DISCRIMINATIVE TRAINING OF HMM STREAM EXPONENTS FOR AUDIO-VISUAL SPEECH RECOGNITION

Gerasimos Potamianos and *Hans Peter Graf**

AT&T Labs–Research, 180 Park Ave, Florham Park, NJ 07932-0971, U.S.A.
*AT&T Labs–Research, 100 Schulz Drive, Red Bank, NJ 07701-7033, U.S.A.
email: {makis,hpg}@research.att.com

ABSTRACT

We propose the use of discriminative training by means of the generalized probabilistic descent (GPD) algorithm to estimate hidden Markov model (HMM) stream exponents for audio-visual speech recognition. Synchronized audio and visual features are used to respectively train audio-only and visual-only single-stream HMMs of identical topology by maximum likelihood. A two-stream HMM is then obtained by combining the two single-stream HMMs and introducing exponents that weigh the log-likelihood of each stream. We present the GPD algorithm for stream exponent estimation, consider a possible initialization, and apply it to the single speaker connected letters task of the AT&T bimodal database. We demonstrate the superior performance of the resulting multi-stream HMM to the audio-only, visual-only, and audio-visual single-stream HMMs.

1. INTRODUCTION

Recently, there has been increasing interest in enhancing *automatic speech recognition* (ASR) by using, in addition to audio, *visual* information derived from the speaker's lips or oral cavity measurements [1]-[4]. One of the many challenges facing an *audio-visual* (*bimodal*) ASR system is the *integration* strategy of the audio and visual information.

We are interested in integration strategies that can be easily and successfully extended to large vocabulary continuous ASR. The *early* integration strategy [1]-[4] constitutes a good candidate. In particular, the use of *multi-stream hidden Markov models* (HMMs) [5] with trainable *stream exponents* is a promising approach [3]. The two *single-stream* HMM components separately model the audio and visual modalities, whereas the introduced exponents capture the reliability of each modality, by appropriately weighting the *log-likelihood* of each single-stream HMM. Exponent *discriminative* training under *matching training* and *testing* conditions, by means of the *generalized probabilistic descent* algorithm (GPD), is the subject of this paper.

Exponent training has received some attention in the ASR community, when two streams are used to model static and dynamic audio features [6]-[9]. Exponent training by means of *maximum likelihood* (ML) is inappropriate [3]. Instead, *discriminative* training techniques, such as maximum mutual information [6] and string *minimum classification error* (MCE) by means of the GPD algorithm [7], have been successfully used. In the audio-visual ASR literature, exponents have been trained in [3], where though, a non-differentiable classification error count is minimized and the segmentations of the correct and recognized hypotheses of the training data are not updated. The MCE based GPD algorithm [10] addresses both these shortcomings.

In Section 2, we establish the multi-stream HMM notation. In Section 3, we discuss stream exponent initialization, and, in Section 4, we present the GPD based exponent training algorithm. In Section 5, we describe our bimodal ASR system, and, in Section 6, we present our results on a single speaker connected letters bimodal ASR task.

2. THE MULTI-STREAM HMM

Let us consider S information sources (streams, or modalities) that provide *time synchronous* D_s-dimensional observation vectors $\underline{O}_s^{(t)}$, $s = 1,...,S$, at each time instance t. In our audio-visual framework, $S = 2$, with $s = 1, 2$, representing the audio and visual modalities, respectively. The D-dimensional *multimodal* observation vector at time t is

$$\underline{O}^{(t)} = [\underline{O}_1^{(t)}, \underline{O}_2^{(t)}, ..., \underline{O}_S^{(t)}] \in \mathsf{R}^D, \quad \text{where } \underline{O}_s^{(t)} \in \mathsf{R}^{D_s},$$

for all $s=1,...,S$, and $D=\Sigma_{s=1}^{S} D_s$. Each observation vector time sequence provides information about a sequence of hidden class labels (*states*) $j \in \mathcal{J} = \{1,...,J\}$. We assume that each unimodal observation sequence $\{\underline{O}_s^{(t)}\}$ is modeled by a single-stream unimodal HMM of *identical topology*, *initial* and *transition probabilities*, over all modalities[1], but with modality dependent *emission* probabilities [5]

$$Pr[\underline{O}_s^{(t)}|j] = b_{js}[\underline{O}_s^{(t)}] = \sum_{m=1}^{M_{js}} w_{jms}\, \mathcal{N}_{D_s}(\underline{O}_s^{(t)}; \underline{\mu}_{jms}, \mathbf{\Sigma}_{jms}), \quad (1)$$

for all $j \in \mathcal{J}$, $s=1,...,S$. In (1), *mixture weights* w_{jms} are positive adding up to one, M_{js} denotes the number of mixtures, and $\mathcal{N}_{D_s}(\underline{x}; \underline{\mu}, \mathbf{\Sigma})$ is the D_s–variate normal distribution with mean $\underline{\mu}$ and covariance matrix $\mathbf{\Sigma}$.

Similarly to (1), we consider the single-stream multimodal HMM of $\{\underline{O}^{(t)}\}$ with emission probabilities

$$Pr[\underline{O}^{(t)}|j] = b_j^{(1)}[\underline{O}^{(t)}] = \sum_{m=1}^{M_j} w_{jm}\, \mathcal{N}_D(\underline{O}^{(t)}; \underline{\mu}_{jm}, \mathbf{\Sigma}_{jm}), \quad (2)$$

[1] Initial and transition probabilities are omitted throughout our derivations, since, in practice, the observation sequence log-likelihood is dominated by the emission probability contribution.

for all $j \in \mathcal{J}$, and the multi-stream HMM of $\{\underline{O}^{(t)}\}$ with emission "probabilities" (see also (1))

$$b_j^{(S)}[\underline{O}^{(t)}] = \prod_{s=1}^{S} Pr\,[\underline{O}_s^{(t)} | j\,]^{\gamma_{js}} = \prod_{s=1}^{S} b_{js}[\underline{O}_s^{(t)}]^{\gamma_{js}} . \qquad (3)$$

In (3), γ_{js} denote the stream exponents. In this work, we assume that the stream exponents satisfy the constraints

$$0 \le \gamma_{js} \le 1 \;\text{ and }\; \sum_{s=1}^{S} \gamma_{js} = 1\,, \;\text{ for all }\; j \in \mathcal{J}\,. \qquad (4)$$

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*.

In the following, we concentrate on the issue of stream exponent training. The remaining multi-stream HMM parameters can be estimated by means of traditional maximum likelihood techniques [5] applied on the single-stream HMMs (1). Exponents γ_{js} can be *tied* [5]; i.e., $\gamma_{js} = \gamma_{Cs}$ for all $j \in C$, where $C \in \mathcal{C}$ and $\mathcal{C}$ *partitions* the set of states $\mathcal{J}$. Three possible ways of exponent tying are: (a) At the *global* level; (b) at the HMM *unit*[2] level; and (c) at the HMM *state* level (no tying). Condition (c) is assumed throughout our derivations.

3. STREAM EXPONENT INITIALIZATION

Let us assume that L multimodal observation training sequences $\mathbf{O}^{(l)} = [\underline{O}'^{(1,l)}, ..., \underline{O}'^{(T_l,l)}]$ of duration T_l, $l = 1,...,L$, are available, and let $\mathbf{O} = [\mathbf{O}^{(1)}, ..., \mathbf{O}^{(L)}]$. Following the *Viterbi* training procedure [5], and given $\mathbf{O}$ and a current HMM model[3], we obtain the *forced segmentation* [5] of the correct sentence hypotheses, $\mathcal{F} = \{j_{\mathcal{F}}(t,l);\ t = 1,...,T_l, l = 1,...,L\}$. The log-likelihood of the training data correct hypotheses is (see also footnote 1, (1), and (3))

$$\mathcal{L}_{\mathcal{F}} = \sum_{j=1}^{J}\sum_{s=1}^{S} \gamma_{js}\mathcal{L}_{js\mathcal{F}}\,, \;\text{ where }\; \mathcal{L}_{js\mathcal{F}} = \sum_{l=1}^{L} T_l\, \mathcal{L}_{js\mathcal{F}}^{(l)}, \quad (5)$$

and

$$\mathcal{L}_{js\mathcal{F}}^{(l)} = \frac{1}{T_l}\sum_{t=1}^{T_l} \Delta^{j}_{j_{\mathcal{F}}(t,l)} \log b_{js}[\underline{O}_s^{(t,l)}]\,. \qquad (6)$$

In (6), $\Delta_i^j = 1$, iff $i = j$, $\Delta_i^j = 0$, otherwise.

Maximizing (5) under constraint (4) yields

$$\hat{\gamma}_{js} = \begin{cases} 1, & \text{if } s = \arg_s \max\{\mathcal{L}_{js\mathcal{F}}\,;\, s = 1,...,S\}\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

Clearly, ML exponent estimation fails. Alternative ML exponent estimates appear in [8], under constraints that differ from (4). However, the performance of the resulting HMM in our audio-visual ASR experiments was not satisfactory.

Instead, in addition to (4), we choose to require

$$\gamma_{j1}\mathcal{L}_{j1\mathcal{F}} = \gamma_{js}\mathcal{L}_{js\mathcal{F}}\,, \;\text{ for all }\; s = 2,...,S\,,\; j \in \mathcal{J}\,.$$

[2] In this work all HMM units are *context independent words*.

[3] Such a model could be HMM (3) with all $\gamma_{js} = 1/S$.

The solution to the resulting set of equations is

$$\hat{\gamma}_{js} = \mathcal{L}_{js\mathcal{F}}^{-1}\,(\sum_{s'=1}^{S} \mathcal{L}_{js'\mathcal{F}}^{-1}\,)^{-1}\,. \qquad (7)$$

When estimating tied exponents γ_{Cs}, $C \in \mathcal{C}$, we must replace $\mathcal{L}_{js\mathcal{F}}$ by $\Sigma_{j\in C}\mathcal{L}_{js\mathcal{F}}$ in (7). Choice (7) is suggested in [9] as an initialization of a discriminative exponent training algorithm. In our experiments, (7) provides a good choice when all exponents are globally tied, i.e., $\mathcal{C} = \{\mathcal{J}\}$.

4. STREAM EXPONENT GPD TRAINING

4.1. The GPD algorithm.

Let us denote the *N-best recognized hypotheses* [6], given training data $\mathbf{O}$ and a current HMM model, by $\mathcal{R}_n = \{j_{\mathcal{R}_n}(t,l), t = 1,...,T_l, l = 1,...,L\}$, $n = 1,...,N$. The log-likelihood of the n^{th}-best candidate of sentence l, normalized by the sentence length T_l, is

$$\mathcal{L}_{\mathcal{R}_n}^{(l)} = \sum_{j=1}^{J}\sum_{s=1}^{S} \gamma_{js}\mathcal{L}_{js\mathcal{R}_n}^{(l)}\,, \qquad (8)$$

where (compare to (6))

$$\mathcal{L}_{js\mathcal{R}_n}^{(l)} = \frac{1}{T_l}\sum_{t=1}^{T_l} \Delta^{j}_{j_{\mathcal{R}_n}(t,l)} \log b_{js}[\underline{O}_s^{(t,l)}]\,, \qquad (9)$$

and the normalized log-likelihood of its forced segmentation is (see also (6))

$$\mathcal{L}_{\mathcal{F}}^{(l)} = \sum_{j=1}^{J}\sum_{s=1}^{S} \gamma_{js}\mathcal{L}_{js\mathcal{F}}^{(l)}\,. \qquad (10)$$

Let the misrecognition measure of sentence l be [5], [7], [10],

$$d^{(l)} = -\mathcal{L}_{\mathcal{F}}^{(l)} + \log\left[\frac{1}{N_l'}\sum_{n=1}^{N} \delta_{\mathcal{F}l}^{\mathcal{R}n} \exp\left[\eta\mathcal{L}_{\mathcal{R}_n}^{(l)}\right]\right]^{\frac{1}{\eta}}, \qquad (11)$$

where η is a smoothing parameter, $N_l' = \Sigma_{n=1}^{N}\delta_{\mathcal{F}l}^{\mathcal{R}n}$, and $\delta_{\mathcal{F}l}^{\mathcal{R}n} = 1$ (0), iff the n^{th}-best hypothesis and correct label of sentence l differ (are the same). The loss function

$$\mathcal{E}^{(l)} = \frac{1}{1 + \exp\left[-\alpha\,(d^{(l)} + \beta)\right]}\,, \;\text{ with } \alpha > 0\,, \qquad (12)$$

is a figure of merit of the discrimination between the correct and the recognized hypotheses for sentence l.

The goal in discriminative training is minimizing $\mathcal{E}$, the expected value (average) of $\mathcal{E}^{(l)}$ over all $l = 1,...,L$. This is achieved by means of the GPD algorithm, which updates the HMM parameter vector $\underline{\Lambda}$ according to

$$\underline{\Lambda_{k+1}} = \underline{\Lambda_k} - \epsilon_k \mathbf{U}_k \underline{\nabla\mathcal{E}^{(k')}}\,, \;\text{ for }\; k = 1, 2, \ldots\,, \qquad (13)$$

where $\epsilon_k > 0$, $\lim_{K\to\infty}\Sigma_{k=1}^{K}\epsilon_k = \infty$, $\sup_{K\to\infty}\Sigma_{k=1}^{K}\epsilon_k^2 < \infty$, $\{\mathbf{U}_k\}$ is a sequence of positive definite matrices, and $k' = (k \bmod L) + 1$. The algorithm converges with probability one to a local minimum of $\mathcal{E}$, as $k \to \infty$ [10].

Part	Subjects	Task	Voc.	Words
S-1	1	connected digits	11	500 × 5
S-2	1	connected letters	26	2500 × 4
M-1	50	isolated words	123	1250 × 1
M-2	50	connected letters	26	1250 × 4

Table 1: Current tasks of the AT&T bimodal database.

4.2. Stream exponent update formulas.

We now consider the GPD minimization of $\mathcal{E}$ with respect to γ_{js}, under (4). We introduce parameters $\overline{\gamma}_{js}$, such that

$$\gamma_{js} = (\exp \overline{\gamma}_{js}) (\sum_{s'=1}^{S} \exp \overline{\gamma}_{js'})^{-1} \tag{14}$$

is satisfied, for example

$$\overline{\gamma}_{js} = \log \gamma_{js} , \tag{15}$$

if $\gamma_{js} > 0$. Notice that the optimization problem in the transformed parameter domain is unconstrained, iff $\gamma_{js} \neq 0, 1$. This will hold, if γ_{js} are not initialized to $0, 1$.

From (12) we obtain

$$\frac{\partial \mathcal{E}^{(l)}}{\partial \overline{\gamma}_{js}} = \alpha \, \mathcal{E}^{(l)} \, (1 - \mathcal{E}^{(l)}) \, \frac{\partial d^{(l)}}{\partial \overline{\gamma}_{js}} , \tag{16}$$

where (see also (6) and (8)-(11))

$$\frac{\partial d^{(l)}}{\partial \overline{\gamma}_{js}} = - \frac{\partial \mathcal{L}_{\mathcal{F}}^{(l)}}{\partial \overline{\gamma}_{js}} + \frac{\sum_{n=1}^{N} \delta_{\mathcal{F}l}^{\mathcal{R}n} \frac{\partial \mathcal{L}_{\mathcal{R}n}^{(l)}}{\partial \overline{\gamma}_{js}} \exp [\, \eta \mathcal{L}_{\mathcal{R}n}^{(l)}]}{\sum_{n=1}^{N} \delta_{\mathcal{F}l}^{\mathcal{R}n} \exp [\, \eta \mathcal{L}_{\mathcal{R}n}^{(l)}]} , \tag{17}$$

$$\frac{\partial \mathcal{L}_{\mathcal{M}}^{(l)}}{\partial \overline{\gamma}_{js}} = \frac{\partial \mathcal{L}_{j\mathcal{M}}^{(l)}}{\partial \overline{\gamma}_{js}} = \gamma_{js} [\mathcal{L}_{js\mathcal{M}}^{(l)} - \sum_{s'=1}^{S} \gamma_{js'} \mathcal{L}_{js'\mathcal{M}}^{(l)}] , \tag{18}$$

and $\mathcal{M} = \mathcal{F}, \mathcal{R}_n$. Thus, given a training sentence l, updating γ_{js} involves four steps. I: Computation of $\overline{\gamma}_{js}$ by (15); II: computation of (16), by means of (6), (8)-(12), and (16)-(18); III: update of $\overline{\gamma}_{js}$ by using (13); and IV: computation of γ_{js} by (14). In the tied exponent case, we need to replace $\overline{\gamma}_{js}$ with $\overline{\gamma}_{Cs}$ in (14)-(18) and set $\partial \mathcal{L}_{\mathcal{M}}^{(l)} / \partial \overline{\gamma}_{js} = \Sigma_{j \in C} \partial \mathcal{L}_{\mathcal{M}}^{(l)} / \partial \overline{\gamma}_{js}$ in (18).

4.3. Implementation details.

The GPD algorithm convergence pattern greatly depends on the choice of a variety of parameters, most notably ϵ_k and $\mathbf{U}_k$. In our simulations we use $\mathbf{U}_k = \mathrm{diag}(1,...,1)$ and

$$\epsilon_k = \frac{\epsilon_1}{1 + \lfloor (k-1)/K_o \rfloor} , \text{ for } k = 1, 2, \ldots , \tag{19}$$

where $K_o \geq 1$, i.e., the value of ϵ is updated every K_o training set sentences. Clearly, (19) meets the GPD convergence conditions. In our experiments, and for globally tied stream exponent estimation, two iterations over our whole training set of $L = 2000$ sentences suffice. Values $\eta = \alpha = 1$, $\beta = 0$, $N = 3$, $K_o = 100$, and $\epsilon_1 = 10$ are used.

Figure 1: Visual front end: (a) Original Y-band frame; (b) histogram equalized frame; (c) thresholded frame; (d) mouth center and frame region boundary where DWT is applied.

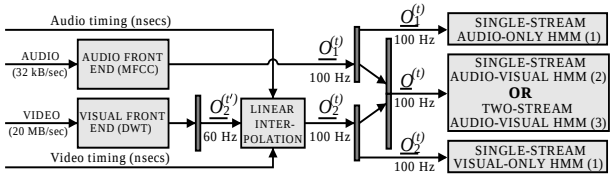

Figure 2: HMM based audio, visual, and audio-visual ASR.

5. THE AUDIO-VISUAL ASR SYSTEM

Our bimodal database [4] addresses various ASR tasks (see Table 1). In this paper, we consider the single speaker connected letters recognition task (part S-2), a challenging task due to the highly confusable *E-set* [5]. Part S-2 consists of 2500 strings (letter four-tuples), randomly partitioned into a 2000-string training and a 500-string test set. The string length is considered unknown at recognition.

The audio front end produces a 39-dimensional *mel-frequency cepstral* coefficient based feature vector, at a 100 Hz rate, with *cepstral mean subtraction* applied to it [5].

The visual front end is a simplified version of our system in [4], and works sufficiently well for the speaker dependent recognition task at hand (Fig. 1). Each 96×80 pixel, YUV 4:2:2, video frame of the speaker's captured frontal face is histogram equalized and further processed by means of simple thresholding for center of mouth location estimation. A *discrete wavelet transform* (DWT) of a Y-band, 16×16 pixel, subsampled image of the area around the mouth center is then performed. The visual feature vector consists of 15 of the resulting wavelet coefficients, as well as their first and second derivatives, over time [4]. Visual features are available at 60 Hz. Linear interpolation is used to align them to the audio features at 100 Hz (see Fig. 2).

All single-stream HMMs (audio-, visual-only, and audio-visual) consist of 26 left-to-right context independent word models of 6-10 states, 8 mixtures per state, and diagonal covariances, and one 32-mixture, single state silence model. They are all trained by using Viterbi (ML) training and the *segmental K-means* algorithm [4], [5]. The two-stream audio-visual HMM is obtained from the audio- and visual-only single-stream HMMs (see (3)). The HMM stream exponents are initialized by (7), and subsequently trained under global, HMM unit, and state level tying.

6. EXPERIMENTS

Various experiments have been conducted to investigate the HMM exponent training algorithm behavior and the relative merits of the multi- vs. single-stream bimodal HMM.

HMM type	Training	Acc. (a)	Acc. (b)
1-stream audio	ML-Viterbi	96.0(85.2)	18.8(3.2)
1-stream visual	ML-Viterbi	35.5(4.6)	35.5(4.6)
1-stream bimodal	ML-Viterbi	92.5(74.4)	71.8(32.6)
2-stream bimodal	Expon. (7)	96.4(86.6)	83.0(50.8)
2-stream bimodal	MCE-GPD	96.4(86.6)	86.4(55.4)

Table 2: Test set word (string) % recognition accuracies (Acc.) by means of various HMMs for two ASR tasks: (a) noise free audio; (b) 12 dB SNR background connected letter noise. The two-stream HMM exponents are globally tied.

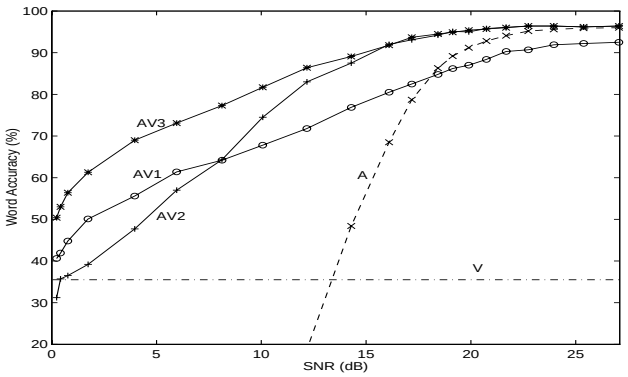

Figure 3: Test set word recognition accuracy of various HMMs (A: audio-only; V: visual-only; AV1: single-stream audio-visual; AV2: two-stream with globally tied exponents given by (7); AV3: two-stream with GPD trained, globally tied exponents), as a function of audio SNR for background connected letters noise.

In Table 2, we depict test set recognition results for the connected letters bimodal ASR task, described in Section 5. Notice that inclusion of the visual modality results in improved ASR when the two-stream audio-visual HMM is used, even in the noise free audio case. Not surprisingly, the ASR improvement is more dramatic in the case of noisy audio. In Figure 3, we depict such a scenario, where the noise consists of connected letters, spoken in the background by the same speaker, and for various SNRs. Clearly, the multi-stream HMM with trained exponents is significantly more robust to noise than the single-stream audio-visual HMM, and exhibits better performance than both audio-only and visual-only HMMs. Setting globally tied exponents to values given by (7) results in HMMs that perform surprisingly better than the audio-only HMM, but, as expected, worse than the multi-stream HMM with GPD trained exponents. As the audio SNR decreases, the relative reliability of the visual modality increases, and, therefore, the estimated audio stream exponent value decreases (from $\gamma_{\mathcal{J}1} \approx 0.9$ in the noise free audio case, to $\gamma_{\mathcal{J}1} \approx 0.45$ in the 0.17 dB SNR case). Finally, in Figure 4, we depict typical convergence behavior of the HMM exponent training algorithm. The algorithm converges to $\gamma_{\mathcal{J}1} \approx 0.9$ (noise free audio case), regardless of initialization. In this specific case, (7) provides an almost identical exponent value. Of course, in general this is not true, as it becomes clear from Figure 3.

The above results refer to exponents that have been tied at a global level. Our experimental results on HMM unit and state level tying have been inconclusive as to whether additional performance gains can be achieved.

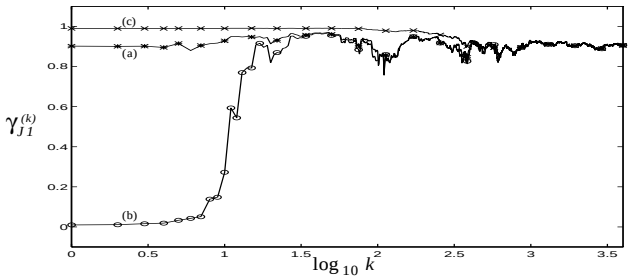

Figure 4: Convergence example of GPD trained, globally tied exponent $\gamma_{\mathcal{J}1}$ for two iterations over the training set and initialization: (a) By (7); (b) 0.01; (c) 0.99.

7. CONCLUSIONS AND FUTURE WORK

We considered the problem of multi-stream HMM exponent training in the context of audio-visual ASR. We proposed the use of the GPD algorithm for discriminative training of such exponents, and we discussed exponent initialization. We achieved significant performance gains in bimodal ASR for a single-speaker connected letters recognition task, when using a two-stream HMM with trained exponents over a single-stream bimodal HMM. Global exponent tying suffices to achieve such gains. Additional work is currently under way to further investigate stream exponent tying strategies.

8. REFERENCES

[1] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 461-471, 1996.

[2] A. Rogozan, P. Deléglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," *Proc. Europ. Tut. Work. Audio-Visual Speech Process. (AVSP)*, Rhodes, pp. 61-64, 1997.

[3] P. Jourlin, "Word dependent acoustic-labial weights in HMM-based speech recognition," *Proc. AVSP*, Rhodes, pp. 69-72, 1997.

[4] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe, "Speaker independent audio-visual database for bimodal ASR," *Proc. AVSP*, Rhodes, pp. 65-68, 1997.

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

[6] Y.-Lu Chow, "Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm," *Proc. ICASSP*, Albuquerque, Vol. 1, pp. 701-704, 1990.

[7] J. Hernando, J. Ayarte, and E. Monte, "Optimization of speech parameter weighting for CDHMM word recognition," *Proc. Eurospeech*, Madrid, Vol. 1, pp. 105-108, 1995.

[8] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," *Proc. ICASSP*, Munich, Vol. 2, pp. 1267-1270, 1997.

[9] C.M. del Álamo, F.J. Caminero-Gil, C. de la Torre-Munilla, and L. Hernández-Gómez, "Codebook weights adaptation for discriminative training of SCHMM-based speech recognition systems," *Proc. Eurospeech*, Madrid, Vol. 1, pp. 93-96, 1995.

[10] W. Chou, B.-H. Juang, C.-H. Lee, and F.K. Soong, "A minimum error rate pattern recognition approach to speech recognition," *J. Pattern Recog. Art. Intell.*, Vol. VIII, pp. 5-31, 1994.