ASYNCHRONY MODELING FOR AUDIO-VISUAL SPEECH RECOGNITION

Guillaume Gravier, Gerasimos Potamianos, and Chalapathy Neti

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

E-mails: {ggravier,gpotam,cneti}@us.ibm.com

ABSTRACT

We investigate the use of a particular type of multi-stream HMM, known as the product HMM, for the automatic recognition of audio-visual speech. Such a model allows the modeling of asynchrony between the audio and visual state sequences at a variety of levels (phone, syllable, word, etc). In this paper, we investigate a product model that is synchronized at the phone boundary level, allowing limited degree of state sequence asynchrony. Furthermore, we investigate joint training of all product HMM parameters at once, instead of composing the model from separately trained audio- and visual-only HMMs. We demonstrate that the resulting product HMM reduces WER on a multi-subject connected digit recognition task by up to 32% relative over the separately trained product HMM. Compared to the audio-only performance at 10 dB SNR, the new model achieves an effective SNR gain of 9 dB, about 1.5 dB more than the separately trained model. We also demonstrate that it outperforms a number of common audio-visual combination techniques both in high-noise and relatively clean environments (19.5 dB SNR). Partial results are also presented for a speaker-independent LVCSR task, to be completed in the final paper.

1. INTRODUCTION

We have made significant progress in automatic speech recognition (ASR) for well-defined applications like dictation and medium vocabulary transaction processing tasks in relatively controlled environments. However, ASR performance has yet to reach the level required for speech to become a truly pervasive user interface. Indeed, even in "clean" acoustic environments, and for a variety of tasks, state of the art ASR system performance lags human speech perception by up to an order of magnitude [1]. In addition, current systems are quite sensitive to channel, environment, and style of speech variations, as a number of techniques for improving ASR robustness have met limited success in severely degraded environments, mismatched to system training [2]–[4]. Clearly, novel, non-traditional approaches, that use orthogonal sources of information to the acoustic input, are needed to achieve ASR performance closer to the human speech perception level, and robust enough to be deployable in field applications. Visual speech constitutes a promising such source, obviously much less affected by the acoustic environment and noise.

Both human speech production and perception are bimodal in nature [5]. This fact has recently motivated significant interest in automatic recognition of visual speech, formally known as automatic lipreading, or speechreading. Work in this field aims at improving ASR by exploring the visual modality of the speaker's mouth region, in addition to the traditional audio modality, thus giving rise to audio-visual automatic speech recognition (AVASR) systems [6]–[13]. There are two main problems in achieving this goal: First, the design of the visual front end, i.e. how to obtain informative visual features given the video of the speaker's face, and, second, the combination of such features with the traditional audio features. In this paper, we concentrate on the latter issue.

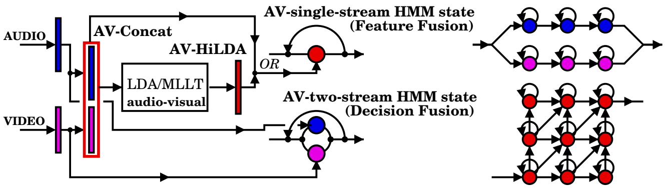

Fig. 1. Left: Two feature fusion methods versus multi-stream HMM based decision fusion. Right: Phone-synchronous (state-asynchronous) multi-stream HMM with three states per phone and modality, and its equivalent product HMM; audio stream emission probabilities are tied along rows, visual ones along columns.

A number of audio-visual integration strategies appear in the literature that can be grouped into two categories (see Fig 1): Feature fusion methods, which combine single modality features into audio-visual features and use a classical classifier to recognize the joined audio-visual features [10, 11, 14], and decision fusion methods that instead combine single-modality (audio-only and visual-only) classifier decisions [8]–[13]. Typically, decision fusion methods linearly combine the log-likelihoods of the audio and visual stream observations, employing the multi-stream hidden Markov model (MSHMM) framework, a popular model for multi-band ASR [15]. In most cases, such combination occurs at the HMM state level, thus forcing synchrony between the audio and visual sub-phonetic classes. It is well known, however, that visual speech activity precedes the audio signal by as much as 120 ms [6]. To model this phenomenon, many researchers have proposed combining the single-modality log-likelihoods at a coarser level, such as the phone (or, word) level [8, 9, 10, 12], giving rise to the composite, or product HMM [16] (see also Fig 1). However, in all these works, the product HMM has either been composed by audio-only and visual-only HMMs that have been independently trained for each stream [8, 9, 12], or has been jointly trained without the appropriate stream ty-

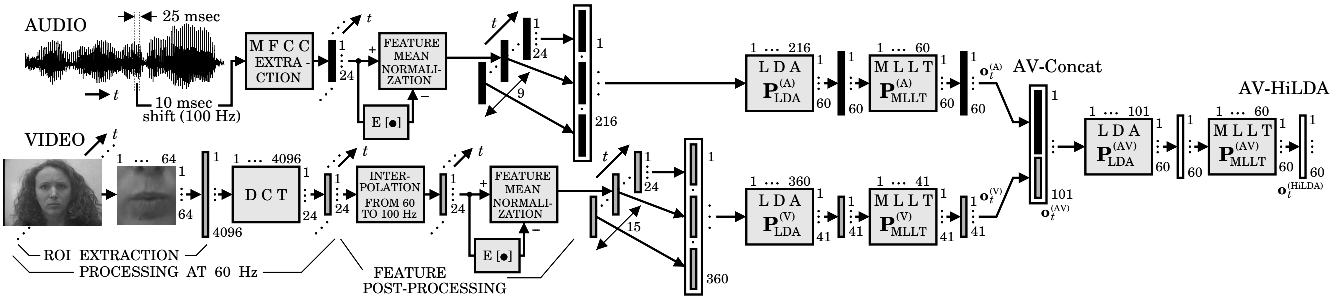

Fig. 2. Feature extraction for audio-visual ASR.

ing [10]. As a result, to date, comparisons between state- and phone-synchronous multi-stream HMMs have been inconclusive and incomplete.

In this paper, we proceed to jointly estimate all the product HMM parameters in a single step, using the appropriate stream density tying across states, by employing the standard EM HMM parameter estimation algorithm. Such a scheme ensures that the audio-visual state asynchrony within each phone is modeled at training, whereas stream tying guarantees that the new model has the same number of HMM density parameters as the state-synchronous MSHMM. We demonstrate that the new training scheme significantly outperforms the product HMM composed from independently trained single-stream HMMs, by conducting audio-visual recognition experiments on a 50-subject connected digits database, over a range of acoustic signal-to-noise ratios (SNR). We also show that it improves ASR over alternative decision and feature fusion techniques, thus providing a comprehensive study on the relative performance of a number of audio-visual integration algorithms. Results on a speaker-independent large-vocabulary continuous speech recognition (LVCSR) task are also reported, using some of the fusion techniques mentioned above. State-asynchronous HMM results for LVCSR will be reported in the final paper.

2. FEATURE EXTRACTION AND STATISTICAL MODELING FOR AUDIO-VISUAL ASR

Given an audio-visual utterance, let us denote its extracted audio- and visual-only features by $\{\mathbf{o}_t^{(s)}\}$, where $s = \mathrm{A}, \mathrm{V}$, respectively. Fig. 1 summarizes the feature extraction process. From the audio signal, 24 mel-frequency cepstral coefficients are retained as audio features, after feature mean normalization. At every frame, 9 consecutive audio features are concatenated, subsequently projected to a lower dimensional space using linear discriminant analysis (LDA), and rotated by a maximum likelihood linear transform (MLLT). The audio features are extracted with a 100 Hz frame rate and the final dimension is $D_{\mathrm{A}} = 60$.

Given the video of the speaker's face, sampled at 60 Hz, a normalized mouth region-of-interest is first extracted, and subsequently compressed using a discrete cosine transform (DCT). The 24 highest energy DCT coefficients are retained, and after linear interpolation sub-sampling from 60 to 100 Hz and mean normalization, they result to static visual features. As for the audio features, dynamic information is obtained by concatenating consecutive static features over 15 frames, and by applying the LDA and MLLT transforms. The final visual feature dimension is $D_{\mathrm{V}} = 41$.

Classical single-stream HMMs are used to model sequences of audio-only or visual-only features, where the state conditional density function is a Gaussian mixture defined as

$$P[\mathbf{o}_t^{(s)}|\,\mathsf{i}\,] = \sum_{k=1}^{K_{s\,\mathsf{i}}} w_{s\,\mathsf{i}\,k}\,\mathcal{N}_{D_s}\,(\,\mathbf{o}_t^{(s)}; \mathbf{m}_{s\,\mathsf{i}\,k}\,, \mathbf{s}_{s\,\mathsf{i}\,k}\,)\,. \qquad (1)$$

In audio-visual feature fusion, single stream HMMs are also used to model sequences of concatenated audio-visual features

$$\mathbf{o}_t^{(\mathrm{AV})} = [\,\mathbf{o}_t^{(\mathrm{A})\,\top}, \mathbf{o}_t^{(\mathrm{V})\,\top}\,]^\top \in \mathsf{R}^{D} \qquad (2)$$

of dimension $D = D_{\mathrm{A}} + D_{\mathrm{V}}$ (concatenative feature fusion: AV-Concat), or any transformation of (2), such as the hierarchical discriminant (AV-HiLDA) feature fusion [14]. Such features are obtained after a second LDA-MLLT transform is applied on $\mathbf{o}_t^{(\mathrm{AV})}$ (see also Figs.1 and 2).

In multi-stream HMM based decision fusion, the most likely speech class or word sequence is obtained by linearly combining the log-likelihoods of the audio- and visual-only single-stream HMMs using appropriate weights. This is equivalent to an HMM with emission "score" given by

$$Pr[\mathbf{o}_t^{(\mathrm{AV})}|\mathsf{i}_{\mathrm{A}},\mathsf{i}_{\mathrm{V}}] = \prod_{s\in\{\mathrm{A},\mathrm{V}\}} \Big[\sum_{k=1}^{K_{s\mathsf{i}_s}} w_{s\mathsf{i}_s k}\,\mathcal{N}_{D_s}(\mathbf{o}_t^{(s)}; \mathbf{m}_{s\mathsf{i}_s k}, \mathbf{s}_{s\mathsf{i}_s k})\Big]^{\lambda_s}, \qquad (3)$$

for audio state i_{A} and visual state i_{V}. For the state-synchronous multi-stream HMM, $\mathsf{c}_{\mathrm{A}} = \mathsf{c}_{\mathrm{V}} \leftarrow \mathsf{c}$ holds, i.e., the HMM states are the same as the ones of the single-stream HMMs. In the case of the product HMM however, (3) gives rise to composite states where c_{A} and c_{V} are restricted to belong to the same phone (see also Fig.1). Note that although the number of states in the product HMM increases, the number of emission density parameters do not increase compared to the state synchronous model, due to the stream parameter tying, inherent in (3).

With the exception of exponents λ_s, maximum-likelihood estimates of all other HMM parameters in (3) can be obtained using the EM algorithm, either separately per-stream, or at once, using joint training. Subsequently, exponent values can be obtained using discriminative training [12, 13], or by minimizing the word error rate (WER) on a held-out set [9, 10], as in this paper. For state-synchronous HMMs, both separate [9, 10], and joint training schemes [10,

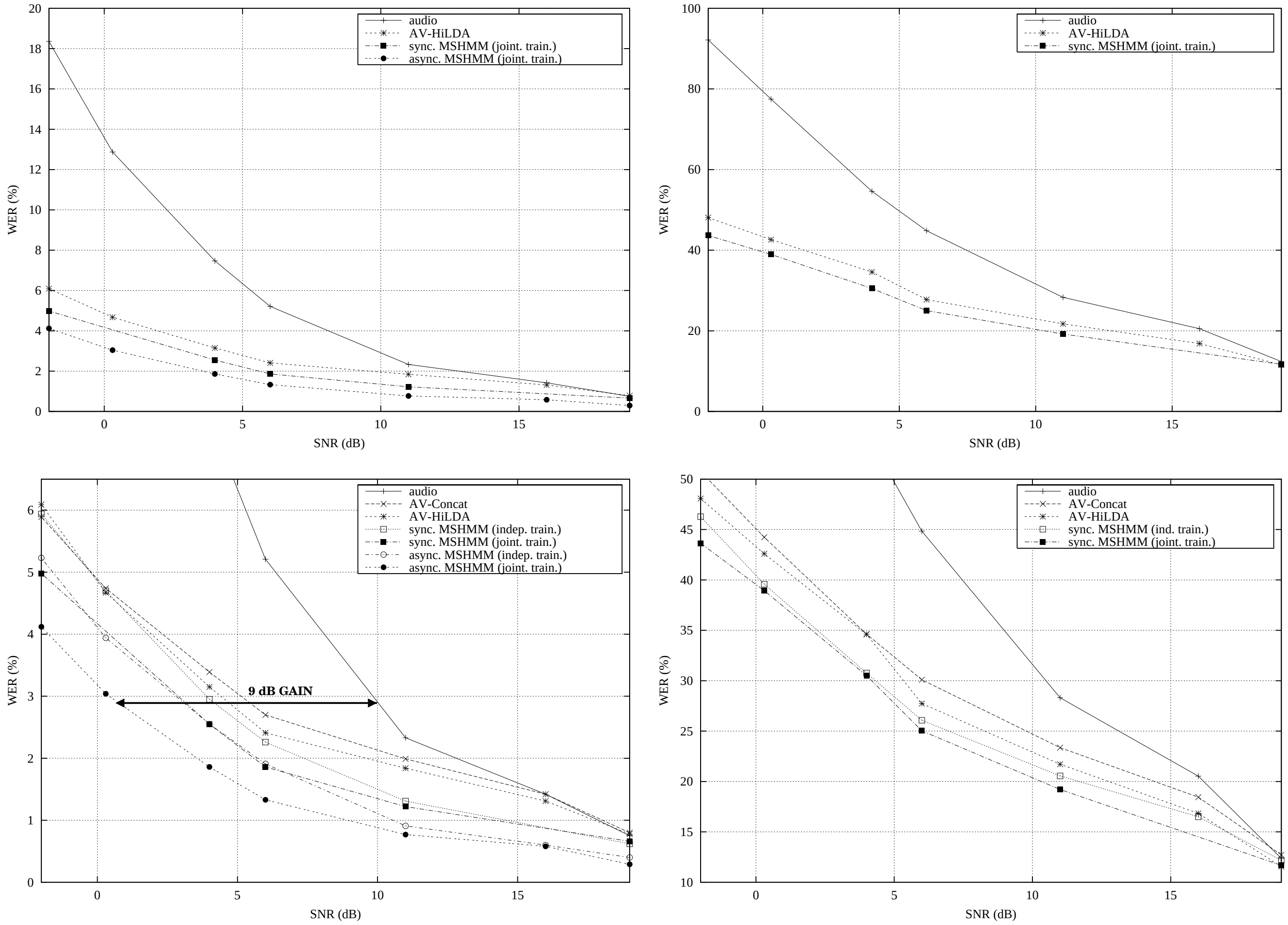

Fig. 3. Test set audio-only and audio-visual WERs using various fusion methods, depicted vs. audio channel SNR, for connected digit recognition (left) and LVCSR (right). Plots on the top row show the full-scale results while plots on the bottom row focus on the AV method performances.

13] have been previously considered. For state-asynchronous HMMs only separate training has been considered [8, 9, 10, 12], whereas in [10], appropriate stream tying has not been employed. Proper joint product HMM training is employed for the first time in this paper.

3. RESULTS

We close this extended abstract with a brief presentation of our experimental results on audio-visual ASR. We report a comparison of the jointly trained state-asynchronous MSHMM to the separately trained one, as well as to the two feature fusion techniques mentioned above and both jointly and separately trained state-synchronous multi-stream HMMs. Results are given first for a 50-subject connected digit recognition task corresponding to a 10-hour audio-visual database collected similarly to the IBM ViaVoice dataset [10]. Test set audio-only and audio-visual WERs on this task are depicted in Fig. 2 for various SNR values of the audio channel. Clearly, the jointly trained product HMM outperforms the separately trained one at all SNRs. For example, at 19 dB, it results to a 0.29% WER compared to 0.40% of the latter, a 32% relative WER reduction, whereas at -2 dB SNR, it reaches 4.12% compared to 5.23%, a 27% relative reduction. Note that the audio-only performance is respectively 0.75% and 18.4% for these two conditions, therefore pointing out that the visual modality benefit to ASR is dramatic. This benefit translates to a 9 dB effective SNR gain compared to the audio-only performance at 10 dB SNR (see Fig. 2). It is also worth mentioning that the product HMM outperforms all other fusion techniques considered in the previous section, and that multi-stream HMMs in general outperform both feature fusion methods. Joint stream training is also superior to separate training in the case of the state-synchronous multi-stream HMM. Similar observations hold in the case of the speaker-independent LVCSR task based on the IBM ViaVoice database [10] (see also Fig. 2). Product HMM results on LVCSR will be reported in the final paper.

4. REFERENCES

[1] R.P. Lippmann, "Speech recognition by machines and humans," Speech Comm., 22(1):1–15, 1997.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Process., 2(4):578–589, 1994.

[3] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comp. Speech and Lang., 9:171–185, 1995.

[4] R. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," In C.-H. Lee, F.K. Soong, and K.K. Paliwal, Eds., Automatic Speech and Speaker Recognition. Advanced Topics. Kluwer Academic Pub., pp. 357–384, 1997.

[5] R. Campbell, R., B. Dodd, and D. Burnham, Eds., Hearing by Eye II. Psychology Press Ltd. Pub., 1998.

[6] C. Bregler and Y. Konig, "Eigenlips" for robust speech recognition. Proc. IEEE Intl. Conf. Acous. Speech Sig. Process., pp. 669–672, 1994.

[7] D.G. Stork and M.E. Hennecke, Eds., Speechreading by Humans and Machines. Springer, 1996.

[8] M.J. Tomlinson, M.J. Russell, and N.M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," Proc. IEEE Intl. Conf. Acous. Speech Sig. Process., pp. 821–824, 1996.

[9] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Trans. Multimedia, 2:141–151, 2000.

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

[11] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization." IEEE Sig. Process. Mag., 18:9–21, 2001.

[12] S. Nakamura, "Fusion of audio-visual information for integrated speech processing," In J. Bigun and F. Smeraldi, Eds., Audio-and Video-Based Biometric Person Authentication, Springer-Verlag, pp. 127–143, 2001.

[13] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," To appear: Int. Conf. Acous. Speech Sig. Process., 2002.

[14] G. Potamianos, J. Luettin, and C. Neti, Hierarchical discriminant features for audio-visual LVCSR. Int. Conf. Acous. Speech Sig. Process., pp. 165–168, 2001.

[15] H. Bourlard, and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," Proc. Int. Conf. Spoken Lang. Process., pp. 426–429, 1996.

[16] P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise," Proc. Int. Conf. Acous. Speech Sig. Process., pp. 845–848, 1990.