the IR(WV) data. This is probably due to the fact that all systems are bootstrapped from a clean digit system which matches more to the STUDIO data than to the IR data sets. In addition, the possibly worse quality of the audio signal due to its wireless transmission and presence of breathing noise may contribute to performance degradation.

As expected, the significantly improved visual-only ASR on the headset data translates to increased relative benefit of the visual modality to bimodal recognition, both in the clean and noisy acoustic conditions. Indeed, in the former, by using audio-visual feature fusion, there is no improvement for the STUDIO data (the results for STUDIO AV feature fusion are different in [6] when more training data are available). However, the improvement in the IR(CU) data is dramatically high (3.36%→2.50%), i.e., 25.6% relative, whereas the improvement in the IR(WV) data is from 2.02% to 1.95%, i.e., 2.5% relative. A similar observation holds in the noisy condition.

Compared to the OFFICE data, the benefit of using the infrared headset is even greater. Instead of gaining from the visual information, combining inaccurate visual features of the OFFICE data with audio features by means of feature fusion degrades ASR performance in the clean condition [6]. However, the visual modality remains of benefit to ASR in the noisy condition (improves the WER from 25.58% to 15.92%). But the relative improvement from audio-only to AV-ASR remains significantly smaller than on both IR datasets.

5. Conclusions

We investigated how a specially designed infrared headset would benefit audio-visual ASR. The headset bypasses the need for face tracking, thus significantly improving the robustness of visual speech feature extraction, while reducing computation and increasing the speed in AV-ASR implementations. We benchmarked the visual front end performance against a typical "visually clean" data domain, as well as a "visually challenging" data set with face tracking employed. Our results demonstrate that the infrared headset dramatically improves visual-only ASR and boosts the contribution of the visual modality to audio-visual recognition. In the future, we plan to investigate the ASR benefit of additional visual features, such as mouth opening and tongue visibility, that we believe can be easily extracted from the high-resolution ROI images provided by the infrared headset.

6. Acknowledgements

7. References

[1] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," Proc. Int. Conf. Spoken Lang. Process., pp. 547–550, 1994.

[2] M.E. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in Speechreading by Humans and Machines, D.G. Stork and M.E. Hennecke, Eds. Berlin: Springer, pp. 331–349, 1996.

[3] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Trans. Multimedia, 2(3):141–151, 2000.

[4] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," IEEE Signal Processing Mag., 18(1):9–21, 2001.

[5] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," EURASIP J. Appl. Signal Process., 2002(11):1260–1273, 2002.

[6] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," Europ. Conf. Speech Commun. Technol., 2003, to appear.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," To Appear: Proc. IEEE, 2003.

[8] A. Adjoudani, T. Guiard-Marigny, B. Le Goff, L. Reveret, and C. Benoît, "A multimedia platform for audio-visual speech processing," Proc. Europ. Conf. Speech Commun. Technol., pp. 1671–1674, 1997.