DYNAMIC STREAM WEIGHT MODELING FOR AUDIO-VISUAL SPEECH RECOGNITION

Etienne Marcheret, Vit Libal, Gerasimos Potamianos

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{etiennem, libalvit, gpotam}@us.ibm.com

ABSTRACT

To generate optimal multi-stream audio-visual speech recognition performance, appropriate dynamic weighting of each modality is desired. In this paper, we propose to estimate such weights based on a combination of acoustic signal space observations and single-modality audio and visual speech model likelihoods. Two modeling approaches are investigated for such weight estimation: one based on a sigmoid fitting function, the other employing Gaussian mixture models. Reported experiments demonstrate that the later approach outperforms sigmoid based modeling, and is dramatically superior to the static weighting scheme.

Index Terms— Speech Processing, Audio-Visual Speech Recognition, Multi-Modal Fusion, Multi-Stream HMM.

1. INTRODUCTION

Recently, there has been significant interest in the use of multi-stream hidden Markov models (HMMs) for automatic speech recognition (ASR) [9]. For example, such models have been successfully considered for multi-band ASR [1], separate static and dynamic acoustic feature modeling [12], as well as for audio-visual ASR [3], [11].

In its application in audio-visual speech recognition, the multi-stream approach gives rise to an effective paradigm to fuse and model the two separate information sources carried in the audio and visual observations. Specifically, it has been demonstrated that multi-stream decision fusion attains significant improvement in recognition accuracy over the state-of-the-art single-stream based fusion methods, e.g., hierarchical linear discriminant analysis (HiLDA) [11].

To deal effectively with varying noise conditions on either or both the audio and visual channels a number of approaches can be taken. The expected noisy conditions can be accounted for in the training step, which has the practical limitation of trying to cover all expected conditions. Keeping the training step fixed, one can compensate the audio and visual feature spaces to fit the trained models at test time. One such approach using multiple stream feature space maximum likelihood linear regression (FMLLR) is discussed in [7].

Even with the effectiveness of the multi-stream FMLLR in compensating for the training/testing mismatch, the issue of accurate and robust computation of the multi-stream HMM (MSHMM) observation probabilities remains. Modeling the audio/visual observations jointly is difficult. Instead, modeling the streams independently and assigning stream weights that capture the reliability of the individual channels to the computed HMM observation probabilities has been shown to be effective [6], [4], [5]. These studies however have limitations that stem from either the particular features used to estimate the channel reliability (eg. derived strictly from the acoustic feature space), or the underlying model that maps such features to stream weights (lookup table or sigmoid fitting function).

In this paper, we attempt to address some of these limitations by focusing on two areas, robust features to capture channel reliabilities, and a flexible framework to estimate weights from such features. Our work is based on the state synchronous multi-stream HMM (MSHMM) applied to the audio/visual speech recognition problem. With the synchronous assumption and given the audio visual observation $x_t = [x_{a,t}, x_{v,t}]$ for frame time t the MSHMM state s conditional likelihood assuming stream independence is given by

$$P(x_t|s) = \prod_{m \in a,v} P(x_{m,t}|s)^{\lambda_{m,t}} \quad (1)$$

where state s denotes the context dependent phoneme, and stream weights $\lambda_{m,t}$ capture the stream reliabilities, and are assumed to satisfy $\lambda_{a,t} + \lambda_{v,t} = 1$. To estimate these weights, we propose to use a combination of acoustic signal space observations and single-modality audio and visual speech model likelihoods, in conjunction with a novel statistical modeling technique based on a full covariance GMM. Experiments conducted on an appropriate audio-visual database demonstrate the robustness of the proposed approach over a wide range of environments.

The rest of the paper is organized as follows. In section 2 we describe features usable for capturing the reliability of a stream. Novel features to capture this reliability are described in section 2.2. Section 3 discusses two techniques for weight computation, one using a function fitting, the other based on a statistical framework. The audio-visual ASR system is described in Section 4, experimental results are presented in Section 5, with conclusions in Section 6.

2. STREAM RELIABILITY FEATURES

As discussed in the Introduction, the initial step to weight estimation is to select informative features about the reliability of the two streams of interest. In this section we propose two types of features for this purpose. The first are based on speech model likelihoods trained on the audio and visual modalities, whereas the second are based on the acoustic signal alone. The proposed approach is novel in two ways; the particular signal space features used, as well as their combination with the former. Next we briefly describe the two types of features.

2.1. Audio-Visual Likelihood Based Features

In [6] and [4] dispersion measures either at the phoneme level or HMM state level are employed. In [6] the dispersion is based on the phoneme posterior probabilities. Here we choose the dispersion

measures as defined in [4] and computed at the MSHMM state level. For each stream m, we have at frame time t the N-best log-likelihood difference $\mathcal{D}^m_{1,t}$ and the N-best log-likelihood dispersion $\mathcal{D}^m_{2,t}$

$$\begin{aligned}\mathcal{D}^m_{1,t} &= \frac{1}{N-1}\sum_{n=2}^{N} log\frac{P(x_{m,t}|s_{t,1})}{P(x_{m,t}|s_{t,n})} \\ \mathcal{D}^m_{2,t} &= \frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{n'=n+1}^{N} log\frac{P(x_{m,t}|s_{t,n})}{P(x_{m,t}|s_{t,n'})},\end{aligned} \quad (2)$$

where the MSHMM state $s_{t,n}$ represents a ranking of the state conditional likelihoods. In the following N=5. Equation (2) captures the reliability of the stream through the entropy of the class conditional distribution, the more "peaked" the distribution, the more confident one can be that the true state at time t is the state with the top rank. Equation (2) defines a 4 dimensional speech model likelihood based feature space (two features for each of the two modalities).

2.2. Acoustic Signal Based Features

The energy based observations discussed in this section are based on features used in our speech activity detection work [10]. Among the features proposed there, we use two only features here: the so called "low energy track" and the "mid to low energy track". These features capture the signal to noise ratio (SNR) and the non-speech signal energy.

These features are based on the instantaneous root mean square (rms) of the acoustic signal and are computed iteratively on a left to right (causal) fashion, as follows

$$\begin{aligned}lt(t) &= (1-\alpha_{l,t})\times lt(t-1) + \alpha_{l,t}\times rms(t) \\ mt(t) &= (1-\alpha_m)\times mt(t-1) + \alpha_m \times rms(t),\end{aligned} \quad (3)$$

where the "filter" coefficients are given by

$$\alpha_m = 0.1 \quad \text{and} \quad \alpha_{l,t} = \left(\frac{lt(t-1)}{rms(t)}\right)^2 \quad (4)$$

The final features are obtained as the logarithm of (3). More details can be found in [10]

3. DYNAMIC STREAM WEIGHT MODELING

With the observations discussed in sections (2.1) and (2.2) the problem now becomes one of determining the appropriate stream weight for the corresponding observation. In our work we investigate two approaches for this purpose discussed next.

3.1. Sigmoid Function Fitting

Optimal weight selection can be viewed as a function $\lambda_t = F_\lambda(w_1(t), w_2(t)...w_N(t))$ of $N \geq 0$ variables $w_i(t)$ – selected features that characterize the audio or visual stream reliability at a given time instant. Then, training the optimal weight classifier means inferring the function F_λ. While the function values fall into the interval $< 0, 1 >$, the shape of the function F_λ is generally unknown. For sake of simplicity we will assume that the function is smooth and has monotonic dependency on each of the variables $w_i(t)$. Indeed, the initial experiments revealed that the relation between the weight λ and e.g. audio stream's SNR can be modeled by such function. One of possible functions that fulfill requirements of smoothness, monotonicity and output values from interval $< 0, 1 >$ is a sigmoid function:

$$F_\lambda(w_1, w_2..w_N) = \frac{1}{1+\exp(a_0 + \sum_{i=1}^{N} a_i w_i)} \quad (5)$$

where a_0, a_i are unknown parameters. Sigmoid function corresponds very well to the initial experimental measurements of the optimal weight dependency on acoustic stream SNR. Also this function has been previously used for similar purposes in [6], [4].

Having selected a sigmoid as a the function to model the relation between the optimal weight λ and the stream reliability features w_i, all we need is to estimate optimal values of a_i, $i = 0..N$ from equation (5) using the data from the training set. The ultimate objective is to minimize an error rate of the speech recognition using the HMM stream weighting delivered by $F_\lambda(t)$ over the training set. To achieve this, the feature space is first clustered into the N_r regions r_j, $j = 1..N_r$, each containing roughly the same amount of samples from the training set and then for each cluster a best scoring weight $\lambda^\circ_{c_j}$ is looked up from the weight sweep measurement on the training set. The value of $\lambda^\circ_{c_j}$ is then considered as a representative value of F_λ at the feature space point $\bar{w}^{r_j} = \{w_1^{r_j}, w_2^{r_j}..w_N^{r_j}\}$ that corresponds to the center of gravity of the region r_j. The nonlinear least squares regression may be used to find the best fit of F_λ to the set of points $\{\lambda^\circ_{c_j}, \bar{w}^{r_j}\}$.

3.2. Statistical Weight Modeling

In this framework, the weight posterior is given by Bayes' rule

$$p(\lambda_t/\bar{w}_t) = \frac{p(\bar{w}_t/\lambda_t)p(\lambda_t)}{p(\bar{w}_t)}, \quad (6)$$

where $\bar{w}_t$ denotes the stream reliability informative feature vector. The problem now becomes modeling the posterior shown in (6). From the posterior we may compute the expected λ_t which is optimal under some chosen criterion. In this work the objective function to be optimized is the WER of the ASR process. This optimization will yield the λ posterior distributions, from which we may compute the optimal weight with the expectation operation

$$\lambda^*_t = E\{\lambda/\bar{w}_t\} = \int_\lambda \lambda p(\lambda/\bar{w}_t)d\lambda. \quad (7)$$

To make the problem tractable we quantize the weights, chosen empirically at quantization levels 0.05, we have $\lambda^{(j)} = j * 0.05$, for $j = 0, .., 20$. The expectation operation becomes:

$$\lambda^*_t = E\{\lambda/\bar{w}_t\} = \sum_j \lambda^{(j)} p(\lambda^{(j)}/\bar{w}_t). \quad (8)$$

Note that this quantization allows us to treat the λ computation as an M-ary hypothesis test:

$$\lambda^*_t = \arg\max_j p(\lambda^{(j)}/\bar{w}_t). \quad (9)$$

In the equations above, $p(\lambda^{(j)}/\bar{w}_t)$ is estimated using a full covariance GMM model (FCGMM) as discussed next. Therefore, the two methods of computing λ^*_t using (8) and (9) will be referred to in our experiments as "GMM avg" and "GMM max" respectively.

As mentioned, we train the λ distributions based on the minimum WER criterion. We mix clean training data with noise, such that we get a resulting sweep of average SNR's, here chosen between

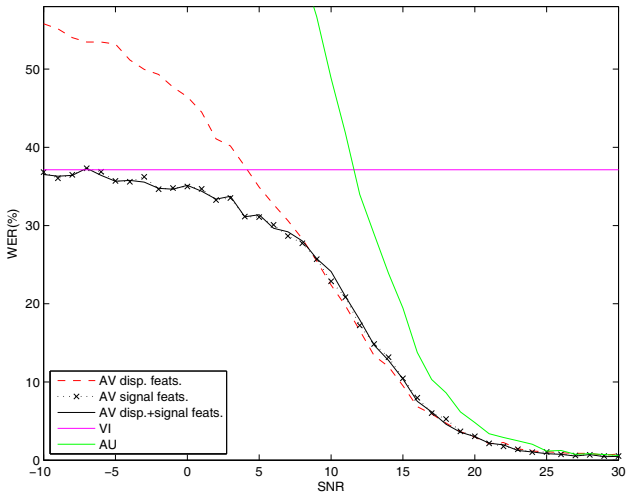

Fig. 1. *Impact of proposed features on audio-visual ASR performance, using the average GMM derived weights. Test set WER, %, is depicted over a range of SNRs for dispersion only features, signal based ones, as well as their combination. Audio and visual only WERs are also shown.*

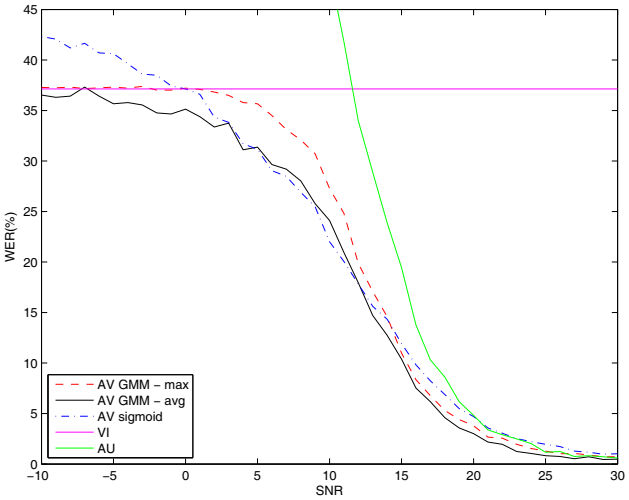

Fig. 2. *Comparison of proposed dynamic weight modeling techniques. Audio-visual WER, %, is shown vs. SNR for the sigmoid and the two GMM based proposed models. Audio and visual only WERs are also depicted.*

-10 dB and 30 dB (see Experiments Section below). The recognizer is then run over all quantized $\lambda^{(j)}$ values, and the stream reliability informative features are pooled in the $\lambda^{(j)}$ bins of minimum WER. Around each of these $\lambda^{(j)}$ feature pools we estimate a full covariance Gaussian mixture model with two mixture components.

4. AUDIO-VISUAL ASR SYSTEM AND DATA

The visual stream of our system is generated from the IBM infrared headset [8]. The infrared headset is specially designed equipment that captures the video of the speaker's mouth region, independently of the speaker's movement and head pose. It reduces environmental lighting effect on captured images, allowing good visibility of the mouth ROI even in a dark room. Since the headset consistently focuses on the desired mouth region, face tracking is no longer required. Given video from this device, the visual front-end component extracts appearance-based features within a region of interest (ROI) defined on the mouth area of the speaker. The ROI extraction on headset captured video is based on tracking two mouth corners of the recorded subject. This allows correcting slight positioning errors, boom rotation, and rescaling of the physical mouth size. The visual features are computed by applying a two-dimensional separable DCT to the sub-image defined by a 64×64 pixel ROI, and retaining the top 100 coefficients with respect to energy. The resulting vectors pass though a pipeline consisting 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 (VI stream) with 41 dimensions. Details can be found in [8].

The 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, generating the AU feature stream.

The system is built on 22kHz audio and 720x480 pixel resolution at 30 Hz video. The MSHMM training database consists of 87 speakers each uttering approximately 35 random length connected digit sequences, comprising approximately 4 hours of speech. The training data has an average SNR of 20dB.

The recognition system uses three-state, left-to-right phonetic HMMs with 105 context-dependent states (the context is cross-word, spanning up to 5 phones to either side) and $3,200$ Gaussian mixture components with diagonal covariances.

The stream weight training database consist of sentences with digitally mixed approximately uniform-level babble noise. The noise mixing was performed in a controlled fashion to produce subsets that differ from each other only by the SNR level of its sentences. As a result, we have 41 subsets, each subset containing sentences with approximately equal SNR value, the SNR values being from the set: $\{-10.0, -9.0, ...30.0dB\}$. The speech signal for all subsets is formed by taking randomly selected sentences of the MSHMM training data (1/10th of the whole MSHMM training database). As a result, each stream weight training subset contains 71 speakers, and total of 368 sentences. Similarly to the training set, the test set consists of sentences with digitally mixed, approximately uniform-level babble noise. The set consists of 8 speakers (none of which is part of the training database), each uttering 39 random length digit sequences for a total of 312 testing sentences.

5. EXPERIMENTS

We now proceed to compare the various methods discussed in the previous sections concerning feature selection and modeling for stream weight estimation.

Figure 1 shows the performance of the AVASR system based on the proposed features: likelihood, acoustic signal based as well as their combination. In all three cases the "GMM avg" is used for obtaining the weight It is quite clear from the figure that in the decreasing SNR condition the acoustic signal based features result in a better λ estimate, in the clean condition a statistical difference is

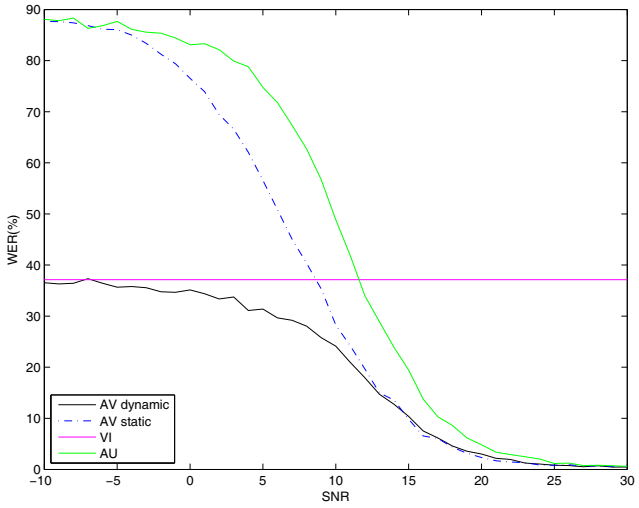

Fig. 3. *Dynamic vs. static weight audio-visual ASR performance. Combined dispersion and signal based features are used for dynamic weighting, utilizing the average GMM approach, whereas fixed values of 0.7, 0.3 are used as static weights (best scoring weight setting across entire test set). Audio and visual only performance is also depicted.*

not observed. Although difficult to discern from this figure use of the dispersions and energy features together provide on average a 7.25% reduction in WER for SNRs $>$ 15 dB compared to use of the signal features only.

It is clear from figure 1 that the signal space observations appear to be more stable than likelihood dispersions across the range of test SNR's. Presently we are not distinguishing between speech and silence segments, so this could possibly be an issue for the likelihood dispersions ([6] removes the silence state from the dispersion computation and comments that at lower SNR's an increasing number of states are confused with silence). In order to avoid the additional constraint of dealing with speech and silence states, we prefer using the dispersions directly and hope for an overall gain.

In figure 2 we compare various modeling techniques for weight estimation: the sigmoid and the two GMM techniques, "GMM avg" and "GMM max". In all three cases the two-dimensional acoustic signal features are used. Clearly the "GMM avg" technique achieves the best performance, always with a WER less than the minimum of the audio only and visual only systems (non-catastrophic fusion). The main disadvantage of the weight modeling by sigmoid function fitting is that adding more dimension means increasing complexity of the algorithm and feasibility issues. Therefore, we have modeled the sigmoid function dynamic weighting for two dimensions.

Finally in figure 3 we demonstrate the usefulness of the dynamic weighting scheme for robust AVASR, by comparing our best proposed algorithm (joint likelihood and signal features with "GMM avg" weight estimation) to an AVASR system that uses static, constant weights for all conditions. Clearly the proposed system dramatically outperforms the static weighting scheme.

6. CONCLUSIONS

In this paper we investigated dynamic weight estimation techniques for multi-stream HMM based audio-visual speech recognition. In particular, we focused on two aspects of this process. Feature selection for capturing the reliability of the audio and visual streams, and weight estimation based on such features. We considered likelihood and signal space features, whereas for weight modeling we investigated a sigmoid function fitting and two variants of GMM estimation. The best results were obtained using both types of stream reliability features and "GMM avg" based weight estimation. The proposed technique dramatically outperforms static weighting schemes. The algorithm provides a flexible framework for integrating instantaneously changes in the environment, thus providing a means to fully exploit the visual modality benefit for robust AVASR.

7. REFERENCES

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

[2] Connell, J.H., Haas, N., Marcheret, E., Neti, C., Potamianos, G., and Velipasalar, S., "A real-time prototype for small-vocabulary audio-visual ASR," *Proc. Int. Conf. Multimedia Expo*, pp. 469–472, 2003.

[3] Dupont, S. and Luettin, J., "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, 2(3): 141–151, 2000.

[4] Garg, A., Potamianos, G., Neti, C., and Huang, T., "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," *Proc. ICASSP*, pp. 24–27, 2003.

[5] Gravier, G., Axelrod, S., Potamianos, G., and Neti, C., "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," *Proc. ICASSP*, pp. 853–856, 2002.

[6] Heckmann, M., Berthommier, F., and Kroschel, K., "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, 2002(11): 1260–1273, 2002.

[7] Huang, J., Marcheret, E., and Visweswariah, K., "Rapid feature space adaptation for multi-stream HMM-based audio-visual speech recognition," *Proc. Int. Conf. Multimedia Expo*, pp. 338–341, 2005.

[8] Huang, J., Potamianos, G., Connell, J., and Neti, C., "Audio-visual speech recognition using an infrared headset," *Speech Communication*, 44(4): 83–96, 2004.

[9] Janin, A., Ellis, D., and Morgan, N., "Multi-stream speech recognition: Ready for prime time?," *Proc. Eurospeech*, pp. 591–594, 1999.

[10] Marcheret, E., Visweswariah, K., and Potamianos, G., "Speech activity detection fusing acoustic phonetic and energy features," *Proc. Interspeech*, 2005.

[11] Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A.W., "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, 91(9): 1306–1326, 2003.

[12] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., *The HTK Book*. United Kingdom: Entropic Ltd., 1999.