TOWARDS PRACTICAL DEPLOYMENT OF AUDIO-VISUAL SPEECH RECOGNITION

G. Potamianos, C. Neti, J. Huang, J.H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, J. Jiang *

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
http://www.research.ibm.com/AVSTG

ABSTRACT

Much progress has been achieved during the past two decades in audio-visual automatic speech recognition (AVASR). However, challenges persist that hinder AVASR deployment in practical situations, most notably, robust and fast extraction of visual speech features. We review our effort in overcoming this problem, based on an appearance-based visual feature representation of the speaker's mouth region. In particular: (a) We discuss AVASR in realistic, visually challenging domains, where lighting, background, and head-pose vary significantly. To enhance visual-front-end robustness in such environments, we employ an improved statistical-based face detection algorithm, that significantly outperforms our baseline scheme. However, visual-only recognition remains inferior to visually "clean" (studio-like) data, thus demonstrating the importance of accurate mouth region extraction. (b) We then consider a wearable audio-visual sensor to directly capture the mouth region, thus eliminating face detection. Its use improves visual-only recognition, even over full-face videos recorded in the studio-like environment. (c) Finally, we address the speed issue in visual feature extraction, by discussing our real-time AVASR prototype implementation. The reported progress demonstrates the feasibility of practical AVASR.

1. INTRODUCTION

Motivated by the bimodality of human speech perception and the need for robust automatic speech recognition in noisy environments, significant research effort has recently been directed into the study of *audio-visual automatic speech recognition* (AVASR) [1–5]. In addition to the acoustic input, AVASR utilizes speech information present in the speaker's mouth region, and has been successfully demonstrated to improve the accuracy and noise robustness of ASR systems for both small- and large-vocabulary tasks. For example, bimodal ASR of a small-vocabulary task under speech babble noise at 0 dB *signal-to-noise ratio* (SNR) is reported in [5] to achieve the performance of an acoustic-only recognizer at 10 dB, i.e., to provide an "effective" SNR gain of 10 dB in the ASR "usable" range. Significant gains are also reported on the same task even in clean acoustic conditions. Furthermore, an 8 dB "effective" SNR gain is demonstrated for *large-vocabulary continuous speech recognition* (LVCSR) [5].

In spite of such impressive benefits however, AVASR systems have yet to be deployed in real-life applications. This is mainly due to issues related to the extraction of visual speech information, most notably *robustness* and *computational complexity* of visual front end processing. Regarding the former, most research work has concentrated and reported on databases recorded under controlled, visually "clean" conditions [1–4]. Such sets contain high-resolution video of the subjects' full frontal face, with very limited variation in head pose and subject-camera distance, rather uniform lighting, and, in most cases, constant background. In contrast, little is known about AVASR performance in realistic, non-ideal environments, where the visual channel quality is poor, thus presenting challenges to speech-informative visual feature extraction. Preliminary experiments reported in [6] show significant degradation of the visual modality ASR benefit in such visually challenging domains, for example videos recorded in moving automobiles, or by low quality web-cams. Concerning the computational requirements of visual front end processing, many approaches such as lip-contour geometric or statistical representations are intensive, thus not being amenable to real-time AVASR implementation [2].

In this paper, we review our ongoing research effort to overcome the above mentioned challenges, employing an appearance-based feature representation of the visual *region-of-interest* (ROI), which contains speech information [5]. With respect to visual front end robustness, we present two directions of work: The first assumes that full-face video data are available, and, therefore, it requires face and facial feature detection in order to accurately extract the ROI. In particular, we consider three audio-visual domains that represent increasing challenges to robust ROI estimation: A database collected in a studio-like setting, data recorded in typical offices using an inexpensive web-cam, and a set collected in stationary and moving automobiles. We compare the visual front end performance of our AVASR system on the three sets, as measured by face detection accuracy, visual-only ASR accuracy, and bimodal ASR improvement. To enhance visual front end robustness, we utilize a recently improved statistical-based face and facial feature detection algorithm [7]. Our experiments show it to significantly benefit visual front end performance compared to a baseline scheme [8], however ROI extraction remains unrobust in the most challenging automobile environment. This motivates a second direction of work, that utilizes a specially designed audio-visual headset to capture the video of the speaker's mouth region, independently of subject movement and head pose [9]. Such headset eliminates the need for face detection, thus providing the desired visual front end robustness and computational efficiency [10]. Not surprisingly, its use improves visual-only recognition even over studio-like full-face videos.

With respect to the visual front end computational requirements, we review the real-time implementation of our AVASR prototype [11]. On the average, the required visual processing computations utilize approximately 67% of a Pentium 4, 1.8 GHz processor, thus achieving better than real-time performance. In the case of the audio-visual headset, such load is significantly less.

The remainder of the paper is structured as follows: Section 2 reviews the main AVASR components, mostly focusing on visual front end processing for full-face and headset-captured videos. Section 3 discusses its real-time implementation, whereas Section 4 compares AVASR across a number of audio-visual databases. Finally, Section 5 concludes the paper.

2. THE AVASR COMPONENTS

Compared to audio-only speech recognition, AVASR introduces two new tasks: Extraction of speech informative visual features, followed by their integration with the traditional acoustic features into the ASR framework. In this section, we review both components, with our primary focus on the visual front end.

2.1. The visual front end

Three are the main visual speech representation approaches in the literature [2]: Appearance-based features that typically seek a suitable transform of the pixel values within a visual region-of-interest (ROI) [1,5], shape-based features that consist of a geometric or statistical representation of the lip contours [3], and combination of the two strategies [3]. In our AVASR work [5–7, 9–11], we utilize visual features that belong in the first category. We believe that such features successfully capture visual speech activity present in the mouth cavity and surrounding face region, which is difficult to describe by means of high-level (shape-based) features and

* With the House Ear Institute, Los Angeles, CA 90057, USA.

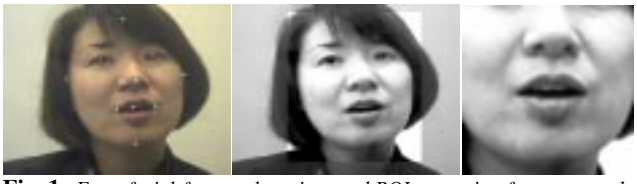

Fig. 1. *Face, facial-feature detection, and ROI extraction for an example full-face video frame. Left-to-right: Original frame with eleven detected facial features super-imposed; face-area enhanced frame; normalized ROI.*

a lip-contour encoding alone. In addition, appearance feature extraction requires only a gross estimation of the visual ROI, with obvious advantages in speed and robustness, as compared to more involved facial shape tracking. ROI extraction is the central part of the visual front end in our AVASR system, and we describe it in more detail next. As we mentioned in the introduction, we consider two types of visual data as input to our bimodal recognizer: Frontal, full-face videos, captured by a single camera in various environments, for example, a studio-like setting, typical offices, and automobiles, and a second scenario, where videos that only contain the speaker's lower face are available, as captured by a specially designed wearable audio-visual headset. In both cases, improved algorithms for ROI extraction are presented, compared to previously used baselines [5, 9].

2.1.1. Full-face data processing

Preceding ROI extraction, the estimation of the speaker's face location and of landmark facial features is required, for example mouth corners and eyes that can provide head-pose information. In our work, we employ a statistical approach to this problem, reported in [8], as our *baseline* system.

In more detail, given a video frame, face detection is first performed by searching for face candidates that contain a relatively high proportion of skin-tone pixels over an image "pyramid" of possible locations and scales. Each candidate is size-normalized to a chosen template size (here, an 11×11 square), and its greyscale pixel values are placed into a 121-dimensional face candidate vector. Every vector is given a score based on the combination of the two-class (face versus non-face) Fisher linear discriminant, as well as its "distance from face space" (DFFS), i.e., the face vector projection error onto a lower, 40-dimensional space, obtained by means of *principal components analysis* (PCA). Candidate regions exceeding a threshold score are considered as faces. Once a face has been detected, an ensemble of facial feature detectors are used to estimate the locations of 26 facial features (eleven such facial features are depicted in Fig.1). Each feature location is determined using a within-face restricted search, based on the feature linear discriminant and "distance from feature space" (similar to the DFFS discussed above) scores of 11×11 square candidate features. A training step is required to estimate the Fisher discriminant and PCA eigenvectors for both stages, utilizing a number of manually annotated video frames [8].

An *improved* version of this algorithm appears in an accompanying paper [7]. It considers a compressed representation of the candidate face or feature vectors by means of their *discrete cosine transform* (DCT). The vector of the top 50 or 32 coefficients, obtained by a zig-zag scan on the face or facial feature template, respectively, is scored using a two-class *Gaussian mixture model* (GMM) classifier with up to 50 mixtures. If the score is sufficiently high, DFFS is also utilized in order to reduce false detects. The process is repeated for facial feature detection. In contrast to the pre-set skintone map of the baseline algorithm, the new approach uses a single-mixture full-covariance GMM of the (r,g,b) skin chromatic space, estimated over face training data. Further enhancements include a rectangular, instead of square, face template of size 11×14 pixels, and a new scheme for generating additional training samples for the facial feature GMMs [7]. The performance of the two algorithms is studied in Section 4.

Fig. 2. *Subject wearing the audio-visual headset, shown also in close-up.*

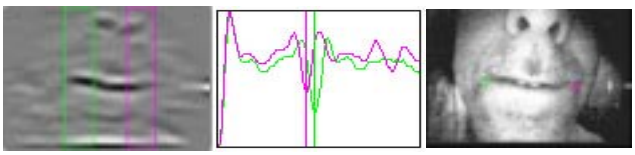

Fig. 3. *Some mouth-corner y-coordinate estimation steps for an example frame captured by the audio-visual headset. Left-to-right: Mouth corner search regions super-imposed on barmask-enhanced frame; vertical image projection within search regions; located mouth corners.*

Face and facial feature tracking provides mouth location, size, and orientation estimates. These are subsequently smoothed over time to improve robustness. Based on the result, a 64×64 pixel ROI is obtained for every video frame. This contains the lower face around the speaker's mouth, and is properly normalized to compensate for rotation, size, and lighting variations, the latter by using histogram equalization of the face, as depicted in Fig.1.

2.1.2. Headset-captured data processing

Much of the processing presented above can be simplified in the case of video captured by a suitably head-mounted camera, so that it mostly contains the speaker's mouth region. Such an audio-visual sensor, depicted in Fig.2, has been reported in [9]. There, a *baseline* visual front end is presented, that extracts the 64×64 pixel ROI by simply truncating and subsampling the original 720×480 pixel monochrome frame.

An *improved* ROI estimation method appears in an accompanying paper [10]. There, in order to compensate for inter-speaker, and headset positioning variability, ROI normalization is employed based on mouth size and orientation, driven on basis of detected mouth corners. In addition, the ROI is histogram equalized to compensate for variation in illumination. Estimation of the mouth corners is fairly simple, operating under the assumption that the camera is already aimed nearly directly at the mouth. The algorithm employs a number of histogram equalization steps, image thresholding operations, as well as extrema detection of the image histograms along the horizontal and vertical axes, and is described in detail in [10] (see also Fig.3). The performance of the two schemes is reported in Section 4.

2.1.3. Visual speech features

Once the visual ROI is extracted, a two-dimensional, separable DCT is applied to it, and the 100 highest-energy transform coefficients are retained. Subsequently, an *intra-frame* cascade of a *linear discriminant analysis* (LDA) projection and a rotation by means of a *maximum likelihood linear transformation* (MLLT) is used, resulting in a 30-dimensional feature vector [5]. To facilitate audio-visual fusion, visual features are linearly interpolated to

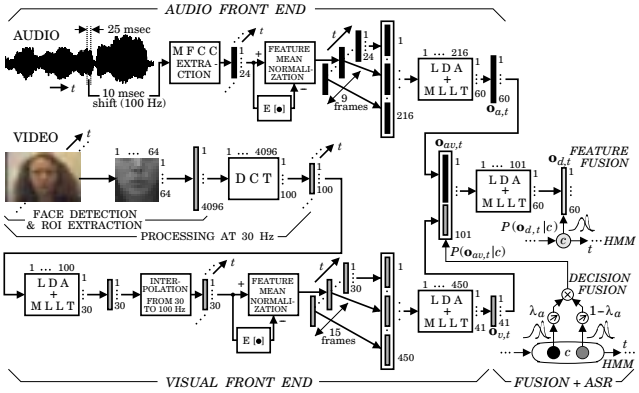

Fig. 4. *Block diagram of the AVASR system. Time-synchronous, 60-dimensional audio feature vectors, $\mathbf{o}_{a,t}$, and 41-dimensional visual, $\mathbf{o}_{v,t}$, are extracted, both at 100 Hz. Subsequently, a feature fusion and a decision fusion strategy are considered for bimodal speech recognition.*

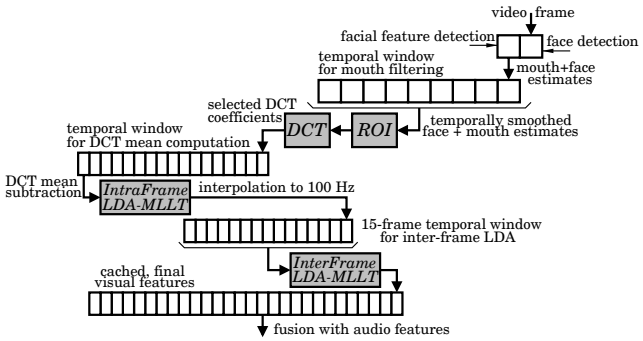

Fig. 5. *Real-time processing employs several circular buffers for implementing the required visual front end steps of Fig.4.*

synchronize with the audio feature frame rate of 100 Hz. This step is followed by *feature mean normalization* to partially compensate for lighting variations. Fifteen consecutive feature vectors 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 Fig.4).

2.2. Audio-visual fusion

In addition to visual features, time-synchronous audio features, $\mathbf{o}_{a,t}$, are extracted. They consist of 24 mel-frequency cepstral coefficients, mean normalized, concatenated over 9 frames, and projected by an LDA/MLLT cascade onto a 60-dimensional space (see Fig.4). They can then be combined in various ways with the visual features for bimodal ASR [5]. Two such integration strategies are considered here: (a) *Feature fusion*, by 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, and considering a single-stream *hidden Markov model* (HMM) as the generative model of the resulting features; and (b) *Decision fusion*, where a two-stream HMM is used to provide the class-conditional score for the concatenated vector $\mathbf{o}_{av,t}$, as the product of the class-conditional probabilities of single-modality classifiers, raised to appropriate stream exponents. In both schemes, HMM parameters are obtained by the traditional maximum likelihood approach, based on available training data. For decision fusion in particular, the HMM stream component parameters are separately trained and subsequently joined, with the stream exponents estimated to minimize the *word error rate* (WER) on held-out data [5].

3. REAL-TIME AVASR

The appearance-based visual front end discussed in the previous section is amenable to real-time processing. Indeed, we have recently presented an AVASR prototype [11] that operates on both full-face videos (using the baseline face detection algorithm of [8]) and videos captured by the audio-visual headset (the improved front end of [10] is implemented). The prototype functions both in batch mode on pre-recorded audio-visual data, as well as in live mode, i.e., full-frame video input captured through a USB 2.0 or Firewire interface. The whole implementation is an addition to the basic IBM ViaVoice architecture, as discussed in [11].

To achieve real-time performance, a number of modifications to the visual front end have been employed [11]. First, face and facial feature detection alternate at every second frame, hence introducing a 2-frame latency in processing. Subsequently, the computation of the smoothed mouth geometry estimates is altered: Size, rotation, and face boundaries, all used in ROI extraction, are averaged over an appropriate temporal window, requiring a limited look-ahead. A further latency is introduced due to DCT coefficient mean computation: In order to obtain reliable estimates of such means, some temporal look-ahead is required, especially at the beginning of the utterance, where little data is available. In addition to the above, the intra-frame LDA/MLLT requires the availability of few future frames at the 100 Hz rate (see Fig.4). All these steps add up to a latency of approximately 0.8 secs, for a 30 frame/sec input video rate. Implementation of the modified visual front end requires the use of circular buffers for holding sufficient number of video frames and visual features at various stages of processing, as shown in Fig.5. On the average, for full-face videos, the entire visual front end, including fusion, utilizes 67% of the processor in a Pentium 4, 1.8 GHz desktop, thus achieving better than real-time performance. An exact break-down per visual front end stage is depicted in Table 1. For headset-captured data, this time is considerably less. The original version of the AVASR prototype version utilized feature fusion for audio-visual integration, currently however decision fusion is also available.

Computation type	Time (ms)	Fraction of total
Frame grab	3.3	15 %
Face Finding	10.5	24 %
Feature Localization	20.9	48 %
Image Buffer and ROI	2.0	9 %
DCT and LDAs	0.9	4 %
Total	21.9	100 %

Table 1. *Average processing load of the AVASR prototype visual front end, depicted per video frame available at 33.3 ms.*

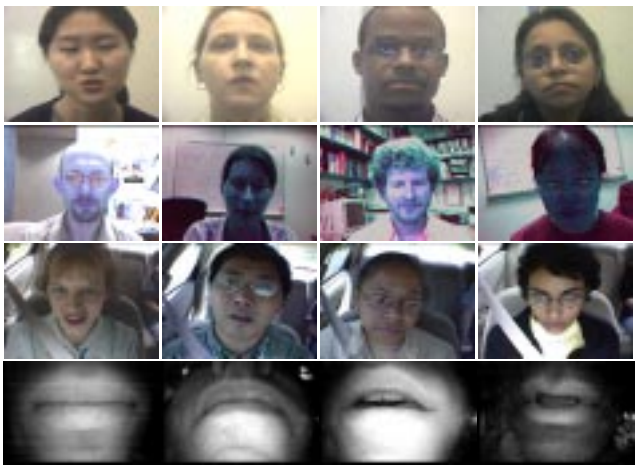

Fig. 6. *Example frames from each of the four datasets considered in this paper for AVASR. Top-to-bottom: Studio, office, car, and headset data.*

4. EXPERIMENTS

We now discuss the performance of the visual front end algorithms discussed above. We are interested in both their robustness to visually challenging domains, as well as in quantifying any degradation due to their real-time implementation.

4.1. Databases

We consider three audio-visual databases that contain frontal full-face videos. The three sets are chosen to correspond to increasingly more challenging visual domains, as a means to investigate robustness of the visual front end. The first corpus, referred to as "STU", was recorded in a quiet studio-like environment, using a high-quality camera, uniform lighting and background, and relatively stable frontal subject head pose, due to the use of a teleprompter. The second corpus was recorded using a portable collection system on a laptop, with video captured via an inexpensive USB 2.0 web-cam and audio by the built-in PC microphone. The database subjects were typically recorded in their own offices without the use of a teleprompter, therefore the lighting, background, and head-pose vary greatly. This set is denoted by "OFF". The third set has been recorded in an automobile, both stationary and moving at approximately a 30 or 60 mph speed. The vehicle was equipped with a wideband microphone and a lipstick-style camera, mounted on the middle of the passenger-side overhead visor. Compared to the previous two databases, the lighting, background, and head-pose vary significantly, therefore this "CAR" database consists the most challenging set.

In order to study the benefits of direct visual ROI capture, we also consider a fourth database, referred to as "IR". The set was

DB	Sp.	Set	Fr.	Utter.	Dur.	DB	Sp.	Set	Fr.	Utter.	Dur.
S		Tr	1000	5403	7:53	C		Tr	2254	1209	1:04
T	50	Ch	—	663	0:58	A	87	Ch	—	139	0:07
U		Ts	100	623	0:55	R		Ts	287	137	0:07
O		Tr	1368	4591	6:07	I		Tr	—	3000	4:24
F	101	Ch	—	549	0:44	R	90	Ch	—	300	0:26
F		Ts	253	537	0:43			Ts	—	340	0:30

Table 2. *Partitioning of the four audio-visual databases (DB: studio (STU), office (OFF), automobile (CAR), and headset (IR)) into training (Tr), held-out (Ch: check), and test (Ts) sets (number of utterances and duration (in hours) are shown). The number of database subjects (Sp) and number of face-annotated video frames (Fr) used for face/facial feature detection (for the three full-face sets) are also depicted. Note that the CAR database is significantly smaller in duration compared to the other three.*

recorded by means of the specially designed audio-visual wearable sensor, depicted in Fig.2. In addition to the microphone, the headset boom contains an infrared camera. The system is thus relatively insensitive to head-pose and lighting variations, providing high-resolution video of the speaker's mouth region. Both audio and visual signals are carried wirelessly (through an RF transmitter located at the headset earpiece) to a base station, connected to the database collection computer via a Firewire interface [9].

All four sets contain a large number of subjects uttering both large-vocabulary continuous speech, as well as connected-digit utterances. In the following, we only consider the small-vocabulary (digit recognition) task, since it provides meaningful comparisons for visual-only recognition. Example frames of the four sets are shown in Fig.6. More database information is given in Table 2 (see also [6, 10]).

4.2. Visual-only and AVASR performance

A summary of our experimental results on the four databases is reported in Table 3. In particular, we use the multi-subject training/testing scenario (as defined in Table 2) to compare visual-only ASR of connected digit sequences, as well as AVASR (relative to audio-only performance) by means of the baseline (B) and improved (N) algorithms discussed in Section 2.1.

We consider the three full-face sets, concentrating on the face detection error rate first. Clearly, this increases dramatically as the task becomes more challenging (from no errors in the STU data, to about 25% in the CAR domain, i.e., 72 out of the 287 test faces being incorrectly detected). To improve performance to acceptable levels, *speaker-dependent* (SD) face detectors are also considered using the baseline algorithm [8]. Performance improves significantly for both OFF and CAR sets, reaching a 16% error in the latter case. By employing the improved face detector of [7] though, an even better performance (6.3%) is attained using only *multi-speaker* (MS) face tracking! In general, improved face detection leads to improved recognition, at least when comparing MS tracked data (B_{MS} vs. N_{MS}), as depicted by the consistently better visual-only and AVASR at both clean and noisy acoustic conditions. A fair comparison to the B_{SD} tracked data is more complicated, since the facial feature detection accuracy, not discussed here, plays also a role. It is interesting to note, that the real-time visual front end implementation degrades performance only moderately (see N^{rt}_{MS} vs N_{MS}).

Although the face detection algorithm of [7] improves performance across all three tasks, it is clear from Table 3 that, as the challenge of the visual domain increases, the gain in ASR due to the visual modality is reduced (see the right-most table column, for example). This is obviously due to the visual front end processing, as demonstrated by the increasing face detection error and visual-only WER across tasks. This fact (together with eliminating face detection computations) has motivated us to develop the audio-visual headset. Indeed, visual-only WER is significantly lower even compared to the STU data (20.7% vs. 27.4%), with obvious advantages in AVASR, resulting in the largest percentage gains in both acoustic conditions considered. Note that both the visual processing algorithms of Section 2.1.2 perform similarly well in the multi-speaker training/testing scenario of Table 2. Speaker-independent recognition however (not reported in Table 3) clearly points to the superiority of the improved algorithm (N) over the baseline (B). For example, its use reduces visual-only WER from 46.5% to 35.3% (a similar experiment is also reported in [10]).

DB	Alg. mode	Face error	VI	Clean AU	Clean AVf	Clean AVd	Noisy AU	Noisy AVf	Noisy AVd
S	B_{MS}	0.0	28.26	0.84	0.82	0.64	24.56	12.17	10.86
T	B_{SD}	0.0	29.84	0.84	0.81	0.69	24.56	11.69	10.15
U	N_{MS}	0.0	27.44	0.84	0.82	0.66	24.56	11.69	10.36
	N^{rt}_{MS}	0.0	29.73	0.84	0.86	0.71	24.56	12.72	10.66
O	B_{MS}	10.9	48.18	2.51	2.53	2.07	24.91	14.53	16.00
F	B_{SD}	2.8	43.44	2.51	2.31	1.98	24.91	12.69	14.47
F	N_{MS}	2.8	43.33	2.51	2.49	1.96	24.91	12.64	14.73
C	B_{MS}	25.1	69.79	2.83	4.02	2.68	25.89	20.09	16.37
A	B_{SD}	16.0	64.58	2.83	3.72	2.38	25.89	17.26	16.37
R	N_{MS}	6.3	68.75	2.83	3.13	2.38	25.89	14.14	16.22
I	B_{MS}	n/a	20.69	1.33	1.43	1.01	25.23	9.28	7.92
R	N_{MS}	n/a	21.35	1.33	1.40	0.94	25.23	10.26	7.75

Table 3. *Comparisons of the baseline (B) and improved (N) visual front end algorithms on the four audio-visual databases, in terms of test set visual-only (VI) and audio-visual (AV) word error rate (WER), %, by means of feature (AVf) and decision (AVd) fusion, for connected-digit ASR. Two acoustic conditions are considered: The original database audio (clean), and a degraded version (noisy) by additive babble noise. Audio-only (AU) WER is also shown in these two conditions. For full-face videos (studio, office, and car datasets), face detection is performed in a speaker-dependent (SD), or multi-speaker (MS) mode, with the face detection error rate, %, also depicted. Performance of the real-time (rt) visual front end is shown for the studio data.*

5. SUMMARY

We have presented recent progress in our research towards practical deployment of AVASR. Improved and fast visual front end processing algorithms allow us today to integrate the visual modality into real-time systems that are highly beneficial to ASR performance in relatively controlled visual environments. Both speed and performance can be further enhanced by using a wearable audio-visual sensor.

6. 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 T. Multimedia*, 2:141–151, 2000.

[4] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Signal Process. Mag.*, 10(1): 9–21, 2001.

[5] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, 91(9): 1306-1326, 2003.

[6] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," *Proc. Europ. Conf. Speech Technol.*, pp. 1293–1296, 2003.

[7] 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," Submitted to: *Int. Conf. Acoust. Speech Signal Process.*, 2004.

[8] A.W. Senior, "Face and feature finding for a face recognition system," *Proc. Int. Conf. Audio Video-based Biometric Person Authent.*, pp. 154–159, 1999.

[9] J. Huang, G. Potamianos, and C. Neti, "Improving audio-visual speech recognition with an infrared headset," *Proc. Work. Audio-Visual Speech Process.*, pp. 175–178, 2003.

[10] J. Huang, G. Potamianos, J. Connell, and C. Neti, "Audio-visual speech recognition using an infrared headset," Submitted to: *Int. Conf. Acoust. Speech Signal Process.*, 2004.

[11] J.H. Connell, N. Haas, E. Marcheret, C. Neti, G. Potamianos, and S. Velipasalar, "A real-time prototype for small-vocabulary audio-visual ASR," *Proc. Int. Conf. Multimedia Expo*, pp. 469–472, 2003.