Modelling and Analyzing Multimodal Dyadic Interactions Using Social Networks

Sergio Escalera, Petia Radeva, Jordi Vitrià
Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona
Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain
Computer Vision Center, Universitat Autònoma de Barcelona
Campus UAB, Edifici 0, 08193, Bellaterra, Barcelona, Spain
{sergio,petia,jordi}@maia.ub.es

Xavier Baró
Department of Computer Science, Multimedia, and Telecomunications, Universitat Oberta de Catalunya
Rambla del Poblenou 156, 08018, Barcelona, Spain
Computer Vision Center, Universitat Autònoma de Barcelona
Campus UAB, Edifici 0, 08193, Bellaterra, Barcelona, Spain
xbaro@uoc.edu

Bogdan Raducanu
Computer Vision Center, Universitat Autònoma de Barcelona
Campus UAB, Edifici 0, 08193, Bellaterra, Barcelona, Spain
bogdan@cvc.uab.es

ABSTRACT

Social network analysis became a common technique used to model and quantify the properties of social interactions. In this paper, we propose an integrated framework to explore the characteristics of a social network extracted from multimodal dyadic interactions. First, speech detection is performed through an audio/visual fusion scheme based on stacked sequential learning. In the audio domain, speech is detected through clusterization of audio features. Clusters are modelled by means of an One-state Hidden Markov Model containing a diagonal covariance Gaussian Mixture Model. In the visual domain, speech detection is performed through differential-based feature extraction from the segmented mouth region, and a dynamic programming matching procedure. Second, in order to model the dyadic interactions, we employed the Influence Model whose states encode the previous integrated audio/visual data. Third, the social network is extracted based on the estimated influences. For our study, we used a set of videos belonging to New York Times' Blogging Heads opinion blog. The results are reported both in terms of accuracy of the audio/visual data fusion and centrality measures used to characterize the social network.

Categories and Subject Descriptors

[**Social Signal Processing**]: Social Networks

General Terms

Application

Keywords

Social interaction, multimodal fusion, influence model, social network analysis

ICMI-MLMI'10, November 8-12, 2010, Beijing, China.

1. INTRODUCTION

Social interactions play an important role in people's daily lives. Either if they are face-to-face or electronic (via e-mails, SMS, online communities, etc.), they represent the main communication channel people use to strengthen their inter-personal ties. As stated in [16], social interactions can be applicable in two main fields. One of them comes from linguistics and addresses the problem of social interactions from the perspective of dialog understanding. The other one comes from nonverbal communication. Within this latter framework, nonverbal communication is used in order to get hints about personal behavior. Facial expression, gaze, voice prosody, or body gestures provide powerful cues to display and perceive engagement, persuasion, mirroring, or status, just to mention a few. At the same time, social signals provide powerful hints regarding the human behavior, personality, hierarchical position in a group and to predict the outcome of the interaction process.

Social Network Analysis (SNA) [18] has its roots in social psychology and has been developed as a tool to model the social interactions in terms of a graph-based structure. 'Nodes' represent the 'actors' who make the subject of the analysis (persons, communities, institutions, corporate data) and the 'links' represent the specific type of interdependencies (friendship, familiarity, knowledge exchange, financial transactions) that exist between 'nodes'. SNA uncovers the implicit relationships between 'actors' and provides understanding of the underlying social processes and behaviors. For instance, we can identify central nodes, which can have roles such as leaders or serving as hubs.

In [6], authors used SNA for the purpose of analyzing the structure of online hate group blogs. In order to build the social network, a rule-based algorithm was applied. In [7], authors examine how communities can be discovered through interconnected blogs as a form of social hypertext. From the model created, they measure community in the blogs by aligning centrality measures from SNA with measures of sense of community obtained using behavioral surveys. Regarding the application of SNA to workgroups, some results are reported in [3]. The model proposed allowed not only the study of direct effects of functional role, status, and

communication role on individual performance, but also indirect effects through individual centrality. In [17], authors offer a different application of SNA to workgroups: speaker role recognition in meetings. The study of SNA-based role identification has also been reported in [19]. In this work, the authors build a social network by manually annotating the interactions between different players in a movie. The objective is to identify those actors holding a leading role and the groups created around them. Based on the network topology, they could infer the movie's storyline.

Our current paper presents an integrated framework for extraction and analysis of a social network from multimodal dyadic interactions:

• First, speech detection is performed through an audio/visual fusion scheme. Audio cues are the primary source of information that can provide useful evidence for speech production. However, they cannot easily distinguish a user who is speaking from the others in a noisy environment, where several speakers talk simultaneously. Alternatively, visual cues can be useful in deciding whether somebody is the active talking person by analyzing his/her lips movement. However, visual cues alone cannot easily distinguish a speaker from an active listener, who may be just smiling or nodding without saying anything. For this reason, by combining both audio and visual cues, we expect to obtain an increased robustness in the speaker diarization process. In the audio domain, speech is detected through clusterization of audio features. Clusters are modelled by means of an One-state Hidden Markov Model (HMM) containing a diagonal covariance Gaussian Mixture Model (GMM). In the visual domain, speech detection is performed through differential-based feature extraction from the segmented mouth region, and a dynamic programming matching procedure. Finally, both cues are fused based on stacked sequential learning [9]. The basic idea of stacked sequential learning is to create an extended data set that joins the original training data features with the predicted labels considering a neighborhood around the example.

• Second, in order to model the relationship in the dyadic interactions, we employed the Influence Model [5] whose states encode the integrated audio/visual data.

• Third, the social network is represented as a directed graph, whose links are estimated based on the influence one person has over the other. The dyadic interactions which are the object of our study belong to publicly available New York Times' Blogging Heads opinion blog [1]. The whole process is depicted in figure 1.

An important remark needs to be made at this point. The use of the term 'influence' in the context of the present paper was inspired by previous work. Choudhury [8] was able to show how turn-taking in face-to-face conversation could be a measure of 'influence', which was shown to have an extremely high correlation with one measure of the social novelty (and thus presumably the interestingness) of the information being presented.

The paper is structured as follows. Section 2 presents our approach for audio/visual feature extraction and fusion, based on the particularities of our problem. Section 3 presents the extraction of the social network, including a brief recall of the Influence Model. In section 4, we present the experimental results both in terms of accuracy of the audio/visual data fusion and centrality measures used to characterize the social network. Finally, section 5 concludes the paper.

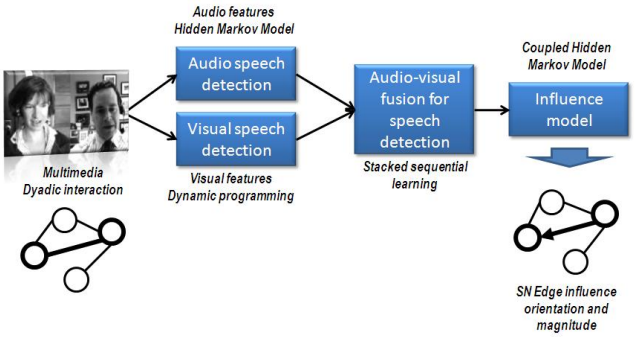


Figure 1: Block diagram of our integrated framework for Social Network extraction and analysis.

2. AUDIO-VISUAL CUES EXTRACTION AND FUSION

In this section, we describe the audio/video feature extraction and fusion methodologies.

2.1 Audio Cue

In order to obtain the audio structure, we use a diarization scheme based on the approach presented in [11]. According to this, the mono audio input channel is segmented into a set of different blocks according to their speech characteristics. In this section, we introduce the audio processing and the diarization methodology.

Classical diarization approaches in the literature are only based on acoustic features [15], and no prior information of the number of speakers or their voices is necessary. The final result of the diarization process is a segmentation of the input audio in several clusters where the speaker characteristics are homogeneous. Those clusters must be assigned later using the visual cue to each one of the speakers.

2.1.1 Speech Feature extraction

The speech signal in the audio domain consists of tones with different frequencies. The first step in order to process an audio stream is to create a good representation of this signal. Classical state-of-the-art representation for audio is based on Mel Frequency Cepstrum Coefficients (MFCC), which used to outperform other feature representation techniques. In addition to the MFCC coefficients, the dynamic features *delta* and *delta-delta*, which correspond to the first and second time-derivatives of cepstral coefficients are added to enrich the descriptor.

MFCC extraction: The input audio is analyzed using a sliding-window of 25ms, with an overlap of 10ms between two consecutive windows, and each window is processed using a short-time Discrete Fourier Transform. As common in speech recognition, frequencies are mapped to the mel scale, using the classical correspondence:

$$mel(f) = 2595 \log_{10}\left(\frac{f}{700}+1\right) = 1127 \log_{e}\left(\frac{f}{700}+1\right) \quad (1)$$

where f is a frequency in Hz. Finally, the log mel frequency is converted back to time using the Discrete

Cosine Transform (DCT), obtaining as many MFCC coefficients as necessary. We extract the first 12 components C_i:

$$C_i = \sum_{k=1}^{K} (\log S_k) \left[i \left(k - \frac{1}{2} \right) \frac{\pi}{K} \right] \quad (2)$$

for $i = 1, \ldots, 12$, excluding the C_0 coefficient, which corresponds to the mean value of the input signal. The value S_k corresponds to the k component of the FFT mapped to the mel scale.

Energy Computation: In the literature, it is common to complement the MFCC with the normalized log of the raw signal energy, used as the energy coefficient. The energy is computed as the logarithm of the signal energy:

$$E = \log \sum_{n=1}^{N} S^2(n) \quad (3)$$

Δ MFCC and Δ^2 MFCC computation: Finally, in order to improve the audio description, the speech spectrum is extended by the temporal cepstral derivative. The first derivative (Δ MFCC) is computed by the equation:

$$\Delta C_m(t) = \frac{\partial C_m(t)}{\partial t} \approx \mu \sum_{k=-K}^{K} C_m(t+k) \quad (4)$$

where μ and K are proper normalization constants. The approximation of the second derivative is computed as:

$$\Delta^2 C_m(t) = \Delta C_m(t+1) - \Delta C_m(t) \quad (5)$$

2.1.2 Speaker segmentation

Once the audio data is properly codified by means of the extended feature set, an audio-segmentation procedure based on Bayesian Information Criterion (BIC), following the steps proposed in [11], is performed.

First, a coarse segmentation is generated according to a Generalized Likelihood Ratio, computed over two consecutive windows of 2.5s. Each block is represented using a full covariance Gaussian over the extracted features. This process produces an over segmentation of the audio into homogeneous small blocks.

The second step consists of a hierarchical clustering over the segments. Initially, each segment is considered a cluster and at each iteration of the clustering process, the two most similar clusters are merged, until the BIC stopping criterion is met. As in the previous step, each cluster is modelled by means of a full covariance Gaussian. A first pass of clustering is performed only over adjacent windows in order to minimize the clustering time.

Finally, a Viterbi decoding is performed in order to adjust the segments boundaries. Clusters are modelled by means of an One-state HMM containing a diagonal covariance GMM of 8 components.

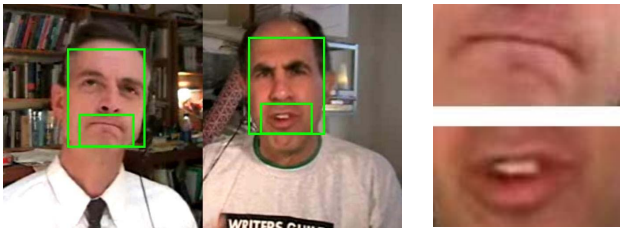
2.2 Visual Cue

Visual cue extraction refers to the segmentation of the mouth region and classification of the mouth-area appearance in speaking and non-speaking patterns.

2.2.1 Feature extraction

The data we are working with consists of dyadic conversations in which the speakers are in near-frontal view with respect to the camera. For this kind of scenarios, we first apply a face detector (as the one proposed by Viola & Jones [14]) in order to limit the search area for the mouth. Once we have localized the face, we search for the mouth region in the lower part of the segmented image. An example of a detected face, together with its corresponding mouth region, is shown in Figure 2(a) and Figure 2(b), respectively.

The next step consists in the extraction of discriminative features from the mouth region. In this case, the Histogram of Oriented Gradients (HOG) [10] is one of the preferred descriptors because of its tolerance to illumination changes and noise.



(a) (b)

Figure 2: (a) Face and mouth detection and (b) segmented mouth regions.

2.2.2 Speech classification

Once we have segmented and described the mouth regions over time, we define the visual speaker diarization as an one-class classification problem. In other words, we are interested in modelling the non-speaking patterns, since they present a smaller intra-class variability, and then classify those temporal patterns that are far away in terms of distance as speaking patterns. For this purpose we take advantage of the dynamic programming paradigm to match temporal series, and in particular, we use the Dynamic Time Warping (DTW) technique [4].

The goal of DTW is to find an warping path that aligns two time series $Q = \{q_1, .., q_n\}$ and $C = \{c_1, .., c_m\}$. In order to align these two sequences, a $n \times m$ matrix is designed, where the position (i, j) of the matrix contains the distance between q_i and c_j. The Euclidean distance is the most frequently applied. Then, a warping path $W = \{w_1, .., w_T\}, \max(m, n) \leq T < m + n + 1$ is defined as a set of "contiguous" matrix elements that defines a mapping between Q and C. This warping path is typically subjected to several constraints:

Boundary conditions: $w_1 = (1, 1)$ and $w_T = (m, n)$.

Continuity: Given $w_{t-1} = (a', b')$, then $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$.

Monotonicity: Given $w_{t-1} = (a', b')$, $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$, this forces the points in W to be monotonically spaced in time.

We are interested in the final warping path, which, satisfying these conditions, minimizes the warping cost:

$$DTW(Q, C) = \min \left\{ \frac{1}{T} \sqrt{\sum_{t=1}^{T} w_t} \right\} \quad (6)$$

where T compensates for the different lengths of the warping paths. This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i,j)$ as the distance $d(i,j)$ found in the current cell and the minimum of the cumulative distance of the adjacent elements: [1]

$$\gamma(i,j) = d(i,j) + \min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (7)$$

In our case, the set of model sequences $\{C\}$ is computed using the HOG descriptors of non-speaking mouth regions in m consecutive frames. The set $\{C\}$ is obtained by supervised labelling using a reduced set of training samples. Afterwards, the classification of samples as speaking/non-speaking patterns is performed by the extraction of the set of queries $\{Q\}$ (each query fixed to length m in our case) from the test samples with some overlapping percentage among consecutive queries, and their alignment to all non-speaking time series $\{C\}$. If a minimum of k samples from $\{C\}$ has a warping cost inferior to a given cost threshold T for a given query Q, then we classify Q as a non-speaking pattern, otherwise, Q is classified as speaking pattern. Once the vector of label prediction is obtained, we filter the output vector using a median filtering in order to correct isolated miss-classifications.

In our case, the selection of a good value of T is critical for the good performance of the classification task. In order to look for the optimal choice of T, we perform cross-validation over the training sequences, selecting the value that maximizes the classification task over a validation subset. Figure 3 shows an example of the data sequences $\{Q\}$ classification for a speaker within a conversation. The square points are the true labels corresponding to the speaking patterns, and the triangle points to the non-speaking ones. A threshold near 0.01 in the training step correctly splits both patterns in the testing step.

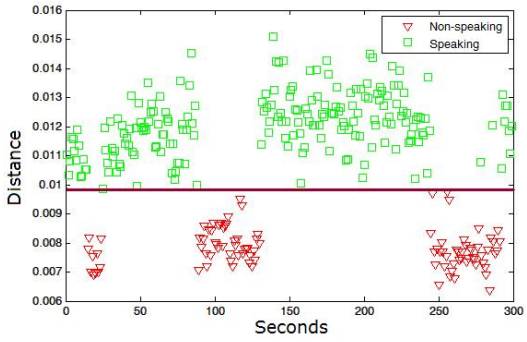


Figure 3: Result of an one-class classification process for an excerpt of five minutes conversation. The legend shows the true label of the samples. Samples are linearly separable using the DTW-based one-classifier.

2.3 Audio-Visual Data Fusion: Stacked Sequential Learning

Once we have performed the classification of audio and video sequences in speaking and non-speaking patterns, we want to integrate both cues in order to improve the performance of the speaker diarization process. Since our data is characterized by temporal coherence, we use sequential learning, which can deal with the fusion of audio-video features at the same time that includes temporal knowledge in the classification process. This is done by considering the predicted labels of the neighborhood samples as new features for a second learning procedure. In this way, useful temporal relations help the audio-visual fusion to improve final speech classification process.

Sequential learning deals with temporal coherence among data samples. In literature, sequential learning was addressed from different perspectives. From the point of view of meta-learning, by means of sliding window techniques, recurrent sliding windows [12], or stacked sequential learning (SSL) [9] were proposed. From the point of view of graphical models, HMM and Conditional Random Fields (CRF) were used to infer the joint or conditional probability of the sequence. Recently, Cohen et al. [9] showed that Stacked Sequential Learning performs better than CRF and HMM on a subset of problems called "sequential partitioning problems". These problems are characterized by long runs of identical labels. Figure 4 shows a block diagram of the SSL method. The basic SSL method uses a five-fold cross-validation on the training set to obtain the predicted set Y' and considers a sliding window of length w with origin in the prediction of the current example to extend its features. That is, for each example in the training set $x_i \in \mathbf{X}$, the predicted values $y'_i \in Y'$ are obtained and joined creating an extended example $x_i^{ext} = (x_i, y'_{i-w_a}, \ldots, y'_{i+w_b}) \in \mathbf{X^{ext}}$, where the number of added features is $w = w_a + w_b + 1$. The extended training set is used to train a second classifier that is expected to capture the sequentiality of the data.

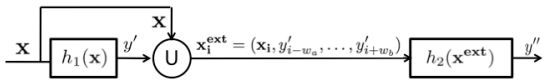


Figure 4: Stacked Sequential Learning scheme.

In our case, we resize the vector of visual features to the audio sampling size. Once the vectors fit in size, the combined feature vector is used to train the first classifier h_1. From the output of this classifier over the training data, a neighborhood w of predicted labels is included as extra feature for each data point, and a second classifier h_2 is trained. As a result of this procedure, we take into account both audio and visual features together and their temporal relations in the training stage.

3. SOCIAL NETWORK EXTRACTION AND ANALYSIS

The social network is represented as a directed graph whose links are estimated by the influence model, based on the integrated audio/visual features. Afterwards, the properties of the social network are studied using several centrality measures.

3.1 Network Extraction: The Influence Model

The Influence Model (InfModel) [5] is a tool developed to quantitatively analyze a group of interacting agents. In

[1]Note that though different adjacency elements can be considered varying the warping normalization factor T, here we follow the present adjacency rule as the most extended one.

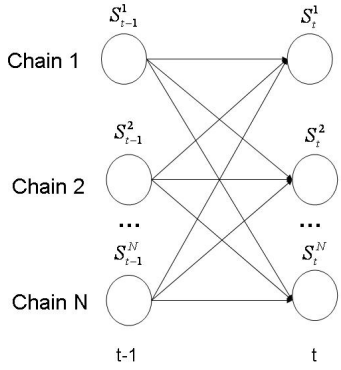


Figure 5: The Influence Model architecture.

particular, it can be used to model human behavior in a conversational setting. In this context, the participants and their corresponding interactions are modelled through a coupled HMM. In figure 5, we offer a visual representation of this architecture.

The model is completely defined by a parametrization scheme that represent the influence of one chain over the others. More concrete, given N participants, the multi-process transition probability $P(S_t^i|S_{t-1}^1, ..., S_{t-1}^N)$ is approximated only by the transition probability $P(S_t^i|S_{t-1}^j)$, where t represents the time stamp. With this convention, the multi-process transition could be expressed now as:

$$P(S_t^i|S_{t-1}^1, ..., S_{t-1}^N) = \sum_j \alpha_{ij} P(S_t^i|S_{t-1}^j) \quad (8)$$

In other words, the state of chain i at time t is conditioned only by the state of chain j at time $t-1$. The α_{ij} parameters that appear in the equation above are referred as 'influences', because they are constant factors that tell us how much the state transitions of a given chain depend on a given neighbor. A more intuitive interpretation could be the following: the amount of influence from a neighbor is constant, but how this influence is reflected, depends on its state.

In its current implementation, the InfModel is able to model interactions between pairs of participants, but it is not able to model the joint effect of several chains together. The learning algorithm for the InfModel is based on constrained gradient descent. For our experiment, we estimated the InfModel based on voicing features (i.e. speaking-non speaking segments), resulted from the integration of audio/video modalities. Thus, our social network will be created by estimating the influence one person has over the other from the dyadic interactions.

3.2 Network Analysis

In social network analysis, a common measure to assess a person's position in a group is centrality [18]. Several centrality measures exist, which are used to quantify different aspects of a network. Some of them take into account only if there is a link between two nodes. Others are based on the links' weight (as a way to quantify the 'distance' between two nodes). In our case, the weight values are given by the α coefficients from eq.(8). The most common centrality measures are: degree, closeness, betweenness, and eigenvector.

Degree centrality: refers to which person is more active by counting the number of connections to other persons. In other words, this means which person is able to communicate directly with the others. In directed graphs, this centrality measure has two components: **in-degree centrality** (number of incoming connections) and **out-degree centrality** (number of outgoing connections). A high in-degree value reflects a person's availability to receive more information (to be influenced) by others. On the other hand, a high out-degree value reflects a person's ability to influence the others.

Closeness centrality: is based on the geodesic distance between one person and the other in the network. It shows the facility of one person to communicate with the other. Nodes with small centrality values mean that they are 'close' to each other. In other words, we expect that the smaller the centrality value is, the higher the influence of the node in the network is.

Betweenness centrality: measures how important a person is in bridging two different parts of a network. The removal of such a person (node) could create a breach in the network, which will ultimately lead to a loss of network cohesion. This kind of nodes are very influential in the network topology.

Eigenvector centrality: is a measure of the importance of a node in a network. A person's position in the network is influenced by the other persons position. In other words, a person's position (influence) in the network increases due to people with high position (influence).

4. EXPERIMENTAL RESULTS

In this section we report some experimental results for the integrated framework previously described. We first assess the performance of our proposed scheme for audio/visual feature fusion. Afterwards, we present some centrality measures computed for the social network extracted using the InfModel based on audio/visual features.

Before presenting the results, we make a brief description of the data, methods, and validation protocol used in the experiments.

Data: The data used consists of dyadic interactions from the publicly available New York Times' Blogging Head opinion blog [1]. The videos show close-ups of two persons talking in front of a webcam about a given topic (most common, politics). One of the persons is the 'anchor' and the other one the 'guest'. The character of the conversation is totally informal, so the audio signal is somehow monotonic and there are no significant variations in voice energy and speaking rate. In a limited number of videos, we could see that speakers interrupt each other quite often. In most of the cases, however, the conversations are characterized by long turn-takings and almost the absence of overlapping speech fragments. The average duration of the analyzed conversations is 30 minutes.

From the whole set of dyadic conversations in the blog, we collected a subset of 17 videos from 15 different people. It is important to remark that this selection has been done taking into account the most active people of the blog. Moreover, the number of conversations selected for each speaker is proportional to his/her activity in the blog. The people featuring in the videos also are somehow connected. This se-

lection criteria is important since it shows the general structure of the most active people in the blog. The remaining participants who do not appear in our selection have a very sporadic participation, and form small isolated non-connected sub-graphs in the social network. This selection criteria is important in order to apply the centrality measures described in the previous section.

Methods:

· *Audio parameters*: The size of the sliding window is fixed to 25ms, with an overlap of 10ms. Each window is represented by means of 13 (12MFCC+E) ceptral features and complemented with the Δ and Δ^2, using only frequencies in the band (130Hz - 6800Hz). The minimum considered audio segment is fixed to 2.5s.

·*Video parameters*: 32 orientation features are computed using the HOG descriptor over the segmented mouth regions. The length of the DTW sequences is fixed to 18 region descriptions, which correspond to 1.5 seconds with an overlapping of 0.5 seconds among regions. 10% of the video samples are used for extracting the ground truth of non-speaking patterns and computing the non-speaking cost threshold, and the remaining 90% of the samples are used for testing the visual speaking methodology.

· *Fusion parameters*: Adaboost with 50 iterations of decision stumps is used as the classifier for stacked sequential learning [13]. For the second classifier of the stacked methodology, the same classifier is used extending the feature space for each point in a neighborhood of 10 label predictions before and after each prediction. These values have been experimentally tested to be suitable for our application.

Validation protocol: We used 90% of the data for training and the remaining for test. For each video, we show the speaker diarization performance by comparing the visual cue alone with the audio/visual feature fusion process. The comparison is done taking into account the ground truth segmented from the audio data. Centrality measures are also computed over the extracted social network.

4.1 Audio-video fusion results

The audio clustering methodology returns a vector of labels, where each label corresponds to a possible speaking cluster, including the non-speaking cluster, and thus obtaining different number of clusters for each conversation. For this reason, we can not obtain a direct performance for speech segmentation in the audio domain. However, the fusion scheme using the stacked sequential learning associates the audio cluster labels to the corresponding speakers or non-speaking patterns based on the information provided by visual features.

Table 1 shows the visual and audio-visual speaker diarization results by comparison with the ground truth data. Each row of the table corresponds to a conversation. The first column identifies the subjects that participate in each conversation (see figure 6). The labels identifying the subjects and the conversations are graphically shown in the social network of Figure 6. The best performance of both speakers in each conversation is marked in bold. Note that in most of the cases the fusion methodology considerably improves the video classification, obtaining high accurate predictions. Only in three of the 34 speakers classification, the fusion is not able to improve the video results.

In the case of conversation (1-3), though the fusion is not improved for the second speaker, the final performances are very similar. In the case of conversation (13-14), the visual features for the first speaker do not split speech and non-speech patterns properly. In this case, the 61.99 of accuracy is obtained because of a high number of non-speech classifications (near 80%), meanwhile the speech detection is near 40%. These unbalanced results make the fusion a hard task even with the combination of audio cluster labels. On the other hand, the case of conversation (7-14) has the opposite problem. In this case, the audio methodology assigned similar labels to clusters of different speakers, which makes difficult the fusion with visual features. Finally, significant performance improvements are obtained with the fusion methodology in more of the 90% of the cases (31 of 34 subjects). An interesting point is that those people who appear in different conversations used to maintain their speech/non-speed discriminability, since they used to act in a similar way.

Table 1: Visual and Audio-Visual Speaker Diarization Accuracy

Conversation	Video		A-V	
	Left	Right	Left	Right
1 - 2	66.58	58.36	**81.90**	**78.87**
1 - 3	58.01	**75.82**	**72.90**	72.52
1 - 4	68.37	78.50	**85.48**	**79.16**
1 - 5	88.99	72.50	**89.02**	**84.02**
1 - 6	69.51	61.86	**91.47**	**90.72**
9 - 3	82.63	61.95	**97.88**	**80.75**
9 - 10	65.01	63.71	**96.92**	**65.44**
3 - 11	65.77	74.91	**92.40**	**93.58**
4 - 3	75.35	64.09	**80.05**	**91.73**
4 - 12	94.13	75.36	**94.21**	**93.36**
13 - 15	70.96	71.95	**97.36**	**95.18**
13 - 14	**61.99**	65.11	56.56	**95.24**
12 - 14	86.20	64.02	**90.25**	**88.31**
12 - 7	97.75	85.26	**97.82**	**98.51**
8 - 10	61.44	55.93	**92.56**	**95.09**
9 - 11	67.09	66.88	**97.80**	**94.98**
7 - 14	55.88	**63.54**	**96.09**	60.12

4.2 Centrality measures results

As a result of the audio/visual fusion scheme previously introduced, we obtained a binary vector whose values represent the speaking/non-speaking states for each person. These vectors are fed into the InfModel in order to get the α coefficients (from equation (8)) which encode the influence values for each person. We used the InfModel implementation which comes with the MIT library for speech processing [2]. Based on the α coefficients, we extracted the graph of inter-personal relations: the direction of links reflects the influence of one person over the other. In other words, the links are weighted by the α coefficients. An 'A$\rightarrow$B' link can be interpreted as 'A has influence over B'. The resulting graph is depicted in figure 6.

Note that in this graph only the links with the highest weights (α coefficients) are represented. The lack of a link between two persons means that these persons do not interact at all. The number which appears on the upper-right part of each node (face circle) represents person's number.

Based on the α coefficients, we constructed the associated sociomatrix which has been subsequently used in the computation of several centrality measures: degree (with its two versions, in-degree and out-degree), closeness, be-

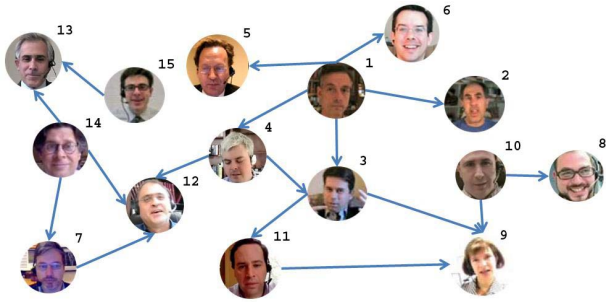


Figure 6: Social Network showing participant's labels and influence directions.

tweenness, and eigencentrality. The computed measures are summarized in table 2. We represented in bold characters the highest values of these measures.

The conclusions we can extract from this table, in some cases, are straightforward. If we are interested, for instance, in the most important node (person) in the network, this is by far the node 1. This is confirmed by out-degree, in-degree, and eigenvector centrality measures. At the same time, we could say that the person designated by node 1 is the most influential person in the group: he influences up to four persons, but nobody else is able to influence him.

If we are interested in the person acting as a cohesion factor for the group (or a hub), this is represented by node 4. This fact is confirmed by the betweenness centrality. We can identify up to 3 subgroups that are bridged by this node: (1,2,5,6), (7,12,13,14,15) and (3,8,9,10,11). For this reason, the role of node 4 in this network is crucial, but his relevance is different from the node 1. For instance, if he is removed from the group, the network structure is severely damaged, practically being split in 3 parts. On the other hand, if we remove node 1, indeed, we face obviously a loss, but the network structure is not affected too much.

From the same table, we can also deduce that node 8 is the most irrelevant component of the group, fact confirmed by all centrality measures we used. If we remove it from the network, the effect is null. On the other hand, nodes 9 and 12 are the weakest, in the sense that they are influenced by all the surrounding neighbors (in-degree centrality measure).

Based on the closeness centrality measures, we can infer that person 9 is the most influenced. On the other hand, person 15 has the most influencing potential.

In all the other cases, some aspects might not be that obvious and a disagreement between measures might happen. This can be explained due to the unsupervised nature of the network extraction process. As we have mentioned before, the dyadic interaction takes place in an informal manner. For this reason, it is impossible to establish, even using human annotators, an absolute ground truth of the interactional process (who has more influence over whom), because in our approach we take into account only the non-verbal analysis, making total abstraction of the conversation content.

As a final remark it is important to discuss the general applicability and usefulness of the proposed methodology. A potential application of our approach is its use in collaborative working environments. They could offer very useful insights of the actual role each person is playing in the group (based on their activity level or involvement degree) and the group structure per se (is it a coherent group or not). On the other hand, social networks can be used to study how the informational flow propagates in a group or small communities. This will allow distant users to become better integrated into ongoing discussions, and thus improve distance-separated social interaction, teamwork, and social networking. In this context, it is more difficult to quantify these aspects based on explicit input (exchange of messages) due to privacy concerns. But an implicit input analysis would offer a more clear perspective of the interactional patterns.

5. CONCLUSIONS

In this paper, we presented an integrated framework for automatic extraction and analysis of a social network from implicit input (multimodal dyadic interactions), based on the integration of audio/visual features.

After the extraction of audio and visual cues, we performed a fusion of audio/video data, based on Stacked Sequential Learning, in order to increase the robustness of the speaker segmentation process. Subsequently, the feature vector consisting of the audio/video data has been fed into InfModel in order to determine the influence between persons in the dyadic interaction. Based on the resulting relationships, we built the structure of the social network. Finally, we applied some SNA specific measures (different types of centrality) in order to extract some characteristics of the discovered network.

In the future, we plan to extend the current approach and study the problem of social interactions to a larger scale and in different scenarios. Starting from the premise that people's lives are more structured than it might seem a priori, we plan to study long-term interactions between persons, with the aim to discover underlying behavioral patterns present in our day-to-day existence. For this purpose, we plan to use some wearable, sophisticated sociometric devices, able to record audio, video, and location of the subjects.

6. ACKNOWLEDGMENTS

This work has been supported in part by the projects TIN2009-14404-C02 and CONSOLIDER-INGENIO CSD 2007-00018, Ministerio de Educación y Ciencia, Spain.

7. REFERENCES

[1] Blogging heads: New York Times public opinion blog. *http://video.nytimes.com/*.

[2] Speech feature extraction library. *http://groupmedia.media.mit.edu/*.

[3] M. Ahuja, D. Galletta, and K. Carley. Individual centrality and performance in virtual r&d groups: An empirical study. *Management Science*, 49(1):21–38, 2003.

[4] J. Alon, V. Athistos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1685–1699, 2009.

[5] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Learning human interactions with the influence model. In *Tech Report 539*, MIT Media Lab, June 2001.

[6] M. Chau and J. Xu. Mining communities and their relationships in blogs: A study of online hate groups.

Table 2: Centrality Measures

Node No.	Out-degree	In-degree	Closeness	Betweenness	Eigencentrality
1	**5**	0	0.5344	0.7582	**0.7264**
2	0	1	0.4628	0	0.1278
3	2	2	0.7898	0.7692	0.3241
4	2	1	0.7510	**0.9341**	0.4321
5	0	1	0.5339	0	0.0152
6	0	1	0.3985	0	0.2832
7	1	1	0.5052	0	0.1224
8	0	1	0	0	0
9	0	**3**	**0.8228**	0.5165	0.0437
10	2	0	0.5375	0.1429	0.0187
11	1	1	0.6177	0	0.1084
12	0	**3**	0.6453	0.8352	0.2159
13	0	2	0.4897	0.2747	0.0131
14	3	0	0.5344	0.6264	0.0896
15	1	0	0.3434	0	0.0063

Int'l. Journal of Human-Computer Studies, 65:57–70, 2007.

[7] A. Chind and M. Chignell. A social hypertext model for finding community in blogs. In *Proc. of the seventeenth conference on Hypertext and hypermedia*, pages 11–22, Odense, Denmark, 2006.

[8] T. Choudhury. Sensing and modeling human networks. MIT Media Lab, 2003. PhD thesis.

[9] W. W. Cohen and V. R. de Carvalho. Stacked sequential learning. In *Proc. of IJCAI 2005*, pages 671–676, 2005.

[10] N. Dalai and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of CVPR2005*, pages 886–893, San Diego, USA, 2005.

[11] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. The LIUM speech transcription system: a CMU sphinx iii-based system for french broadcast news. In *Proc. of INTERSPEECH'05*, pages 1653–1656, Lisbon, Portugal, September 2005.

[12] T. G. Dietterich. Machine learning for sequential data: A review. In *Proc. on Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, 2002.

[13] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. In *The annals of statistics*, volume 38, pages 337–374, 1998.

[14] M. Jones and P. Viola. Robust real-time face detection. In *International Journal of Computer Vision*, volume 57, pages 137–154, 2004.

[15] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech and Language*, (2–3):303–330, 2006.

[16] A. Pentland and A. Madan. Perception of social interest. In *Proc. IEEE Intl. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*, Beijing, China, October 2005.

[17] H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Trans. on Multimedia*, 11(7):1373–1380, 2009.

[18] S. Wasserman and K. Faust. Social network analysis: Methods and applications. New York, 1994. Cambridge University Press.

[19] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Movies analysis based on roles social network. In *Proc. of Int'l. Conference on Multimedia and Expo (ICME)*, pages 1403–1406, Beijing, China, 2007.