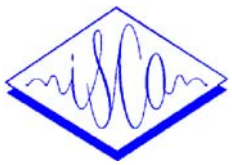

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEAKER ADAPTATION FOR AUDIO-VISUAL SPEECH RECOGNITION

Gerasimos Potamianos and *Alexandros Potamianos*†

AT&T Labs–Research, 180 Park Ave, Florham Park, NJ 07932-0971, U.S.A.
email: makis@research.att.com

ABSTRACT

In this paper, speaker adaptation is investigated for audio-visual automatic speech recognition (ASR) using the multi-stream hidden Markov model (HMM). First, audio-only and visual-only HMM parameters are adapted by combining maximum a posteriori and maximum likelihood linear regression adaptation. Subsequently, the audio-visual HMM stream exponents are adapted to better capture the reliability of each modality for the specific speaker, by means of discriminative training. Various visual feature sets are compared, and features based on linear discriminant analysis are demonstrated to result in superior multi-speaker and speaker-adapted recognition performance. In addition, visual feature mean normalization is shown to significantly improve visual-only and audio-visual ASR performance. Adaptation experiments on a 49-subject database are reported. On average, a 28% relative word error reduction is achieved by adapting the multi-speaker audio-visual HMM to each subject in the database.

1. INTRODUCTION

Audio-visual (AV) *automatic speech recognition* (ASR) has attracted significant interest as a means of improving performance and robustness over audio-only ASR [1-5]. Indeed, a relative *word error rate* (WER) reduction of up to 30% over audio-only ASR has been reported in the clean-audio case [3]. Bimodal ASR systems extract relevant features from the video sequence of the speaker's face and combine them with acoustic features using an appropriate *hidden Markov model* (HMM) classifier, such as the *multi-stream* HMM. The observation likelihood of the multi-stream HMM is the product of the observation likelihoods of its audio-only and visual-only stream components, raised to appropriate *stream exponents* that model the reliability of each modality [3].

Speaker adaptation is successfully used in practical audio-only ASR systems to improve *speaker-independent* (SI) system performance, when few data from a speaker of interest is available [6-8]. In this paper, we propose the use of speaker adaptation as a means of improving performance of visual-only and audio-visual ASR. To our knowledge, speaker adaptation has never before been considered for this application, although visual front-end normalization has been introduced in [2], [5]. In general, speaker adaptation algorithms can be applied to the features (front-end), as well as to the HMM parameters. For example, *feature mean normalization* (FMN) has been successfully applied to cepstral audio features to improve audio-only ASR robustness [9]. Various algorithms exist for HMM parameter adaptation, such as *maximum a posteriori* (MAP) adaptation [6-7] and *maximum likelihood linear regression* (MLLR) [7-8]. In this work, we consider both methods, and we demonstrate that visual-only and audio-visual HMM adaptation result in significant improvements over the SI baseline HMMs. We also propose speaker adaptation for the audio-visual multi-stream HMM exponents.

The paper is structured as follows: Section 2 describes the audio and various visual front-ends, as well as, feature mean normalization. The multi-stream HMM is discussed in Section 3, and algorithms for its parameter adaptation are presented in Section 4. Adaptation experiments on the AT&T internal bimodal database are reported in Section 5, and our conclusions are drawn in Section 6.

2. FEATURES FOR AUDIO-VISUAL ASR

To perform audio-visual ASR given the recorded bimodal speaker utterance, we extract the *time-synchronous* audio-visual feature observation sequence, denoted by

$$\{\, \underline{O}^{(t)} = [\underline{O}_A^{(t)}, \underline{O}_V^{(t)}] \in \mathsf{R}^D,\ 1 \leq t \leq T \,\}, \qquad (1)$$

where $D = D_A + D_V$ is the bimodal feature vector size, and $\underline{O}_s^{(t)} \in \mathsf{R}^{D_s}$, for $s = A, V$, represent the unimodal (audio- and visual-only) feature vectors. Audio and visual feature vectors are originally extracted at frequencies $f_A = 100\,\mathsf{Hz}$ and $f_V = 60\,\mathsf{Hz}$, respectively. *Linear interpolation* of the visual features is then used to time-align them to $\underline{O}_A^{(t)}$ and obtain (1) at $f = 100\,\mathsf{Hz}$ [3].

2.1. Audio features

The audio front-end produces a 12-dimensional (12-dim) *mel-frequency cepstral coefficient* (MFCC) feature vector, augmented by the normalized logarithm of the signal energy within the 25 msec speech analysis window [10]. The resulting 13-dim feature vector, $\underline{o}_A^{(t)}$, is augmented by its first- and second-order time derivatives, yielding a 39-dim vector $[\underline{o}_A^{(t)}, \underline{\Delta} * \underline{o}_A^{(t)}, \underline{\Delta} * \underline{\Delta} * \underline{o}_A^{(t)}]$, where $*$ denotes convolution, and $\underline{\Delta}$ is an appropriate 5 sample kernel.

2.2. Visual features

Given the video sequence of the speaker's face, $\{V(x, y; t): 1 \leq t \leq T\}$, we obtain for every video frame the speaker inner and outer lip contours, denoted by $\mathcal{C}_{\mathcal{I}}^{(t)}$ and $\mathcal{C}_{\mathcal{O}}^{(t)}$, respectively [2-4]. Given these, we can obtain *lip contour based* visual features [1-2], or estimate a *region of interest* (ROI) of the video sequence containing the speaker's mouth and extract *pixel based* visual features [1-5]. The resulting "*static*" visual features, $\underline{o}_V^{(t)}$, are aligned to $\underline{o}_A^{(t)}$, and augmented by their first- and second-order time derivatives, yielding the feature vector $[\underline{o}_V^{(t)}, \underline{\Delta} * \underline{o}_V^{(t)}, \underline{\Delta} * \underline{\Delta} * \underline{o}_V^{(t)}]$.

2.2.1. Lip contour based visual features

Based on a series of experiments in [2], we choose

$$\underline{o}_V^{(t)} = [\mathbf{h}(\mathcal{C}_{\mathcal{I}}^{(t)}), \mathbf{w}(\mathcal{C}_{\mathcal{I}}^{(t)}), \mathbf{p}(\mathcal{C}_{\mathcal{I}}^{(t)}), \mathbf{a}(\mathcal{C}_{\mathcal{I}}^{(t)}), \mathbf{h}(\mathcal{C}_{\mathcal{O}}^{(t)}), \mathbf{w}(\mathcal{C}_{\mathcal{O}}^{(t)}), \mathbf{p}(\mathcal{C}_{\mathcal{O}}^{(t)}), \mathbf{a}(\mathcal{C}_{\mathcal{O}}^{(t)}), \{\, \mathcal{FD}_i(\mathcal{C}_{\mathcal{O}}^{(t)}),\ 2 \leq i \leq 5 \,\}],$$

where $\mathbf{h}(\mathcal{C})$, $\mathbf{w}(\mathcal{C})$, $\mathbf{p}(\mathcal{C})$, $\mathbf{a}(\mathcal{C})$ denote height, width, perimeter, area inside contour $\mathcal{C}$, and $\mathcal{FD}_i(\mathcal{C})$ denotes the magnitude of the i-th order Fourier descriptors of $\mathcal{C}$ [2].

† Currently with Bell Labs, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.; e-mail: potam@research.bell-labs.com

2.2.2. Pixel based visual features

Given the outer lip contour sequence, $\{\mathcal{C}_\mathcal{O}^{(\mathrm{t})}\}$, we obtain the speaker's mouth center of mass sequence, $\{(c_x^{(\mathrm{t})}, c_y^{(\mathrm{t})})\}$. At every time instant t, we consider as data relevant to *lipreading* the *MNK* pixels in $\{V(x,y;z)\colon\ c_x^{(\mathrm{t})} - \lfloor M/2 \rfloor \le x < c_x^{(\mathrm{t})} + \lceil M/2 \rceil, c_y^{(\mathrm{t})} - \lfloor N/2 \rfloor \le y < c_y^{(\mathrm{t})} + \lceil N/2 \rceil,\ \mathrm{t}-t_L \le z \le \mathrm{t}+t_R\}$, where $t_L, t_R \ge 0$, and $K = t_L + t_R + 1$, ordered into an *MNK*-dim data vector $\underline{V}^{(\mathrm{t})}$. We then seek a matrix[1] $\mathbf{P} = [\underline{P}_1^\top, ..., \underline{P}_L^\top]$ of size $MNK \times L$, where $L \ll MNK$, such that the L-dim feature vector $\underline{o}_V^{(\mathrm{t})} = \underline{V}^{(\mathrm{t})}\mathbf{P}$ contains the most relevant information for discriminating among phonemes. Matrix $\mathbf{P}$ can be obtained by applying an image transform, such as the *discrete wavelet transform* (DWT) [2], or by *linear discriminant analysis* (LDA) [4].

In our case, the DWT is implemented by means of the separable Daubechies class wavelet filter of order 3. Transform coefficients on $L = 17$ chosen lattice locations are retained as $\underline{o}_V^{(\mathrm{t})}$ [2]. Values $M = N = 16$, $K = 1$ are used.

In order to obtain $\mathbf{P}$ using LDA, we are given I training examples $\{\underline{V}^{(i)},\ i = 1,...,I\}$. LDA [4] assumes that a set of classes, $\mathcal{J} = \{1,...,J\}$, is given a-priori (here, they coincide with the HMM states), as well as, that the training set data vectors, $\underline{V}^{(i)}$, are labeled as $j(i) \in \mathcal{J}$ (this is automatically done by forced segmentation using either the audio or visual streams [4], [10]). Let $\mathbf{S_W}$, $\mathbf{S_B}$ be the *within-class scatter* and *between-class scatter* matrices of the training sample [4], and let the *generalized eigenvalues* and right *eigenvectors* of the matrix pair $(\mathbf{S_B}, \mathbf{S_W})$ be computed, that satisfy $\mathbf{S_B A} = \mathbf{S_W A \Lambda}$. Let the L largest generalized eigenvalues in $\mathbf{\Lambda}$ be $\{\lambda_{k_1}, ..., \lambda_{k_L}\}$, and let matrix $\mathbf{A} = [\underline{a}_1^\top, \underline{a}_2^\top, ..., \underline{a}_{MNK}^\top]$ have as columns the generalized eigenvectors. Then, $\mathbf{P} = [\underline{a}_{k_1}^\top, ..., \underline{a}_{k_L}^\top]$. Given $\underline{V}^{(\mathrm{t})}$, we extract its L-dim feature vector

$$\underline{o}_V^{(\mathrm{t})} = [\, o_{V,1}^{(\mathrm{t})}, ..., o_{V,L}^{(\mathrm{t})} \,], \quad \text{as} \quad o_{V,l}^{(\mathrm{t})} = < \underline{V}^{(\mathrm{t})}, \underline{a}_{k_l} > .$$

In our case, $L = 18$, $M = N = 16$, $K = 5$, and $I = \mathrm{O}(10^5)$.

2.3. Feature mean normalization

FMN is a simple technique used to improve robustness of audio ASR in mismatched training and testing conditions by subtracting, at each time instant, the feature vector mean, computed over the entire utterance, from the original feature vector [9]. In previous work [2], we have observed that FMN dramatically improves *multi-speaker* visual-only ASR performance, when pixel-based visual features are used. Here, we independently apply FMN to all "static" audio and visual features, obtaining, for each modality, the observation vector (see also (1))

$$\underline{O}_s^{(\mathrm{t})} = \Big[\ \underline{o}_s^{(\mathrm{t})} - \frac{1}{T}\Sigma_{\mathrm{t}'=1}^{T}\, \underline{o}_s^{(\mathrm{t}')},\ \underline{\Delta} * \underline{o}_s^{(\mathrm{t})},\ \underline{\Delta} * \underline{\Delta} * \underline{o}_s^{(\mathrm{t})}\ \Big], \quad (2)$$

where $s = A, V$. Note that intensity normalization of $\{\underline{V}^{(\mathrm{t})},\ 1 \le \mathrm{t} \le \mathrm{T}\}$, over all pixels and frames, proposed in [5], constitutes a restricted case of (2).

3. THE MULTI-STREAM HMM

The audio-visual observation vector sequence provides information about a sequence of hidden class labels (*states*) $j \in \mathcal{J} = \{1,...,J\}$, modeled by a multi-stream HMM with *emission* "probabilities" (see also (1)) [3]

$$Pr[\underline{O}^{(\mathrm{t})} | j] = \prod_{s \in \{A,V\}} \Big[\sum_{m=1}^{M_{js}} w_{jms}\, \mathcal{N}_{D_s}(\underline{O}_s^{(\mathrm{t})}; \underline{\mu}_{jms}, \mathbf{\Sigma}_{jms}) \Big]^{\gamma_{js}}. \quad (3)$$

In (3), *mixture weights* w_{jms} are positive adding up to one, M_{js} denotes the number of mixtures, and $\mathcal{N}_d(\underline{x}; \underline{\mu}, \mathbf{\Sigma})$ is the d–variate normal distribution with mean $\underline{\mu}$ and *diagonal* covariance matrix $\mathbf{\Sigma}$, with its diagonal denoted by $\underline{\sigma}$. In addition, γ_{js} denote the stream exponents that model the reliability of each modality (stream), and satisfy [3]

$$0 \le \gamma_{jA}, \gamma_{jV} \le 1, \ \text{and} \ \ \gamma_{jA} + \gamma_{jV} = 1, \ \text{for all} \ j \in \mathcal{J}. \quad (4)$$

Notice that, in general, (3) does not represent a probability mass function. Our references to log-likelihoods should therefore be broadly interpreted as references to recognition *scores*. The audio-only and visual-only parameters of HMM (3) are given by[2],

$$\underline{\theta}_s = [\, (w_{jms}, \underline{\mu}_{jms}, \underline{\sigma}_{jms}), m = 1,...,M_{js},\ j \in \mathcal{J} \,], \quad (5)$$

for $s = A, V$, and the complete HMM parameter set is

$$\underline{\theta} = [\, \underline{\theta}_A, \{\gamma_{jA}, j \in \mathcal{J}\}, \underline{\theta}_V, \{\gamma_{jV}, j \in \mathcal{J}\} \,]. \quad (6)$$

3.1. Multi-stream HMM parameter estimation

To estimate $\underline{\theta}$, we first obtain *maximum likelihood* (ML) estimates of parameters $\underline{\theta}_A, \underline{\theta}_V$, separately for each modality (stream) using the *expectation-maximization* (EM) algorithm [10]. Subsequently, we estimate γ_{js}, using the *generalized probabilistic descent* (GPD) algorithm [10], since ML estimation of stream exponents is inappropriate [3].

Let us assume that we are given I audio-visual observation training sequences $\mathbf{O}^{(i)} = [\underline{O}^{(1,i)\top}, ..., \underline{O}^{(T_i,i)\top}]$ of duration T_i, $i = 1,...,I$, and let the entire training set observation be $\mathbf{O} = [\mathbf{O}^{(1)}, ..., \mathbf{O}^{(I)}]$. Let us also denote by $\underline{j(i)} = \{\, j(t,i) \in \mathcal{J},\ t = 1,...,T_i \,\}$ any HMM state sequence for utterance i. Given two HMM parameter vectors $\underline{\theta}'$, $\underline{\theta}''$, and denoting by $\underline{\theta}_{\mathrm{tr}}$ the vector of HMM initial and transition probabilities, we define the *auxiliary function* [10]

$$Q(\underline{\theta}', \underline{\theta}'' | \mathbf{O}) = \sum_{i=1}^{I} \sum_{\underline{j(i)}} Pr[\mathbf{O}^{(i)}, \underline{j(i)} | \underline{\theta}', \underline{\theta}_{\mathrm{tr}}] \log Pr[\mathbf{O}^{(i)} | \underline{j(i)}, \underline{\theta}''] . \quad (7)$$

Then, given a current HMM parameter vector at iteration k, $\underline{\theta}_s^{(k)}$, we obtain the reestimated parameter vector [10]

$$\underline{\theta}_s^{(k+1)} = \arg\max_{\underline{\theta}_s} Q(\, \underline{\theta}_s^{(k)}, \underline{\theta}_s \,|\, \mathbf{O}\,), \quad \text{for} \ \ s = A, V. \quad (8)$$

Clearly (see also (3)-(7)), (8) implies separate audio-only and visual-only HMM parameter estimation, based on unimodal observations. Alternatively, and if a current estimate of the stream exponents is available, we can estimate

$$\underline{\theta}_s^{(k+1)} = \arg\max_{\underline{\theta}_s} Q(\, \underline{\theta}^{(k)}, \underline{\theta}_s \,|\, \mathbf{O}\,), \quad \text{for} \ \ s = A, V, \quad (9)$$

based on the bimodal observations.

Stream exponent estimates can be obtained at a second stage, following the estimation of $\underline{\theta}_A$, $\underline{\theta}_V$ by (7)-(9). Given training data $\mathbf{O}^{(i)}$ and the current HMM parameters $\underline{\theta}$, let us denote the *N-best recognized hypotheses* [10] of utterance i by $\mathcal{R}_n^{(i)}$, $n = 1,...,N$, and its *forced segmentation* [10] by $\mathcal{F}^{(i)}$. Let us also denote the corresponding HMM state sequences by $\mathcal{G} = \{j_\mathcal{G}(t,i), t = 1,...,T_i\}$, where $\mathcal{G} = \mathcal{R}_n^{(i)}, \mathcal{F}^{(i)}$. The log-likelihoods of the N-best recognized hypotheses and forced segmentation (ignoring $\underline{\theta}_{\mathrm{tr}}$), normalized by the utterance length T_i, are then given by

$$\mathcal{L}_\mathcal{G} = \frac{1}{T_i} \sum_{t=1}^{T_i} \log Pr\,[\, \underline{O}^{(t,i)} | \, j_\mathcal{G}(t,i)\,,\, \underline{\theta}\,], \quad (10)$$

[1] In this paper, $\bullet^\top$ denotes vector (or, matrix) *transposition*. In addition, $<\bullet,\bullet>$ denotes vector *inner product*.

[2] HMM initial and transition probabilities are omitted in our derivations, since, in practice, the observation sequence likelihood is dominated by the emission probability contribution.

where $\mathcal{G}=\mathcal{R}_n^{(i)},\mathcal{F}^{(i)}$. Let us define the loss function [3],

$$d^{(i)}=\Big[1+\frac{\exp[\mathcal{L}_{\mathcal{F}^{(i)}}]}{\frac{1}{N_i'}\sum_{n=1}^{N}\delta(\mathcal{F}^{(i)},\mathcal{R}_n^{(i)})\exp[\mathcal{L}_{\mathcal{R}_n^{(i)}}]}\Big]^{-1}, \quad (11)$$

as a figure of merit of the discrimination between the correct and the recognized hypotheses for sentence i. In (11), $N_i'=\Sigma_{n=1}^{N}\delta(\mathcal{F}^{(i)},\mathcal{R}_n^{(i)})$, and $\delta(\mathcal{F}^{(i)},\mathcal{R}_n^{(i)})=1\,(0)$, iff the n^{th}-best hypothesis and correct labels of sentence i differ (are the same). In addition, let us define $\underline{\theta}_{\rm str}=\{\overline{\gamma}_{js},\ j\in\mathcal{J},\ s=A,V\}$, where (see also (4) and [3])

$$\gamma_{js}=\frac{\exp\overline{\gamma}_{js}}{\exp\overline{\gamma}_{jA}+\exp\overline{\gamma}_{jV}},\ \text{ and }\ \overline{\gamma}_{js}=\log\gamma_{js}. \quad (12)$$

The GPD algorithm reestimates $\underline{\theta}_{\rm str}$ by considering a single training utterance at a time, as [3], [10]

$$\underline{\theta}_{\rm str}^{(k+1)}=\underline{\theta}_{\rm str}^{(k)}-\frac{\epsilon_1}{1+\lfloor\frac{k-1}{K_o}\rfloor}\frac{\partial}{\partial\underline{\theta}_{\rm str}}d^{(k\bmod I+1)}\Big|_{\underline{\theta}_{\rm str}=\underline{\theta}_{\rm str}^{(k)}}, \quad (13)$$

where $K_o\geq 1$ and $\epsilon_1>0$.

In our work, a simplified version of (7), (8) is used for estimating $\underline{\theta}_A$, $\underline{\theta}_V$, namely *Viterbi training*, preceded by the *segmental K-means algorithm* [10]. Five iterations of the algorithm are used. At the second stage, two passes over the training set with values $N=3$, $K_o=100$, and $\epsilon_1=10$ are used in (10)-(13) to estimate stream exponents tied over all states ($\gamma_{js}=\gamma_s$, for all $j\in\mathcal{J}$).

4. AUDIO-VISUAL SPEAKER ADAPTATION

Given few bimodal adaptation data $\mathbf{O}$ from a particular speaker, and a baseline speaker-independent multi-stream HMM with parameters $\underline{\theta}^{(\rm SI)}$, we wish to estimate adapted HMM parameters $\underline{\theta}^{(\rm AD)}$ that better model the audio-visual observations of the particular speaker.

Two popular algorithms for speaker adaptation are maximum likelihood linear regression (MLLR) [7-8] and maximum a posteriori adaptation [6-7]. MLLR obtains a maximum likelihood estimate of a *linear transformation* of the HMM means, while leaving covariance matrices and mixture weights unchanged. With appropriate tying of the transformation parameters, MLLR is known to achieve *rapid* adaptation, i.e., successful adaptation with a small amount of adaptation data $\mathbf{O}$. On the other hand, MAP follows the *Bayesian* paradigm for estimating the mean, covariance, and mixture weight parameters, given $\mathbf{O}$. MAP estimates of HMM parameters are known to converge to their ML counterparts as the amount of training (here, adaptation) data becomes large. However, such convergence is slow, and therefore, MAP is not suitable for rapid adaptation. In practice, MAP is often used in conjunction with MLLR [7].

Clearly, both MLLR and MAP are likelihood based techniques, and neither method can be used to obtain adapted stream exponents $\{\gamma_{js}^{(\rm AD)},\ j\in\mathcal{J},\ s=A,V\}$. Similarly to Section 3, we therefore propose a two stage adaptation algorithm, namely MLLR and/or MAP adaptation of HMM parameters $\underline{\theta}_s$, for $s=A,V$, followed by the application of the GPD algorithm on the adaptation data to estimate the stream exponents.

4.1. MLLR adaptation

Let $\mathcal{P}_s$ be a *partition* (clustering)[3] of the set of all Gaussian mixture components of HMM (3) in each of the two streams $s\in\{A,V\}$, $\{(j,m,s),\ m=1,...,M_{js},\ j\in\mathcal{J}\}$, and let $p_s\in\mathcal{P}_s$ denote any member of this partition. Then, for each stream s, we seek MLLR adapted HMM parameters

$$\underline{\theta}_s^{(\rm MLLR)}=[(w_{jms},\underline{\mu}_{jms}^{(\rm MLLR)},\underline{\sigma}_{jms}),\ m=1,...,M_{js},\ j\in\mathcal{J}], \quad (14)$$

where the means are linearly transformed as

$$\underline{\mu}_{jms}^{(\rm MLLR)}=[1,\underline{\mu}_{jms}]\,\mathbf{W}_{p_s}, \quad (15)$$

where $(j,m,s)\in p_s$, and $\mathbf{W}_{p_s}$, $p_s=1,...,|\mathcal{P}_s|$ are matrices of dimension $(D_s+1)\times D_s$, that are estimated on basis of the adaptation data $\mathbf{O}$ [8].

The transformation matrices are estimated by means of the EM algorithm, similarly to (8), (9); indeed, we can obtain the MLLR-adapted parameters as

$$\underline{\theta}_s^{(\rm MLLR)}=\underset{\underline{\theta}_s\ \text{satisfy (14), (15)}}{\arg\max}\,Q(\underline{\theta}_s^{(\rm SI)},\underline{\theta}_s\,|\,\mathbf{O}),\ \text{ for }\ s=A,V, \quad (16)$$

by using the single modality, or as

$$\underline{\theta}_s^{(\rm MLLR)}=\underset{\underline{\theta}_s\ \text{satisfy (14), (15)}}{\arg\max}\,Q(\underline{\theta}^{(\rm SI)},\underline{\theta}_s\,|\,\mathbf{O}),\ \text{ for }\ s=A,V, \quad (17)$$

by using the bimodal SI HMM. In both cases, closed form solutions for the unknown matrices exist, if the HMM covariances are diagonal [8].

In this paper, we use a small number of MLLR classes, $|\mathcal{P}_s|=8$, $s=A,V$, and estimate only 23 diagonals of the matrices (recall that $D_A=39$, $D_V=54$). This choice of matrix topology and number of classes gave the best results in our adaptation experiments. Notice that, in general, partitions $\mathcal{P}_A$, $\mathcal{P}_V$ differ. In addition, we use a single iteration of (16), or (17), to estimate $\underline{\theta}_s^{(\rm MLLR)}$, since more iterations do not improve recognition performance.

4.2. MAP adaptation

Our MAP implementation is similar to the *approximate* MAP adaptation algorithm (AMAP) [7]. AMAP interpolates the "counts" of the speaker independent training data and the adaptation data. If $\mathbf{O}^{(\rm SI)}$ denotes the training data observations for the SI HMM, and $\mathbf{O}^{(\rm AD)}$ denotes the adaptation data observations, we obtain the training data $\mathbf{O}$ of the adapted HMM as

$$\mathbf{O}=[\mathbf{O}^{(\rm SI)},\underbrace{\mathbf{O}^{(\rm AD)},...,\mathbf{O}^{(\rm AD)}}_{n\ \text{times}}].$$

Subsequently, (8), or (9), is used to estimate the adapted HMM parameters. In this work, we use $n=5$, and Viterbi training to obtain $\underline{\theta}_s^{(\rm MAP)}$. Single-stream parameters of all mixtures are adapted, provided the adaptation data contain instances of the mixture component in question.

4.3. Stream exponent adaptation

Similarly to Section 3.1, we apply the GPD algorithm (10)-(13) to obtain adapted estimates of the stream exponents, based on I adaptation utterances. Eq. (13) gets initialized with the SI HMM stream exponents. Due to the small amount of adaptation data in our experiments, we use a relatively small $\epsilon_1=0.01$ (thus disallowing large changes away from the SI stream exponents), $K_o=I$, $N=3$, and random visits of the adaptation utterances, for a total of 10 iterations through the adaptation data. Similarly to Section 3.1, the stream exponents remain tied over all states ($\gamma_{js}^{(\rm AD)}=\gamma_s^{(\rm AD)}$, for all $j\in\mathcal{J}$).

5. EXPERIMENTS

Adaptation experiments are performed on the 49-speaker connected spoken letter ('A'–'Z') AT&T bimodal database [2-4]. Each speaker utters 25 random four-letter sequences. In this work, whole word, 6-10 state, left-to-right HMMs

[3] Obtained by K-means clustering [10], for example.

Features-HMM	Audio	Visual	AV
Lips-MS	84.9 (61.2)	22.4 (0.4)	85.9 (63.7)
Lips-MLLR	88.4 (69.8)	23.9 (0.4)	89.4 (72.2)
DWT-MS	84.9 (61.2)	26.4 (0.4)	86.5 (64.9)
DWT-MLLR	88.4 (69.8)	32.7 (2.0)	89.5 (72.2)
LDA-MS	84.9 (61.2)	36.6 (3.3)	86.8 (67.8)
LDA-MLLR	88.5 (70.2)	40.5 (3.3)	90.2 (74.3)

Table 1: Supervised adaptation for various visual front-ends.

with 16 mixtures per state and stream, and a single-state silence HMM with 32 mixtures per state and stream are used. No grammar constraints are used during recognition (free word loop of unknown length). Unless otherwise stated, FMN is applied to both feature streams, and LDA visual features are used.

Our adaptation scenario is the following: For each speaker, twenty sequences are used as a training and adaptation set (a total of 980 sequences), whereas, the remaining 5 sequences per speaker are used as a test set (a total of 245 sequences). The training set is first used to obtain *multi-speaker* (MS) audio-visual HMMs, as described in Section 3.1. Subsequently, and for each speaker, we consider the twenty sequences as *supervised* adaptation sequences (i.e., with known transcriptions) on which audio-visual HMMs are adapted. For each speaker, the adapted HMMs are tested on the remaining five sequences, and the results are accumulated for all 49 speakers, thus being reported on the same 245 utterance test set. This scenario deviates somewhat from the traditional speaker adaptation scenarios [6-8], however it is warranted due to the small size of our database.

In Table 1, we investigate MLLR only adaptation for various visual front-ends. Clearly, the pixel based visual front-ends perform significantly better than the lip contour based visual front-end, both in the multi-speaker and speaker-adapted cases, with LDA based features giving the best results.

In Table 2, we investigate the performance of the various adaptation algorithms discussed in this paper using the LDA based visual features. The performance of the original multi-speaker audio-only, visual-only, and audio-visual HMMs is depicted in the first two lines of Table 2, with and without FMN. Notice the dramatic improvement in the visual-only recognition, achieved due to FMN. Subsequently, MLLR and MAP adaptation are considered for the audio-only and visual-only HMM parameters. Overall, due to the small amount of adaptation data, MLLR adaptation gives superior results. Notice also that MLLR performance is slightly better when using the bimodal multi-stream HMM (see (17)) instead of the unimodal HMMs (see (16)) in the E-step of the MLLR adaptation algorithm. In addition, a combination of MAP and MLLR improves both audio-only and visual-only ASR. A similar result has been observed in [7], for audio-only ASR, although, in our implementation, MAP precedes the use of MLLR. In all three adaptation experiments, the adapted audio-visual HMM performance is significantly improved over the multi-speaker HMM, reaching a relative word error rate reduction of 25% in the case of MAP/MLLR adaptation. In addition, the inclusion of the visual modality reduces word error rate by 12% over audio-only adaptation. Finally, when HMM stream exponents are discriminatively trained, the relative word error rate reduction due to audio-visual adaptation becomes 28% (see Table 2).

6. CONCLUSION

In this paper, we have investigated speaker adaptation techniques for audio-visual ASR, as a means of improving recognition performance of multi-speaker audio-visual HMMs to specific speakers. We have used the multi-stream HMM for audio-visual ASR, and we have proposed a two-stage adaptation algorithm of its parameters, namely, adaptation of its audio-only and visual-only parameters, followed by adaptation of its stream exponents. Both MLLR and MAP adaptation can be used in the first stage for best results. Discriminative training of the HMM stream exponents on the adaptation training data further improved performance of the audio-visual ASR system. Overall, we obtained a 28% relative word error reduction over the multi-speaker audio-visual HMM at a clean-audio environment. In addition, we have demonstrated the importance of visual front-end feature mean normalization for improved audio-visual ASR. Finally, an investigation of lip contour vs. pixel based visual features (DWT and LDA) has indicated that the latter give superior performance both for multi-speaker and speaker-adapted visual-only HMMs, with LDA performing the best.

HMM	Audio	Visual	AV
MS (NO FMN)	84.1 (59.8)	18.2 (0.0)	84.2 (60.1)
MS (FMN)	84.9 (61.2)	36.6 (3.3)	86.8 (67.8)
MLLR, eq. (16)	88.2 (69.4)	39.9 (2.9)	89.8 (72.5)
MLLR, eq. (17)	88.5 (70.2)	40.5 (3.3)	90.2 (74.3)
MAP	85.5 (61.6)	39.7 (3.7)	87.6 (67.8)
MAP/MLLR	88.8 (70.6)	41.0 (4.6)	90.1 (75.1)
MAP/MLLR/GPD	88.8 (70.6)	41.0 (4.6)	90.5 (75.8)

Table 2: Supervised adaptation experiments for LDA visual features. Results are depicted in word (string) accuracy %.

Acknowledgements: The authors would like to thank Hans Peter Graf and Eric Cosatto for providing the lip contour extraction algorithm, as well as, Enrico Bocchieri and Richard Rose for help and useful discussions concerning the MLLR adaptation algorithm implementation.

7. REFERENCES

[1] 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. Stork and M. Hennecke eds., Springer, pp. 331-349, 1996.

[2] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," *Proc. ICIP*, pp. 173-177, 1998.

[3] G. Potamianos and H.P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," *Proc. ICASSP*, pp. 3733-3736, 1998.

[4] G. Potamianos and H.P. Graf, "Linear discriminant analysis for speechreading," *Proc. Work. Multimedia Signal Process.*, pp. 221-226, 1998.

[5] O. Vanegas, A. Tanaka, K. Tokuda, and T. Kitamura, "HMM-based visual speech recognition using intensity and location normalization," *Proc. ICSLP*, pp. 289-292, 1998.

[6] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, v. 2, pp. 291-298, 1994.

[7] L. Neumeyer, A. Sankar, and V. Digalakis, "A comparative study of speaker adaptation techniques," *Proc. EUROSPEECH*, pp. 1127-1130, 1995.

[8] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech Lang.*, v. 9, pp. 171-185, 1995.

[9] F. Liu, R. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," *Proc. ARPA Human Lang. Tech. Work.*, 1993.

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