

Audio-Visual Speech Recognition in Challenging Environments

Gerasimos Potamianos, Chalapathy Neti

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA

{gpotam,cneti}@us.ibm.com

Abstract

Visual speech information is known to improve accuracy and noise robustness of automatic speech recognizers. However, to-date, all audio-visual ASR work has concentrated on “visually clean” data with limited variation in the speaker’s frontal pose, lighting, and background. In this paper, we investigate audio-visual ASR in two practical environments that present significant challenges to robust visual processing: (a) Typical offices, where data are recorded by means of a portable PC equipped with an inexpensive web camera, and (b) automobiles, with data collected at three approximate speeds. The performance of all components of a state-of-the-art audio-visual ASR system is reported on these two sets and benchmarked against “visually clean” data recorded in a studio-like environment. Not surprisingly, both audio- and visual-only ASR degrade, more than doubling their respective word error rates. Nevertheless, visual speech remains beneficial to ASR.

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, to-date, all research work on *audio-visual ASR* (AV-ASR) has concentrated and reported 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. In contrast, the audio channel is usually corrupted artificially by additive noise, with limited work reported on naturally degraded speech [5, 6]. Therefore, little is known about AV-ASR performance in realistic, non-ideal environments, where in addition to possibly noisy audio, the visual channel quality is poor, thus presenting significant challenges to speech-informative visual feature extraction. Clearly, for the visual modality to become a component of main-stream ASR, its benefits need to be demonstrated in such non-ideal domains.

This paper reports our first attempt to investigate AV-ASR in “visually challenging” environments. In particular, we consider two such domains: Typical offices, where data are recorded by means of a portable PC equipped with an inexpensive web camera, and automobiles, both stationary and moving, at two approximate speeds. These represent increasingly challenging visual environments, compared to studio-quality data, exclusively considered in our prior work [4, 5]. Indeed, both office and automobile data are characterized by varying head pose, lighting, and background, with the additional challenges of poor quality video capture in the former, and presence of extreme shadows and head movement in the latter. Of course, in addition to visual channel degradation, the audio is also nois-

ier than the studio-quality dataset, thus providing two realistic, non-ideal audio-visual environments, suitable for investigating any visual modality benefits to AV-ASR.

In this work, we utilize our state-of-the-art AV-ASR system [7] to study the performance of all its components on the two challenging datasets, benchmarking the results against the studio-quality corpus. 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 the face detection and visual-only ASR accuracies, 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, with a large number of subjects available in each database.

The paper is structured as follows: Section 2 reviews the main components of the AV-ASR system, with emphasis placed on visual processing and audio-visual integration, whereas Section 3 briefly describes the three audio-visual corpora considered in this work. Section 4 is devoted to the experimental study of the performance of our AV-ASR system across the three environments, which is subsequently summarized in Section 5.

2. Audio-visual ASR system components

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]: Appearance-based features that typically seek a suitable transform of the pixel values within a visual *region-of-interest* (ROI) [1, 4], shape-based features that consist of a geometric or statistical representation of the lip contours [3], 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], whereas the latter combines the two single-modality classification decisions typically at the likelihood level [3]. Finally, a *hidden Markov model* (HMM) with Gaussian mixture emission probabilities [3, 4], or alternatively, an artificial neural network classifier [1] can be used for AV-ASR. The system considered in this work employs appearance-based features based on the ROI *discrete cosine transform* (DCT), as in [1, 4], HMMs for ASR, and two possible fusion techniques, as discussed below (see also Fig. 1).

2.1. Visual feature extraction

In more detail, given the video of a spoken utterance, a two-stage statistical face tracking algorithm is first used to detect the speaker’s face and subsequently locate 26 facial features

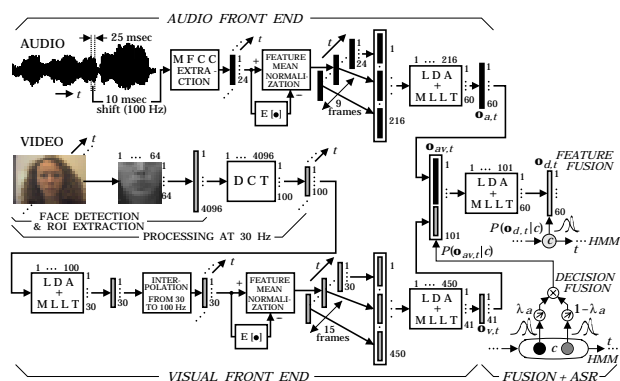


Figure 1: Block diagram of the AV-ASR system, employed in this paper. Time-synchronous, 60-dimensional audio feature vectors, $\mathbf{o}_{a,t}$, and 41-dimensional visual observations, $\mathbf{o}_{v,t}$, are extracted, both at a 100 Hz rate. Subsequently, a feature fusion and a decision fusion strategy are considered for speech recognition by means of hidden Markov models.

(eleven such features are depicted in Fig.2). At each stage, normalized face (or facial feature) candidate vectors are scored by a two-class Fisher discriminant and the projection residual onto an appropriately defined eigenspace [8]. The highest score candidates are retained as detected faces (or facial features). The algorithm requires training on a small number of manually annotated faces, as discussed in Section 4.

Tracking provides the mouth location, size, and orientation, which are then smoothed over a temporal window to improve robustness. Based on the resulting estimates, a 64×64 pixel region-of-interest (ROI) is obtained for every video frame. The ROI contains the lower face around the speaker's mouth, including the jaw and cheeks, and is properly normalized to compensate for rotation, size, and lighting variations (see also Fig.2).

Subsequently, 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) [4], that improves maximum likelihood based statistical data modeling. 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 Fig.1).

2.2. Audio-visual fusion and ASR

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 Fig.1).

Following feature extraction, we consider two simple, well-known integration strategies for AV-ASR [7]: (a) *Feature fusion*, by projecting the 101-dimensional concatenated audio-

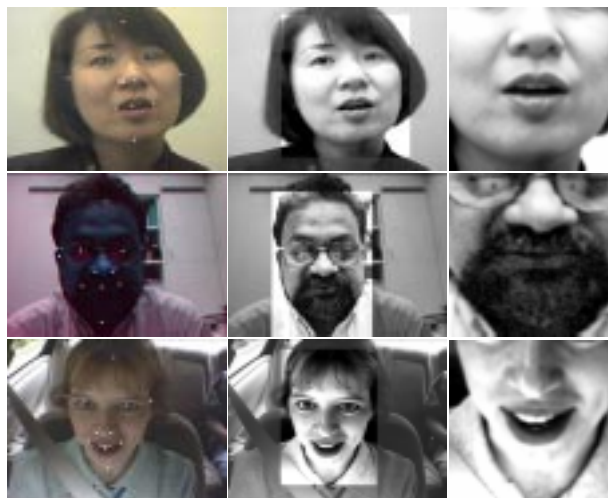


Figure 2: Face, facial part detection, and region-of-interest (ROI) extraction for one example video frame from each of the three corpora considered in this paper (top-to-bottom: studio, office, and automobile data - see also Fig.3). The following are depicted for each set, left-to-right: Original frame with eleven detected facial parts super-imposed; face-area enhanced frame; size-, rotation-, and lighting-normalized ROI.

visual vectors $\mathbf{o}_{av,t} = [\mathbf{o}_{a,t}, \mathbf{o}_{v,t}]$ onto a 60-dimensional space by an LDA/MLLT [4], and considering a single-stream 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:

$$P(\mathbf{o}_{av,t} | c) = P(\mathbf{o}_{a,t} | c)^{\lambda_a} P(\mathbf{o}_{v,t} | c)^{1-\lambda_a}, \quad (1)$$

where c denotes an HMM state, and exponent λ_a is set to a global, database-dependent value within $[0,1]$ (see also Fig.1).

In both schemes, HMM parameters are obtained by the traditional maximum likelihood approach, based on available training data. In particular, the stream component parameters of (1) are separately trained, and joined in (1), with λ_a estimated to minimize the *word error rate* (WER) on a held-out dataset. More sophisticated schemes for training (1) are also possible, as well as introducing state asynchrony, additional streams, or time dependence of λ_a in (1) [7], but this is not our focus here.

3. Three audio-visual databases

As mentioned in the introduction, all AV-ASR research has concentrated on databases collected in ideal visual conditions. One such multi-speaker set is the IBM ViaVoiceTM 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 60 fields/sec and a 704×480 pixel frame size, thus providing satisfactory mouth region resolution (see Fig.3). In addition to video, high quality wideband audio is synchronously collected at a rate of 16 kHz. A 50-subject subset of this database ("STU"), containing connected-digit utterances, is used in this work as a reference for audio-visual ASR under ideal conditions.

In this paper, we deviate from ideal environments, and consider more realistic and challenging data. The first such corpus

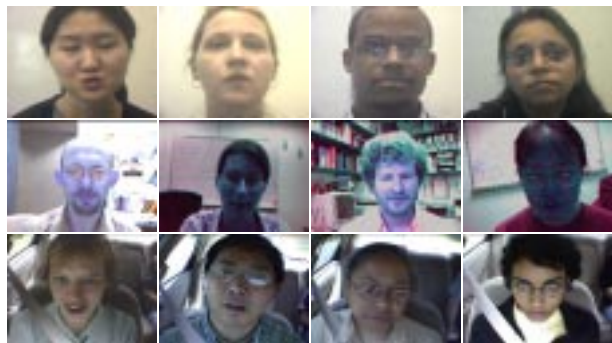


Figure 3: Example frames of four subjects from each of the three domains, considered in this paper for AV-ASR. Upper row: “Visually clean” data, recorded using a teleprompter in a controlled studio environment. Middle row: “Visually challenging” data, captured using a laptop based AV-ASR prototype at an uncontrolled office environment. Lower row: “Visually challenging” data, collected in stationary and moving automobiles.

is captured using a laptop-based audio-visual data collection prototype. The system records wideband 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 previous database, the video quality is now poorer, with automatic gain control present, and only 30 frames/sec available at 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. The significantly more “visually challenging” nature of this office environment set becomes clear from Fig.3. A total of 109 subjects uttering connected digit strings are available in this set, which is referred to in Tables 1 and 2 as “OFF”.

The second challenging database has been recorded in an automobile, both stationary and moving at approximately a 30 or 60 mph speed (split at about 24%, 39%, and 37% between the three conditions). The vehicle is equipped with a wideband microphone and a lipstick-style camera, mounted to the middle of the passenger-side overhead visor. The recorded video is of high quality, and similarly to the studio-data, it is MPEG2 encoded at 60 fields/sec and a 704×480 pixel frame size. However, compared to the previous two databases, the lighting, background,

Data	Subj.	Set	Frames	Utter.	Dur.
S	50	Train	1000	5403	7:53
T		Check	—	663	0:58
U		Test	100	623	0:55
O	109	Train	1368	5054	6:42
F		Check	—	604	0:48
F		Test	253	587	0:46
C	87	Train	2254	1209	1:04
A		Check	—	139	0:07
R		Test	287	137	0:07

Table 1: The three audio-visual databases used in this paper for connected-digit ASR (top-to-bottom: studio, office, and automobile environments). Their partitioning into training, held-out (check), and test sets is depicted (number of utterances and duration (in hours) are shown for each set). The number of database subjects and the number of face-annotated video frames used for face/facial feature detection are also given.

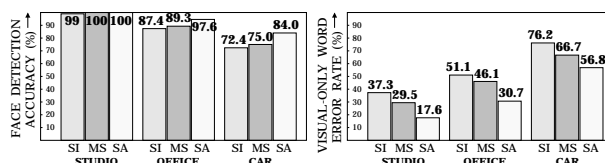


Figure 4: Visual front end performance on the three corpora of Table 1 under speaker-independent (SI), multi-speaker (MS), and speaker-adapted (SA) training/testing. Left: Face detection accuracy, %. Right: Visual-only word error rate, %, for connected-digit ASR.

and head-pose vary significantly in the automobile corpus, as it is apparent from Fig.3. The 87-subject subset of this “CAR” database, containing connected-digit utterances, is used here. A summary of all three corpora is provided in Table 1.

4. Recognition experiments

We now proceed to report a number of ASR experiments on the three databases of Table 1, using the algorithms discussed in Section 2. We first briefly introduce the experimental paradigm adopted, followed by a more detailed presentation of our results.

4.1. The experimental paradigm

Each database is split into three sets, one for training HMMs, a held-out set for optimizing parameter λ_a of (1), and the remaining for testing ASR performance (see Table 1). Three training/testing paradigms are considered: A *speaker-independent* one (SI); a *multi-speaker* (MS) scenario, where separate data from *all* subjects are used for both training and testing; and a *speaker-adapted* one (SA), where, for each subject, MS-trained HMMs are adapted (on the particular subject training data) by maximum-a-posteriori adaptation, followed by maximum likelihood linear regression [9]. A similar data split and paradigm are used for face detection training and testing, as depicted in Table 1. Manual facial feature annotation of the training set frames is carried out.

In addition to the original database acoustic signal, audio-only and AV-ASR are also considered on artificially corrupted audio by additive “speech babble”, using HMMs trained on the original data. The degradation is at a level that results in approximately a 25% audio-only MS WER on each test set.

For connected digit ASR (11-word vocabulary), a two-stage stack decoding algorithm is employed, with unknown digit-string length. HMMs with 159 context dependent states and approximately 3.2k Gaussian mixture components per stream are used (a smaller, 101 state, 1.7k mixture system is utilized for the “sparse” automobile data domain).

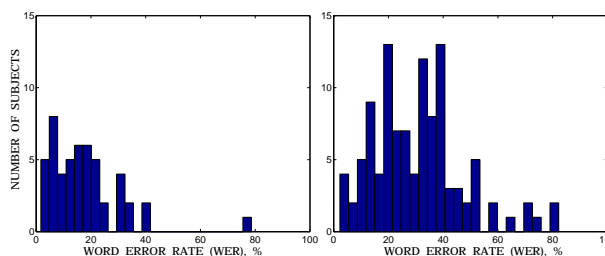


Figure 5: Speaker-adapted, visual-only word error rate histogram of the 50 subjects of the studio quality database (left) and of the 109 subjects in the office environment dataset (right).

Data-base	Tr/Ts parad.	VI	Clean			Noisy		
			AU	AVf	AVd	AU	AVf	AVd
S	SI	37.31	1.52	1.46	1.29	26.54	13.75	13.79
T	MS	29.46	0.84	0.81	0.71	24.56	11.44	10.06
U	SA	17.64	0.64	0.54	0.37	22.01	7.96	5.71
O	SI	51.07	3.23	3.29	2.74	27.93	16.41	16.75
F	MS	46.07	2.54	2.62	2.28	26.43	14.60	13.88
F	SA	30.66	1.40	2.01	1.42	21.47	8.90	8.42
C	SI	76.19	5.51	6.85	5.36	30.06	23.66	22.47
A	MS	66.67	2.83	3.42	2.83	25.89	16.67	14.43
R	SA	56.85	1.93	3.72	1.79	20.54	14.88	13.10

Table 2: Single-modality (visual-only (VI), audio-only (AU)) and audio-visual ASR performance by means of feature (AVf) and decision (AVd) fusion on the test sets of the three audio-visual databases (top-to-bottom: studio, office, and automobile environments). Results (in word error rate, %) for the speaker-independent (SI), multi-speaker (MS), and speaker-adapted (SA) training/testing paradigms are depicted. 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.

4.2. Visual front end performance

The visual front end performance across the three databases is depicted in Fig.4. We are interested in two system components that allow comparisons. The first one is the face detection accuracy, expressed as a percentage of detected faces within 20% of their manually annotated location, orientation, and scale. Clearly, the tracking performance deteriorates significantly as the visual environment becomes more challenging. Thus, for the automobile domain, SA face detection reaches only an 84% accuracy, compared to 97.6% in the office domain, and 100% on the studio data (notice the small test set sizes in Table 1). MS and SI results degrade further, dropping to 72.4% for the latter on the automobile data.

It is only natural to expect that such degradation results in inaccurate ROI extraction, and thus inferior visual-only ASR. Indeed (see Fig.4 and Table 2), visual-only WER approximately doubles when moving from the studio to the office data (SA: 17.6% \rightarrow 30.7%) and triples on the automobile data (SA: 17.6% \rightarrow 56.8%). Fig.5 demonstrates the shift in the subject SA WER histogram for the first two databases of Table 1. Notice also that the benefits of adaptation are reduced as the domain becomes more challenging, reaching only a 15% improvement (66.7% \rightarrow 56.8%) in the car environment, compared to 40% (29.5% \rightarrow 17.6%) for the studio data, relative to the MS WER (see also Table 2). It is also interesting to observe, that in the automobile data, visual-only recognition is superior when the car is stopped, compared to both 30 and 60 mph speed conditions (see Table 3). This could be attributed to the absence of severe lighting and background variations in the 0 mph case.

4.3. Audio-visual recognition

In addition to the visual-only WER degradation, it is not surprising that audio-only ASR also suffers in the challenging domains (see Table 2). For example, the office set exhibits double the WER of the studio data (SA: 1.40%, compared to 0.64%), whereas, in the automobile domain, the WER triples (SA: 1.93%). There, higher speed conditions result in increased WERs (see Table 3).

Speed	VI	AU	AVf	AVd
0 mph	56.81	1.78	1.18	1.18
30 mph	72.95	2.05	3.28	2.87
60 mph	67.18	4.25	5.02	3.86
Combined	66.67	2.83	3.42	2.83

Table 3: Multi-speaker visual-, audio-only, and audio-visual ASR performance (in word error rate, %) on the automobile data, depicted as a function of the automobile speed.

Due to the simultaneous degradation in both audio and visual channels, the visual modality remains of benefit to ASR. However, this benefit is reduced as a result of the severe visual-only WER increase in the visually challenging domains. Indeed, as it is clear from Table 2, the relative improvement in MS WER in the clean condition is 15% for the studio data (0.84% \rightarrow 0.71%), but only 10% for the office data (2.54% \rightarrow 2.28%), with no apparent improvement in the automobile domain. Similarly, in the MS (SA) noisy case, the approximately 25% (22%) audio-only WER reduces, when incorporating the visual modality, to 10% (5.7%) for the studio dataset, but only to 13.9% (8.4%) for the office data and 14.4% (13.1%) on the automobile set. The above results are achieved using decision fusion, which, in general, is superior to feature fusion (see Table 2).

5. Conclusions

We investigated audio-visual ASR in two “challenging” environments that present difficulties for robust processing in both modalities. We particularly concentrated on benchmarking the visual front end performance and the visual modality ASR benefit of a state-of-the-art AV-ASR system against a typical “visually clean” data domain. We demonstrated that increased visual data challenges have a negative effect on both metrics, thus highlighting the need for improving the robustness of the visual front end algorithms employed in the system. Nevertheless, visual speech remains beneficial to ASR in both challenging environments considered.

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. Luetin, “Audio-visual speech modeling for continuous speech recognition,” IEEE Trans. Multimedia, 2(3):141–151, 2000.
- [4] G. Potamianos, J. Luetin, and C. Neti, “Hierarchical discriminant features for audio-visual LVCSR,” Proc. Int. Conf. Acoust., Speech, Signal Process., pp. 165–168, 2001.
- [5] G. Potamianos and C. Neti, “Automatic speechreading of impaired speech,” Proc. Conf. Audio-Visual Speech Process., pp. 177–182, 2001.
- [6] F.J. Huang and T. Chen, “Consideration of Lombard effect for speechreading,” Proc. Works. Multimedia Signal Process., pp. 613–618, 2001.
- [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.W. Senior, “Face and feature finding for a face recognition system,” Proc. Int. Conf. Audio Video-based Biometric Person Auth., pp. 154–159, 1999.
- [9] L. Neumeyer, A. Sankar, and V. Digalakis, “A comparative study of speaker adaptation techniques,” Proc. Europ. Conf. Speech Commun. Technol., pp. 1127–1130, 1995.