

RAPID FEATURE SPACE SPEAKER ADAPTATION FOR MULTI-STREAM HMM-BASED AUDIO-VISUAL SPEECH RECOGNITION

Jing Huang, Etienne Marcheret, Karthik Visweswariah

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

{jghg,etiennem,kv1}@us.ibm.com

ABSTRACT

Multi-stream hidden Markov models (HMMs) have recently been very successful in audio-visual speech recognition, where the audio and visual streams are fused at the final decision level. In this paper we investigate fast feature space speaker adaptation using multi-stream HMMs for audio-visual speech recognition. In particular, we focus on studying the performance of feature-space maximum likelihood linear regression (fMLLR), a fast and effective method for estimating feature space transforms. Unlike the common speaker adaptation techniques of MAP or MLLR, fMLLR does not change the audio or visual HMM parameters, but simply applies a single transform to the testing features. We also address the problem of fast and robust on-line fMLLR adaptation using feature space maximum a posterior linear regression (fMAPLR). Adaptation experiments are reported on the IBM infrared headset audio-visual database. On average for a 20-speaker 1 hour independent test set, the multi-stream fMLLR achieves 31% relative gain on the clean audio condition, and 59% relative gain on the noisy audio condition (approximately 7dB) as compared to the baseline multi-stream system.

1. INTRODUCTION

Recently audio-visual speech recognition (AVSR) has attracted significant interest as a means of improving performance and robustness over audio-only speech recognition (ASR) [1, 2, 3], especially in real-life applications [4, 5]. The most successful AVSR systems extract visual features from the facial region of interest and combine them with acoustic features using multi-stream HMMs. It has been demonstrated that multi-stream decision fusion attains significant improvement in recognition accuracy over the single-stream based fusion methods [6]. The observation likelihood of the multi-stream HMM is the product of the likelihood values from audio and visual streams, raised to appropriate stream exponents that model the reliability of each stream [7].

Speaker adaptation is naturally extended to multi-stream AVSR systems to improve speaker-independent (SI) system performance as it is applied successfully in practical ASR systems [8]. Common adaptation techniques, such as maximum a posterior (MAP) adaptation and maximum likelihood linear regression (MLLR) are exploited in [8] in a supervised way, which uses correct transcripts of the adaptation data. First, audio-only and visual-only HMM parameters are adapted separately by MAP and MLLR. Subsequently the audio-visual HMM stream exponents are adapted by means of discriminative training on N-best recognized hypothesis. This adaptation of the stream exponents may not be effective for on-line adaptation since the correct hypotheses is not available at the adaptation stage.

In this paper we focus on unsupervised on-line feature adaptation technique such as feature-space maximum likelihood linear regression (fMLLR, also known as constrained MLLR [9]). Unlike model adaptation techniques MAP or MLLR, fMLLR does not change the audio and visual HMM parameters, but simply applies a single transform to the testing features. Effectively, fMLLR adapts the means and variances of the HMM model at the same time without the MLLR requirement of saving the speaker adapted HMMs (details see Section 3). This is one reason why fMLLR is preferable over MLLR for fast on-line speaker adaptation. In addition, since only a single transform is estimated for fMLLR, it needs less adaptation data than MAP, which needs to adapt all HMM parameters. Therefore fMLLR is preferable over MAP and MLLR for fast on-line speaker adaptation, which is the focus of this paper.

Moreover when only a small amount of on-line adaptation data is available, we could estimate the fMLLR transform to maximize the a posteriori probability as opposed to just the likelihood, we call this method fMAPLR. The basic idea is to use a prior estimated from the training data and use this prior in the estimation process: the objective function is essentially the product of the likelihood of the data and the prior probability of the transform. Transforms that have not been seen in training are given low likelihood and the transform is constrained by the prior.

The paper is structured as follows: The multi-stream HMM for AVSR is discussed in Section 2. Section 3 briefly describes multi-stream fMLLR and fMAPLR. Adaptation experimental setup and results are reported in Section 4, and conclusions are drawn in Section 5.

2. THE MULTI-STREAM AVSR SYSTEM

There are three main areas that differentiate AVSR systems [10]: the speech recognition method used, the visual front end design and the audio-visual integration strategy. Our AVSR system is an HMM-based speech recognizer, appearance-based visual features and decision fusion for the audio and visual streams (usually referred to as multi-stream HMMs). We briefly describe each part in the following.

The visual features are extracted from the region of interest (ROI). We first estimate the location of the ROI, which contains the area around the speaker's mouth (see Section 4.1). Following ROI extraction, 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 consisting of intra-frame linear discriminant analysis (LDA) and maximum likelihood linear transform (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 [5].

In parallel to the visual feature extraction, audio features are also obtained, time synchronously, at 100 Hz. First, 24 mel frequency cepstral coefficients of the speech signal are computed over a sliding window of 25 msec, and are mean normalized to provide static features. Then, nine consecutive such frames are concatenated and projected by means of LDA/MLLT onto a 60-dimensional space, producing dynamic audio features.

In the multi-stream HMM decision fusion approach, the single-modality observations are assumed generated by audio-only and visual-only HMMs of identical topologies with class-conditional emission probabilities $P_a(\mathbf{o}_{a,t} | c)$ and $P_v(\mathbf{o}_{v,t} | c)$, respectively, where $c \in C$ denotes the speech classes of interest such as context-dependent sub-phonetic units. Both are modeled as mixtures of Gaussian densities. Based on the assumption that audio and visual streams are independent, we compute the joint probability $P_{av}(\mathbf{o}_{av,t} | c)$ as follows [2]:

$$P_{av}(\mathbf{o}_{av,t} | c) = P_a(\mathbf{o}_{a,t} | c)^\lambda \times P_v(\mathbf{o}_{v,t} | c)^{1-\lambda} \quad (1)$$

Exponent λ is used to appropriately weigh the contribution of each stream, depending on the ‘‘relative confidence’’ on each modality. Exponents can be fixed or time dependent [7]. The use of stream exponents are critical to the robust operation of an AVSR system. Failure of either channel can be expected in any practical application, with the visual channel being more prone to failures.

3. MULTI-STREAM FMLLR AND FMAPLR

fMLLR is a widely used and effective technique for the reduction of the mismatch between training and test conditions. In fMLLR the feature x is transformed linearly to maximize the likelihood of the testing data. We will describe the application of this technique to the multi-stream audio-visual speech recognition. Let x_a denote the audio feature vector and x_v denote the video feature vector. In the most general case we could consider the transform:

$$\begin{pmatrix} y_a \\ y_v \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_a \\ x_v \end{pmatrix} + \begin{pmatrix} b_a \\ b_v \end{pmatrix} \quad (2)$$

where A and D are square matrices that match the dimension of x_a and x_v respectively. Some preliminary experiments showed that this general form of the transform performed worse than a separate linear transform for the audio and visual streams respectively. We thus restrict discussion in the rest of the paper to the case where B and C are zero. In this paper we also consider the use of MAP estimation for fMLLR to reduce the amount of data required to reliably estimate the transform.

In standard single stream fMLLR (the top block of (2) expressed as $y = Ax + b$) the objective function is [9]

$$Q(\mathbf{W}) = \log |\det A| - 1/2 \sum_i w_i^T G_i w_i + k_i^T w_i \quad (3)$$

where the mean and variance statistics k_i and G_i respectively are gathered from the adaptation data and $w_i = [a_i b_i]$ is a vector made of the i th row of the transform A and the i th element of b . In the case of multi-stream HMMs we have the same objective function

except the statistics G_i and k_i are gathered with posterior calculated *jointly* using both the audio and visual streams. This multi-stream posterior takes the form of (1).

To estimate the parameters of the transform with small amounts of data we use the MAP objective function instead of just the likelihood. The MAP objective functions is preferable with small amounts of data since the parameters are constrained by the prior, in this case the prior is learned from the training data. We assume that the prior distribution of the transforms is a single full covariance Gaussian with the mean being the identity matrix. The covariance of the prior is estimated on the training data. Since this is the covariance of the transform the number of parameters is roughly d^4 where d is the dimension of the feature space. To reduce the amount of data required to estimate the prior we use Factor Analysis to approximate the covariance of the prior:

$$\Sigma = D + \Lambda \Lambda^T. \quad (4)$$

We estimate the covariance by first estimating the transform for each training speaker. We then write each of the transforms as a vector and estimate the covariance. We use standard EM technique [11] to estimate the Factor Analyzed priors. The Factor Analysis estimation is initialized using probabilistic PCA [12].

Given the assumption of a single Gaussian for the prior, the auxiliary function for MAP estimation is changed by the addition of a quadratic term to the standard ML auxiliary function:

$$-1/2a^T \Sigma_a^{-1} a + \mu_a^T \Sigma_a^{-1} a + \log |\det A| - 1/2 \sum_i w_i^T G_i w_i + k_i^T w_i$$

where a is $\text{vec}(A)$. For the experiments in this paper the estimation of the transform is performed element by element. Each estimation step then has a closed form solution [13].

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

Experiments are conducted on the audio-visual database collected with the IBM infrared headset [5]. 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. Eliminating this step improves the visual front end robustness and reduces CPU requirements by approximately 40% [14].

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. Extracting the two mouth corners turns out to be fairly simple in the headset scenario, since it is assumed that the camera is already aimed nearly directly at the mouth. Because the types of features seen in the captured images are tightly constrained (i.e., no confusing background objects are expected in the scene), the algorithms can use very weak models and hence run quickly. The algorithm first estimates the position of the mouth in the image, then determines the mouth corners. Finally it outputs the normalized mouth image: a 64x64 pixel ROI with an aspect ratio of about 1.7 covering the mouth. Details can be found in [5].

The system is built on 22kHz audio and 720x480 pixel resolution at 30 Hz video. The database consists of 107 subjects each uttering approximately 35 random length connected digit sequences.

The 107 speakers are split into training and testing sets: 87 speakers are used for training, and the remaining 20 speakers are used for testing, there is no overlap in training and testing sets. The training data has about 4 hours of speech, and the test data has around 1 hour speech. Both training and testing data have an average SNR of 20dB. In addition to this clean test data which matches the training data, another noisy test set is built by artificially corrupting the test set with additive “speech babble” noise resulting in an average SNR of 7dB. Recognition results are presented on both clean and noisy test sets.

The recognition system uses three-state, left-to-right phonetic HMMs with 166 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.

Our fMLLR adaptation is unsupervised: for each speaker, we use the baseline speaker independent audio and visual models to get initial multi-stream/single-stream decoding transcripts, and use these transcripts to compute multi-stream/single-stream fMLLR transforms. Then the transformed testing features are used to get the final adapted results. Unlike [8], we keep the stream weights fixed, 0.7 for audio stream, and 0.3 for video stream. To show the effectiveness of fMLLR on multi-stream HMMs, we also present the fMLLR results on individual single stream, audio-only and visual-only.

4.2. Results

The results are presented as word error rate (WER) for visual-only (V), audio-only (A) and multi-stream audio-visual (AV) recognition. These recognition results are run by the standard IBM stack decoder, recently modified to accommodate multi-stream HMM based decision fusion.

In Table 1, we compare the results of multi-stream fMLLR with results of single-stream fMLLR. The second row shows the improvement from single-stream fMLLR: in clean condition, fMLLR gives 22% relative improvement (from 1.8 \rightarrow 1.4) on audio, and 22% relative improvement (from 34.4 \rightarrow 26.9) on video also; in noisy condition, fMLLR shows more gain: from 21.4 \rightarrow 9.9 on audio, relative 54% improvement.

The third row shows the improvement from multi-stream fMLLR: when we use the multi-stream fMLLR to decode the single stream, we get better results than the single-stream fMLLR. In clean condition, there are 0.1% absolute gain on audio-only WER, and 7.5% absolute gain on visual-only WER from multi-stream fMLLR; in noisy condition, there are 2.0% absolute gain on audio-only WER, and 6.3% absolute gain on visual-only WER from multi-stream fMLLR. The improvement indicates that multi-stream fMLLR is better estimated than single-stream fMLLR, and it also shows even more improvement for multi-stream decoding: in clean condition, fMLLR gives 31% relative improvement (from 1.6 \rightarrow 1.1); in noisy condition, fMLLR shows more gain: from 12.9 \rightarrow 5.3 on audio, relative 59% improvement.

The above results are obtained using all test utterances as unsupervised adaptation data. In practice when on-line adaptation is required, the adaptation data is usually very little. Here we investigate how fMAPLR helps with small amounts of adaptation data: for each test speaker, we take only n utterances as adaptation data, n is taken as 1, 2, 3, 4, 5, 10, and all utterances. On average each utterance is 5 seconds long.

Table 2 shows the results from these seven adaptation experiments. In the case of 1-utterance adaptation data, the fMAPLR

System	Clean			Noisy		
	A	V	AV	A	V	AV
baseline	1.8	34.4	1.6	21.4	34.4	12.9
fMLLR(single)	1.4	26.9	1.2	9.9	26.9	5.5
fMLLR(multi)	1.3	19.4	1.1	7.9	20.6	5.3

Table 1. Comparison of fMLLR results on audio-only, visual-only, and audio-visual speech recognition. Transforms estimated from all data.

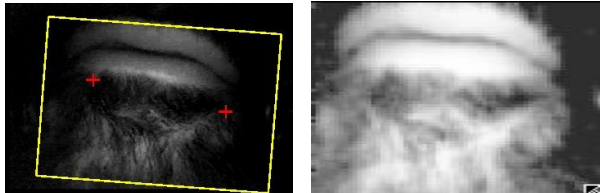


Fig. 1. An example of the bad speaker and his mouth ROI image

performance is worse than baseline. When we examine the 20 test speakers carefully, we find out that one particular speaker contributes half of errors of visual-only and audio-visual recognition. The visual-only WER of this speaker is as high as 80%. This is due to the fact that this speaker has dark dense beard around his mouth. The mouth corners are not correctly located due to the interference of the dark beard. As result the mouth is not centered but rather being push up (see Figure 1, the rectangle forms a bounding box around the mouth region, two crosses mark the mouth corners). When visual-only WER is 80% high, 5 seconds adaptation data is certainly not enough for fMAPLR. Therefore the fMAPLR transform is wrongly estimated and gives worse results than the baseline.

Starting with 2-utterance adaptation data, fMAPLR shows improvement over the baseline results both in clean and noisy conditions. As the adaptation data increases, the fMAPLR performance gets better (see Figure 2 and Figure 3). If fMLLR is used instead, we see from Table 3 that fMLLR fails in the case of 2-utterance, and improves a little less amount than fMAPLR in the case of 5-utterance. This proves the effectiveness of fMAPLR when little adaptation data is available.

When all utterances are used, we notice the results of fMAPLR are not as good as those of fMLLR in Table 1. This might be do to the fact that we only have a small number (87) of training speakers to estimate priors on the fMLLR matrices. Hence when the adaptation data is sufficient, using fMAPLR is less optimal than computing fMLLR from data alone.

5. CONCLUSIONS

In this paper, we have investigated feature space speaker adaptation using multi-stream HMMs for audio-visual speech recognition, as a means of fast and robust on-line adaptation for real-time AVSR applications. We studied the performance of multi-stream fMLLR, which simply applies a single transform to each testing audio/visual features. We also addressed the problem of robust on-line fMLLR adaptation with little adaptation data using maximum a posterior linear regression (fMAPLR). Adaptation experiments are reported on the IBM infrared headset audio-visual database. On average of 20-speaker 1 hour speaker independent test data, the

adaption sentences per speaker	Clean			Noisy		
	A	V	AV	A	V	AV
0	1.8	34.4	1.6	21.4	34.4	12.9
1	2.1	39.2	2.9	17.5	40.3	15.2
2	1.7	29.4	1.5	13.7	30.9	10.2
3	1.6	27.1	1.4	12.8	28.9	9.2
4	1.6	25.5	1.3	11.2	27.0	8.0
5	1.6	24.3	1.3	10.6	25.9	7.6
10	1.5	22.5	1.2	9.6	23.8	6.4
all	1.4	20.0	1.2	8.3	21.2	5.6

Table 2. Effect of adaptation data on fMAPLR

adaption sentences per speaker	Clean			Noisy		
	A	V	AV	A	V	AV
0	1.8	34.4	1.6	21.4	34.4	12.9
2	2.2	35.2	1.9	37.4	43.9	30.2
5	1.5	23.5	1.4	10.9	26.0	8.3

Table 3. fMLLR performance on varying amounts of adaptation data

multi-stream fMLLR achieves 31% relative gain on the clean audio condition, and 59% relative gain on the noisy audio condition (around 7dB) compared to no fMLLR adaptation on multi-stream HMMs. When only a small adaptation data is available, fMAPLR clearly is more robust and effective than fMLLR. We need an accurate prior estimation for fMAPLR from a large number of training speakers, however. If the prior is not adequately estimated, when the adaptation data is sufficient, fMLLR is preferred over fMAPLR for better performance.

6. REFERENCES

- [1] X. Liu, Y. Zhao, X. Pi, L. Liang and A. V. Nefian, "Audio-visual continuous speech recognition using a coupled hidden Markov model", *IEEE International Conference on Spoken Language Processing*, pp. 213-216, September 2002.
- [2] S. Dupont and J. Luetttin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, 2(3): 141-151, 2000.
- [3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, 91(9): 1306-1326, 2003.
- [4] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," *Europ. Conf. Speech Commun. Technol.*, 2003.
- [5] J. Huang, G. Potamianos, J. Connell and C. Neti, "Audio-Visual Speech Recognition Using an Infrared Headset," *Speech Communication*, Dec. 2004.
- [6] E. Marcheret, S. Chu, V. Goel, G. Potamianos, "Efficient Likelihood Computation in Multi-Stream HMM Based Audio-Visual Speech Recognition," *Int. Conf. Speech and Language Processing*, 2004.
- [7] A. Garg, G. Potamianos, C. Neti, T. Huang, "Frame-Dependent Multi-Stream Reliability Indicators for Audio-Visual Speech Recognition," *Int. Conf. Acoustic Speech and Signal Processing*, 2003.

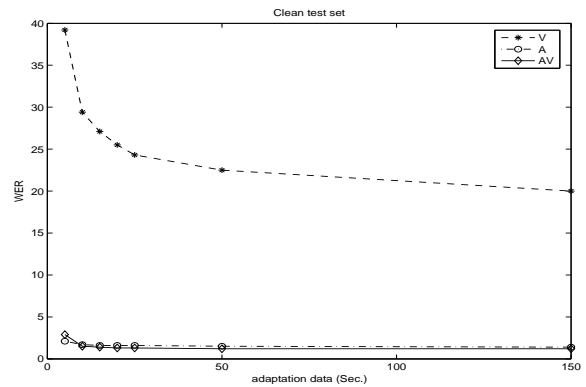


Fig. 2. Effect of the amount of clean adaptation data on fMAPLR

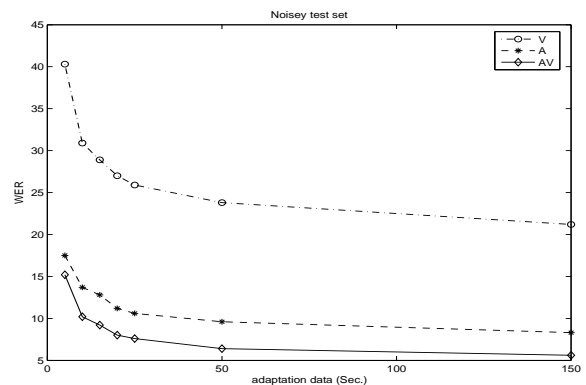


Fig. 3. Effect of the amount of noisy adaptation data on fMAPLR

- [8] G. Potamianos and A. Potamianos, "Speaker Adaptation for Audio-Visual Speech Recognition," *Europ. Conf. Speech Commun. Technol.*, 1999.
- [9] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Technical report, TR 291, Cambridge University*, 1997.
- [10] M.E. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke, Eds. Berlin: Springer, pp. 331-349, 1996.
- [11] D. B. Rubin, D. T. Thayer, "EM algorithms for ML factor analysis," *Psychometrika*, 47(1), March 1982.
- [12] M. Tipping, C. Bishop, "Probabilistic principal component analysis," *Technical Report Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, Birmingham*, 1997.
- [13] K. Visweswariah, V. Goel, and R.A. Gopinath, "Structuring Linear Transformations For Adaptation Using Training Time Information," *Int. Conf. Acoustic Speech and Signal Processing*, 2002.
- [14] J. Connell, N. Haas, E. Marcheret, C. Neti, G. Potamianos, S. Velipasalar, "A Real-Time Prototype for Small-Vocabulary Audio-Visual ASR," *IEEE Int. Conf. on Multimedia & Expo*, 2003.