Crossmodal Matching of Speakers using Lip and Voice Features in Temporally Non-overlapping Audio and Video Streams

Anindya Roy[1,2] and Sébastien Marcel[2]
[1]*Idiap Research Institute, Martigny, Switzerland*
[2]*Ecole Polytechnique Fédérale de Lausanne, Switzerland*
Email: {aroy, marcel}@idiap.ch

***Abstract*—Person identification using audio (speech) and visual (facial appearance, static or dynamic) modalities, either independently or jointly, is a thoroughly investigated problem in pattern recognition. In this work, we explore a novel task : person identification in a cross-modal scenario, i.e., matching the speaker in an audio recording to the same speaker in a video recording, where the two recordings have been made during different sessions, using speaker specific information which is common to both the audio and video modalities. Several recent psychological studies have shown how humans can indeed perform this task with an accuracy significantly higher than chance. Here we propose two systems which can solve this task comparably well, using purely pattern recognition techniques. We hypothesize that such systems could be put to practical use in multimodal biometric and surveillance systems.**

***Keywords*-Multi-modal biometrics, audio-visual speaker recognition, crossmodal matching, audio and video classification.**

I. Introduction

We often create a mental image of a person whose voice is familiar (from telephone conversations, for example) but whom we have never seen. We often also create a mental "voice model" from visual information (either static or dynamic) of persons we have never heard. Recent studies have investigated these phenomena scientifically [1] [2] [3] [4], asking human observers to match an audio recording of an unknown voice X to two video recordings of two unknown speakers, A and B, one of which is X, and vice versa. It was found that humans performed in this task with an accuracy significantly above chance. Let us define this crossmodal matching task, termed as the XAB task [1], as follows.

The XAB task has two stages : (1) the learning stage and (2) the matching stage. In the learning stage, joint audio and video information is available in the form of synchronized audio and video (dynamic facial appearance) recordings of persons speaking. The purpose of this stage is to extract or store knowledge required to map speaker identities between audio and video modalities. In the matching stage, there are two cases, the Audio-to-Video (a-v) matching task and the Video-to-Audio (v-a) matching task. In the a-v task, an audio recording of a person X speaking, and two video recordings showing two different persons speaking, A and B, are provided. Given that exactly one out of A and B is X, the task is to decide which one it is. For all the speakers in the matching stage, it is critical that no joint (synchronized) audio and video information be available. We term this the Audio-Video Mismatch criterion. This causes the XAB task to be distinct from a simple audio-to-video synchronization task where both modalities capture the same event in time [5]. To ensure this, the audio and video recordings in the matching stage should be temporally non-overlapping, i.e., they should be made during different sessions, and speakers in the matching stage should be all distinct from speakers in the learning stage. The converse v-a task is exactly the same as the a-v task with the roles of the modalities reversed.

There are several studies with human observers performing the XAB task.[1] Lachs et al. [2] and Kamachi et al. [1] reported human observers correctly matching X to A or B around 65% of the times. Krauss et al. have shown similar matching performance using static instead of dynamic visual information [4]. Campanella et al. [3] provide additional insights on cross-modal information transfer in humans.

In this preliminary work, we explore a possible solution to the XAB task by creating modality independent speaker models which can be used equally on both audio and video data. We study two approaches, the K-means clustering approach and the K-nearest neighbour approach. Our methods have shown reasonable results which compare well with that shown by human observers.

The rest of the paper is organized as follows. In Sec.II, we describe the proposed speaker matching framework. We describe our experiments in Sec.III. In Sec.IV, we discuss the results and highlight certain aspects of our method. Finally, Sec.V outlines the main conclusions of our work.

II. The Proposed Framework

A. Feature Extraction

For the video modality, we concentrated on lip appearance features since they have been shown to be robust and efficient [6]. The video frame rate was 25fps. From each

[1]For humans, the learning stage comprises of all speech-related joint audio-visual stimuli received as part of normal day-to-day activities prior to the experiments.

video frame, a 16×16 Region-Of-Interest (ROI) around the lips was extracted using available annotation, followed by geometric normalization and inter-frame alignment. Next, 2D-DCT features [6] were extracted and 3rd to 10th highest energy coefficients were retained [2] to form the video feature vectors. Mean normalization was performed for each video sequence [6]. For the audio modality, the audio data sampled at 8kHz was blocked into frames equal in duration to the video frames (corresponding to 320 samples per frame) and 16 Mel-Frequency Cepstral Coefficients (MFCC) [6] were extracted from each block, out of which 1st to 8th were retained [2] to form the audio feature vectors. For each audio sequence, Cepstral Mean Subtraction [6] was performed. It is to be noted that only voiced frames were used, both for audio and video modalities.

B. Cross-modal Learning and Matching

Let $\mathbf{R}^a$ and $\mathbf{R}^v$ denote the audio and video feature spaces. For the learning stage, synchronized audio and video data is available. Let $\mathbf{S}^a$ and $\mathbf{S}^v$ denote the sets of audio and video feature vectors extracted from this data, i.e. $\mathbf{S}^a \subset \mathbf{R}^a$, $\mathbf{S}^v \subset \mathbf{R}^v$. These sets, termed the audio and video learning sets, are ordered such that the i-th element $\mathbf{x}_i^a \in \mathbf{S}^a$ is synchronous to the i-th element, $\mathbf{x}_i^v \in \mathbf{S}^v$. For the matching stage, let X, A and B also denote the respective recordings as well as the persons X, A and B. Let $\mathbf{S}_{\mathrm{X}}^m, \mathbf{S}_{\mathrm{A}}^m, \mathbf{S}_{\mathrm{B}}^m$ denote the feature vectors extracted from X, A and B, where m can indicate either the audio (a) or the video (v) modality depending on whether it is an (a-v) or (v-a) task. Let $|\cdot|$ denote the size of a countable set, and $\mathbf{1}_{\mathbf{S}}(\mathbf{x})$, the indicator function of any set $\mathbf{S}$, i.e. $\mathbf{1}_{\mathbf{S}}(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathbf{S}$ and is zero otherwise.

1) K-means Clustering (KMC) Approach: In the learning stage, the learning sets $\mathbf{S}^a$ and $\mathbf{S}^v$ are independently clustered into K clusters, $\{\mathbf{S}_k^a\}_{k=1}^K$ and $\{\mathbf{S}_k^v\}_{k=1}^K$, using K-means algorithm [7] with squared-Euclidean distance. Let $\{\mathbf{R}_k^a\}_{k=1}^K$ and $\{\mathbf{R}_k^v\}_{k=1}^K$ denote the corresponding Voronoi cells formed by segmenting the spaces $\mathbf{R}^a$ and $\mathbf{R}^v$ according to these clusters, i.e, $\mathbf{S}_k^a \subset \mathbf{R}_k^a$, $\mathbf{S}_k^v \subset \mathbf{R}_k^v$ for $1 \leq k \leq K$. Let $\mathbf{H}^{va}$ denote the $K \times K$ Hebbian projection matrix [8], each of whose elements $\mathbf{H}^{va}(k_a, k_v)$ estimates the probability that an audio vector $\mathbf{x}^a$ belongs to a particular cell $\mathbf{R}_{k_a}^a$ in the audio feature space, given that its synchronous video vector $\mathbf{x}^v$ belongs to the cell $\mathbf{R}_{k_v}^v$ in the video feature space, i.e. $\mathbf{H}^{va}(k_a, k_v) = \Pr(\mathbf{x}^a \in \mathbf{R}_{k_a}^a | \mathbf{x}^v \in \mathbf{R}_{k_v}^v)$. It is estimated as

$$\mathbf{H}^{va}(k_a, k_v) = \frac{1}{|\mathbf{S}_{k_v}^v|} \sum_{\mathbf{x}^v \in \mathbf{S}_{k_v}^v} \mathbf{1}_{\mathbf{S}_{k_a}^a}(\mathbf{x}^a) \tag{1}$$

where $1 \leq k_a, k_v \leq K$, $\mathbf{x}^a$ is the audio vector synchronous with video vector $\mathbf{x}^v$ and $|\cdot|$ denotes the size of a countable set. The inverse Hebbian projection, $\mathbf{H}^{av}$ can be calculated as in Eqn. 1 by interchanging the audio and video modalities. The matrices $\mathbf{H}^{av}$ and $\mathbf{H}^{va}$ are the outputs of the learning stage.

For the matching stage, let us consider the (a-v) task. Let $\mathbf{p}_{\mathrm{X}}^a$,$\mathbf{p}_{\mathrm{A}}^v$ and $\mathbf{p}_{\mathrm{B}}^v$ be the probability mass functions (PMF) of the feature vectors extracted from X, A and B, i.e. $\mathbf{S}_{\mathrm{X}}^a, \mathbf{S}_{\mathrm{A}}^v$ and $\mathbf{S}_{\mathrm{B}}^v$ respectively, based on the K clusters formed in the learning stage. Thus, $\mathbf{p}_{\mathrm{X}}^a(k) = \Pr(\mathbf{x}^a \in \mathbf{R}_k^a | \mathbf{x}^a \in \mathbf{S}_{\mathrm{X}}^a)$, $\mathbf{p}_{\mathrm{A}}^v(k) = \Pr(\mathbf{x}^v \in \mathbf{R}_k^v | \mathbf{x}^v \in \mathbf{S}_{\mathrm{A}}^v)$ and $\mathbf{p}_{\mathrm{B}}^v(k) = \Pr(\mathbf{x}^v \in \mathbf{R}_k^v | \mathbf{x}^v \in \mathbf{S}_{\mathrm{B}}^v)$. These PMFs are estimated as,

$$\mathbf{p}_{\mathrm{X}}^a(k) = \frac{1}{|\mathbf{S}_{\mathrm{X}}^a|} \sum_{\mathbf{x}^a \in \mathbf{S}_{\mathrm{X}}^a} \mathbf{1}_{\mathbf{R}_k^a}(\mathbf{x}^a) \tag{2}$$

$$\mathbf{p}_{\mathrm{A}}^v(k) = \frac{1}{|\mathbf{S}_{\mathrm{A}}^v|} \sum_{\mathbf{x}^v \in \mathbf{S}_{\mathrm{A}}^v} \mathbf{1}_{\mathbf{R}_k^v}(\mathbf{x}^v) \tag{3}$$

$$\mathbf{p}_{\mathrm{B}}^v(k) = \frac{1}{|\mathbf{S}_{\mathrm{B}}^v|} \sum_{\mathbf{x}^v \in \mathbf{S}_{\mathrm{B}}^v} \mathbf{1}_{\mathbf{R}_k^v}(\mathbf{x}^v) \tag{4}$$

where $1 \leq k \leq K$. Next, we use the Hebbian projection matrix, $\mathbf{H}^{va}$ to project the two PMFs in the video space, $\mathbf{p}_{\mathrm{A}}^v, \mathbf{p}_{\mathrm{B}}^v$ to the audio space, as follows,

$$\tilde{\mathbf{p}}_{\mathrm{A}}^a = \mathbf{H}^{va} \mathbf{p}_{\mathrm{A}}^v \tag{5}$$

$$\tilde{\mathbf{p}}_{\mathrm{B}}^a = \mathbf{H}^{va} \mathbf{p}_{\mathrm{B}}^v \tag{6}$$

These two PMFs (which we term as pseudo-PMFs) are used to approximate the true PMFs of the unavailable audio feature vectors corresponding to the video-only recordings A and B [8]. For the matching task, we consider these PMFs as speaker specific models and decide,

$$\mathrm{X} \equiv \begin{cases} \mathrm{A} & \text{if } \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{A}}^a) \geq \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{B}}^a), \\ \mathrm{B} & \text{if } \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{A}}^a) < \rho_{\mathrm{B}}(\mathbf{p}_{\mathrm{X}}^a, \tilde{\mathbf{p}}_{\mathrm{B}}^a) \end{cases} \tag{7}$$

where ρ_B denotes the Bhattacharyya coefficient [7] between two PMFs $\mathbf{p}_1, \mathbf{p}_2$ and is calculated as, $\rho_B(\mathbf{p}_1, \mathbf{p}_2) = \sum_{\forall k} \mathbf{p}_1(k)^{\frac{1}{2}} \mathbf{p}_2(k)^{\frac{1}{2}}$. For the (v-a) task, a similar procedure was followed, interchanging the roles of the audio and video modalities.

2) K-Nearest Neighbours (KNN) Approach: There is no separate learning stage in this approach. Information in the audio and video learning sets $\mathbf{S}^a, \mathbf{S}^v$ (ref. Sec.II-B) is directly used in the matching stage. For the matching stage, let us again consider the (a-v) task. For each audio vector $\mathbf{x}_{\mathrm{X},i}^a \in \mathbf{S}_{\mathrm{X}}^a$ extracted from X, we form the set $\mathbf{\Psi}_{X,i}$ of the indices of K_a-nearest neighbours [7] of $\mathbf{x}_{\mathrm{X},i}^a$ in $\mathbf{S}^a$, the audio learning set. Similarly, we form sets of indices of K_v-nearest neighbours $\{\mathbf{\Psi}_{\mathrm{A},i}\}, \{\mathbf{\Psi}_{\mathrm{B},i}\}$ for each vector in $\mathbf{S}_{\mathrm{A}}^v, \mathbf{S}_{\mathrm{B}}^v$, the video vectors extracted from A and B respectively, from $\mathbf{S}^v$, the video learning set. These nearest neighbour sets are independent of modalities since each element in $\mathbf{S}^v$ has a corresponding element in $\mathbf{S}^a$ (ref. Sec.II-B). This forms the basis of the cross-modal mapping in this approach. To match X to A or B, we use the sum of the sizes of intersections $\mathbf{s}_I$

[2]These coefficients have been selected by trial-and-error to give best performance.

between the nearest neighbour sets of X and those of A,B, as follows,

$$\mathrm{X} \equiv \begin{cases} \mathrm{A} & \text{if } \mathbf{s}_I(\mathrm{X},\mathrm{A}) \geq \mathbf{s}_I(\mathrm{X},\mathrm{B}), \\ \mathrm{B} & \mathbf{s}_I(\mathrm{X},\mathrm{A}) < \mathbf{s}_I(\mathrm{X},\mathrm{B}) \end{cases} \quad (8)$$

where $\mathbf{s}_I(\mathrm{X},\mathrm{A}), \mathbf{s}_I(\mathrm{X},\mathrm{B})$ are defined as follows,

$$\mathbf{s}_I(\mathrm{X},\mathrm{A}) = \frac{1}{|\mathbf{S}^a_{\mathrm{X}}||\mathbf{S}^v_{\mathrm{A}}|} \sum_{\mathbf{x}^a_{\mathrm{X},i} \in \mathbf{S}^a_{\mathrm{X}}} \sum_{\mathbf{x}^v_{\mathrm{A},j} \in \mathbf{S}^v_{\mathrm{A}}} |\mathbf{\Psi}_{\mathrm{X},i} \cap \mathbf{\Psi}_{\mathrm{A},j}| \quad (9)$$

$$\mathbf{s}_I(\mathrm{X},\mathrm{B}) = \frac{1}{|\mathbf{S}^a_{\mathrm{X}}||\mathbf{S}^v_{\mathrm{B}}|} \sum_{\mathbf{x}^a_{\mathrm{X},i} \in \mathbf{S}^a_{\mathrm{X}}} \sum_{\mathbf{x}^v_{\mathrm{B},j} \in \mathbf{S}^v_{\mathrm{B}}} |\mathbf{\Psi}_{\mathrm{X},i} \cap \mathbf{\Psi}_{\mathrm{B},j}| \quad (10)$$

For the (v-a) task, a similar procedure was followed, interchanging the role of the audio and video modalities. It can be shown that the sums $\mathbf{s}_I(\mathrm{X},\mathrm{A}), \mathbf{s}_I(\mathrm{X},\mathrm{B})$ can be equivalently expressed as approximations to the $\mathbf{L}^2$-inner product of the PMFs corresponding to the audio and video data. However, compared to Sec.II-B1, the feature space is now subdivided much more minutely, each vector in the learning sets $\mathbf{S}^a, \mathbf{S}^v$ forming its own cell. This amounts to exploiting maximally the information available for cross-modal matching. Our proposed matching criterion based on comparing the $\mathbf{s}_I$ values is motivated by the use of the $\mathbf{L}^2$ inner product kernel in state-of-the-art speaker verification systems [9].

III. Experiments

All experiments were performed on the M2VTS audio-visual database [10] with 24 male and 10 female speakers. Synchronized audio and video data was recorded in a controlled environment across multiple sessions separated by one week intervals. Lip annotations were obtained from http://www.ee.surrey.ac.uk/Projects/M2VTS/experiments/lip_tracking/. We tested our approach on two conditions : (1) lexically matched and (2) lexically mismatched. For condition (1), speech content in X, A and B were lexically matched. Recordings from the database were used as it is : in each recording, the speaker counted from '0' to '9' in their native language. For the second (more difficult) condition, the recordings were rearranged so that segments used for X were lexically mismatched with A and B : if X contained '0' to '4', A and B contained '5' to '9' and vice-versa. Ofcourse, the Audio-Video Mismatch criterion (ref. Sec.I) was always maintained in both conditions. X, A and B consisted of around 4.5 seconds of data each. Separate experiments were performed on only male (M), only female (F) and both male and female (F+M) speakers. For each XAB task, two speakers were separated from the complete set, these two were used in the matching stage, while all the remaining speakers were used in the learning stage. For one complete experiment, the XAB task was repeated for all possible pairs of speakers in the matching stage. Considering all possible combinations, the total number of times the XAB task (a-v and v-a each) was independently evaluated is 2208 for the M case, 360 for the F case and 4488 for the F+M case. The match score for each experiment is calculated as,

$$\text{Match score} = \frac{\text{No. of succesful matches}}{\text{Total no. of XAB tasks}} \times 100\% \quad (11)$$

Since each task has two alternatives only one out of which is correct, the expected score for a random classifier would be 50%. Each experiment was repeated for different values of K, the number of clusters, and K_a, K_v, the number of nearest neighbours, for the KMC and KNN approaches respectively. Optimal value of K was 64, while for K_a, K_v it varied from 2 to 256 according to the conditions tested. Table 1 gives the results of our experiments in terms of the match scores obtained, using the optimal parameter values.

Proposed Approach	XAB task type		Lex. matched	Lex. mismatched
KMC	a-v	M	66.6	*
		F	79.4	*
		F+M	66.4	*
	v-a	M	65.1	*
		F	60.0	*
		F+M	64.9	*
KNN	a-v	M	68.9	56.0
		F	64.2	57.8
		F+M	66.4	56.6
	v-a	M	66.0	55.6
		F	61.9	60.6
		F+M	63.4	56.1

Table I
MATCH SCORES (%) FOR THE XAB TASK USING THE PROPOSED APPROACHES. AN ASTERISK (*) DENOTES THAT A MATCH SCORE BETTER THAN RANDOM CHANCE (50%) COULD NOT BE OBTAINED.

	XAB task type	Lex. matched	Lex. mismatched
Kamachi et al. [1]	a-v	69.0	59.0
	v-a	66.0	60.0
Lachs et al. [2]	a-v	60.7	n.a.
	v-a	65.1	n.a.

Table II
MATCH SCORES (%) FOR THE XAB TASK PERFORMED BY HUMAN OBSERVERS.

IV. Discussions

For the lexically matched case, both the KMC and KNN approaches give match scores around 65%. This is statistically significant, given the total number of times the XAB task was evaluated (ref. Sec.III). For the lexically mismatched case, the performance of KNN drops by 10% but is still significant; KMC is unable to perform at more than chance level. This shows the relative robustness of the KNN approach. Our method compares well with results reported by studies using human observers on the XAB task [1] [2] as shown in Table 2, although it is to be noted that

these studies used different databases. It is to be noted that human performance fell drastically for time-reversed stimuli [1] [2]; our method is unaffected by this, being based on static feature vectors only. Furthermore, human observers had information from the entire face available to them, while our method uses information exclusively from the lip region.

In future, we aim to develop our method further, using this preliminary study as a basis, and improve the match scores so that it can be used in practical applications, such as (1) a cross-modal surveillance scenario where prior speech data (but no visual data, for example via telephone conversations) about a person X has been collected and presently it is required to identify this person out of a group which is under video surveillance (but no audio data is currently available, for example due to distance from group or noisy environment), and (2) a multimodal biometric system which uses cross-modalities (a-v, v-a) to augment the normal audio and video modalities and make it more reliable.

V. Conclusion

In this work, we explored a novel pattern recognition task: crossmodal person identification, where the identity of a speaker X in an audio recording is matched with one of two speakers A and B in two video recordings, and vice-versa. The recordings are temporally non-overlapping. The basis of our idea is to form modality independent speaker models which can be used on either audio or video data independently. We have proposed two approaches, the K-nearest neighbour approach and the K-means clustering approach, both of which have shown performance better than chance.

Acknowledgment

The authors would like to thank the Swiss National Science Foundation, projects MultiModal Interaction and MultiMedia Data Mining (MULTI, 200020-122062) and Interactive Multimodal Information Management (IM2, 51NF40-111401) and the FP7 European MOBIO project (IST-214324) for their financial support. The authors would also like to thank Francesco Orabona, Mathew Magimai.-Doss and the anonymous reviewers for their helpful comments and advice.

References

[1] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "'Putting the Face to the Voice': Matching Identity across Modality," *Current Biology*, vol. 13, pp. 1709–1714, 2003.

[2] L. Lachs and P. Pisoni, "Crossmodal source identification in speech perception," *Ecological Psychology*, vol. 16, no. 3, pp. 159–187, 2004.

[3] S. Campanella and P. Bellin, "Integrating face and voice in person perception," *Trends in Cognitive Science*, vol. 11, no. 12, 2007.

[4] R. Krauss, R. Freyberg, and E. Morsella, "Inferring speakers' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, pp. 618–625, 2002.

[5] K. Kumar and et al., "Audio-Visual Speech Synchronization Detection Using a Bimodal Linear Prediction Model," in *CVPR*, 2009.

[6] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. MIT Press, 2004.

[7] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. John Wiley and Sons, 2000.

[8] M. Coen, "Multimodal Dynamics : Self-Supervised Learning in Perceptual and Motor Systems," Massachusetts Institute of Technology, PhD Thesis, 2006.

[9] W. Campbell, D. Sturim, and D. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, 2006.

[10] "M2VTS Multimodal Face Database, Release 1.00," http://www.tele.ucl.ac.be/PROJECTS/M2VTS/m2fdb.html.