A NEW EM ESTIMATION OF DYNAMIC STREAM WEIGHTS FOR COUPLED-HMM-BASED AUDIO-VISUAL ASR

Ahmed Hussen Abdelaziz, Steffen Zeiler, Dorothea Kolossa*

Institute of Communication Acoustics, Digital Signal Processing Group,
Ruhr-Universität Bochum, 44801 Bochum
{Ahmed.HussenAbdelAziz, Steffen.Zeiler, Dorothea.Kolossa}@rub.de

ABSTRACT

Mutually deploying visual and acoustical information in automatic speech recognition systems increases their robustness against acoustical environmental effects like additive noise and reverberation. Optimal fusion of the audio and video streams requires dynamic adaptation of the relative contribution of each modality. This can be achieved by weighting each stream according to its reliability by an appropriate *stream weight*. In this paper we propose a new expectation maximization algorithm that estimates oracle frame-dependent stream weights for coupled-HMM-based audio-visual speech recognition. Moreover, we introduce a greedy optimization approach that reasonably initializes this algorithm. The proposed approach is evaluated on the Grid audio-visual database and results in an average relative word error rate reduction of 38% and 58% compared to grid search and Bayes fusion, respectively. The estimated oracle stream weights can be used instead of the conventional global fixed stream weights to improve the supervised training of stream weight estimators.

Index Terms— AVASR, CHMM, Stream weight

1. INTRODUCTION & RELATION TO PRIOR WORK

The performance of automatic speech recognition (ASR) systems under lab conditions has recently become very accurate. However, in noisy and reverberant environments, this performance degrades rapidly. In such cases, any additional features that are independent of the acoustical environment while relevant to the speech production process can be useful for achieving a good level of robustness against these effects. Visual features that encode the appearance and the shape of the speaker's mouth are good candidates for such features.

Many models have been proposed to fuse the audio and video information in one audio-visual (AV) ASR system. The difference between these models depends on where this fusion takes place. The fusion can be applied on the feature level, referred to as direct integration (DI), by simply concatenating the audio and visual features [1, 2] or by combining the features in a more complex manner using techniques like dominant or motor recording [3, 4]. Alternatively, the fusion between audio and video modality can be applied at the classifier output level, which is called separate integration (SI). The fusion level in SI techniques varies according to the classifier output definition, e.g., word, phoneme, or state level, [5, 6] and the classifier type, e.g., artificial neural network (ANN) [7] or hidden Markov model (HMM) [8].

In many studies, e.g. [4], it has been shown that SI models outperform DI models. This can be attributed to their capability of modelling asynchrony at different levels as appropriate, ranging from completely asynchronous models like independent HMMs [9] to fully synchronous models like the state-dependent multi-stream HMM (MSHMM) [10]. In this paper, we use the so-called coupled HMM (CHMM) [5, 8, 10, 11], which has the advantage of allowing asynchrony on the state level while preserving the natural dependency between audio and video modalities by forcing synchronization at certain speech units (here, at word boundaries).

Another reason for the superior performance of SI models compared to DI models is their capability to deploy so-called stream weights (SWs). Depending on the reliability of each modality, which varies according to its information content and the time-varying environmental influences, the SWs control the contribution of each modality to the final decision. While in some prior works, the SW for the whole data-set has been set to a fixed value, which was found using grid search, e.g., [10, 12], or using other tuning algorithms, e.g., [13], some authors have assumed that the SW is a model parameter and have estimated it using generative [14] or discriminative [15,16] criteria. In real scenarios, however, the reliability of the audio and video modality can vary quickly, even on the frame level, and such fixed or model-dependent estimation might lead to worse results than using Bayes fusion [7], i.e., equal weights. The main question to be addressed is how to estimate the oracle dynamic (frame-dependent) stream weights (ODSWs) that achieve best recognition accuracy. The estimated ODSWs can then be used as target values when train-

*This work has been supported by the Ministry of Economic Affairs and Energy of the State of North Rhine- Westphalia, Grant IV.5-43-02/2-005-WFBO-009.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

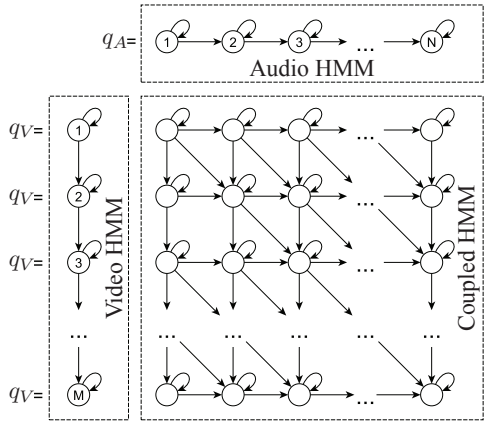

Fig. 1. Coupled HMM with $M \times N$ composite states.

ing blind SW estimators in a supervised manner.

Frame-dependent SW estimators have been reported, using unsupervised or supervised techniques [17, 18]. However, in [18], training has been done using global fixed SWs. In this paper, we propose an expectation maximization (EM) framework that estimates ODSWs for CHMM-based AVASR by maximizing their joint probability with the observable audio-visual features, given the correctly labelled word sequence.

The proposed algorithm differs from the approach in [19], which also estimates frame-dependent SWs, in various aspects. For example, unlike [19], we split the problem into two smaller problems: (1) finding ODSWs, and (2) mapping these OSDWs to reliability measures. Moreover, our approach does not need frame-level labels for audio-visual observations, and uses a different objective function than [19].

The remainder of the paper is organized as follows: After a brief overview of the CHMM, we will introduce the proposed EM algorithm for estimating ODSWs in Sec. 3. Next, a greedy optimization algorithm that estimates an initialization ODSW vector is introduced and a summary of the whole algorithm is given in Sec. 4 and 5, respectively. In Sec. 6, we use the Grid audio-visual database [20] to evaluate the proposed algorithm. Finally, after some conclusions, we give an outlook on further work in Sec. 7.

2. COUPLED HMMS

As mentioned above, the CHMM is an audio-visual fusion model that takes into account the asynchronicities between articulator movements and voice production while enforcing synchronization at word boundaries. A CHMM consists of a 2-dimensional matrix of composite states q, tuples of the marginal audio state q_A and the marginal video state q_V, cf. Fig. 1. Each composite state q has an emission score

$$b_{q,\lambda}(O) = \mathrm{p}(O_A|q_A)^{\lambda}\, \mathrm{p}(O_V|q_V)^{1-\lambda}, \tag{1}$$

where $\mathrm{p}(O_A|q_A)$ and $\mathrm{p}(O_V|q_V)$ are the audio- and video-only state-conditional feature (observation) likelihoods, respectively, and $0 \leq \lambda \leq 1$ is the SW that reflects the relative reliability of the acoustical observation O_A compared to the visual observation O_V. Note that the audio-visual observation vector is defined as $O = \{O_A, O_V\}$. The transition probability between two composite states $i = (i_A, i_V)$ and $j = (j_A, j_V)$ in a CHMM can be written as

$$a_{i,j} = \mathrm{p}\left(q = i | q = j\right) = \prod_{s \in \{A,V\}} a^s_{i_s, j_s}, \tag{2}$$

where $a^s_{i_s,j_s}$ is the transition probability of the corresponding states of the single-stream HMMs.

3. ORACLE DYNAMIC STREAM WEIGHT

We estimate the ODSWs $\boldsymbol{\lambda} = \{\lambda_t\}_{t=1}^T$ given an audio-visual observation sequence $\boldsymbol{O} = \{O_t\}_{t=1}^T$ and the corresponding true word sequence $\boldsymbol{w}$ as

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \left\{\mathcal{F}\left(\boldsymbol{\lambda}\right) = \mathrm{p}\left(\boldsymbol{O}, \boldsymbol{\lambda} | \boldsymbol{w}\right)\right\} \tag{3}$$

$$= \arg\max_{\boldsymbol{\lambda}} \left\{ \log\left(\sum_{\boldsymbol{q} \in \mathcal{Q}} \mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda} | \boldsymbol{w}\right)\right)\right\}, \tag{4}$$

where $\mathcal{Q}$ is the space of all possible 2-dimensional state sequences and $\boldsymbol{q} = \{q_t\}_{t=1}^T$ is one particular state sequence.

Following the technique in [21], it can be shown that

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda'}) = \sum_{\boldsymbol{q}} \frac{\mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda'} | \boldsymbol{w}\right)}{\sum_{\overline{\boldsymbol{q}}} \mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda'} | \boldsymbol{w}\right)} \log\left(\mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda} | \boldsymbol{w}\right)\right) \tag{5}$$

is a strong-sense auxiliary function for the objective function $\mathcal{F}\left(\boldsymbol{\lambda}\right)$ in (3). This means that by iteratively maximizing $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda'})$, a local maximum $\hat{\boldsymbol{\lambda}}$ of $\mathcal{F}\left(\boldsymbol{\lambda}\right)$ can be found.

The joint log-probability in (5) can be factored as:

$$\log(\mathrm{p}(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda} | \boldsymbol{w})) = \log\left(\mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q} | \boldsymbol{w}, \boldsymbol{\lambda}\right) \mathrm{p}(\boldsymbol{\lambda} | \boldsymbol{w})\right) \tag{6}$$

$$= \log\left(\pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}, q_t} b_{q_t, \lambda_t}(O_t) \mathrm{p}(\lambda_t)\right), \tag{7}$$

where π_{q_0} is the initial state probability. Here, we assume that

$$\mathrm{p}(\boldsymbol{\lambda} | \boldsymbol{w}) = \mathrm{p}(\boldsymbol{\lambda}) = \prod_{t=1}^{T} \mathrm{p}(\lambda_t). \tag{8}$$

In (8), the vector $\boldsymbol{\lambda}$ is assumed to be independent of the word sequence $\boldsymbol{w}$ and consists of identical and independently distributed (iid) random variables λ_t. Inserting (7) in (5), we get

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda'}) = \sum_{\boldsymbol{q}} \frac{\mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda'} | \boldsymbol{w}\right)}{\sum_{\overline{\boldsymbol{q}}} \mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda'} | \boldsymbol{w}\right)} \sum_{t=1}^{T} \log\left(b_{q_t, \lambda_t}(\boldsymbol{O}_t) \mathrm{p}(\lambda_t)\right)$$

$$+ \underbrace{\sum_{\boldsymbol{q}} \frac{\mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda'} | \boldsymbol{w}\right)}{\sum_{\overline{\boldsymbol{q}}} \mathrm{p}\left(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{\lambda'} | \boldsymbol{w}\right)} \log\left(\pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}, q_t}\right)}_{\text{independent of } \boldsymbol{\lambda}}. \tag{9}$$

Neglecting the independent terms of $\boldsymbol{\lambda}$ in (9), we get

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda'}) = \sum_{t=1}^{T} \sum_{i=(1,1)}^{(N,M)} \gamma'_t(i) \log \left(b_{q_t=i,\lambda_t}(\boldsymbol{O}_t) \mathrm{p}(\lambda_t) \right), \quad (10)$$

where

$$\gamma'_t(i) = \frac{\mathrm{p}\left(\boldsymbol{O}, q_t = i, \boldsymbol{\lambda'} | \boldsymbol{w}\right)}{\sum_{j=(1,1)}^{(N,M)} \mathrm{p}\left(\boldsymbol{O}, q_t = j, \boldsymbol{\lambda'} | \boldsymbol{w}\right)} \quad (11)$$

is the probability of being in the 2-D state $q = i$ at time t, which is calculated using the state emission scores defined in (1) and the predefined stream weight vector $\boldsymbol{\lambda'}$. The auxiliary function in (10) can now be optimized by separately optimizing the summation terms for each time t.

Thus, by applying (1) to (10) we get the following expression for the auxiliary function at each time point t:

$$Q(\lambda_t, \boldsymbol{\lambda'}) = \left[\sum_{i=(1,1)}^{(N,M)} \gamma'_t(i) \log \left(\frac{\mathrm{p}(O_{At} | q_t = i)}{\mathrm{p}(O_{Vt} | q_t = i)} \right) \right] \lambda_t + \log \mathrm{p}(\lambda_t) + \underbrace{\sum_{i=(1,1)}^{(N,M)} \gamma'_t(i) \log \mathrm{p}(O_{Vt} | q_t)}_{\text{independent of } \lambda_t}. \quad (12)$$

Neglecting the independent terms in (12), we get

$$Q(\lambda_t, \boldsymbol{\lambda'}) = \left[\sum_{i=(1,1)}^{(N,M)} \gamma'_t(i) \log \left(\frac{\mathrm{p}(O_{At} | q_t)}{\mathrm{p}(O_{Vt} | q_t)} \right) \right] \lambda_t + \log \mathrm{p}(\lambda_t). \quad (13)$$

If λ_t is assumed to be uniformly distributed, the auxiliary function (13) will be a linear function of λ_t. Optimizing this function leads to the problem already reported in [22,23], that λ_t can take only the boundary values, i.e., $\lambda_t \in \{0, 1\}$. In other words, at each time t, one stream should be turned off. Which stream is unnecessary depends only on the sign - not on the magnitude - of the derivative of (13).

However, if λ_t is assumed to be normally distributed with mean μ_λ and standard deviation σ_λ, the auxiliary function will be a quadratic function and we need to find

$$\hat{\lambda}_t = \arg\max_{\lambda_t} \left\{ \left[\sum_{i=(1,1)}^{(N,M)} \gamma'_t(i) \log \left(\frac{\mathrm{p}(O_{At} | q_t)}{\mathrm{p}(O_{Vt} | q_t)} \right) \right] \lambda_t - 0.5 \frac{(\lambda_t - \mu_\lambda)^2}{\sigma_\lambda^2} \right\} \quad \text{s.t.} \quad 0 \leq \lambda_t \leq 1. \quad (14)$$

The solution of this problem can be expressed as follows:

$$\hat{\lambda}_t = \begin{cases} \hat{\lambda}_{t_u} & \text{for} \quad 0 \leq \hat{\lambda}_{t_u} \leq 1 \\ 1 & \text{for} \quad \hat{\lambda}_{t_u} > 1 \\ 0 & \text{for} \quad \hat{\lambda}_{t_u} < 0, \end{cases} \quad (15)$$

where

$$\hat{\lambda}_{t_u} = \mu_\lambda + \sigma_\lambda^2 \sum_{i=(1,1)}^{(N,M)} \gamma'_t(i) \log \left(\frac{\mathrm{p}(O_{At} | q_t)}{\mathrm{p}(O_{Vt} | q_t)} \right) \quad (16)$$

is the solution of (14), but without constraints.

Since the EM algorithm only estimates a local maximum near a predefined vector $\boldsymbol{\lambda'}$, in the next section we propose a greedy optimization algorithm to suitably initialize $\boldsymbol{\lambda'}$.

4. INITIALIZATION

In order to find an initial vector $\boldsymbol{\lambda'}$, we apply a step-wise greedy optimization algorithm as follows: At time t, we estimate λ'_t as

$$\lambda'_t = \arg\max_{\lambda_t} \left\{ \mathrm{p}\left(\boldsymbol{O}_{1,\cdots,t}, \lambda_t | \boldsymbol{\lambda'}_{1,\cdots,t-1}, \boldsymbol{w}\right) \right\}. \quad (17)$$

The objective function (17) can be reformulated as follows:

$$\mathrm{p}(\boldsymbol{O}_{1\cdots t}, \lambda_t | \boldsymbol{\lambda'}_{1\cdots t-1}, \boldsymbol{w}) = \mathrm{p}\left(\lambda_t\right) \sum_{i=(1,1)}^{(N,M)} w_i \left(\frac{\mathrm{p}(O_{A_t} | q_t = i)}{\mathrm{p}(O_{V_t} | q_t = i)} \right)^{\lambda_t} \mathrm{p}(O_{V_t} | q_t = i), \quad (18)$$

where

$$w_i = \begin{cases} \pi_i & \text{for} \quad t = 1 \\ \sum_{j=(1,1)}^{(N,M)} a_{i,j} \alpha'_{t-1}(j) & \text{otherwise.} \end{cases} \quad (19)$$

In (19),

$$\alpha'_{t-1}(j) = \mathrm{p}\left(\boldsymbol{O}_{1\cdots t-1}, q_{t-1} = j | \boldsymbol{\lambda'}_{1\cdots t-1}, \boldsymbol{w}\right) \quad (20)$$

is the joint probability of the partial observation sequence $\boldsymbol{O}_{1\cdots t-1}$ and state $q_{t-1} = j$ given the correct word sequence. The values of α' are computed using the state emission scores defined in (1) and the optimized SWs of the preceding time frames $\boldsymbol{\lambda}_{1\cdots t-1}$.

Assuming Gaussian priors $\mathrm{p}\left(\lambda_t\right) = \mathcal{N}\left(\lambda_t, \mu_\lambda, \sigma_\lambda^2\right)$, as discussed in Sec. 3, the objective functions in (18) are convex in the feasible region, i.e. $0 \leq \lambda_t \leq 1$, and can be optimized by gradient ascent. The inequality constraints can be taken into account by adding a logarithmic barrier function [24].

5. SUMMARY

The last piece of the puzzle is how to set the prior parameters μ_λ and σ_λ. A plausible choice of the mean μ_λ, which we call the bias parameter, is the global fixed stream weight, which can be found experimentally by grid search [10]. For the prior's standard deviation σ_λ, which we call the sensitivity parameter, we start with a small value, e.g., 0.1, and increase it iteratively until the estimated SW vector $\hat{\boldsymbol{\lambda}}$ contains only binary values. Indeed, increasing σ_λ more than that does not make any sense. Finally, we choose the weight vector $\hat{\boldsymbol{\lambda}}$ that achieves best accuracy. The entire approach is summarized in Algorithm 1.

Algorithm 1 ODSW for CHMM-based AVASR

A. Set the prior parameters

(1) Set μ_λ to the global fixed stream weight

(2) Initialize σ_λ with a small value.

B. Initialization

(3) Find $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}'$ using the greedy algorithm, cf. Sec. 4.

C. EM Algorithm

(4) Calculate $P = \mathrm{p}\left(\boldsymbol{O}, \hat{\boldsymbol{\lambda}} | \boldsymbol{w}\right)$

E step

(5) Use $\hat{\boldsymbol{\lambda}}$ to calculate $\gamma'_t(i)$ for all times and states.

M step

(6) Update $\hat{\boldsymbol{\lambda}}$ using the procedure in Sec. 3.

Convergence test

(7) Calculate $P^* = \mathrm{p}\left(\boldsymbol{O}, \hat{\boldsymbol{\lambda}} | \boldsymbol{w}\right)$

(8) If $P^* - P > \epsilon \Longrightarrow P = P^*$, go to (5), else go to (9).

D. Recognition

(9) Use the estimated $\hat{\boldsymbol{\lambda}}$ to recognize the training utterance.

(10) Calculate the accuracy A and increase σ_λ.

E. Iteration

(11) Repeat (3-10) until all values of $\hat{\boldsymbol{\lambda}}$ are 0 or 1.

(12) Choose $\hat{\boldsymbol{\lambda}}$ that achieves best accuracy.

6. EXPERIMENTS AND RESULTS

6.1. Data set

Two male and two female speakers have been chosen from the Grid audio-visual database [20] to evaluate the presented framework. The single-stream HMMs have been trained with 3600 clean utterances, i.e., 90% of the speakers' utterances. We have considered 188 speech signals a test set and have artificially distorted them with three types of noise: white noise, buccaneer jet cockpit noise and speech babble. The noise signals, taken from the NOISEX-92 database [25], have been added to the speech data at signal-to-noise ratios (SNRs) between 0 and 15 dB according to ITU-T P.56 [26].

6.2. Experimental setup

All speech signals have been down-sampled from $f_s = 25$kHz – the sampling frequency of the Grid database – to 8kHz.

We have used the first 13 static MFCCs [27] extracted by the ETSI advanced front-end (AFE) [28] along with the 26 corresponding Δ and $\Delta\Delta$ coefficients as the audio features. The video features are the 64-dimensional DCT coefficients, encoding the appearance and shape of the speaker's mouth. The corresponding mouth region was determined automatically by Viola-Jones face and mouth detector [29]. The dimension of both audio and video features has been reduced to 31 by linear discriminant analysis (LDA) [30].

Noise Type	SNR [dB]	A-only	AV Bayes	AV Fix	AV Prop.
Babble	0	27.39	76.60	86.52	**90.34** *
	5	52.30	87.68	89.27	**94.77** *
	10	73.49	93.44	93.44	**97.25** *
	15	91.58	96.54	97.25	**98.76** *
White	0	23.49	79.61	87.32	**90.25** *
	5	43.62	87.23	88.48	**93.44** *
	10	69.95	92.55	93.17	**96.45** *
	15	90.96	95.48	96.81	**98.08** *
J-Slow	0	25.53	70.12	86.26	**89.27** *
	5	39.01	79.52	87.59	**91.76** *
	10	49.20	86.79	89.18	**94.59** *
	15	66.31	91.13	92.29	**96.10** *
Clean	-	98.58	97.96	99.02	**99.47**
Avg.	-	57.80	87.28	91.28	**94.66** *

Table 1. Word accuracies in different environments.

The marginal single-stream HMMs have been generatively trained using the clean signals. Each HMM set consists of 51 whole-word HMMs and one silence HMM. The word HMMs are left-to-right linear models, with three states per phone in audio HMMs and one state per phone in video HMMs. The state conditional probabilities are 4-component diagonal covariance GMMs. All experiments have been conducted using our own AVASR system JASPER, see, e.g., [31].

6.3. Results

Table 1 shows the recognition performance in terms of word accuracies of the audio-only ASR and the AVASR when Bayes fusion [7], the best global fixed SW, and the proposed estimated SWs are used. The visual ASR accuracy in this experiment was 85.9%. The global stream weights are obtained by grid search to maximize the recognition accuracy for each data set separately, i.e. for each noise type and at each noise level, one fixed weight is found. Using the proposed weights improves the recognition performance of the AVASR in every considered case by on average 3.38%. The results with asterisks achieved statistically significant improvements relative to the results obtained using the best global SWs, according to Fisher's exact test applied at $\mathrm{p} = 0.05$.

7. CONCLUSIONS

In this paper, we have introduced a new EM algorithm that estimates ODSWs for CHMM-based AVASR. The estimated ODSWs can be used to train mapping functions that map reliability measures like the signal to noise ratio or the entropy to the optimal dynamic rather than fixed global global SWs. Moreover, due to the numerous data points provided by the proposed algorithm more complex mapping functions such as DNNs can be applied.

8. REFERENCES

[1] P. L. Silsbee and Q. Su, "Audio-visual sensory integration using hidden Markov models," in *NATO ASI Conference on Speech reading by Man and Machine: Models, Systems and Applications*, Berlin, Germany, 1996.

[2] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," in *NATO ASI Conference on Speech reading by Man and Machine: Models, Systems and Applications*, Berlin, Germany, 1996.

[3] T. Watanabe and M. Kohda, "Lip-reading of Japanese vowels using neural networks," in *ICSLP*, Kobe, Japan, 1990.

[4] P. Teissier, J. Robert-Ribes, J.-L. Schwartz, and A. Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 6, pp. 629–642, 1999.

[5] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *ICASSP*, Salt Lake City, Utah, USA, 2001.

[6] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *ICSLP*, Philadelphia, Pennsylvania, USA, 1996.

[7] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURAS*, vol. 2002, pp. 1260–1273, 2002.

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[9] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large- vocabulary audio-visual speech recognition: a summary of the Johns Hopkins summer 2000 workshop," in *IEEE Workshop on Multimedia Signal Processing*, Cannes, France, 2001.

[10] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1274–1288, 2002.

[11] M. Tomlinson, M. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *ICASSP*, Atlanta, Georgia, USA, 1996.

[12] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[13] A. Kantor and M. Hasegawa-Johnson, "Stream weight tuning in dynamic Bayesian networks," in *ICASSP*, Las Vegas, Nevada, USA, 2008.

[14] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," in *ICASSP*, Munich, Germany, 1997.

[15] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *ICASSP*, Orlando, Florida, USA, 2002.

[16] L. Peng and W. Zuoying, "Stream weight training based on MCE for audio-visual LVCSR," *Tsinghua Science and Technology*, vol. 10, no. 2, pp. 141–144, 2005.

[17] M. Gurban and J.-P. Thiran, "Using entropy as a stream reliability estimate for audio-visual speech recognition," in *European Signal Processing Conference*, Lausanne, Switzerland, 2008.

[18] V. Estellers, M. Gurban, and J.-P. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1145–1157, 2012.

[19] A. Garg, G. Potamianos, C. Neti, and T. S. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," in *International Conference on Multimedia and Expo*, Baltimore, Maryland, USA, 2003.

[20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[22] P. Jourlin, "Word-dependent acoustic-labial weights in HMM-based speech recognition," in *European Tutorial Workshop on Audio-visual Speech Recognition*, Rhodes, Greece, 1997.

[23] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *ICASSP*, Seattle, Washington, USA, 1998.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*, 7th ed. Cambridge University Press, 2009.

[25] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[26] *Objective measurement of active speech level*, International Telecommunication Union Std., 1993.

[27] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.

[28] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, ETSI, ES.202.050 Std., 2003.

[29] G. Bradski and A. Kaehler, *Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

[30] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Pearson, 2007.

[31] A. Vorwerk, S. Zeiler, D. Kolossa, R. F. Astudillo, and D. Lerch, *Robust Speech Recognition of Uncertain or Missing Data*. Springer, 2011, ch. Use of Missing and Unreliable Data for Audiovisual Speech Recognition, pp. 345–373.