ASYNCHRONOUS STREAM MODELING FOR LARGE VOCABULARY AUDIO-VISUAL SPEECH RECOGNITION

Juergen Luettin

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

Gerasimos Potamianos, Chalapathy Neti

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
{gpotam,cneti}@us.ibm.com

ABSTRACT

This paper addresses the problem of audio-visual information fusion to provide highly robust speech recognition. We investigate methods that make different assumptions about asynchrony and conditional dependence across streams and propose a technique based on composite HMMs that can account for stream asynchrony and different levels of information integration. We show how these models can be trained jointly based on maximum likelihood estimation. Experiments, performed for a speaker-independent large vocabulary continuous speech recognition task and different integration methods, show that best performance is obtained by asynchronous stream integration. This system reduces the error rate at a 8.5 dB SNR with additive speech "babble" noise by 27 % relative over audio-only models and by 12 % relative over traditional audio-visual models using concatenative feature fusion.

1. INTRODUCTION

Automatic speech recognition systems that use visual information from the speaker's mouth, so-called *lipreading* or *speechreading*, have been shown to improve the word recognition rate over audio-only systems, especially in noisy audio conditions. One of the main challenges in audio-visual speech recognition (AVSR) systems is the audio-visual information integration problem. The main issues in information integration are, (a) the class conditional dependence assumption made across streams, (b) the level (e.g. frame, phone, word) of integration, and (c) the kind (e.g. feature, partial likelihood, partial decision) of integration. Mainly two different integration methods have been reported in the literature [1]: *Feature fusion* and *decision fusion*. *Feature fusion* assumes class-conditional dependence between streams and frame synchronous information integration. In *decision fusion*, class-conditional independence is assumed and integration is typically done at the phrase level by integrating hypotheses of both streams.

Here, we describe integration techniques based on *multi-stream HMMs* that can be placed between these two extreme cases. These models allow for different assumptions about the level of integration and the degree of asynchrony to be made. We show how these models can be trained jointly using maximum likelihood training and report results for a large vocabulary continuous audio-visual speech database. Related work based on *feature fusion* and *decision fusion* is presented in [2] and [3], respectively.

2. DATABASE AND RECOGNITION TASK

All experiments have been performed on a continuous, large vocabulary, speaker independent database that has been collected at IBM Thomas J. Watson Research Center [2, 4]. The database consists of full-face frontal video and audio of 290 subjects, uttering ViaVoice^{TM} training scripts, i.e., continuous read speech with mostly verbalized punctuation (dictation style), and a vocabulary size of approximately 10,500 words. The duration of the entire database is approximately 50 hours, thus it is the largest audio-visual database collected to date.

The database has been partitioned into a number of disjoint sets of which we have used three for our experiments: a training set (35 hours, 239 subjects), a held-out data set (5 hours, 25 subjects) to train parameters for decision fusion, and a test set (2.5 hours, 26 subjects).

To assess the benefits of the visual modality to LVCSR for both clean and noisy audio, two different audio conditions were considered: The original clean wideband audio, and audio that is artificially corrupted by additive "*babble*" noise resulting in a 8.5 dB SNR. Experiments were performed according to matched audio conditions, i.e. using the same audio conditions for training and testing, which can be considered a "best case" scenario.

3. AUDIO-VISUAL FEATURES

The acoustic feature vectors are of dimension 60 and are extracted for both clean and noisy conditions at a rate of 100 Hz [4]. These features are obtained by a *linear discriminant analysis* (LDA) data projection, applied on a concatenation of nine consecutive feature frames consisting of a 24-dimensional *discrete cosine transform* (DCT) of mel-scale filter bank energies. LDA is followed by a *maximum likelihood linear transform* (MLLT) based data rotation. *Cepstral mean subtraction* (CMS) and *energy normalization* are applied to the DCT features at the utterance level, prior to the LDA/MLLT feature projection. For both clean and noisy audio, the LDA and MLLT matrices are estimated using the training set data in the *matched* condition.

Visual feature extraction is based on pure video pixel, appearance based features of the mouth region [2, 4]. This is achieved by the subsequent processes of face detection, mouth detection, and discrete cosine image transform of the subject's mouth area. The resulting features are further processed by an LDA projection and an MLLT feature rotation. The 41-dimensional feature vectors, extracted at 60 Hz, are linearly interpolated to obtain a frame rate of 100 Hz, synchronous to the audio features.

4. BASELINE RECOGNITION SYSTEM

A baseline speech recognition system has been implemented using the HTK toolkit [5]. Cross-word context dependent phoneme models are used as speech units and are modeled with HMMs with Gaussian mixture class-conditional observation probabilities. These are trained based on maximum likelihood estimation using embedded training by means of the Expectation-Maximization (EM) algorithm. Context-dependent phone models are obtained by decision-tree based clustering. The training procedure has been the same for all parameter sets, whether audio-only, visual-only, or synchronous audio-visual.

All decoding experiments were performed by lattice rescoring. Lattices were generated off line using the IBM LVCSR decoder with a trigram language model and an IBM trained HMM system. The lattices are rescored using different models and integration strategies. The language model score and the word insertion penalty are roughly optimized during rescoring. Three lattices were generated based on either clean audio, noisy audio, or noisy audio-visual features. The visual-only system has been rescored using the lattices from the noisy audio features. As these lattices contain some audio information, it's performance can not be considered as the real visual-only performance.

5. FRAME SYNCHRONOUS INTEGRATION

This feature fusion method, also referred to as *early integration*, is based on time-synchronous integration of the audio and visual features and makes the assumption of class-conditional dependence between the two streams.

The audio- and video-only feature vectors at instant t, are denoted by $\mathbf{o}_s^{(t)} \in \mathsf{R}^{D_s}$, of dimension D_s, where $s = \mathrm{A}, \mathrm{V}$, respectively. The joint audio-visual feature vector is the concatenation of the two, namely

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

where $D = D_{\mathrm{A}} + D_{\mathrm{V}}$.

The class conditional observation probabilities, of a sequence of such features is given by

$$P[\mathbf{o}^{(t)}|\,\mathsf{c}] = \sum_{j=1}^{J_{\mathsf{c}}} w_{\mathsf{c}j}\, \mathcal{N}_D(\mathbf{o}^{(t)}; \mathbf{m}_{\mathsf{c}j}, \mathbf{s}_{\mathsf{c}j}), \tag{2}$$

where $\mathsf{c} \in \mathcal{C}$ denote the HMM context dependent states (classes). In addition, mixture weights $w_{\mathsf{c}j}$ are positive adding up to one, J_{c} denotes the number of mixtures, and $\mathcal{N}_d(\mathbf{o}; \mathbf{m}, \mathbf{s})$ is the d-variate normal distribution with mean $\mathbf{m}$ and a diagonal covariance matrix, its diagonal being denoted by $\mathbf{s}$. The dimension of the integrated feature vector was 101 for our experiments.

6. MULTI-STREAM HMM

If we relax the assumption of class-conditional dependence between the streams and model each observation stream with a single-stream HMM we obtain the general form of a multi-stream HMM (Fig. 1). The class conditional observation likelihood of the multi-stream HMM is the product of the observation likelihoods of its single-stream components, raised to appropriate *stream exponents* that capture the reliability of each modality, or, equivalently, the confidence of each single-stream classifier. Such model has been

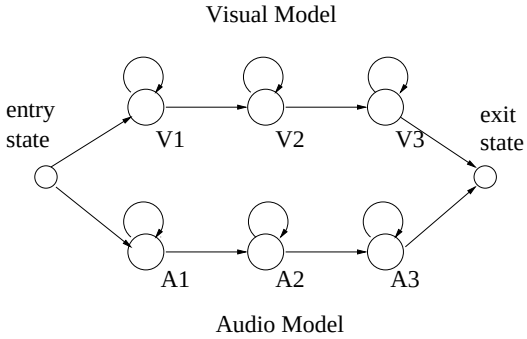

Fig. 1: Example of a multi-stream HMM with 2 streams and 3 states in each stream.

considered in speech-noise decomposition [6], multi-band audio-only ASR [7] and in small vocabulary audio-visual ASR [8–12]. Here, we extend that work in several ways: we describe a method for training the streams jointly, we apply it to the LVCSR domain, and we perform comparisons with other fusion algorithms.

Given the bimodal observation vector $\mathbf{o}^{(t)}$, the state emission (class conditional) probability of the multi-stream HMM is,

$$P[\mathbf{o}^{(t)}|\,\mathsf{c}] = \prod_{s \in \{\mathrm{A},\mathrm{V}\}} \Big[\sum_{j=1}^{J_{s\mathsf{c}}} w_{s\mathsf{c}j}\, \mathcal{N}_{D_s}(\mathbf{o}_s^{(t)}; \mathbf{m}_{s\mathsf{c}j}, \mathbf{s}_{s\mathsf{c}j}) \Big]^{\lambda_{s\mathsf{c}t}}, \tag{3}$$

where $\lambda_{s\mathsf{c}t}$ are the stream exponents, that are non-negative, and, in general, depend on the modality s, the HMM state (class) $\mathsf{c} \in \mathcal{C}$, and, locally, on the utterance frame (time) t. Such time-dependence can be used to capture the "local" reliability of each stream, and can be estimated on basis of stream confidences or acoustic signal characteristics.

In this work, we consider global, modality-dependent weights, i.e., two stream exponents constant over the entire database

$$\lambda_s = \lambda_{s\mathsf{c}t}, \quad \text{for all } \mathsf{c} \in \mathcal{C}, \quad \text{all } t, \quad \text{and } s = \mathrm{A}, \mathrm{V}. \tag{4}$$

Exponents λ_{A} and λ_{V} are constrained to satisfy

$$0 \le \lambda_{\mathrm{A}}, \lambda_{\mathrm{V}} \le 1, \quad \text{and} \quad \lambda_{\mathrm{A}} + \lambda_{\mathrm{V}} = 1. \tag{5}$$

6.1. State Synchronous Integration

Since in (3), c denote the HMM context dependent states, the states across the audio- and video-stream are constrained to be synchronous. We therefore denote this integration method *state synchronous integration.*

Training the multi-stream HMM consists of two tasks: First, estimation of its stream component parameters (mixture weights, means, variances, and state transition probabilities) and estimation of appropriate stream exponents (4) that satisfy (5).

Maximum likelihood parameter estimation by means of the EM algorithm [13] can be used in a straightforward manner to train the first set of parameters. This can be done in two ways: Either train each stream component parameter set separately, based on single-stream observations, and subsequently combine the resulting single-stream HMMs as in (3), or, train the entire parameter set (excluding the exponents) at once using the bimodal observations.

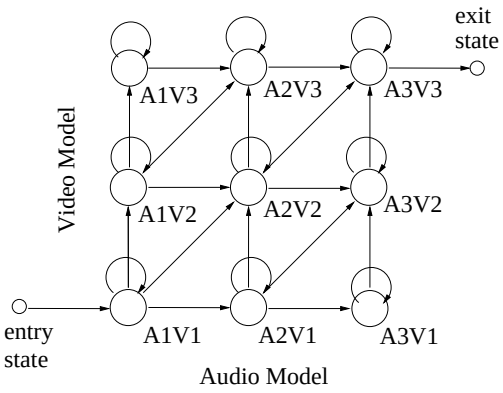

Fig. 2: Audio-visual product HMM, equivalent to the multi-stream HMM shown in Fig. 1. The emission probabilities of the audio- and video-stream are denoted by A_n and V_m, where n and m is the state of the audio and visual stream model, respectively.

An obvious drawback of the first approach is that the two single-modality HMMs are trained asynchronously (i.e., using different forced alignments), whereas (3) assumes that the HMM stream components are state synchronous. The alternative is to train the whole model at once, in order to enforce state synchrony. This approach requires an a-priori choice of stream exponents. Such stream exponents cannot be obtained by maximum likelihood estimation [10, 11]. A simple technique consists in directly minimizing the WER on a held-out data set, which was used here.

6.2. Model Synchronous Integration – The Product HMM

It is well known that visual speech activity usually precedes the audio signal by as much as 120 ms [14], which is close to the average duration of a phoneme. The multi-stream HMM discussed above, however, enforces state synchrony between the audio and visual streams. It is therefore of interest to relax the assumption of state synchronous integration, and instead allow some degree of asynchrony between the audio and visual streams.

An extension of the multi-stream HMM allows the single-stream HMMs to be in asynchrony within a model but forces them to be in synchrony at the model boundaries. Single-stream log-likelihoods are linearly combined at such boundaries using stream weights. A reasonable choice for forcing synchrony constitute the phone boundaries.

Decoding based on this integration method requires to individually compute the best state sequences for both audio and visual streams. To avoid the computation of two best state paths, the model can be formulated as a composite, or product, HMM [6, 8, 12]. Decoding under such a model requires to calculate a single best path. The product HMM consists of composite states that have audio-visual emission probabilities of the form (3), with audio and visual stream components that correspond to the emission probabilities of certain audio and visual-only HMM states (Fig. 2).

As depicted in Fig. 3, the single-stream emission probabilities can be tied across states, therefore the original number of mixture weight, mean, and variance parameters can be kept in the new model. The transition probabilities of the single-modality HMMs are now shared by several transition probabilities in the composite model.

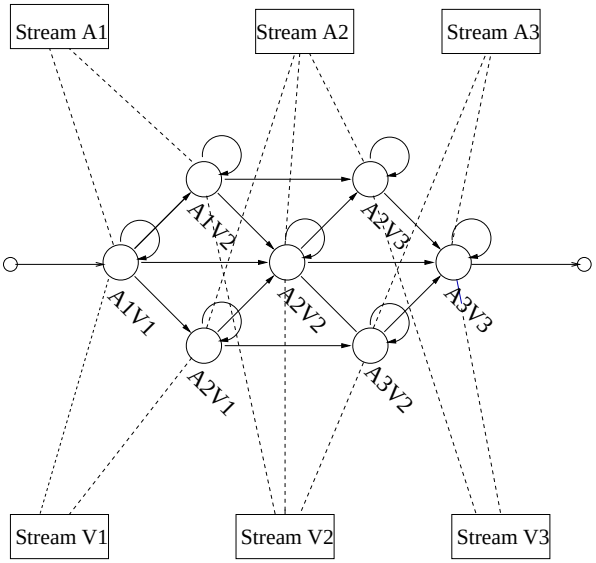

Fig. 3: Stream tying in a product HMM with limited state asynchrony.

The product HMM allows to restrict the degree of asynchrony between the two streams, by excluding certain composite states in the model topology (Fig. 3). As the number of states in the composite HMM is the product of the number of states of all its individual streams, such restrictions can reduce this number considerably, and speed up decoding. In the extreme case, when only the states that lie in its "diagonal" are kept, the model becomes equivalent to the state synchronous model.

Similarly to the multi-stream HMM, training of the product HMM can be done separately, or jointly. In joint training, all product HMM parameters (with the exception of the stream exponents) are trained at once, by means of the EM algorithm, and using the audio-visual training data as shown in Fig. 3.

7. EXPERIMENTS

Systems for audio-only, visual-only, and frame synchronous integration (AV-FSI) features were trained separately and their recognition results are depicted in Table 1. The performance of the audio-only system deteriorates considerably in the noisy condition and is only slightly better than the performance of the visual-only system. Audio-visual features based on synchronous integration (AV-FSI) result in higher performance for the noisy audio case but increase the error rate for clean audio.

We have trained two multi-stream HMMs using the training procedures described in the previous section: First, we obtained a multi-stream HMM, referred to as AV-MS-2, by separately training two single-stream models, and subsequently combining them. A second multi-stream HMM, denoted by AV-MS-1, was jointly trained as a single model. For both models, the stream exponents were estimated to values $\lambda_{\mathrm{A}} = 0.7$, $\lambda_{\mathrm{V}} = 0.3$, in the clean audio case, and $\lambda_{\mathrm{A}} = 0.6$, $\lambda_{\mathrm{V}} = 0.4$, in the noisy audio one. These values were obtained by minimizing the WER of various AV-MS-1 trained models on the held-out data set. The audio-visual recognition results on the test set for both clean and noisy audio environments are depicted in Table 1. All results are in word error rate

	Clean audio	Noisy audio
Visual-only		*51.08*
Audio-only	14.44	48.10
AV-FSI	16.00	40.00
AV-MS-1	14.62	36.61
AV-MS-2	14.92	38.38
AV-PROD	14.19	35.21

Table 1: Recognition performance for different features and audio-visual integration methods. All results are in WER (%).

(WER) (%). As expected, the AV-MS-1 models outperformed the AV-MS-2 ones, but the AV-MS-1 HMM was unable to improve the clean audio-only system. This is somewhat surprising, and could indicate an inappropriate choice of stream exponents in this case. On the other hand, in the noisy audio case, the AV-MS-1 based decision fusion significantly outperformed the audio-only baseline.

In our experiments using the product HMM, and in view of the results in the multi-stream HMM case, we have only considered the training approach where both streams are trained jointly. We have limited the degree of asynchrony allowed to one state only. Lattice rescoring experiments were conducted on the test set for both clean and audio conditions, using the jointly trained product HMM (AV-PROD). Stream exponents $\lambda_{\mathrm{A}} = 0.6$, $\lambda_{\mathrm{V}} = 0.4$, were used in the clean audio case, and $\lambda_{\mathrm{A}} = 0.7$, $\lambda_{\mathrm{V}} = 0.3$, in the noisy audio one. The obtained results are depicted in Table 1. Clearly, the product HMM consistently exhibits superior performance to both audio-only and AV-MS-1 models. Overall, it achieves a 2% WER relative reduction in the clean audio case and a 27% one in noisy audio, over the corresponding audio-only system. Its WER reduction over the frame synchronous integration is 12% and over state synchronous integration by AV-MS-1 models is 3.8%.

8. CONCLUSIONS

We have proposed and evaluated different information integration techniques for audio-visual speech recognition. Frame synchronous integration results in improved performance at a 8.5 dB SNR with additive speech "babble" noise but increases the error rate in the clean audio case which might be due to the high dimension of the resulting feature vector. State synchronous integration using multi-stream HMM degrades performance only slightly in the clean case but improves the performance considerably in the noisy scenario. Better results are achieved if the multi-stream HMMs are trained jointly. Overall best performance is obtained by model-synchronous integration. This approach results in a significant improvement of 27 % relative WER reduction over audio-only matched models for the noisy audio. This method even decreases the WER in the clean audio case by 2%. This result indicates that audio-visual speech is modeled more accurately with asynchronous stream models.

9. ACKNOWLEDGMENTS

This work has been performed as part of the Johns Hopkins University CLSP Workshop 2000. We would like to thank the other AVSR team members, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou for their collaboration as well as F. Jelinek for hosting the workshop.

10. REFERENCES

[1] Marcus E. Hennecke, David G. Stork, and K. Venkatesh Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines: Models, Systems and Applications*, David G. Stork and Marcus E. Hennecke, Eds., vol. 150 of *NATO ASI Series F: Computer and Systems Sciences*, pp. 331–349. Springer-Verlag, Berlin, 1996.

[2] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2001.

[3] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2001.

[4] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Tech. Rep., Johns Hopkins University, Baltimore, 2000, http://www.clsp.jhu.edu/ws2000/final_reports/avsr.

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

[6] P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1990, pp. 845–848.

[7] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. Int. Conf. on Spoken Language Processing*, 1996, vol. 1, pp. 426–429.

[8] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1996, pp. 821–824.

[9] A. Rogozan, P. Deléglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," in *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, 1997, pp. 61–64.

[10] P. Jourlin, "Word dependent acoustic-labial weights in HMM-based speech recognition," in *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, 1997, pp. 69–72.

[11] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1998, vol. 6, pp. 3733–3736.

[12] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, 1993.

[14] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, no. 3, pp. 236–244, 1998.