To Appear: Workshop on Multimedia Signal Processing, Special Session on Joint Audio-Visual Processing, Cannes, October 3-5, 2001

LARGE-VOCABULARY AUDIO-VISUAL SPEECH RECOGNITION: A SUMMARY OF THE JOHNS HOPKINS SUMMER 2000 WORKSHOP

Chalapathy Neti **Gerasimos Potamianos**
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{cneti,gpotam}@us.ibm.com

Juergen Luettin
Ascom Systec AG
5506 Maegenwil, Switzerland
Juergen.Luettin@ascom.ch

Iain Matthews
Robotics Institute, CMU
Pittsburgh, PA 15213, USA
iainm@cs.cmu.edu

Hervé Glotin
ICP, Grenoble, France; and
IDIAP, Martigny, Switzerland
glotin@idiap.ch

Dimitra Vergyri
SRI International
Menlo Park, CA 94025, USA
dverg@speech.sri.com

Abstract - We report a summary of the Johns Hopkins Summer 2000 Workshop on audio-visual automatic speech recognition (ASR) in the large-vocabulary, continuous speech domain. Two problems of audio-visual ASR were mainly addressed: Visual feature extraction and audio-visual information fusion. First, image transform and model-based visual features were considered, obtained by means of the discrete cosine transform (DCT) and active appearance models, respectively. The former were demonstrated to yield superior automatic speechreading. Subsequently, a number of feature fusion and decision fusion techniques for combining the DCT visual features with traditional acoustic ones were implemented and compared. Hierarchical discriminant feature fusion and asynchronous decision fusion by means of the multi-stream hidden Markov model consistently improved ASR for both clean and noisy speech. Compared to an equivalent audio-only recognizer, introducing the visual modality reduced ASR word error rate by 7% relative in clean speech, and by 27% relative at an 8.5 dB SNR audio condition.

INTRODUCTION

Exploiting visual, lip-region information for improving human speech perception as well as *automatic speech recognition* (ASR) has been well documented in both the psychological [1] and technical literatures [2]. However, until recently, all *automatic speechreading* studies have been limited to small-vocabulary tasks and small subject populations [2], [3]. Thus, no definite answers existed on the two key issues for the design of *speaker-independent*, audio-visual, *large-vocabulary continuous speech recognition* (LVCSR) systems [2]: (a) The choice of appropriate *visual features*, informative about unconstrained, continuous visual speech; and (b) The design of audio-visual information *fusion* algorithms that consistently outperform audio-only LVCSR systems. To address these issues, we participated in the Summer 2000 Workshop at the Johns Hopkins University on audio-visual ASR, seriously tackling the problem of speaker-independent audio-visual LVCSR for the first time [4]. In this paper, we provide a summary of this work and highlight our main contributions and results.

The paper is structured as follows: First, a section is devoted to visual feature extraction, with two visual front ends discussed. The subsequent section presents a number of audio-visual information fusion algorithms, grouped into feature and decision fusion methods. Next, a description of the audio-visual database is provided, together with the experimental design and results. A summary section concludes the paper.

VISUAL FEATURE EXTRACTION

Various sets of visual features for automatic speechreading have been proposed in the literature over the last 25 years. In general, they can be grouped into three categories: High-level *lip contour* (*shape*) based features, low-level *video pixel* (*appearance*) based ones, and features that are a combination of both [2]. In the first approach, a parametric, or statistical lip contour model is fitted to the mouth image, and the model parameters are used as visual features [3].

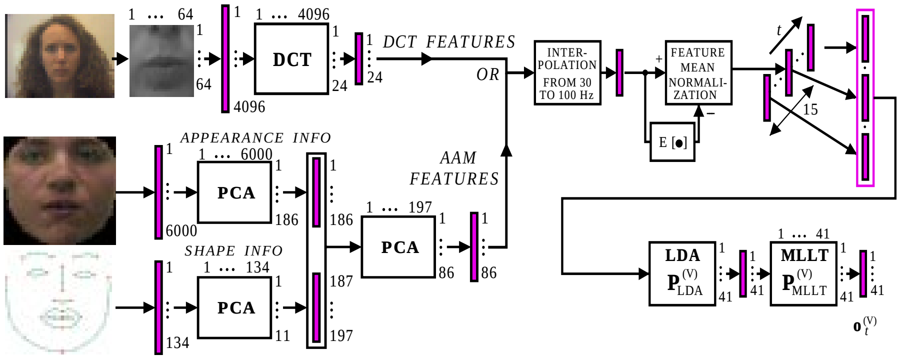

Figure 1: DCT versus AAM based visual feature extraction for automatic speechreading, followed by linear interpolation, feature mean normalization, and the application of LDA and MLLT.

Alternatively, geometric features of an estimate of the inner and/or outer lip contours, such as mouth height and width, are used [2]. In the second approach, the entire image containing the speaker's mouth is considered as informative for speechreading (*region of interest* - ROI), and appropriate *image transformations* of its pixel values are used as visual features [5]. In the third approach, both high- and low-level features are combined to give rise to joint shape and appearance visual features [3], [6]. At the Johns Hopkins Workshop we considered one representative technique from the second and third categories, as discussed next (see Fig.1).

Discrete cosine transform based visual features

In this approach, a statistical face tracker [4], trained on 3744 images (each annotated with 26 facial feature locations), was first employed to detect the speaker's face and facial feature locations, including the mouth corners (see Fig.2). Based on these, the speaker's mouth center and size were estimated, and averaged, for robustness, over a number of neighboring frames. A size-normalized 64×64 pixel ROI, centered around the speaker's mouth was subsequently extracted (see also Fig.2). Finally, a two-dimensional, separable *discrete cosine transform* (DCT) was applied on the ROI pixels, and the 24 highest energy DCT coefficients were retained as visual features [5] (see Fig.1).

Notice that the resulting features can be extracted in real-time, since the DCT allows a fast implementation, whereas it suffices that face and facial feature tracking be performed at a low frame rate. DCT features are also robust to small face tracking inaccuracies, since, in this approach, only a gross estimate of the ROI is required.

Active appearance model based visual features

In this approach, an *active appearance model* (AAM) of the entire face was built from a training set of 4072 facial images, each annotated with 68 landmark points that outlined the eyebrows, eyes, nose, nostrils, jaw, and lip contours. For each image, a 134-dimensional shape vector containing the coordinates of the 68 landmark points (after a *similarity* transformation for normalization to a reference shape), as well as the 6000-dimensional appearance vector of the luminance values of a 6000-pixel normalized face, were obtained (see also Fig.1). The main modes of the shape and appearance vector variation in the training set were independently computed using *principal component analysis* (PCA), and the 11 and 186 largest PCA matrix eigenvalues were retained for the shape- and appearance-only models, respectively. Subsequently, a joint shape and appearance model (AAM) was obtained using a second stage of PCA on the concatenated vector of shape and appearance features, after appropriately rescaling the latter [6]. The 86 largest PCA matrix eigenvalues were retained, giving rise to 86-dimensional AAM based visual features.

To extract AAM visual features from a given facial image, the AAM algorithm of [7] was first employed to fit the appearance model to the image. Subsequently, the joint model was used to extract visual features by hierarchically applying the two stages of PCA [4] (see also Fig.1).

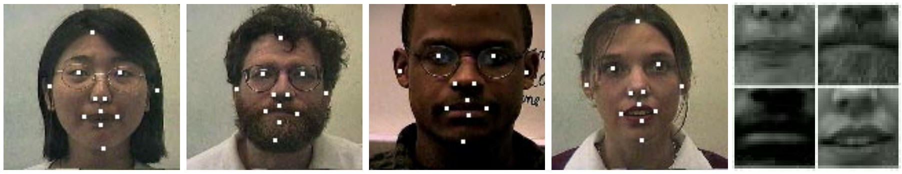

Figure 2: Sample frames from four database subjects, with detected facial features superimposed (*left*), and grey-scale, 64×64 pixel, mouth regions of interest extracted (*right*).

Visual feature postprocessing

Additional processing was applied on both DCT and AAM features to incorporate dynamic visual speech information, and to improve phoneme discrimination and maximum-likelihood based statistical modeling: Fifteen consecutive-frame DCT or AAM visual feature vectors were concatenated, subsequently projected onto a 41-dimensional space using *linear discriminant analysis* (LDA), and then “rotated” by means of a *maximum-likelihood linear transform* (MLLT) [4], [5]. Furthermore, *linear interpolation* was used to align the visual features to the audio feature rate (100 Hz), and visual *feature mean normalization* was employed to compensate for lighting and other variations. The final visual feature vector $\mathbf{o}_t^{(\mathrm{V})}$ dimension was $D_{\mathrm{V}} = 41$, for both DCT and AAM parametrizations (see also Fig.1).

Audio features

In addition to the visual features, static audio features were extracted from the speech signal, consisting of 24 mel-cepstral coefficients, computed over a sliding window of 25 msec and at a rate of 100 Hz. Dynamic audio features were obtained by concatenating 9 consecutive feature frames, and applying LDA and MLLT. The final audio feature $\mathbf{o}_t^{(\mathrm{A})}$ dimension was $D_{\mathrm{A}} = 60$.

AUDIO-VISUAL INFORMATION FUSION

A number of techniques have been suggested for audio-visual fusion [2], which can be broadly grouped into *feature fusion* and *decision fusion* methods. The first are based on training a traditional *hidden Markov model* (HMM) classifier on the concatenated vector of the audio and visual features, or appropriate transformations of it. Decision fusion techniques combine classification decisions based on single modality observations, typically by appropriately weighting their respective log-likelihoods. At the Johns Hopkins Workshop, we considered two feature fusion algorithms and a number of decision fusion approaches, discussed in the following [4].

Feature fusion

We considered two *feature fusion* techniques for audio-visual ASR (see also Fig.3(a)): The first approach was a simple *audio-visual feature concatenation* (**AV-Concat**), giving rise to audio-visual features

$$\mathbf{o}_t^{(\mathrm{AV})} = [\,\mathbf{o}_t^{(\mathrm{A})\,\top}, \mathbf{o}_t^{(\mathrm{V})\,\top}\,]^\top \in \mathsf{R}^{\,D} , \tag{1}$$

where $D = D_{\mathrm{A}} + D_{\mathrm{V}} = 101$. The second approach was aimed at reducing the dimensionality of (1), by means of an LDA projection to a 60-dimensional space, followed by an MLLT rotation,

$$\mathbf{o}_t^{(\mathrm{HiLDA})} \;=\; \mathbf{P}_{\mathrm{MLLT}}^{(\mathrm{AV})}\,\mathbf{P}_{\mathrm{LDA}}^{(\mathrm{AV})}\,\mathbf{o}_t^{(\mathrm{AV})} \;. \tag{2}$$

Since LDA and MLLT were also independently applied on the audio- and visual-only features, this scheme was referred to as *audio-visual hierarchical* LDA (**AV-HiLDA**), or hierarchical discriminant feature fusion. In both cases, we modeled the generation process of a sequence of

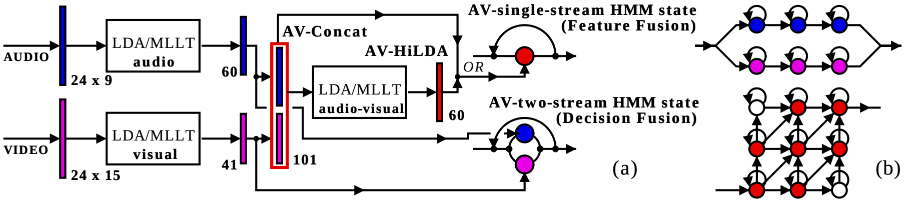

Figure 3: (a): Feature fusion and multi-stream HMM based decision fusion. (b): Phone-synchronous (state-asynchronous) multi-stream HMM with three states per phone and modality, and its equivalent product HMM; white circles denote states that are removed when limiting the degree of allowed asynchrony.

features (1) or (2) by a *single-stream* HMM, with Gaussian mixture *emission* (class c conditional observation) probabilities, given, in the case of (1) (a similar equation holds for (2)), by [8]

$$Pr\,[\,\mathbf{o}_t^{(\mathrm{AV})}\,|\,\mathsf{c}\,] = \sum_{k=1}^{K_{\mathsf{c}}} w_{\mathsf{c}\,k}\,\mathcal{N}_D\,(\,\mathbf{o}_t^{(\mathrm{AV})};\mathbf{m}_{\mathsf{c}\,k}\,,\mathbf{s}_{\mathsf{c}\,k}\,)\;. \tag{3}$$

In (3), $\mathbf{m}_{\mathsf{c}\,k}$ and $\mathbf{s}_{\mathsf{c}\,k}$ denote the D-dimensional *mean* and *diagonal covariance* vectors of the class c, k-th mixture component D-variate normal distribution with *prior* $w_{\mathsf{c}\,k}$.

Decision fusion

The main *decision fusion* approach considered was by means of the *multi-stream* HMM (MS-HMM) [9], with state emission (class c conditional) probability given by

$$Pr\,[\,\mathbf{o}_t^{(\mathrm{AV})}\,|\,\mathsf{c}\,] \;=\; \prod_{s\in\{\mathrm{A}\,,\mathrm{V}\}} \Big[\,\sum_{k=1}^{K_{s\,\mathsf{c}}} w_{s\,\mathsf{c}\,k}\,\mathcal{N}_{D_s}\,(\,\mathbf{o}_t^{(s)};\mathbf{m}_{s\,\mathsf{c}\,k}\,,\mathbf{s}_{s\,\mathsf{c}\,k}\,)\,\Big]^{\lambda_{s\,\mathsf{c}\,t}}\;. \tag{4}$$

In (4), $\lambda_{s\,\mathsf{c}\,t}$ are the stream exponents, that are non-negative, and, in general, depend on the modality s, the HMM state (class) c, and, locally, on the utterance frame (time) t. Initially, we considered *constant* exponents, λ_{A}, λ_{V} constrained to $\lambda_{\mathrm{A}} + \lambda_{\mathrm{V}} = 1$. We then separately trained single-stream audio- and visual-only HMMs, and we combined their emission probabilities in (4), after optimizing the exponents on held-out data. We referred to this scheme as **AV-MS-2**. However, superior results were obtained, by, instead, *jointly* training both streams of (4) using the *expectation-maximization* (EM) algorithm [8], and again optimizing the exponents on held-out data. We referred to this scheme as **AV-MS-1**. Subsequently, we relaxed the constant exponent assumption, considering utterance-level dependent exponents. We set the audio exponents $\lambda_{\mathrm{A}\,t}$ of the jointly trained MS-HMM (AV-MS-1) equal to the normalized (between 0 and 1) estimate of the degree of *voicing* present in the speech signal [10], averaged over the entire utterance, with the visual exponents being $\lambda_{\mathrm{V}t} = 1 - \lambda_{\mathrm{A}t}$. We referred to this scheme as **AV-MS-UTTER**.

Next, we relaxed the *state synchrony* assumption when combining single-stream likelihoods in (4), by enforcing synchrony at the *phone boundaries* only (each phone had three states). This gave rise to the *product* HMM, depicted in Fig.3(b) [4], [9]. In our particular implementation, we allowed asynchrony up to a single audio-visual state, as depicted in Fig.3(b). Similar to the state synchronous case, we jointly trained the resulting HMM parameters. We referred to this method as asynchronous decision fusion (**AV-PROD**).

Finally, we considered a "late" integration method using the *discriminative model combination* (**AV-DMC**) approach of [11]. In this technique, for every utterance, a list of n-best hypotheses $\{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\}$, obtained using an audio-only HMM, were first forced-aligned [8] to their corresponding phone sequences $\mathbf{h}_i = \{\mathsf{c}_{i,1}, \mathsf{c}_{i,2}, \ldots, \mathsf{c}_{i,N_i}\}$ by means of both audio- and visual-only HMMs. The resulting phone $\mathsf{c}_{i,j}$ boundaries were denoted by $[\,t_{i,j,s}^{\mathrm{start}}\,, t_{i,j,s}^{\mathrm{end}}\,]$, for $s \in \{\mathrm{A}\,,\mathrm{V}\}$, $j = 1,...,N_i$, and $i = 1,...,n$. The hypotheses were subsequently rescored by

$$Pr\,[\,\mathbf{h}_i\,] \;\sim\; Pr_{\mathrm{LM}}(\mathbf{h}_i)^{\lambda_{\mathrm{LM}}} \prod_{s\in\{\mathrm{A}\,,\mathrm{V}\}} \prod_{j=1}^{N_i} Pr\,(\mathbf{o}_t^{(s)},\; t \in [\,t_{i,j,s}^{\mathrm{start}}\,, t_{i,j,s}^{\mathrm{end}}\,]\;|\;\mathsf{c}_{i,j}\,)^{\lambda_{s\,\mathsf{c}_{i,j}}}\,,$$

Audio Condition:	Clean	Noisy	Audio Condition:	Clean	Noisy
Audio-only	14.44	48.10	AV-MS-1 (DF)	14.62	36.61
AV-Concat (FF)	16.00	40.00	AV-MS-2 (DF)	14.92	38.38
AV-HiLDA (FF)	13.84	36.99	AV-MS-PROD (DF)	14.19	35.21
AV-DMC (DF)	13.65 → 12.95	—	AV-MS-UTTER (DF)	13.47	35.27

Table 1: Test set audio-only and audio-visual word error rate (WER), %, for both clean and noisy audio. Two feature fusion (FF) and five decision fusion (DF) based audio-visual systems were evaluated.

where $Pr_{\rm LM}(\mathbf{h}_i)$ was the *language model* (LM) probability of hypothesis $\mathbf{h}_i$. All audio-visual phone weights and the LM weight were discriminatively trained on a held-out set [4], [11].

DATABASE AND EXPERIMENTS

The IBM ViaVoice audio-visual database

All experiments were performed on the IBM ViaVoice$^{\rm TM}$ audio-visual database, that consists of full-face frontal video and wideband audio of 290 subjects, uttering continuous read speech with a 10,400 word vocabulary. The database video is of size 704×480 pixels, interlaced, captured in color at a rate of 30 Hz, and it is MPEG2 encoded at the relatively high compression ratio of 50:1. The audio is synchronously collected with the video at a rate of 16 kHz, and at a relatively clean office environment at a 19.5 dB *signal-to-noise ratio* (SNR) [4].

The experimental framework

Approximately 42 hours of data were used in speaker-independent audio-visual ASR experiments, partitioned into the *training* set (239 subjects, 35 hours), used for HMM parameter estimation, the *held-out* set (25 subjects, 5 hours), used for training parameters relevant to audio-visual decision fusion, and the *test* set (26 subjects, 2.5 hours). Two audio conditions were considered: The original database clean audio and a degraded one with an 8.5 dB SNR, where the audio was artificially corrupted by additive "*babble*" speech ("cafeteria"-like) noise. All HMMs, as well as the LDA and MLLT matrices used in feature extraction, were trained and tested in the *matched* condition.

All experiments were conducted using the HTK toolkit [8] and a lattice rescoring strategy: Using the IBM LVCSR decoder with a trigram LM and IBM-trained HMMs, appropriate ASR lattices were generated. These lattices were then rescored using the HTK decoder by various cross-word, context-dependent triphone HTK-trained HMMs, based on a number of feature sets and fusion strategies.

Experimental results

First, baseline audio-only results were obtained for both clean and noisy audio conditions, using HMMs trained in the matched audio condition to rescore IBM-generated audio lattices at the corresponding condition. Performance deteriorated significantly from a 14.44% *word error rate* (WER) for clean audio to a 48.10% WER in the noisy case (see also Table 1). Subsequently, the relative performance of the two visual front ends was investigated. Visual-only HMMs were trained and used to rescore noisy audio lattices. Of course, such lattices contained audio information, therefore the obtained results could not be interpreted as visual-only recognition. The DCT visual features resulted in a 58.14% WER, outperforming the AAM features, that achieved a 64.00% WER. Therefore, the DCT features were exclusively used in the audio-visual fusion experiments, next.

A number of feature and decision fusion techniques were used to train appropriate audio-visual HMMs at the clean (19.5 dB) and noisy (8.5 dB) conditions, that were subsequently evaluated by rescoring clean audio and noisy audio-visual lattices, respectively (the latter were obtained at IBM using HiLDA feature fusion). The performance of all algorithms is depicted in Table 1. Notice that every fusion method considered outperformed audio-only ASR in the noisy audio case. Furthermore, hierarchical discriminant feature fusion (HiLDA) and decision fusion by means of the multi-stream HMM with utterance-based stream exponents (MS-UTTER), as well as by using the product HMM (MS-PROD), improved ASR in the clean audio condition too. The latter achieved a 27% relative reduction in WER in the noisy audio case (from a 48.10% audio-only WER to 35.21% audio-visual), whereas the MS-UTTER method outperformed all fusion methods in clean speech achieving a 7% relative reduction in WER (from 14.44% to 13.47%). The DMC method was only applied to the clean speech case, and it reduced WER to 12.95%, amounting to a 5% reduction from its clean audio-only baseline of 13.65% (this was different from the 14.44% audio-only result due to the use of n-best list instead of lattice rescoring). Overall, decision fusion methods outperformed feature fusion techniques. Further investigation of exponent estimation of the state-synchronous and product multi-stream HMMs is expected to yield additional improvements.

SUMMARY

We provided a summary of our work on speaker-independent audio-visual large-vocabulary, continuous speech recognition, during the Johns Hopkins Summer 2000 Workshop. We studied both image transform and model based visual features, as well as a number of feature fusion and decision fusion techniques for audio-visual integration. In our particular implementation, the DCT based visual front end outperformed the AAM one. Among the audio-visual fusion techniques considered, hierarchical discriminant feature fusion, as well as decision fusion by means of the product HMM with limited state asynchrony, or the state-synchronous multi-stream HMM with utterance dependent audio-visual exponents, consistently improved recognition performance for both clean and noisy audio conditions considered. This constitutes the first time that such improvements have been obtained in the LVCSR domain.

ACKNOWLEDGMENTS

We would like to thank the entire audio-visual ASR research team at the Johns Hopkins summer 2000 workshop for fruitful collaboration, as well as F. Jelinek (CLSP) for hosting the workshop. Furthermore, we would like to acknowledge contributions to this work by G. Iyengar and A.W. Senior (IBM) for help with face and mouth region detection for the IBM ViaVoice^{TM} audio-visual data and by E. Helmuth (IBM) for help in data collection.

References

[1] Campbell, R., Dodd, B., and Burnham, D. eds., *Hearing by Eye II*, Psychology Press, Hove, 1998.

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

[3] Dupont, S. and Luettin, J., “Audio-visual speech modeling for continuous speech recognition,” *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, 2000.

[4] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J., “Audio-visual speech recognition,” *Final Workshop 2000 Report*, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, 2000 (`http://www.clsp.jhu.edu/ws2000/final_reports/avsr/`).

[5] Potamianos, G., Verma, A., Neti, C., Iyengar, G., and Basu, S., “A cascade image transform for speaker independent automatic speechreading,” *Proc. ICME*, vol. II, pp. 1097–1100, 2000.

[6] Matthews, I., Potamianos, G., Neti, C., and Luettin, J., “A comparison of model and transform-based visual features for audio-visual LVCSR,” (In Press), *Proc. ICME*, 2001.

[7] Cootes, T.F., Edwards, G.J., and Taylor, T.J., “Active appearance models,” *Proc. Europ. Conf. Comp. Vision*, pp. 484–498, 1998.

[8] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., *The HTK Book*. Entropic Ltd., Cambridge, 1999.

[9] Bourlard, H. and Dupont, S., “A new ASR approach based on independent processing and recombination of partial frequency bands,” *Proc. ICSLP*, vol. 1, pp. 426–429, 1996.

[10] Berthommier, F. and Glotin, H., “A new SNR-feature mapping for robust multistream speech recognition,” *Proc. Int. Congress Phonetic Sciences*, vol. 1, pp. 711–715, 1999.

[11] Beyerlein, P., “Discriminative model combination,” *Proc. ICASSP*, vol. 1, pp. 481–484, 1998.