

Audio-Visual ASR from Multiple Views inside Smart Rooms

Gerasimos Potamianos*

Human Language Technology Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
gpotam@us.ibm.com

Patrick Lucey**

Speech, Audio, Image and Video Research Lab
Queensland University of Technology
Brisbane, QLD 4001, Australia
p.lucey@qut.edu.au

Abstract— Visual information from a speaker’s mouth region is known to improve automatic speech recognition robustness. However, the vast majority of audio-visual automatic speech recognition (AVASR) studies assume frontal images of the speaker’s face, which is not always the case in realistic human-computer interaction (HCI) scenarios. One such case of interest is HCI inside smart rooms, equipped with pan-tilt-zoom (PTZ) cameras that closely track the subject’s head. Since however these cameras are fixed in space, they cannot necessarily obtain frontal views of the speaker. Clearly, AVASR from non-frontal views is required, as well as fusion of multiple camera views, if available. In this paper, we report our very preliminary work on this subject. In particular, we concentrate on two topics: First, the design of an AVASR system that operates on profile face views and its comparison with a traditional frontal-view AVASR system, and second, the fusion of the two systems into a multi-view frontal/profile system. We in particular describe our visual front end approach for the profile view system, and report experiments on a multi-subject, small-vocabulary, bimodal, multi-sensory database that contains synchronously captured audio with frontal and profile face video, recorded inside the IBM smart room as part of the CHIL project. Our experiments demonstrate that AVASR is possible from profile views, however the visual modality benefit is decreased compared to frontal video data.

I. INTRODUCTION

Over the past two decades, considerable research activity has concentrated on utilizing visual speech extracted from a speaker’s face in conjunction with the acoustic signal, in order to improve robustness of *automatic speech recognition* (ASR) systems [1]. Even though a great deal of progress has been achieved in *audio-visual ASR* (AVASR), so far the vast majority of works in the field have focussed on using video of a speaker’s fully frontal face. This however is a rather strong assumption, unrealistic in typical human-computer interaction scenarios.

One such scenario is *meeting* or *lecture* events inside *smart rooms* [2], [3] that are equipped with a number of far-field audio-visual sensors, including microphone arrays, fixed and *pan-tilt-zoom* (PTZ) cameras. This scenario is of central interest in the “*Computers in the Human Interaction Loop*” (CHIL) integrated project currently funded by the European Union [4]. A schematic diagram of one of the smart rooms developed for this project, in particular the one located at IBM Research, is depicted in Fig. 1. Clearly, audio-visual

* This work was supported by the European Commission under the integrated project CHIL, “Computers in the Human Interaction Loop”, contract number 506909.

** Work performed during Patrick Lucey’s internship with the IBM T.J. Watson Research Center.

speech technologies, such as speech activity detection, source separation, and speech recognition, are of prime interest in this scenario, due to overlapping and noisy speech, typical in multi-person interaction, captured by far-field microphones. Data from the smart room fixed cameras are of insufficient quality to be used for this purpose, as they typically capture the participants’ faces in low resolution (see also Fig. 2). On the other hand, video captured by the PTZ cameras can provide high resolution data, assuming that successful active camera control is employed, based on tracking the person(s) of interest [5]. Nevertheless, since the PTZ cameras are fixed in space, they cannot necessarily obtain frontal views of the speaker. Clearly therefore, AVASR from non-frontal views is required in this scenario, as well as fusion of multiple camera views, if available. In this paper, we report our very preliminary work on the subject, namely the design of an AVASR system that operates on profile face views instead of the traditional frontal-view AVASR. In addition, we are particularly interested in comparing system performance between the two, as well as fusing the two systems into multi-view frontal/profile AVASR.

In the literature, there is surprisingly little work on the subject of visual speech from side views: we have found only three relevant studies. In the first paper, Yoshinaga et al. [6] extracted lip information from the horizontal and vertical variances of the mouth image optical flow. In this paper, no mouth detection or tracking was performed. In [7], Yoshinaga et al. refined their system by incorporating a mouth tracker, which utilizes Sobel edge detection and binary images, and used the lip angle and its derivative as visual features on a limited data set. The improvement sought from these primitive features was minimal as expected, essentially due to the fact that only two visual features were used, compared to most other frontal-view systems that utilize significantly more features [1]. The third study was a comprehensive psychological study conducted by Jordan and Thomas [8]. Their findings were rather intuitive, as the authors determined that human identification of visual speech became more difficult as the angle (from frontal to profile view) increased. To the best of our knowledge, no other effort to solve this particular problem has been made. As such, we believe our paper to be the first real attempt in determining how much visual speech information can be automatically extracted from profile views, and to compare this with visual speech information obtained from frontal images.

The task of recognizing visual speech from profile views is in principle very similar to that from frontal views, requiring

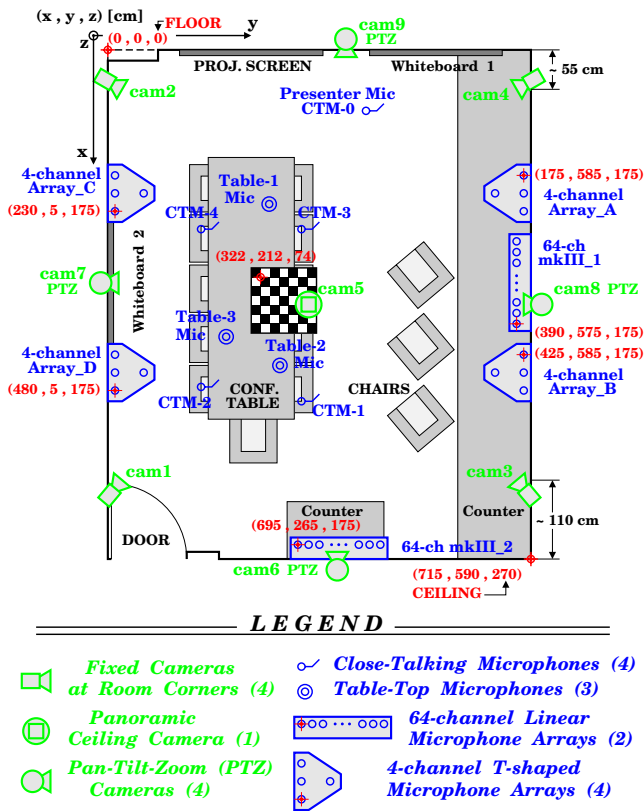


Fig. 1. The IBM smart room developed for the purpose of the CHIL project. Notice the fixed and PTZ cameras, as well as the far-field table-top and array microphones.

to first detect and track the mouth region and subsequently to extract visual features. However, the problem is far more complicated than in the frontal case, because the facial features required to detect and track the mouth lie in a much more limited spatial plane, as can be seen in Fig. 3. Clearly, much less data is available compared to that of a fully frontal face, since many of the facial features of interest (eyes, nostrils, mouth, chin area, etc.) are fully or partially occluded. In addition, the search region for all visible features is approximately halved, as the remaining features are compactly confined within the profile facial region. These facts remove redundancy in the facial feature search problem, and therefore make robust mouth tracking a much more difficult endeavor.

Nevertheless, one can still achieve mouth region tracking by employing techniques similar to frontal facial feature detection. In particular, in the AVASR literature, most systems use appearance-based methods for face and facial feature detection. Some are based on “strong” classifiers, such as neural networks [9], support vector machines (SVMs) [10], eigenfaces [1], hidden Markov models (HMMs) [11], or Gaussian mixture models (GMMs) [12], and others utilize cascades of “weak” classifiers, such as the Adaboost framework of Viola and Jones [13], later extended by Leinhardt and Maydt [14]. In this paper, we use the latter approach to first detect the face and subsequently the facial features in profile views, as described in more detail in Section II.

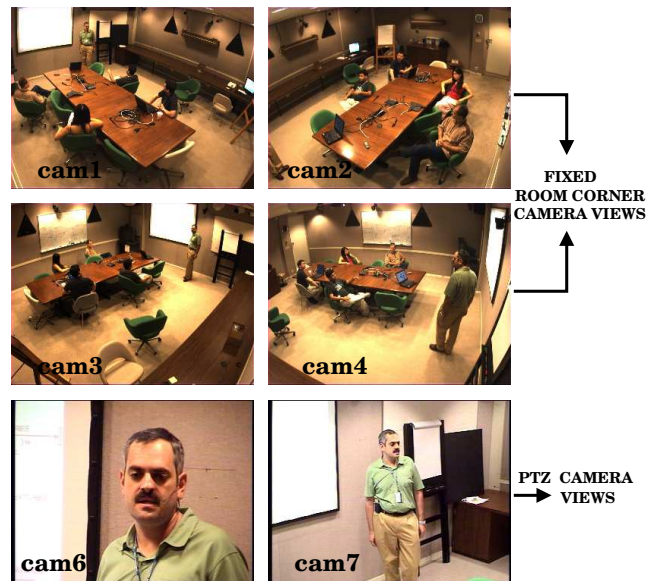


Fig. 2. Examples of image views captured by the IBM smart room cameras. In contrast to the four corner cameras (two upper rows), the two PTZ cameras (lower row) provide closer views of the lecturer, albeit not necessarily frontal (see also Fig. 1).

Following that, Section III focuses on the AVASR system description, namely visual feature extraction based on the tracked mouth region and audio-visual fusion. In addition, details of a number of systems used in our experiments are given, including a baseline frontal-view AVASR system that has been refined in our previous work [1]. Section IV presents our experimental results, and, finally, Section V concludes the paper with a summary and a few remarks.

II. MOUTH DETECTION AND REGION-OF-INTEREST EXTRACTION FROM PROFILE VIEWS

For the task of mouth detection and *region-of-interest* (ROI) extraction, we devised a similar strategy to that of Cristinacce et al. [15], employing a boosted cascade of classifiers based on simple Haar-like features to detect the face and subsequently the facial features. These classifiers were generated using OpenCV libraries [16].

The positive examples used for training these classifiers were obtained from a set of 847 training images, with 17 manually labeled points for each face. Due to the compactness of the facial features within the dataset, we initially

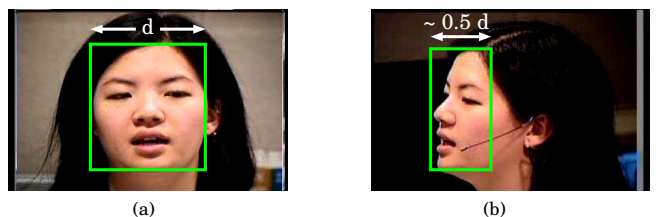


Fig. 3. Synchronous (a) frontal and (b) profile views of a subject recorded in the IBM smart room (see Section IV). In the profile view, visible facial features are “compacted” within approximately half the area compared to the frontal face case, thus making their tracking more difficult.

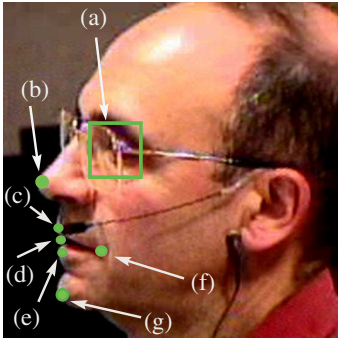


Fig. 4. Labeled facial features: (a) left eye, (b) nose, (c) top mouth, (d) mouth center, (e) bottom mouth, (f) left mouth corner, and (g) chin.

utilized only seven of the 17 annotated points, namely the left eye, nose, top of the mouth, mouth center, bottom of the mouth, left mouth corner, and chin, as depicted in Fig. 4. This provided 847 positive examples for all seven facial features. Approximately 5000 negative examples were used for each facial feature. These negative examples consisted of images of the other facial features that surrounded its location, as these areas would be the most likely to cause false alarms. The face training set was further augmented by including rotations in the image plane by ± 5 and ± 10 degrees, providing 4235 positive examples. A similar amount of negative examples of the background were also employed in the training scheme. For the facial features however, since they were located close to each other, we opted not to include rotated examples in their training.

A dilemma we experienced was on selecting appropriate facial feature points to use for image normalization. In the frontal face scenario, eyes are predominately considered for this task, but in the profile-view case we don't have the luxury of choosing two geometrically aligned features. We instead chose to use the nose and the chin, with a normalized constant distance of $K = 64$ pixels between them. This way, the problem of head pose variation was minimized, compared to the other possibilities (such as employing the eye-to-nose distance, etc.). The top mouth, center mouth, bottom mouth, and left mouth corner were trained on templates of size 10×10 pixels, based on normalized training faces. Both nose and chin classifiers were trained on templates of size 15×15 pixels, whereas the eye templates were somewhat larger, at 20×20 pixels. For face detection, the positive face examples were normalized to 16×16 pixels (see also Fig. 5).

To judge performance of the adopted scheme, all classifiers were tested on a small validation set of 37 images. This provided us with an indication of what particular features can be most reliably tracked. Table I depicts the detection accuracy of the seven facial feature classifiers, with a feature considered detected, if the location error is less than 10% of the annotated nose-to-chin distance. Clearly, the left eye and left mouth corner yielded the best results, therefore we decided to use them for scale normalization during testing. Compared to using the nose and chin (as discussed in the previous paragraph), this amounts to changing the scaling factor K from 64 to 45. Concerning face detection accuracy,

TABLE I
DETECTION ACCURACY, %, FOR SEVEN FACIAL FEATURES OF INTEREST,
REPORTED ON A SMALL VALIDATION SET (37 ANNOTATED IMAGES).

FACIAL FEATURE	ACCURACY (%)
Left Eye	86.49
Nose	81.08
Top Mouth	78.37
Center Mouth	81.08
Bottom Mouth	72.97
Left Mouth Corner	86.49
Chin	62.16

all 37 faces were correctly located in the selected validation set.

The entire profile mouth detection and tracking employed in our AVASR system is outlined in Fig. 5. Given the video of a spoken utterance, face detection is first applied to estimate the location of the speaker's face at different scales, since the face size is unknown. Once the face is detected, the search for the left eye and nose commences over specific regions of the face, based on training data statistics. While developing the system, we observed that the lower boundary of the face bounding box was often inaccurate, being far below or well above the bottom of the subject's actual face. As the face box defines the search region for the various facial features, this caused the system to miss detecting the lower regions of the face. To overcome this, we used the ratio (*metric1*) of the vertical eye-to-nose distance over the vertical nose-to-lower face boundary distance. If *metric1* was lower than a fuzzy threshold, as determined by training statistics, the lower part of the box was lengthened, whereas if it was greater than the threshold, the lower part was shortened. We observed that this greatly improved the detection of the mouth area (trained on normalized 32×32 mouth images), which was located next. This procedure is illustrated in Fig. 6(b).

Once the general mouth region is found, the left mouth corner is detected. The next step is to define a scaling metric, so that all ROI images get normalized to the same size. As mentioned previously, the ratio (*metric2*) of the vertical left eye to left mouth corner distance over constant $K = 45$ is used to achieve this (see Fig. 6). A $(48 \times 48) \cdot \text{metric2}$ pixel normalized ROI, based on the left mouth corner is then extracted (see Fig. 6). The ROI is then subsequently downsampled to 32×32 pixels, for use in the AVASR system (see Section III).

Following ROI detection, the ROI is tracked over consecutive frames. If the detected ROI is located too far away from the previous frame, this is regarded as a detection failure and the previous ROI location is used instead. A mean filter is then used to smooth the tracking. Due to the real-time speed of the boosted cascade of classifiers, this detection and tracking scheme is used for every frame.

Overall, the accuracy of the ROI detection and tracking system was very good, with only a very few number of poorly or mistracked ROIs in the dataset. A major factor affecting performance was due to random head movement

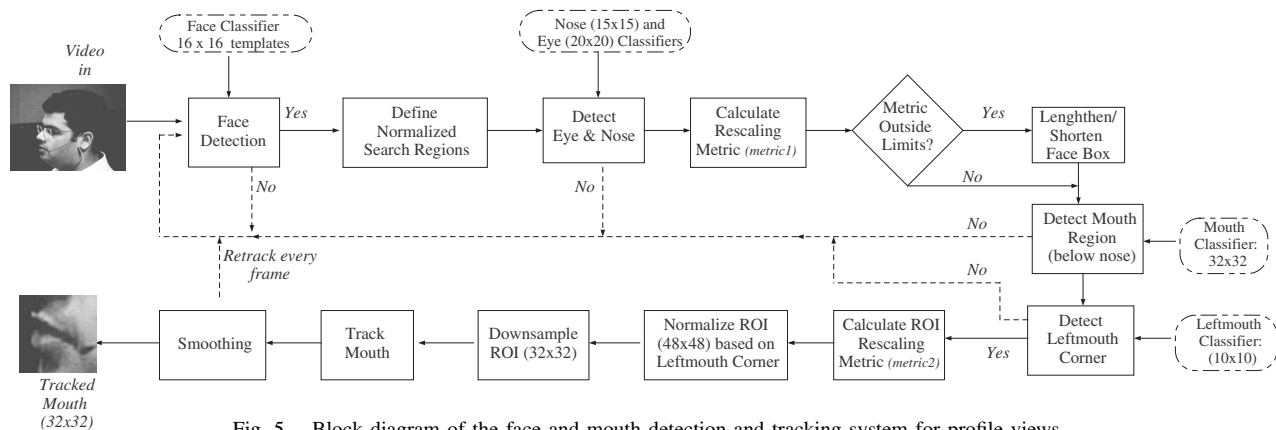


Fig. 5. Block diagram of the face and mouth detection and tracking system for profile views.

and some head pose variability, where subjects exhibit a somewhat more frontal pose than the profile view of the majority of the subjects – see also Fig. 7, where examples of accurately and poorly tracked ROIs are depicted. The latter is also the reason why we were not able to employ any rotation normalization. Many different configurations were experimented with, however they seemed to cause more problems than they solved. For example, we tried rotating the ROI according to the angle between the left eye and the left mouth corner, however the head pose variation in the data made this problematic. Another attempt was made to rotate the ROI employing the angle between the mouth center and the left mouth corner. This also failed, as the distance between these two points was too small (around 20 pixels), and any slight detection inaccuracy caused large rotation errors.

III. THE AVASR SYSTEM

We now proceed to briefly describe the remaining components of the AVASR system, following detection of the mouth ROI. There exist two main such components, overviewed in the next two subsections: (a) feature extraction, which includes the visual features that complete the visual front end sub-system, and of course the audio feature extraction step; and (b): the audio-visual fusion (integration) step. In this work, neither component exhibits significant differences between the introduced profile-view AVASR system and our baseline frontal-view AVASR system refined in previous work [1]. These systems will be compared in Section IV.B. Furthermore, performance of a combined AVASR system that uses *both* profile and frontal views will also be discussed there. Specifics of all three AVASR systems are briefly overviewed in Section III.C.

A. Feature Extraction

Following ROI extraction, a two-dimensional, separable, *discrete cosine transform* (DCT) is applied to it, with the 100 top-energy DCT coefficients retained. The resulting 100-dimensional vectors are available at the video rate (30 Hz). In order to simplify integration with audio and to improve system robustness, the vectors are interpolated to

the audio feature frame rate of 100 Hz, and are mean-normalized, independently over each utterance. Furthermore, for dimensionality reduction, an *intra-frame* cascade of *linear discriminant analysis* (LDA) followed by a *maximum-likelihood linear transform* (MLLT) is applied, resulting to 30-dimensional “static” visual features. Subsequently, to incorporate dynamic speech information, 15 neighboring such features over ± 7 adjacent frames are concatenated, and are projected via an *inter-frame* LDA/MLLT cascade to 41-dimensional “dynamic” visual feature vectors. More details can be found in [1].

In parallel to visual feature extraction, 24-dimensional *mel-frequency cepstral coefficients* (MFCCs) are extracted at a 100 Hz frame rate, based on the audio signal. After mean normalization, the features are processed by an inter-frame LDA/MLLT cascade over ± 5 frames to produce 60-dimensional acoustic features.

B. Audio-Visual Integration

Following feature extraction, time-synchronous audio and visual features are available at 100 Hz with dimensions 60 and 41, respectively. In this work, we consider two commonly used audio-visual integration techniques [1]. The first one is *feature fusion*, where the bimodal feature vectors are concatenated, resulting in our case to 101-dimensional features that are subsequently projected onto 60 dimensions using an LDA/MLLT cascade (note that this equals the audio feature vector dimensionality). The second is *decision fusion*, based on multi-stream HMMs. The latter method typically yields significantly better results than the feature fusion technique, but requires optimizing the modality integration weights (typically on held-out data). Notice that these fusion mechanisms will also be used in our experiments to combine the profile- and frontal-view visual-only ASR (lipreading) systems into a “multi-view” lipreading system, as discussed next.

C. The Speech Recognition Systems

In our experiments below, we will be comparing three AVASR systems: The introduced profile-view AVASR system, a baseline AVASR system based on frontal views [1], and a combination of the two, namely a “multi-view” AVASR

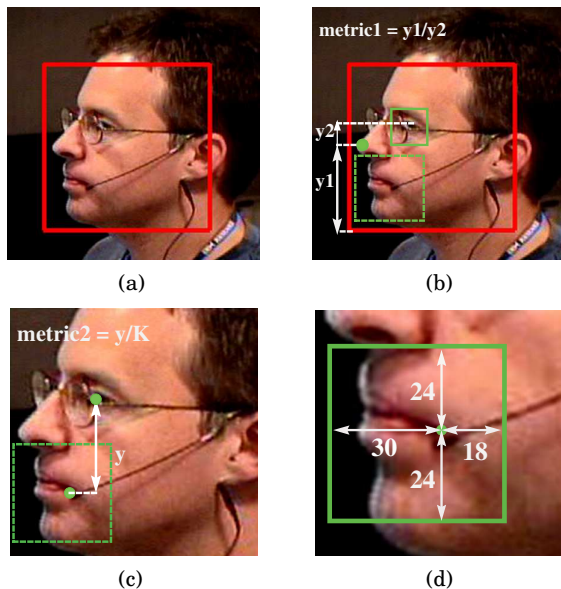


Fig. 6. (a) An example of face detection. (b) Based on the face detection result, a search area is obtained to detect the left eye and nose. The face bounding box is lengthened or shortened according to $metric1$. (c) The left mouth corner is detected within the general mouth region. The ratio ($metric2$) is then used for normalizing the ROI. (d) An example of the scaled normalized detected ROI of size $(48 \times 48) \cdot metric2$ pixels.

system. Furthermore, audio-only and visual-only systems will also be compared. All such systems are designed in this work to recognize connected-digit sequences (10-word vocabulary with no grammar). All single-stream HMMs are trained by employing the expectation-maximization algorithm over an available training set (see Section IV.A), and have an identical topology, containing 104 context-dependent states and approximately 1.7k Gaussian mixture components. For multi-stream HMM based AVASR, the audio and visual stream HMMs are separately trained and then combined using fixed integration weights, that are optimized to minimize the word error rate on held-out data.

Before moving on to our experiments, we should emphasize a few differences between the compared systems: The “multi-view” visual-only system operates on 60-dimensional visual features that result from an LDA/MLLT cascade applied on the concatenated single-view (frontal + profile) visual-only feature vectors having a combined dimension of 82 (=41+41). This is in contrast to the single-view (profile, or frontal) visual-only systems, that use 41-dimensional features. In addition, one should note that the visual front end sub-systems of the frontal- and profile-view AVASR systems differ in two aspects: One concerns the face and mouth region tracking algorithm, where the frontal view system tracking is based on a set of “strong” classifiers, as described in detail in [1], [12]. The second difference lies on the size of the extracted ROIs, before DCT feature extraction is applied. It is 32×32 pixels for the profile-view system, but 64×64 pixels for frontal-view AVASR.

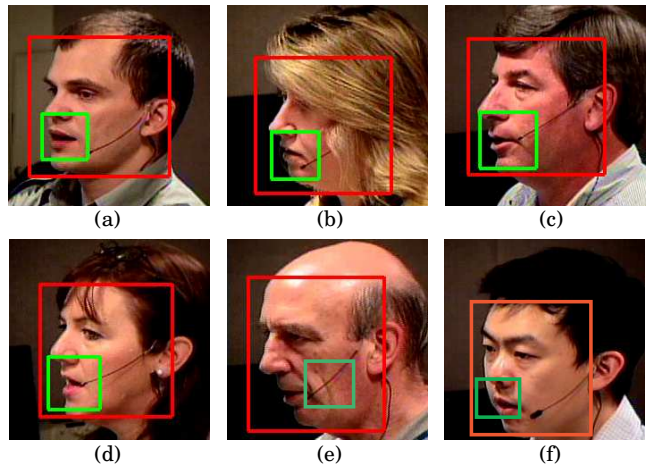


Fig. 7. Examples of accurate (a-d) and inaccurate (e,f) results of the detection and tracking system. In (f), it can be seen that the subject exhibits a somewhat more frontal pose compared to the profile view of the other subjects.

IV. EXPERIMENTAL RESULTS

We now proceed to report a number of experimental results on the performance of the developed profile-view AVASR system. The experiments are conducted on a multi-sensory audio-visual database, recorded in the IBM smart room, that is briefly described next.

A. The Audio-Visual Database

A total of 38 subjects uttering connected digit strings have been recorded inside the IBM smart room, using two microphones and three pan-tilt-zoom (PTZ) cameras. Of the two microphones, one is head-mounted (close-talking channel – see also Fig. 3) and the other is omni-directional, located on a wall close to the recorded subject (far-field channel). The three PTZ cameras record frontal and two side views of the subject, and feed a single video channel into a laptop via a quad-splitter and an S-video-to-DV converter. As a result, two synchronous audio streams at 22kHz and three visual streams at 30 Hz and 368×240 -pixel frames are available. Among these available streams, in the reported experiments we utilize the far-field audio channel and two video views: the frontal and one of the two side views, namely the one that consistently provides views closest to the profile pose (see also Fig. 3). A total of 1661 utterances are used in our experiments, partitioned using a multi-speaker paradigm into 1247 sequences for training (1 hr 51 min in duration), 250 for testing (23 min), and 164 sequences (15 min) that are allocated to a held-out set. The test set reference contains 2155 digit words.

B. Recognition Results

In the first experiment, we report the *visual-only* system performance on this dataset. The *word error rate* (WER) of the baseline frontal-view system [1] on the test set is 25.4%. In contrast, the developed profile-view system achieves a significantly worse performance of 39.9% WER, a relative degradation of about 60% compared to the frontal view

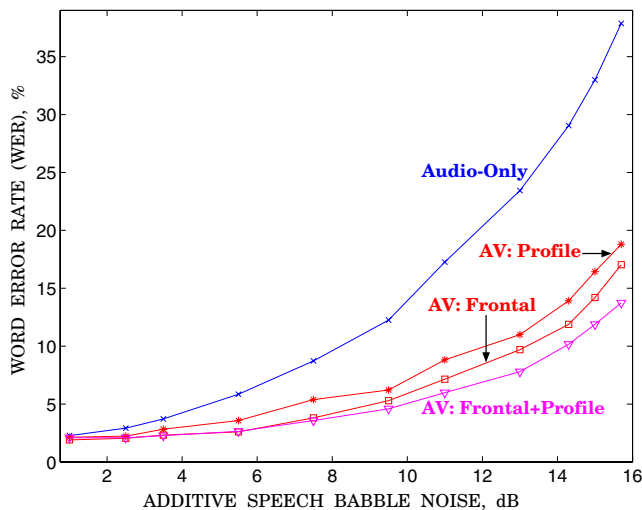


Fig. 8. Test set WER, %, of audio-only and audio-visual ASR. Three AVASR systems are depicted, based on profile view, frontal view, and both views (“multi-view” system). Additive speech babble noise at various dBs has been applied on the far-field audio channel.

results. Nevertheless, the profile system is clearly capable of recognizing visual speech, but of course less so than the frontal system, in line with human lipreading experiments reported in [8]. Interestingly, by combining the two systems using feature fusion, the resulting “multi-view” visual-only performance becomes 23.7%, which demonstrates that there may exist information in the profile view, not captured by the frontal-view system (possibly that of lip protrusion).

When combining the three systems with the far-field audio channel using a two-stream HMM (decision fusion), the *audio-only* system performance of 1.62% WER improves somewhat to an *audio-visual* WER of 1.53%, 1.53%, and 1.48%, when incorporating the frontal-, profile-, and “multi-view” visual information, respectively. These differences are however not significant due to the small database size. Of course, they become more pronounced, if we corrupt the audio channel by additive noise; in our experiments, “speech babble” is used for this purpose. The results are depicted in Fig. 8, and have been obtained using feature fusion to simplify and speed up the experiments (no optimization of integration weights is required). These results further verify the experimental observations of the previous paragraph. As expected, in high noise environments, the visual modality benefit to ASR is dramatic, with the “multi-view” system demonstrating the biggest gains, mostly due to the contribution of the frontal view video – especially for the low noise region. Interestingly, the profile view system, although lagging compared to the frontal view one, is still capable of providing much of the visual modality benefit to ASR. We view this result as very encouraging for AVASR applications in scenarios where a frontal view cannot always be guaranteed.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an AVASR system capable of extracting visual speech information from profile views. To

our knowledge, this is the first serious attempt to lipreading from side views that allows quantifying the performance degradation as compared to lipreading from the traditional frontal view of the speaker’s mouth. In our experiments, we demonstrated that profile views contain significant visual speech information, sufficient to improve ASR robustness to noise. Such benefit is of course less pronounced than when using the frontal view, however is not totally redundant to the frontal video, as the “multi-view” experiments demonstrated.

In further work, we will extend these experiments to more complex recognition tasks, such as connected letters, alpha-digits, and large-vocabulary speech, to verify whether our conclusions generalize well. We have already collected the appropriate multi-view database for such experiments and plan to report results soon. We view such work as the first step towards head-pose independent AVASR, which we believe may open the field to natural human-computer interaction scenarios in environments such as the CHIL smart rooms.

REFERENCES

- [1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, “Recent advances in the automatic recognition of audio-visual speech,” *Proc. of the IEEE*, vol. 91, no. 9, 2003.
- [2] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, “Multimodal multispeaker probabilistic tracking in meetings,” in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, 2005.
- [3] A. Pentland, “Smart rooms, smart clothes,” in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, vol. 2, 1998.
- [4] CHIL: Computers in the Human Interaction Loop, <http://chil.server.de>
- [5] Z. Zhang, G. Potamianos, S. M. Chu, J. Tu, and T. S. Huang, “Person tracking in smart rooms using dynamic programming and adaptive subspace learning,” in *Proc. Int. Conf. on Multimedia and Expo*, 2006, pp. 2061–2064.
- [6] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, “Audio visual speech recognition using lip movement extracted from side-face images,” in *Proc. Auditory Visual Speech Processing (AVSP)*, 2003, pp. 117–120.
- [7] —, “Audio visual speech recognition using new lip features extracted from side-face images,” in *Proc. ROBUST2004*, 2004.
- [8] T. R. Jordan and S. M. Thomas, “Effects of horizontal viewing angle on visual and audiovisual speech recognition,” in *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 6, 2001, pp. 1386–1403.
- [9] H. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, 1998.
- [10] L. Liang, X. Liu, Y. Zhao, X. Pi, and A. Nefian, “Speaker independent audio-visual continuous speech recognition,” in *Proc. Int. Conf. on Multimedia and Expo*, vol. 2, August 2002, pp. 25–28.
- [11] A. Nefian and M. Hayes, “Face detection and recognition using hidden Markov models,” in *Proc. Int. Conf. on Image Processing*, 1998, pp. 141–145.
- [12] J. Jiang, G. Potamianos, H. Nock, G. Iyengar, and C. Neti, “Improved face and feature finding for audio-visual speech recognition in visually challenging environments,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 5, 2004, pp. 873–876.
- [13] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.
- [14] R. Leinhardt and J. Maydt, “An extended set of Haar-like features,” in *Proc. Int. Conf. on Image Processing*, 2002, pp. 900–903.
- [15] D. Cristinacce, T. Cootes, and I. Scott, “A multi-stage approach to facial feature detection,” in *15th British Machine Vision Conference, London, England, 2004*, pp. 277–286.
- [16] Open Source Computer Vision Library, www.intel.com/research/mrl/research/opencv