

MULTISTAGE INFORMATION FUSION FOR AUDIO-VISUAL SPEECH RECOGNITION

S. M. Chu, V. Libal, E. Marcheret, C. Neti, and G. Potamianos

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

This paper looks into the information fusion problem in the context of audio-visual speech recognition. Existing approaches to audio-visual fusion typically address the problem in either the feature domain or the decision domain. In this work, we consider a hybrid approach that aims to take advantages of both the feature fusion and the decision fusion methodologies. We introduce a general formulation to facilitate information fusion at multiple stages, followed by an experimental study of a set of fusion schemes allowed by the framework. The proposed method is implemented on a real-time audio-visual speech recognition system, and evaluated on connected digit recognition tasks under varying acoustic conditions. The results show that the multistage fusion system consistently achieves lower word error rates than the reference feature fusion and decision fusion systems. It is further shown that removing the audio only channel from the multistage system only leads to minimal degradations in recognition performance while providing a noticeable reduction in computational load.

1. INTRODUCTION

Exploiting speech-relevant information in the visual modality, in addition to acoustic speech, has been demonstrated as an effective approach to improve the performance of automatic speech recognition, especially under adverse acoustic environments [1]. One of the challenging and actively pursued topics in audio-visual speech recognition is the search for a fusion scheme that can best utilize the information present in the two individual modalities.

Studies on human perception give strong evidence supporting the bimodal nature of the speech recognition process, and have provided important cues to the research on automatic systems. However, exactly how the fusion of audio and visual information takes place in human perception is not yet answered. In particular, the existing studies do not agree on the central question of at which stage the fusion occurs. The various existing models can be roughly categorized into two groups: the early integration models and the late integration models [2]. In the perspective of the early integration models, fusion takes place before the recognition stage, whereas in the late integration models, the classifications of each modality are performed independently from one another, and the fusion is carried out by integrating the decisions given by the parallel channels.

Similar groupings can also be observed in the different approaches to audio-visual fusion in automatic systems. One

class of methods perform fusion at the feature level. These are usually referred as early integration or feature fusion in the literature [1]. In a system based on feature fusion, the fusion takes place before the classification stage, thus requires only one classifier. The other class of methods carry out fusion in the decision level, which are referred as late integration or decision fusion methods. These methods typically combine the likelihood scores of the single-modality classifiers to recognize audio-visual speech.

In general, feature fusion allows explicit joint modeling of the two modalities, and therefore is capable of capturing the low-level inter-modal dependencies. At the same time, the approach imposes a rigid coupling between the audio and visual information streams, which makes it unable to handle asynchronous audio-visual speech events. Decision fusion, on the other hand, detaches the two channels in the feature level, and permits formulations that directly address channel weighting [3] and audio-visual asynchrony [4]. However, the underlying assumption of conditional independence between the audio and visual observations implied by the decision fusion approach may obscure any inter-modal dependencies that exist in the feature domain.

In this work, we consider a multistage approach to the fusion problem that aims to take advantage of both the feature fusion and the decision fusion methodologies by exploiting the complementarity between the two. The system constitutes a real-time implementation of hybrid fusion, first proposed in [1]. In essence, the method attempts to utilize the audio-visual information from all available perspectives to maximize the possible performance gain brought by the addition of the visual modality. In this paper, we describe a general framework that facilitates multistage fusion, and proceed with an experimental study of a series of fusion schemes allowed by the framework. The rest of the paper is organized as follows. In the next section we discuss the problem formulation and introduce the multi-stage fusion framework. The development of the audio-visual speech recognition system is described in Section 3. We discuss the fusion experiments and present the corresponding results in Section 4.

2. MULTISTAGE INFORMATION FUSION

The fusion of audio-visual speech is an instance of the general sensory fusion problem. The sensory fusion problem arises in the situation when multiple channels carry complementary information about different components of a system. In the case of audio-visual speech, the two modalities manifest two aspects of the same underlying speech production process. From an observer's point of view, the audio channel

and the visual channel represent two interdependent stochastic processes. We seek a framework that can model the two individual processes as well as their low-level dependencies.

2.1. Problem Formulation

Decision fusion, in its classic form, is based on the naïve Bayes paradigm. It makes the assumption that the channels are conditionally independent given the class label, thus the joint likelihood of the observations from the audio and visual channels can be factorized as

$$p(\mathbf{x}_a, \mathbf{x}_v | C) = p(\mathbf{x}_a | C) \cdot p(\mathbf{x}_v | C) \quad (1)$$

where variables \mathbf{x}_a and \mathbf{x}_v are the observations from the audio and the visual channels, respectively; and C is the class label. The independent structure of the variables can be clearly represented using a graphical model as shown in Figure 1(b).

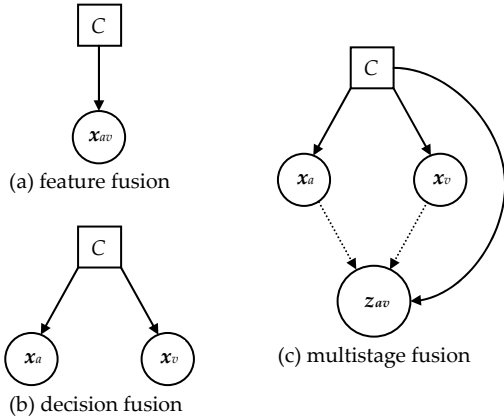


Figure 1. Different approaches to audio-visual fusion represented as graphic models: (a) feature fusion, (b) decision fusion, and (c) multistage fusion.

Feature fusion, shown in Figure 1 (a), makes no assumption with regard to the independence between the audio and visual observations. Indeed, the joint likelihood of the two observation variables is directly characterized by a single density $p(\mathbf{x}_{av} | C)$, where \mathbf{x}_{av} is obtained by joining the audio and visual vectors.

$$\mathbf{x}_{av} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_v \end{pmatrix} \in \mathfrak{R}^{d_{av}}, \quad d_{av} = d_a + d_v \quad (2)$$

The multistage fusion approach, shown in Figure 1 (c), can be viewed as an extension to decision fusion. In addition to the audio and visual observation variables, a low-level fusion node is added to the graph. This new node is derived from both the audio and visual observables through transformation Γ ,

$$\mathbf{z}_{av} = \Gamma(\mathbf{x}_a, \mathbf{x}_v, C) \quad (3)$$

In our implementation, Γ is defined by two consecutive transformations. First, linear discriminate analysis (LDA) is applied to project the joint audio-visual observation vector to a lower dimensional space while seeking the best discrimina-

tion among the speech classes of interest. Second, the lower dimensional feature space is rotated through maximum likelihood linear transform (MLLT) to improve statistical modeling using multivariate Gaussian densities with diagonal covariance matrices. Hence,

$$\mathbf{z}_{av} = \mathbf{M}_{av} \mathbf{L}_{av} \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_v \end{pmatrix} \quad (4)$$

where, \mathbf{M}_{av} is the MLLT matrix and \mathbf{L}_{av} is the LDA matrix. Beyond the transformation, we shall treat the new node as another observation variable. Further, we assert,

$$p(\mathbf{x}_a, \mathbf{x}_v, \mathbf{z}_{av} | C) = p(\mathbf{x}_a | C) \cdot p(\mathbf{x}_v | C) \cdot p(\mathbf{z}_{av} | C) \quad (5)$$

Note that the conditional independence between the new audio-visual node and the other two unimodal nodes implied in (5) does not hold in general. However, this factorization permits a straightforward implementation in a decision fusion system by adding one more channel modeling the joint audio-visual observations.

2.2. Multistage Fusion Implementation

As discussed in the previous section, multistage fusion can be implemented on a decision fusion baseline (Figure 2). Specifically, the information carried in each channel is processed by a dedicated learner/classifier which gives a quantity y_i that approximates the likelihood of the observed data in the given channel:

$$y_i = f_i(\mathbf{o}_i | C, \theta_i) \propto p(\mathbf{o}_i | C) \quad (6)$$

where \mathbf{o}_i is the observation for channel i , and θ_i denotes the particular parameterization scheme used to model the target distribution in this channel. The choice of the domain of class label C determines the exact place where the decision fusion takes places. In this experiment, C is set at the HMM state level. A direct implementation of equation (5) will sim-

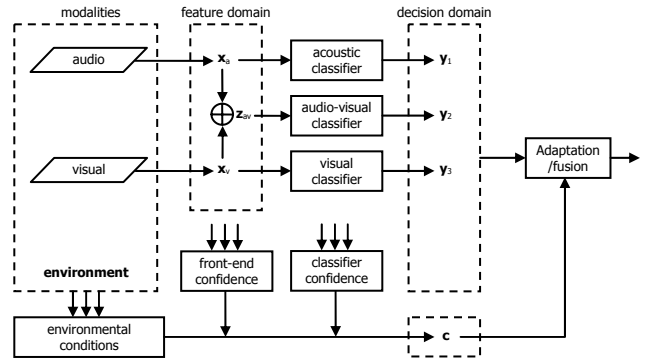


Figure 2. Multistage information fusion allows the integration of audio and visual modalities in both the feature domain as well as the decision domain.

ply be a product of y_i 's from all three channels. This implies that channels are essentially weighted equally under all conditions. Empirically, this orthodox Bayesian approach usually yields sub-optimal classification results. One plausible

way to improve is to adjust the contribution of an individual channel to the overall decision by an exponential weight.

$$g = \prod_i [f_i(\mathbf{o}_i | C, \theta_i)]^{w_i(t)} = \prod_i y_i^{w_i(t)} \quad (7)$$

In practice, the weight w_i for a given channel can be determined based on a number of measures, including environmental conditions, front-end confidence, and classifier confidence. In the scope of this paper, we shall adhere to stationary channel weights obtained through grid search.

3. AVSR SYSTEM DEVELOPMENT

3.1. Audio-Visual Front-End

Visual feature extraction is a crucial component in audio-visual speech recognition systems. The visual front-end in our system extracts appearance-based features within a region of interest (ROI) defined on the mouth area of the speaker.

Given the video input, the system first performs face detection at frame-level, using multi-scale template matching based on a distance measure composed of the two-class (face/non-face) Fisher linear discriminant and the error incurred by projecting the candidate vector to a lower dimensional “face space” obtained through principal component analysis (PCA). Following face detection, 26 key facial points (e.g., eye corners and mouth corners) are tracked using algorithms reported in [9]. The tracking results provide the location, size, and orientation estimates of the mouth. These parameters are subsequently smoothed over time and used to determine a 64×64 -pixel ROI. Notice that the distance scores calculated in the face detection and the key-point tracking steps can be utilized to derive a confidence measure of the visual front-end, as indicated in Figure 2.

The visual features are computed by applying a two-dimensional separable DCT to the sub-image defined by the ROI, and retaining the top 100 coefficients with respect to energy. The resulting vectors then go through a pipeline con-

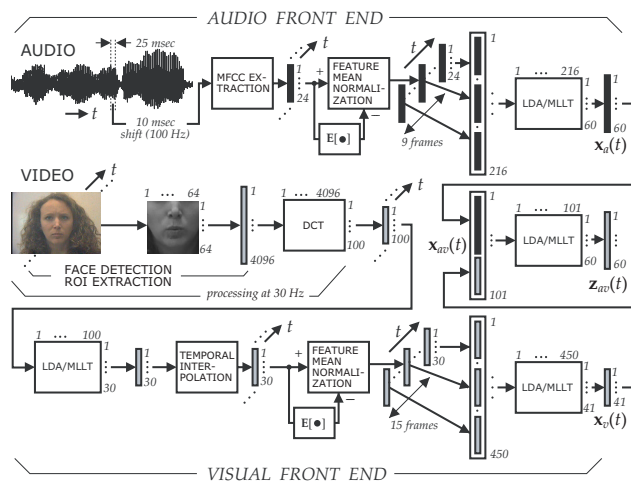


Figure 3. The audio-visual speech front-end outputs three time-synchronous feature streams at 100Hz: 60-dimensional audio features, 41-dimensional visual features, and 60-dimensional fused features

sisting of intra-frame LDA/MLLT, temporal interpolation, and feature mean normalization, producing a 30 dimensional feature stream at 100Hz. To account for inter-frame dynamics, fifteen consecutive frames in the stream are joined and subject to another LDA/MLLT step to give the final visual feature vectors with 41 dimensions.

The basic audio features extracted by the front-end are 24-dimensional Mel-frequency cepstral coefficients. After cepstral mean normalization, nine consecutive frames are concatenated and projected onto a 60 dimensional space through an LDA/MLLT cascade.

The feature-level fusion in the multistage fusion paradigm described in Section 2 is independent from the choice of classifiers, and therefore can be implemented as an additional component in the front-end. The time-synchronous audio and visual feature streams are first concatenated at each frame. The resulting 101-dimensional joint vectors are then transformed using LDA/MLLT to form a new observation stream with 60 dimensions. The complete schematic of the audio-visual front-end is given in Figure 3.

3.2. HMM Implementation

In this work, the decision level fusion is set to take places at the HMM state level. Thus, it can be implemented with a multi-stream HMM. The recognition system uses three-state, left-to-right phonetic HMMs with context-dependent states. The instances of the sub-phonetic states are identified by growing a decision tree that clusters left and right contexts spanning up to five phones on each side. The states are specified by the terminal nodes of the tree, and the corresponding observation streams are modeled by mixtures of Gaussian densities with diagonal covariance matrices.

We have observed in our audio-only speech recognition system based on the same configuration that the evaluation of the state emission probabilities accounts for more than 60% of the CPU time. The use of multiple observation streams in the audio-visual system will considerably increase the computational load. Therefore, special attention must be paid in order to achieve real-time performance. In our implementation, we considered a novel approach in which only one of the observation streams is fully evaluated, while the other two streams are only evaluated at a subset of states dynamically determined by the master stream. The details of this algorithm will be presented in a forthcoming paper.

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental Setup

The audio-visual speech recognition system is evaluated on a connected-digit recognition task using the IBM studio-DIGIT audio-visual database [1]. The corpus consists of full-face frontal video of 50 subjects, uttering 7 and 10-digit strings. A total of 6.7K utterances are recorded in a studio environment with uniform background and lighting. The acoustic signal to noise ratio (SNR) of the recorded data is measured at 19.5 dB.

The dataset is partitioned into three subsets: a training set containing 5.4K utterances, a test set with 623 utterances,

and a held-out set including 663 utterances. The held-out set is used in the grid search for the optimal channel weights defined in (7).

To evaluate the recognition performance in noisy environments, three noisy acoustic conditions were simulated by adding random segments of speech babble recordings to the clean speech samples. The average SNR of the three noisy test sets are 16dB, 8.5dB, and 6dB. The HMMs are trained using the clean data, based on a context tree with 159 leaves modeled by 3.2K Gaussian densities.

4.2. Experimental Results

In addition to the multistage fusion system, we considered four other unimodal and bimodal recognition systems for comparison, including, an audio-only system, a visual-only system, an audio-visual system with feature fusion, and an audio-visual system with decision fusion. They are referred as “a”, “v”, “av”, and “a+v” in the results, respectively. The multistage fusion system is referred as “a+v+av”. To facilitate comparison, all five systems share the same basic HMM topology.

The recognition results measured in word error rate (WER) is summarized in Table 1. We note the following

Table 1. Recognition results measured in word error rate on connected-digits task under different acoustic conditions.

SNR	16 dB	8.5 dB	6 dB
a	4.10%	21.9%	33.1%
v	55.1%	55.1%	55.1%
av	2.81%	15.8%	26.6%
a+v	2.27%	15.5%	22.8%
a+v+av	1.48%	11.8%	17.4%
av+v	3.37%	11.9%	17.6%

observations. First, it is evident that the audio-visual systems outperform the unimodal systems. Further, the classic decision fusion system consistently gives better recognition accuracy than the feature fusion system. Most importantly, we observe that the multistage system achieves significant improvement in WER over both the feature fusion and the decision fusion systems at all three SNR levels. For instance, at 8.5dB, the multistage system attains a WER of 11.8%, which represents a 24% relative reduction over the decision fusion system’s reading of 15.5%. The overall gain in performance is clearly visible in Figure 4.

The results strongly imply that the joint audio-visual observation channel obtained through low-level fusion provides important cues about the speech events which are not captured in the conventional two-channel decision fusion paradigm, despite the seemingly redundant representation of the audio-visual observations. This is indeed in conformity with the reasoning that gives rise to the proposed multistage fusion approach.

Finally, an alternative multistage system was considered, in which the audio observation at the decision fusion stage is removed. The resulting two-stream system is therefore comprised of a joint audio-visual observation node and a visual observation variable in the decision domain (“av+v”). It is

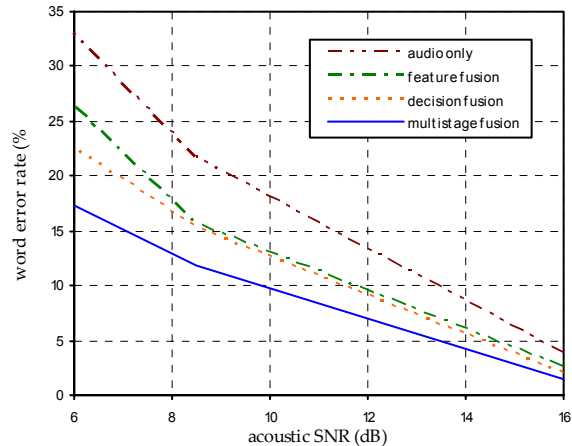


Figure 4. Recognition results on connected-digit task show that the multistage fusion system achieves lower word error rates than other systems at all three noisy conditions.

interesting to see that discarding the audio only observation only has a slight negative effect on recognition performance, with the exception at 16dB, when the audio-channel is less noisy. In a practical system with limited computation resource, the alternative multistage system represents a viable solution that is able to give noticeable improvement in accuracy over a conventional decision fusion system while maintaining comparable computational requirements.

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. IEEE*, vol. 91, no. 9, pp. 1306-1326, 2003.
- [2] A. Q. Summerfield, “Lipreading and audio-visual speech perception,” *Philosophical Transactions of the Royal Society of London, Series B*, vol. 335, pp. 71-78, 1992.
- [3] G. Potamianos and H. P. Graf, “Discriminative training of HMM stream exponents for audio-visual speech recognition,” in *Proc. IEEE ICASSP*, Seattle, WA, 1998, pp. 3733-3736.
- [4] S. M. Chu and T. S. Huang, “Audio-visual speech modeling using coupled hidden Markov models,” in *Proc. IEEE ICASSP*, Orlando, FL, 2002, pp. 2009-2012.
- [5] S. Dupont and J. Luetttin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Trans. Multimedia*, vol. 2, pp. 141-151, 2000.
- [6] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.
- [7] T. Chen, “Audiovisual speech processing,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9-21, 2001.
- [8] S. M. Chu, M. Yeung, L. Liang, and X. Liu, “Environment adaptive multi-channel biometrics,” in *Proc. IEEE ICASSP*, Hong Kong, 2003, vol. 5, pp. 788-791.
- [9] A. W. Senior, “Face and feature finding for a face recognition system,” in *Proc. Int. Conf. Audio Visual-based Biometric Person Authentication*, pp. 154-159, 1999.