HIERARCHICAL DISCRIMINANT FEATURES FOR AUDIO-VISUAL LVCSR

Gerasimos Potamianos,[1] *Juergen Luettin,*[2] and *Chalapathy Neti* [1]

[1] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
[2] Ascom Systec AG, 5506 Maegenwil, Switzerland
E-mails: [1] {gpotam,cneti}@us.ibm.com ; [2] Juergen.Luettin@ascom.ch

ABSTRACT

We propose the use of a hierarchical, two-stage discriminant transformation for obtaining audio-visual features that improve automatic speech recognition. Linear discriminant analysis (LDA), followed by a maximum likelihood linear transform (MLLT) is first applied on MFCC based audio-only features, as well as on visual-only features, obtained by a discrete cosine transform of the video region of interest. Subsequently, a second stage of LDA and MLLT is applied on the concatenation of the resulting single modality features. The obtained audio-visual features are used to train a traditional HMM based speech recognizer. Experiments on the IBM ViaVoice$^{\rm TM}$ audio-visual database demonstrate that the proposed feature fusion method improves speaker-independent, large vocabulary, continuous speech recognition for both clean and noisy audio conditions considered. A 24% relative word error rate reduction over an audio-only system is achieved in the latter case.

1. INTRODUCTION

Improving *automatic speech recognition* (ASR) by exploiting visual, mouth region information has actively been pursued during the last few years [1]-[5]. However, to date, all *automatic speechreading* studies have been limited to small vocabulary tasks, small subject populations, and are hardly ever compared on any common audio-visual database [1]. Thus, no definite answers exist on the two key issues for the design of *speaker-independent*, audio-visual, *large vocabulary continuous speech recognition* (LVCSR) systems: (a) The choice of appropriate *visual features*, informative about unconstrained, continuous visual speech; and (b) The design of audio-visual information *fusion* algorithms that outperform traditional audio-only LVCSR systems, under all possible audio and video channel conditions. To address these issues, we participated in the summer 2000 workshop at the Johns Hopkins University on audio-visual ASR, seriously tackling the problem of speaker-independent audio-visual LVCSR for the first time [6]. Although we did consider a number of different visual feature representations, our main concentration was on fusion algorithms.

Audio-visual fusion is an instance of the general classifier combination problem [7]. Here, two observation streams are available (audio and visual modalities) and provide information about hidden class labels, such as *hidden Markov model* (HMM) states, or, at a higher level, word sequences. Each stream can be used alone to train single modality statistical classifiers to recognize such classes. Combining the two streams can hopefully result in a bimodal classifier that outperforms both single modality ones. A number of techniques have been suggested for audio-visual fusion [1], which can be broadly grouped into *feature fusion* [2], [3] and *decision fusion* [2]-[5] methods. The first are based on training a traditional HMM classifier on the concatenated vector of the audio and visual features. Decision fusion techniques combine classification decisions based on single modality observations, typically by appropriately weighting their respective log-likelihoods.

In this paper, we exclusively consider feature fusion. Decision fusion techniques are presented in two accompanying papers [8], [9]. To eliminate duplication, visual feature extraction, the audio-visual database, and the experimental framework, common to all three papers, are described in most detail here. Specifically, in this work, we propose the use of *linear discriminant analysis* (LDA) [10] to discriminantly reduce the dimensionality of the concatenated audio-visual feature vector, followed by a *maximum likelihood linear transform* (MLLT) [11] to improve data modeling. The method outperforms feature fusion by means of plain audio and visual feature concatenation [2], [3], and it improves LVCSR under both clean and noisy audio conditions. LDA and MLLT are also used to incorporate dynamic information into the audio and visual feature streams, preceding fusion. Hence, the whole scheme amounts to a *hierarchical*, two-stage application of these transforms, and it is referred to as HiLDA (Hierarchical LDA).

Section 2 of the paper reviews LDA and MLLT. Section 3 describes their use in obtaining single modality features. Feature fusion is presented in Section 4. The audio-visual database, experimental paradigm, and LVCSR experiments are reported in Sections 5, 6, and 7, respectively. A summary follows in Section 8.

2. FEATURE TRANSFORMATIONS FOR IMPROVED CLASSIFICATION

LDA and MLLT are used to map features to new spaces for improved classification. Both assume that a set of *classes* $\mathcal{C}$ (such as HMM states) is a-priori given, as well as that the training set data vectors $\mathbf{x}_l$, $l = 1 ,..., L$, of dimension d , are *labeled* as $c(l) \in \mathcal{C}$.

2.1. Linear Discriminant Data Projection

LDA seeks a projection matrix $\mathbf{P}_{\rm LDA}$ of size $D \times d$, where $D < d , |\mathcal{C}|$, such that the projected training sample $\{ \mathbf{P}_{\rm LDA} \, \mathbf{x}_l , \; l = 1 ,..., L \}$ is "well separated" into the set of classes $\mathcal{C}$ according to a function of the training sample *within-class scatter* matrix $\mathbf{S}_W$ and its *between-class scatter* matrix $\mathbf{S}_B$ [10]. These matrices are

$$\mathbf{S}_W = \sum_{\mathsf{c} \in \mathcal{C}} Pr(\mathsf{c}) \Sigma^{(\mathsf{c})}, \;\; \mathbf{S}_B = \sum_{\mathsf{c} \in \mathcal{C}} Pr(\mathsf{c}) (\mathbf{m}^{(\mathsf{c})} - \mathbf{m}) (\mathbf{m}^{(\mathsf{c})} - \mathbf{m})^{\top}, \quad (1)$$

respectively. In (1), $Pr(\mathsf{c}) = L_{\mathsf{c}} / L$, $\mathsf{c} \in \mathcal{C}$, is the class empirical probability mass function, where $L_{\mathsf{c}} = \Sigma_{l=1}^{L} \delta_{c(l),\mathsf{c}}$, and $\delta_{i,j} = 1$, if

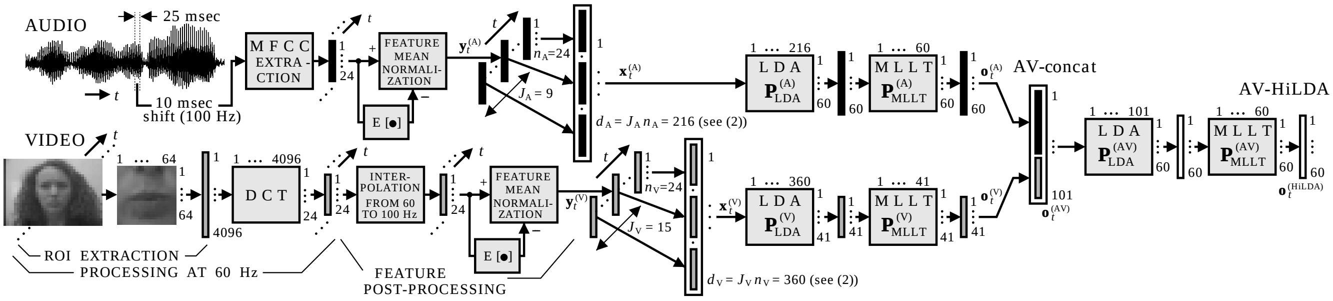

Fig. 1. Feature extraction for audio-visual ASR by a hierarchical, two-stage application of LDA and MLLT (see also (2), (3), (4), and (6)).

$i = j$; 0 , otherwise; $\mathbf{m}^{(c)}$ and $\Sigma^{(c)}$ denote the class sample mean and covariance, respectively; finally, $\mathbf{m} = \Sigma_{c\in\mathcal{C}} \, Pr(c)\, \mathbf{m}^{(c)}$ is the total sample mean.

To estimate $\mathbf{P}_{\rm LDA}$, we compute the *generalized* eigenvalues and *right* eigenvectors of the matrix pair ($\mathbf{S}_B$,$\mathbf{S}_W$) that satisfy $\mathbf{S}_B\mathbf{F} = \mathbf{S}_W\mathbf{F}\Lambda$ [10]. Matrix $\mathbf{F} = [\mathbf{f}_1,..., \mathbf{f}_d]$ has as columns the generalized eigenvectors. Let the D largest eigenvalues be at the $j_1,..., j_D$ diagonal positions of Λ ; then, $\mathbf{P}_{\rm LDA} = [\mathbf{f}_{j_1},..., \mathbf{f}_{j_D}]^\top$.

2.2. Maximum Likelihood Data Rotation

MLLT seeks a square, non-singular, data rotation matrix $\mathbf{P}_{\rm MLLT}$ that maximizes the observation data likelihood in the original feature space, under the assumption of diagonal data covariance in the transformed space. Such a rotation is beneficial, since, in LVCSR, diagonal covariances are typically assumed when modeling the observation class conditional probability distribution. The desired rotation matrix is obtained numerically by solving [11]

$$\mathbf{P}_{\rm MLLT} = \arg\max_{\mathbf{P}} \{ \mathsf{det}(\mathbf{P})^L \prod_{c\in\mathcal{C}} (\mathsf{det}(\mathsf{diag}(\mathbf{P}\Sigma^{(c)}\mathbf{P}^\top)))^{-\frac{L_c}{2}} \},$$

where $\mathsf{diag}(\bullet)$, $\mathsf{det}(\bullet)$, denote matrix *diagonal* and *determinant*.

3. SINGLE MODALITY FEATURES

To obtain single modality (audio- and visual-only) features, first LDA and subsequently MLLT are applied on a concatenation of consecutive *static* features, as a means of incorporating *dynamic* speech information and improving recognition. Let us denote the audio- and visual-only, time-synchronous, static features of dimension n_s , by $\mathbf{y}_t^{(s)} \in \mathsf{R}^{n_s}$, where s = A ,V, respectively. Let us consider J_s consecutive such feature vectors and denote by

$$\mathbf{x}_t^{(s)} = [\, \mathbf{y}_{t-\lfloor J_s/2\rfloor}^{(s)\top}, ..., \mathbf{y}_t^{(s)\top}, ..., \mathbf{y}_{t+\lceil J_s/2\rceil-1}^{(s)\top} \,]^\top, \qquad (2)$$

their concatenation of dimension $d_s = J_s \, n_s$. The final audio- and visual-only feature vectors of dimension D_s are then

$$\mathbf{o}_t^{(s)} = \mathbf{P}_{\rm MLLT}^{(s)} \, \mathbf{P}_{\rm LDA}^{(s)} \, \mathbf{x}_t^{(s)} , \quad \text{where} \quad s = {\rm A}, {\rm V}, \qquad (3)$$

and matrices $\mathbf{P}_{\rm LDA}^{(s)}$ and $\mathbf{P}_{\rm MLLT}^{(s)}$ are of dimensions $D_s \times d_s$ and $D_s \times D_s$, respectively (see also Fig. 1). Values $n_{\rm A} = 24$, $J_{\rm A} = 9$, $D_{\rm A} = 60$, and $n_{\rm V} = 24$, $J_{\rm V} = 15$, $D_{\rm V} = 41$, are used here.

The audio-only static features consist of 24 mel-frequency cepstral coefficients, computed over a sliding window of 25 msec, and at a rate of 100 Hz. *Feature mean normalization* (FMN) is used to obtain $\mathbf{y}_t^{(\rm A)}$ [12]. Static visual features are extracted using an appearance based technique [5], [6]. First, a statistical face tracking algorithm is used to detect the speaker's face and estimate the mouth location and size [13]. Based on these, a size-normalized, 64 × 64 pixel *region of interest* (ROI) is extracted for every video frame at 60 Hz, containing the speaker's mouth. Subsequently, a two-dimensional, separable, *discrete cosine transform* (DCT) is applied to the ROI, and the 24 highest-energy (over all training data) DCT coefficients are retained as static features. To facilitate audio-visual fusion, linear interpolation is used to obtain visual features, time-synchronous to the audio ones at 100 Hz. Finally, FMN is employed to compensate for lighting variations, providing the final visual-only static features $\mathbf{y}_t^{(\rm V)}$ (see Fig. 1).

4. AUDIO-VISUAL FEATURE FUSION

Two feature fusion methods are considered (see also Fig. 1). The first (baseline one) uses the concatenation of the synchronous audio and visual features as the joint bimodal feature vector [2], [3]; the second is the proposed HiLDA technique.

4.1. Concatenative Feature Fusion

The joint, concatenated audio-visual feature vector is (see also (3))

$$\mathbf{o}_t^{(\rm AV)} = [\, \mathbf{o}_t^{(\rm A)\top}, \mathbf{o}_t^{(\rm V)\top} \,]^\top \in \mathsf{R}^D , \qquad (4)$$

where $D = D_{\rm A} + D_{\rm V}$. We model the generation process of a sequence of such features by a *single-stream* HMM, with *emission* (class conditional observation) probabilities, given by [12]

$$Pr\,[\, \mathbf{o}_t^{(\rm AV)} | \, c \,] = \sum_{k=1}^{K_c} w_{c\,k} \, \mathcal{N}_D \, (\, \mathbf{o}_t^{(\rm AV)} ; \mathbf{m}_{c\,k} , \mathbf{s}_{c\,k} \,) \, . \qquad (5)$$

In (5), $c \in \mathcal{C}$ denote the HMM context dependent states (classes). In addition, mixture weights $w_{c\,k}$ are positive adding up to one, K_c denotes the number of mixtures, and $\mathcal{N}_D \, (\mathbf{o} \, ; \mathbf{m} \, , \mathbf{s})$ is the D-variate normal distribution with mean $\mathbf{m}$ and a diagonal covariance matrix, its diagonal being denoted by $\mathbf{s}$.

In our system, the concatenated audio-visual observation vector (4) is of dimension 101. This is rather high, and it can cause inadequate modeling in (5) due to the curse of dimensionality. To avoid this, we seek lower dimensional representations of (4), next.

4.2. Hierarchical Fusion Using Feature Transformations

The visual features currently used contain less speech classification power than audio features, even in the case of extreme noise in

Fig. 2. IBM ViaVoice$^{\mathrm{TM}}$ audio-visual database example subjects.

the audio channel (see Section 7). One would therefore expect that a lower-dimensional representation of (4) could lead to equal, or even better, HMM performance, given the lack of accurate probabilistic modeling in very high dimensional spaces. Similarly to Section 3, we consider LDA, followed by MLLT, as a means of obtaining such a dimensionality reduction. The final audio-visual feature vector is then (see also (3) and (4))

$$\mathbf{o}_t^{(\mathrm{HiLDA})} \;=\; \mathbf{P}_{\mathrm{MLLT}}^{(\mathrm{AV})}\, \mathbf{P}_{\mathrm{LDA}}^{(\mathrm{AV})}\, \mathbf{o}_t^{(\mathrm{AV})} \;. \qquad (6)$$

In our experiments, the LDA matrix $\mathbf{P}_{\mathrm{LDA}}^{(\mathrm{AV})}$ is of size 60×101, giving rise to 60-dimensional HiLDA audio-visual features.

5. THE AUDIO-VISUAL DATABASE

To allow speaker-independent audio-visual LVCSR experiments, a suitable database has been collected at the IBM T.J. Watson Research Center [6]. It consists of full-face frontal video and audio of 290 subjects (see also Fig. 2), uttering ViaVoice$^{\mathrm{TM}}$ scripts (continuous read speech with mostly verbalized punctuation), with a 10,400 word vocabulary. The database video is of size 704×480 pixels, interlaced, captured in color at a rate of 30 Hz (60 fields per second are available at a resolution of 240 lines), and is MPEG2 encoded at the relatively high compression ratio of about 50:1. High quality wideband audio is synchronously collected with the video at a rate of 16 kHz in an office environment at a 19.5 dB SNR. The database duration is close to 50 hours (24,325 utterances). To date, this is the largest audio-visual database collected, and the only one suitable for LVCSR [1]-[5].

6. EXPERIMENTAL FRAMEWORK

Approximately 42 hours of data are used in speaker-independent audio-visual ASR experiments, partitioned into three sets: The *training* set that contains 35 hours of data (17,111 utterances) from 239 subjects, used for HMM parameter estimation. The *held-out* set of close to 5 hours of data (2,277 utterances) from 25 additional subjects, used for roughly optimizing the *language model* (LM) weight and the word insertion penalty during *lattice rescoring*, as well as, in [8], [9], for training parameters relevant to audio-visual decision fusion. Finally, the 2.5-hour *test* set (1,038 utterances) from the 26 remaining subjects is provided for HMM evaluation.

Two audio conditions are considered: The original database clean audio (19.5 dB SNR), and a degraded one, where the audio is artificially corrupted by additive "*babble*" speech noise (8.5 dB SNR). All HMMs, as well as the LDA and MLLT matrices used in feature extraction, are trained in the *matched* condition.

All experiments are conducted using the HTK toolkit [12]. Due to its LVCSR decoding limitations, a lattice rescoring strategy is followed: First, using the IBM LVCSR decoder with a trigram LM and IBM-trained HMM systems, appropriate ASR lattices are generated. These lattices are subsequently rescored using the HTK decoder by various context-dependent triphone HTK-trained HMM systems, based on a number of feature sets and fusion strategies. Three sets of such lattices are generated for both the held-out and test sets (see Table 1). Lattices:

- *Lat* are based on *clean* audio-only features;
- *NLat* on *noisy* audio-only features (matched training); and
- *NAVLat* are based on *noisy* HiLDA audio-visual features.

Lattice rescoring results on the test set, expressed in *word error rate* (WER), % [12], are reported in Section 7.

Lattices	Rescored	Oracle	Anti-oracle	LM-only
"Lat"	14.44	5.53	46.83	29.57
"NLat"	48.10	26.81	96.12	58.31
"NAVLat"	36.99	16.84	103.69	52.02

Table 1. Test set word error rate (WER, %) of the rescored lattices by the corresponding HTK trained system. Oracle, anti-oracle, and language model (LM) only best path WERs are also depicted.

7. EXPERIMENTS

First, baseline audio-only results are obtained for both clean and noisy audio conditions, using HMMs trained in the matched audio condition to rescore lattices "Lat" and "NLat", respectively. Performance deteriorates significantly from a 14.44% WER for clean audio to a 48.10% WER in the noisy case (see also Table 1).

Subsequently, it is investigated whether visual-only features provide useful speech information in the LVCSR domain. Visual-only HMMs are trained and used to rescore lattices "NLat". Of course, such lattices do contain audio information, therefore the obtained results cannot be interpreted as visual-only recognition. As Table 2 shows, in the case where no LM information is used, the performance improves from a 78.14% WER, when no visual feature information is present (random lattice path), to a 61.06% WER, when the visual-only HMM is used in rescoring. Similarly, in the case where the LM scores are utilized, the WER improves from 58.31% (best lattice path based on LM alone) to 51.08%, when the visual-only HMM scores are also employed. Clearly, the visual features do provide useful information for LVCSR.

Next, it is demonstrated that visual speech information improves ASR performance, using the HiLDA fusion method. In Table 3, fusion results are depicted for both clean and noisy audio. In the first case, lattices "Lat" are rescored using HMMs trained on concatenated audio-visual features (AV-concat), as well as on HiLDA features (AV-HiLDA). When using the former, some performance degradation with respect to the baseline clean audio-only WER is observed (from a 14.44% WER to a 16.00% one). However, the HiLDA feature fusion outperforms the audio-only baseline by achieving a 13.84% WER, which amounts to a 4% relative reduction in WER. In the noisy audio case, lattices "NLat" are first rescored. Both fusion techniques give substantial gains over the noisy audio-only baseline performance, with HiLDA being again the best method. When rescoring the "NAVLat" lattices, both results improve significantly. The HiLDA algorithm yields a 36.99% WER, compared to the baseline noisy audio-only 48.10% WER. This amounts to a 24% WER relative reduction. Notice that "NAVLat" lattice rescoring provides the fair result to report for the HiLDA technique. However, the concatenative feature fusion result is "boosted" by its superior "NAVLat" lattices. Its actual, free

Condition	WER (%)
Visual-only (with LM)	51.08
LM-only (no features)	58.31
Visual-only, with no LM	61.06
Random lattice path	78.14
Noisy audio-only	48.10

Table 2. "NLat" lattice rescoring results (in WER, %) on the test set obtained with or without visual-only HMM and language model (LM) scores. Noisy audio-only WER is also shown.

decoding performance is expected to be worse than the 40.00% WER, but better than the 44.97% WER, reported in Table 3.

It is of course not surprising that HiLDA outperforms feature concatenation. In our implementation, concatenated audio-visual features are of dimension 101, compared to audio-only and HiLDA features, that are both of dimension 60. HiLDA uses a discriminative feature projection to efficiently "compact" features (4), and, implicitly, it models the correlation, reliability, and possible asynchrony of the audio- and visual-only feature streams. The curse of dimensionality and undertraining are likely also to blame for the performance degradation compared to the clean audio-only system, when concatenated features are used.

8. SUMMARY

We presented a novel technique for fusing audio and visual information at the feature level in bimodal ASR. The algorithm consists of a two-stage, hierarchical application of LDA and MLLT, first on single modality, static features, and subsequently on their concatenation. The initial LDA captures class-discriminant static and dynamic information within the audio-only and visual-only feature streams, whereas the subsequent LDA provides a discriminative feature projection that compresses the concatenation of the resulting audio and visual features. In both cases, the MLLT provides a data rotation that significantly improves data modeling. We applied the algorithm to speaker-independent, large vocabulary, continuous audio-visual speech recognition, and we demonstrated that it reduced the word error rate of an audio-only state-of-the-art system for both clean and (matched) noisy audio conditions, by about 4% and 24%, relative, respectively. This constitutes the first time that such improvements have been obtained in the LVCSR domain.

In this paper, we also presented the basic framework of our work on audio-visual ASR during the summer 2000 workshop at the Johns Hopkins University, namely visual feature extraction, the audio-visual database, and the experimental paradigm followed. Alternative techniques for incorporating visual information based on decision fusion have also been employed at this workshop. Such algorithms are presented in accompanying papers [8] and [9].

9. ACKNOWLEDGMENTS

We would like to acknowledge contributions to this work by G. Iyengar and A.W. Senior (IBM) for help with face and mouth region detection for the IBM ViaVoice^{TM} audio-visual data and by E. Helmuth (IBM) for help in data collection. Furthermore, we would like to thank the entire audio-visual ASR research team at the Johns Hopkins summer 2000 workshop for fruitful collaboration, as well as F. Jelinek (CLSP) for hosting the workshop.

Audio Condition:	Clean	Noisy	
Rescored Lattices:	"Lat"	"NLat"	"NAVLat"
Audio-only	14.44	48.10	–
AV-concat (4)	16.00	44.97	40.00
AV-HiLDA (6)	13.84	42.86	36.99

Table 3. Audio-visual feature fusion performance (in WER, %) on the test set using concatenated and hierarchical LDA audio-visual features in both clean and noisy audio conditions.

10. 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.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 331-349, 1996.

[2] A. Adjoudani and C. Benoît, "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.

[3] G. Potamianos and H.P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 6, pp. 3733-3736, 1998.

[4] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141-151, 2000.

[5] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A cascade image transform for speaker independent automatic speechreading," *Proc. Int. Conf. Multimedia Expo.*, vol. II, pp. 1097-1100, 2000.

[6] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," *Final Workshop 2000 Report*, Center for Language and Speech Processing, Baltimore, 2000 (`http://www.clsp.jhu.edu/ws2000/final_reports/avsr/`).

[7] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4-37, 2000.

[8] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2001.

[9] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2001.

[10] C.R. Rao, *Linear Statistical Inference and Its Applications*. John Wiley and Sons, New York, 1965.

[11] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 661-664, 1998.

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

[13] A.W. Senior, "Face and feature finding for a face recognition system," *Proc. Int. Conf. Audio and Video-based Biometr. Person Authent.*, pp. 154-159, 1999.