Cross-Modal Analysis of Audio-Visual Programs for Speaker Detection

Dongge Li*, Cuneyt Taskiran*, Nevenka Dimitrova†, Wei Wang*, Mingkun Li‡, and Ishwar Sethi‡

*Multimedia Research Laboratory (MRL), Motorola Labs, 1301 E. Algonquin Rd., Schaumburg, IL 60196
Email: {dongge.li, cuneyt.taskiran, wei.wang}@motorola.com

†Philips Research, 345 Scarborough Rd., Briarcliff Manor, NY 10510
Email: nevenka.dimitrova@philips.com

‡Intelligent Information Engineering Laboratory, Oakland University, Rochester, MI 48309
Email: {sethi, li}@oakland.edu

***Abstract*— This paper describes a speaker detection system using cross-modal association methods. Four association approaches are designed using linear and nonlinear association models. Speaker detection experiments were conducted to compare the approaches.**

I. Introduction

Multimedia content usually contains two or more media streams that share semantic and/or temporal relationships. Examples of such media streams are synchronized video and audio streams for audiovisual programs, a text and its spoken presentation, and images and the captions associated with them. In these examples media streams of different modality jointly contribute to the overall semantics of the multimedia content. In multimedia content analysis and indexing, generally the modalities are processed separately and the outputs of these unimodal systems are fused in a final combination stage. However, in this process of separation of modalities, valuable information is lost about the whole event and/or object that is to be analyzed and detected. Cross-modal multimedia processing systems that share information across all levels of processing will lead to synergistic integration of multiple modalities and thus will have better analysis and detection performance compared with systems that deal with the modalities separately. The cross-modal approach is also well supported biologically, where cross-modal influences between different perceptions, such as visual, auditory, and olfactory inputs, occur at the earliest stages of sensory processing [2].

We refer to the task of employing cross-modality information analysis methods to identify and measure intrinsic associations between media streams of different modalities as *cross-modal association*. In this paper we propose several cross-modal association approaches based on both linear and nonlinear correlation models. Although the proposed approaches are applicable to any cross-modal association task, in this paper we will only consider audiovisual signals with synchronized audio and video. The performance of the proposed approaches are compared on a video analysis task, which is speaker detection when more than one face is present in the video.

There has been several efforts to associate talking heads in video with speech in the audio stream. Slaney and Covell [7] propose FaceSync as an optimal linear detector, which combines the information from all pixels to measure audio-visual synchronization. Fisher et al. [3] present a non-parametric approach to learn the joint distribution of audio and visual features. They first project the data into a maximally informative, low-dimensional subspace, and then model the stochastic relationships using a nonparametric density estimator. Li at al. [5] propose several cross-modal association approaches and compare their performances on retrieval and talking head analysis tasks. Their results show that the cross-modal factor analysis method they propose has the best performance in both tasks.

Cross-modality association problems have been examined for other modalities as well. Barnard et al. [1] propose learning models for the joint statistics of image components, such as segmented regions, and the keywords associated with the images. These models are then used for image retrieval and annotation.

The organization of this paper is as follows: In Section II a talking head detection system based on linear and nonlinear association models is proposed and various cross-modal association approaches are described. Section III presents the experimental results for speaker detection. Conclusions and further work suggestions are given in Section IV.

II. Speaker Detection Using Cross-Modal Association

Figure 1 shows the block diagram of the proposed talking head analysis system. To detect the current speaker, first candidate face regions are located using a face detection module. Facial features are then extracted from each face region. In parallel, audio features are extracted from the audio stream. We have used 12 Mel-Frequency Cepstral Coefficients (MFCCs) as the audio features. Audio classification is performed to determine the speech regions in the audio stream using the algorithm described in [6]. Cross-modal association between audio and video is performed only when speech is

detected in the audio. Finally, cross-modal association analysis is performed on the audio and video features using one of the approaches described in this section.

Given an audiovisual stream consisting of synchronized video and audio streams, we extract n visual features from each frame of the video stream. For the video frame at time $t = t_0$, we also extract m audio features from the audio stream in the window $t \in [t_0 - W/2, t_0 + W/2]$. In our experiments we have used a window size of $W = 22$ms. Assuming there are N frames in the video stream, the video and audio feature matrices may be written as

$$\begin{aligned} \mathbf{V} &= \{v_{i,j}\}, \quad i = 1, \ldots, N,\, j = 1, \ldots, n \\ \mathbf{A} &= \{a_{i,j}\}, \quad i = 1, \ldots, N,\, j = 1, \ldots, m\,. \end{aligned} \tag{1}$$

A. Approaches based on Linear Correlation Models

For many audiovisual data analysis applications, if the analysis time window is small, then the correlation between audio and video features may be approximated by a linear model [7], [5]. Under the linear correlation model, the problem is to find an optimal transformation that can best represent or identify the coupled patterns between audio and video features. Below, we propose three approaches based on the linear correlation model. The difference between the first two approaches is the definition of the optimality criterion used to derive the optimal transformation. For the remainder of the paper we assume that the matrices $\mathbf{V}$ and $\mathbf{A}$ have been processed to have zero mean.

In the *Cross-Modal Factor Analysis* (CFA) approach we seek to find the transformation matrices $\mathbf{X}$ and $\mathbf{Y}$ that minimizes the distance between the projected matrices for the coupled data in $\mathbf{V}$ and $\mathbf{A}$, that is,

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{VX} - \mathbf{AY}\|_F^2\,, \tag{2}$$

where $\mathbf{X}^T\mathbf{X} = \mathbf{Y}^T\mathbf{Y} = \mathbf{I}$, and $\|\cdot\|_F$ is the Frobenius norm defined as

$$\|\mathbf{M}\|_F = \left(\sum_i \sum_j |m_{ij}|^2 \right)^{1/2}.$$

It can be shown [4] that the solution to this minimization problem is given by the matrices

$$\begin{cases} \mathbf{X} = \mathbf{S}_{AV} \\ \mathbf{Y} = \mathbf{D}_{AV}, \end{cases} \tag{3}$$

where $\mathbf{V}^T\mathbf{A} = \mathbf{S}_{AV}\mathbf{\Sigma}_{AV}\mathbf{D}_{AV}$ is the singular value decomposition of $\mathbf{V}^T\mathbf{A}$. Using the optimal transformation matrices $\mathbf{X}$ and $\mathbf{Y}$, the transformed versions of the feature matrices $\mathbf{V}$ and $\mathbf{A}$ are given by

$$\begin{cases} \widetilde{\mathbf{V}} = \mathbf{VX} \\ \widetilde{\mathbf{A}} = \mathbf{AY}. \end{cases} \tag{4}$$

Corresponding vectors in $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{A}}$ matrices are thus optimized to represent the coupled relationships between the two feature subsets without being affected by distribution patterns within each subset. Traditional Pearson correlation or mutual information calculation [3], [5] can then be performed on vectors in $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{A}}$ that correspond to the largest k singular values in $\mathbf{\Sigma}_{AV}$. This reduces the dimension greatly while preserving the principal coupled patterns. Furthermore, noise due to feature variation within video and audio is reduced. In addition to feature dimension reduction and noise removal, feature selection capability is another feature of CFA. The magnitudes of the singular values reflect the significance of corresponding features.

Instead of the optimality criterion given in Equation 2, which minimizes the projected distance, one can choose to maximize the correlation between audio and video features after the transformation. We refer to this approach as *Canonical Correlation Analysis* (CCA). The optimization problem in this case can be stated as

$$\max_{\mathbf{X},\mathbf{Y}} \text{corr}(\mathbf{VX} - \mathbf{AY}), \tag{5}$$

The solution to this maximization problem is given by the matrices [8]

$$\begin{cases} \mathbf{X} = \mathbf{\Sigma}_{VV}^{-1/2}\mathbf{S}_K \\ \mathbf{Y} = \mathbf{\Sigma}_{AA}^{-1/2}\mathbf{D}_K, \end{cases} \tag{6}$$

where

$$\begin{aligned} \mathbf{\Sigma}_{VV} &= E\left\{\mathbf{V}^T\mathbf{V}\right\}, \\ \mathbf{\Sigma}_{AA} &= E\left\{\mathbf{A}^T\mathbf{A}\right\}, \\ \mathbf{\Sigma}_{VA} &= E\left\{\mathbf{V}^T\mathbf{A}\right\}, \text{ and} \\ K &= \mathbf{\Sigma}_{VV}^{-1/2}\mathbf{\Sigma}_{VA}\mathbf{\Sigma}_{AA}^{-1/2} = \mathbf{S}_K\mathbf{V}_K\mathbf{D}_K^T. \end{aligned}$$

The calculation of the inverse matrices in this formulation requires that no linear correlation exists between the vectors in $\mathbf{V}$ or $\mathbf{A}$. Large calculation errors could result even when two vectors are close to linear. This imposes some restrictions on the set of features that can be processed by CCA, especially when the analysis window has to be short to fit the linear model across modalities. As we will see later in our experiments, such restriction of CCA may affect its performance and limit its applications.

Other major differences between CCA and CFA include:

1) The transformations provided by CFA are orthogonal, while this is not necessary true for CCA.
2) CFA favors coupling patterns with high variations (i.e. large amplitude changes), while CCA is more sensitive to highly coupled, but low variation patterns. This is mainly due to the whitening of $\mathbf{V}$ and $\mathbf{A}$ in CCA by the multiplication with $\mathbf{\Sigma}_{VV}^{-1/2}$ or $\mathbf{\Sigma}_{AA}^{-1/2}$.

Rather than treating the audio and video features separately, as is done in CFA and CCA, one can combine them and perform dimensionality reduction on the combined feature matrix,

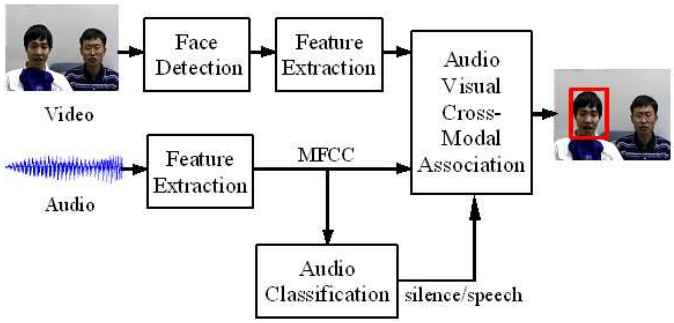


Fig. 1. Block diagram of the talking head analysis system.

which leads to the well-known *latent semantic indexing* (LSI) approach. In this case, we first define the combined feature matrix, $\mathbf{F}$ as

$$\mathbf{F} = [\mathbf{VA}]. \tag{7}$$

After normalizing the entries of $\mathbf{F}$ to the range $[-1, 1]$, singular value decomposition is performed on $\mathbf{F}$ and only the vectors corresponding to the largest k singular values are kept. For details about the LSI approach, refer to [5].

B. GMM-based Speaker Detection

Instead of using linear models, one can use a nonlinear model based on a mixture of Gaussian distributions to represent coupled patterns between audio and visual features. Similar to the LSI approach described above, the audio and feature matrices are combined into one feature matrix $\mathbf{F}$, as specified in Equation 7. Then, each row of this matrix, $\mathbf{f}_i = \{f_{ij}\}, i = 1, \ldots, N, j = 1, \ldots, m + n$ is modeled by *Gaussian mixture model* (GMM) as

$$p(\mathbf{f}_i|\mathbf{\Theta}) = \sum_{k=1}^{K} \alpha_k p_k(\mathbf{f}_i|\boldsymbol{\theta}_k), \tag{8}$$

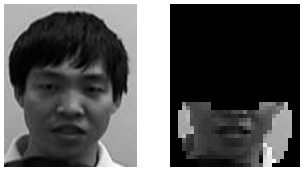
where $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ is the complete set of parameters specifying the model and α_k are the mixture probabilities, subject to the constraint $\sum_{k=1}^{K} \alpha_k = 1$.

The speaker is detected by first extracting the visual features of the face region and audio features from input video frame by frame. Then the joint likelihood of the audio-visual feature pairs are calculated according to the GMM distribution using the equation

$$p(\mathbf{F}|\mathbf{\Theta}) = \prod_{i=1}^{N} p(\mathbf{f}_i|\mathbf{\Theta}) \tag{9}$$

The candidate face with highest joint likelihood is then marked to be the speaker.

Unlike the linear model based approaches described in Section II-A, the GMM based approach cannot provide feature selection and dimensionality reduction. In our current implementation, principal component analysis (PCA) is used to reduce the dimensionality of visual feature space. We choose the most significant 15 eigenfaces as the visual features. Also in some experiments, a visual mask as shown in Figure 2 is used before applying PCA to better preserve speech-related facial features. We refer to this modification as the *masked GMM* approach.



(a) (b)

Fig. 2. (a) A face image and (b) the corresponding mask area used for the masked GMM experiments.

While prior off-line training is not necessary for linear model based approaches, the GMM has to be trained before hand using the Expectation Maximization (EM) algorithm. This process generates a GMM that best fits the joint distribution of visual and audio features given by the training data.

III. Experimental Results

The proposed association methods are tested and compared using video clips collected from 12 different individuals. The video clips are recorded on different days and in different locations to diversify the data. In the experiments, the real talking head competes either with other faces in the video or talking faces chosen from other image sequences in the collection. The total length of the experimental data set is about 20 minutes. Video clips from 4 individuals are used for training and the rest are used for testing. Table I compares the performance of different speaker detection methods. While the GMM-based methods always require prior off-line supervised training, the linear model based methods can also operate in a dynamic processing mode where the transformation matrices are generated on the fly using the input testing video directly. In this case, no prior training is needed and the results are generated based on the Pearson correlation, which will be low if there is no association between the facial features and audio features.

	LSI	CFS	CCA	GMM	GMM with mask
Dynamic Proc.	66.1%	80.4%	73.9%	–	–
Supervised Train.	53.6%	81.1%	70.8%	61.8%	65.2%

TABLE I
SPEAKER DETECTION ACCURACY OF VARIOUS ASSOCIATION APPROACHES.

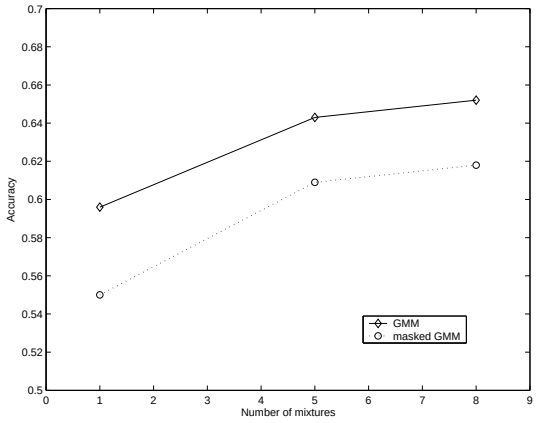


Fig. 3. Comparison of detection accuracy for GMM and masked GMM approaches as a function of the number of mixtures used.

In Figure 3 we show the change on the detection accuracy of the GMM-based methods, with or without the use of a visual mask, when different number of Gaussian mixtures are used. From this figure we observe that the use of a visual mask generally improves the detection accuracy by approximately 4%. Increasing the number of Gaussian mixtures does not significantly improve the accuracy. With a visual mask and 8 Gaussian mixtures, the GMM-based method reaches a detection accuracy of 65.2 percent.

In general, GMM and LSI perform at a comparable level, while CFA and CCA achieve much better accuracy. We believe this difference is mainly due to the optimized feature extraction strategies embedded with CFA and CCA. Although the GMM method uses a more powerful recognition model, its feature selection method is similar to that used by LSI, which leads to a low accuracy similar to that of LSI. The result demonstrate the limitation of traditional feature selection methods in cross-modal association problems. In cross-modal association, it is critical to reveal features that best represent the association between features from different modalities and at the same time be able to remove the intra-modality variations. This is actually a challenging task due to the large amount of intra-modality distribution variations. PCA and many other commonly-used feature selection and dimensionality reduction methods can hardly serve such a purpose. Like CFA and CCA, an effective cross-modal feature selection and dimensionality reduction strategy has to be designed by jointly considering features from both modalities.

IV. CONCLUSIONS

In this paper we described a speaker detection system using cross-modal association methods. Linear and non-linear association models were used to design several cross-modal association approaches. A series of experiments were conducted to test and compare the effectiveness of different methods.

Due to their optimized feature selection capability, CFA and CCA greatly outperform GMM and LSI, with a detection accuracy of over 80%. The experimental results demonstrate the importance of effective feature selection in cross-modal association problems. Like CFA and CCA, an effective cross-modal feature extraction method has to be based on jointly considering features from both modalities in order to remove the noise due to large intra-modality distribution variations and at the same time preserve the association between different modalities.

Our results demonstrate the ineffectiveness of traditional multimedia analysis system architectures, where audio and video features are processed separately. As pointed out in [2] biological evidence suggests the need of a cross-modal processing architecture that can better integrate perceptual processes at all levels in a multimodal analysis system.

Compared to linear situations, it is more difficult to find optimized feature extraction solutions for non-linear models. Solutions in this direction, however, provide more powerful tools that can be used in a wide range of applications. In addition to cross-modal information retrieval and speaker detection, we will further exam the problem of audiovisual compression, facial animation, and audiovisual speech recognition.

REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, February 2003.

[2] M. H. Coen. Multimodal integration a biological view. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 1417–1424, Seattle, WA, August 4–10 2001.

[3] J. W. F. III, T. Darrell, W. T. Freeman, and P. A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Proceedings of Advances in Neural Information Processing Systems 13*, pages 772–778, Denver, CO, Nov 27–Dec 2 2000.

[4] W. Krzanowski. *Principles of Multivariate Analysis*. Oxford University Press, New York, NY, 2000.

[5] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *Proceedings of the ACM Multimedia Conference*, volume 2, pages 604–611, Berkeley, CA, November 2–8 2003.

[6] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, April 2001.

[7] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Proceedings of Advances in Neural Information Processing Systems 13*, pages 814–820, Denver, CO, Nov 27–Dec 2 2000.

[8] B. G. Tabachnick and L. S. Fidell. *Using Multivariate Statistics*. Allyn and Bacon Press, Boston, MA, 2000.