6-30-2010

Detecting Group Turns of Speaker Groups in Meeting Room Conversations Using Audio-Video Change Scale-Space

Ravikiran Krishnan

University of South Florida

Detecting Group Turns of Speaker Groups in Meeting Room Conversations Using Audio-Video Change Scale-Space

by

Ravikiran Krishnan

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Sudeep Sarkar, Ph.D.
Rangachar Kasturi, Ph.D.
Dmitry Goldgof, Ph.D.

Date of Approval:
June 30, 2010

Keywords: Conversation change, Temporal scales, Turn pattern, Multimedia analysis, Taxonomy

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Prof. Sudeep Sarkar, for many insightful conversations during the development of the ideas, for his kindness and most of all, for his patience. He showed me different ways to approach a research problem and the need to be persistent to accomplish any goal. I am also appreciative of his emotional support during difficult times. I would also like to thank my committee members Prof. Dimitry Goldgof and Prof. Rangachar Kasturi, for their time and valuable suggestions during the past two years. I thank the Department of Computer Science and its professors at the University of South Florida for providing me with financial support to pursue my master's degree.

I would like to take this opportunity to thank all my friends who helped me get through two years of graduate school. I thank my parents, who have always been an inspiration to me. Finally, I would like to express my profound appreciation to my brothers Srinivas and Guru. Their continuing encouragement, understanding and guidance have helped me be a better person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Detecting Group Turns of Speaker Groups in Meeting Room Conversations Using Audio-Video Change Scale-Space

Ravikiran Krishnan

ABSTRACT

Automatic analysis of conversations is important for extracting high-level descriptions of meetings. In this work, as an alternative to linguistic approaches, we develop a novel, purely bottom-up representation, constructed from both audio and video signals that help us characterize and build a rich description of the content at multiple temporal scales. Nonverbal communication plays an important role in describing information about the communication and the nature of the conversation. We consider simple audio and video features to extract these changes in conversation. In order to detect these changes, we consider the evolution of the detected change, using the Bayesian Information Criterion (BIC) at multiple temporal scales to build an audio-visual change scale-space. Peaks detected in this representation yields group turn based conversational changes at different temporal scales.

We use the NIST Meeting Room corpus to test our approach. Four clips of eight minutes are extracted from this corpus at random, and the other ten are extracted after 90 seconds of the start of the entire video in the corpus. A single microphone and a single camera are used from the dataset. The group turns detected in this test gave an overall detection result, when compared with different thresholds with fixed group turn scale range, of 82%, and a best result of 91% for a single video.

Conversation overlaps, changes and their inferred models offer an intermediate-level description of meeting videos that are useful in summarization and indexing of meetings. Since the proposed solutions are computationally efficient, require no training and use little domain knowledge, they can be easily added as a feature to other multimedia analysis techniques.

CHAPTER 1

INTRODUCTION

1.1 Overview and Motivation

A meeting, as we know, is a dynamically changing entity. Automatic analysis of a meeting's content and building rich descriptions of it are still difficult problems. Researchers from various fields such as behavioral psychology, human computer interface design, human communication behavior, computer vision and signal processing have focused efforts on analyzing multimedia content, particularly in meetings.

A single meeting is a complex temporal chain of dynamically changing events. For example, a staff meeting, can be looked upon as a discussion when considered in its entirety, but also can be described as a sequence of characterizations appropriate over smaller temporal windows, such as argument, discussion, note-taking, and breaks. Automatic inferring of this wide range of temporally dependent description is still a fundamental problem. Bridging this gap and describing the dynamics of meetings would be an important step toward solving this problem.

Most researchers focus on analyzing meetings by trying to answer questions such as Who? When? Where? Which?. One question that really helps in building an intermediate level description of the meeting conversation is, How? How is the conversation changing? One important property in analyzing how the meeting is occurring, without lexical features, is the turn patterns. Taking turns to talk is a common communication ordering in any type of multimedia content, but it is more prominent in conversations involved in meetings. In order to detect the turn pattern, knowledge of the temporal structure of the each speaker must be identified. This task is not so trivial, as the conversations involves overlap speech.

Turn patterns provide a record of speech transitions in a conversation. According to conversational analysts in [1]-“wherever possible the speaker's current turn will be interpreted as implicating some action by the responder in the immediate next turn. Similarly, the respondent's subsequent talk will, where possible, be interpreted as related to the immediate prior turn". Instances like a presenter with whom a responder agrees or disagrees with what the presenter is saying. These instances are usually detected with the use of lexical features. In this work, non verbal cues are used that enables the detection to be language independent.

Figure 1.1 is the result from the ground truth of a meeting room video. Only five minutes of the meeting are used for better representation purpose. The nodes are the states, the labels are the numbers of people speaking. The state labels are given by binary strings, which specify subjects speaking from right to left. Description of labels 0101 - Subject 1 and Subject 3 are speaking together; 1000 - Only subject 4 is speaking; 0001 - Only subject 1 is speaking. The edges in the below given graph show that there is a transition from one state to another. The solid line here specifies the state change from a lower number when converted to decimal digits, and the dotted lines specify a change in state from a higher number to a lower number. This is done only for visualizing turn patterns.

A single speaker in the meeting can speak at different times in the meeting, and this speaker can speak at the same time as another speaker in the meeting. This is called overlap speech with the current speaker at a particular time in the meeting. Identification of temporal structure depending on scales does not define "what was being spoken", but defines the conversation changes. In [2], speaker turn patterns are used for speaker diarization. The input stream is segmented to get the speaker change point for each speaker. Before segmentation, the overlap and silence frames are dropped. This prevents over segmentation for speaker diarization. But dropping the overlap frames neglects some important group conversational dynamics. In order to incorporate these overlap frames, a higher level abstraction is needed. The focus of this work is to provide an intermediate level description of conversation change in meeting rooms. We use a multi scale approach to segment the input stream, which results in the detection of group turn patterns.

Communication is a dynamic process. Every conversation has its own temporal scale. For example, the amount of time taken for speaker exchanges in an argument may be different from speaker exchanges in a discussion. Information about these unknown temporal changesis

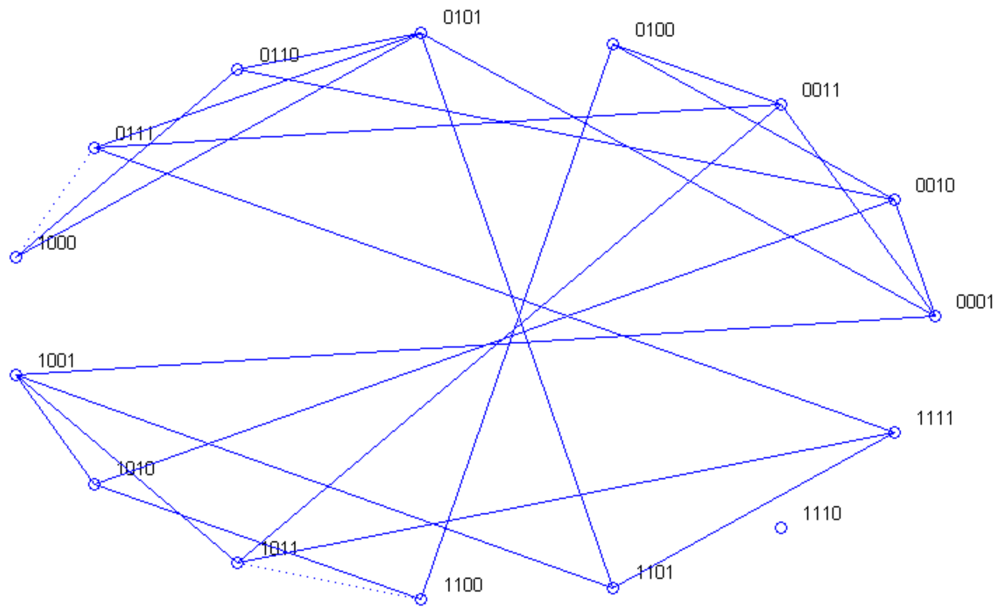


Figure 1.1: Turn pattern for 5-minute video using audio ground truth. Each state in the transition graph represents a state a meeting is in, in terms of how many speakers are speaking at the same time.

important in describing the state of a conversation. To analyze these unknown scale variations in conversations, we consider a multi-scale representation of an audio stream.

In a meeting room, there are always a few dominant group of speakers. These speakers are usually interrupted by other speakers. These interruptions may be considered as speaker changes in single scale BIC, as is the usual practice, but they do not provide us with a high level description of the conversation between the dominant groups. Our method provides a way to detect these longer scale conversation changes called the group turns. In order to detect group turns, which implicitly describe the nature of the conversation, a multi scale approach is considered. The audio-video joint feature set is represented as a scale-space and is called "Audio-video change scale-space". Representation is a scale-dependent description of different turn patterns in conversations, and a change in this scale-space shows the multi scale temporal change point structure of the entire conversation.

These patterns suggest conversation states without considering what was spoken. Many of the algorithms detect the speaker change point based on the BIC using a single scale or one-length window that fits best for a particular dataset. This type of speaker change point detection provides only a single-scale temporal structure of the conversation. Single-scale cannot determine the length of an individual person speaking in a conversation. Our effort focuses on obtaining finer to coarser conversations in a meeting room. Finer conversations are speaker conversations in which the speaker turn is happening at shorter intervals of time. Coarser conversations are for speaker turns at larger intervals. Therefore, determining the length of a conversation is the most important step toward describing speaker turn patterns.

In order to recognize types of conversations, group turn patterns must be identified. This helps us understand the communication order in meetings. This ordering can be classified into groups of similar conversations such as silence, monologue and polylogue conversations. Commonly occurring general conversation models are described, which can be used for classification of conversation changes based on group turns.

In summary, this proposed approach can be used as a playback feature to go to that particular group turn in a meeting and detect dominant groups in the a meeting by clustering. These scale-space representations and detected turn patterns can be used as features in content-based queries of meetings. These inferred group turn models offer an intermediate-level description of meeting videos that can also be useful in summarization and indexing of meetings.

1.2 Assumptions

These are the following assumptions regarding meetings being analyzed:

1. The meeting involves multiple participants, but the number of participants is not known.

2. The meeting is recorded using only a single stationary camera and a single microphone, and these streams are time synchronized.

3. No assumptions are made regarding the placement of the microphone and the camera. In particular, because the participants face each other, all faces may not be visible to the camera.

4. The participants stay seated in their chairs throughout each meeting.

5. The nature of the meetings is unknown - i.e, they may be group discussions, debates, brainstorming sessions or or any such kind of meetings.

1.3 Contribution

This work focuses mainly on three concepts. First, we define two types of turn patterns – speaker and group turn patterns. Based on group turns, we describe general conversational models and their corresponding peaks. Second, we implement a BIC-based multiple scale organization of an audio-video stream, which we call the “Audio-video change Scale-Space”. A scale represents a temporal window that is used to find a conversation change in terms of the two turn patterns. This representation is a scale-dependent description of different turn patterns. Each turn pattern is associated with a temporal scale range in the scale-space. The third aspect of our approach is automatic detection of group turns. Our approach provides a higher level description of the conversation between the dominant groups by detecting group turns. Based on this work, a paper [3] will be presented as an oral presentation at the International Conference on Pattern Recognition (ICPR), 2010.

1.4 Outline of Thesis

The chapters in this thesis are written from a theory of group turn patterns to produce results that indicate the detection of the group turns. The first chapter is an introduction to turn patterns; the second chapter describes the theory of speaker and group turn patterns.

Chapter 2 presents the published prior work to this thesis. The previous work describes different types of turn patterns. The turn patterns is described are categorized into group action turn patterns and conversational turn patterns.

Chapter 3 describes the two turn patterns. The first turn pattern is the speaker turn pattern and its concepts are described using definitions and illustrations. The second turn pattern is the group turn pattern, which is the main focus of this thesis. In addition, the chapter also describes the theory and its conceptual visualization of the most commonly occurring group turn patterns .

Chapter 4 describes the proposed approach for detecting group turn patterns. This chapter starts by describing the features extracted from the video clip. The proposed approach follows the features used. The steps detecting group turn patterns is described in details with importance given to the scale-space creation and the Bayesian Information Criterion.

Chapter 5 describes the dataset used. This is followed by the results of detecting group turn patterns by varying the group turn pattern scale ranges to get the optimum scale for detecting group turn patterns.

CHAPTER 2

PREVIOUS WORK

This chapter reviews previous works that performed automatic analysis of action turn patterns and conversational turn patterns in settings such as meeting rooms. Social interaction understanding has inspired a surge of interest in developing computational techniques for automatic conversational analysis. As an alternative to linguistic approaches, nonverbal cues have been a major domain for automatic analysis of conversational interaction in recent years. A group conversation can proceed through various communication phases in the course of sharing information. Multi party discourse can range from group discussions, presentations, a casual argument, formal meetings and many more [4]. Figure 2.1 categorizes the related approaches into action turn patterns and conversation turn patterns such as speaker and group turn patterns.

In a formal meeting room scenario, a group can assume to be having various conversational activities. Some of these activities involve use of different artifacts such as projector screens, white boards, and notepads for note taking. In [5, 6], a group meeting is segmented into a location specific turn of actions. This is also a type of turn pattern, but does not involve speech in the recognition or segmentation into these turns. This approach used a supervised technique called Hidden Markov Models, or HMMs. The features used in this technique are the simple non-verbal audio and visual cues. Multiple cameras and multiple microphones were used. Some of the simple audio features used were pitch, energy and spectral frequency. The visual features extracted were each participant's skin color, motion blobs and location for the indication of body motion and head position. This approach used a multi-model meeting manager corpus [6]. The results were measured in terms of action recognition rate.

Even though action recognition was convincing and the works showed the value of audio-visual fusion, such methods had a drawback of over fitting the data when learning, as the

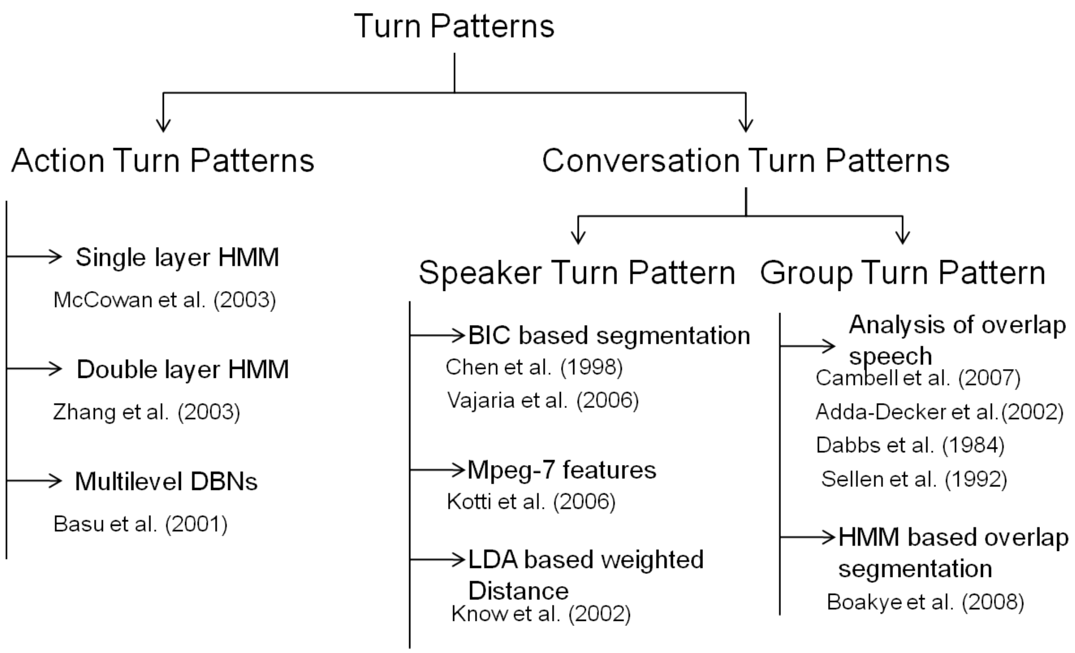


Figure 2.1: Previous work based on action and conversational turn patterns. The action turn pattern is based on classifying the meeting into different types of actions. The conversational turn pattern is divided into two groups, namely speaker turn patterns and group turn patterns.

data available was limited. Some of the segmentation and recognition of the meeting events and actions are given in [7, 8, 9]. These works addressed the concerns by having a multi layer HMM framework. One layer of the HMM framework models basic individual activities from low-level audio-visual features, and the next level of the HMM framework models the interactions. This action performed reasonable well on the M4 corpus. Some of these actions segmented and recognized are the discussion, monologue, note-taking, presentation, white-board. These actions provide a different type of turn taking behavior. This does not involve segmentation of meetings in terms of conversations.

In [10], personal states and meeting states were inferred through Finite State Machines (FSM) separately, where only visual cues were used. States such user's standing-sitting states were also analyzed. The social dynamics of people involved are also described. Detection of actions and activities (sequences of actions) is done using a set of object-oriented finite state machines that run in parallel with one another. The lowest-level inputs to this hierarchy were the observations using visual cues like head gaze and hand position. Similar efforts were pursued in [11, 12]. In [11], the contexts used for interaction detection include head gesture, attention from others, speech tone, speaking time, interaction occasion, and information about previous interaction. The support vector machines (SVM) classifier is adopted to recognize human interactions. The AMI project [13] also deals with interaction issues including turn-taking, gaze behavior, influence and talkativeness. In [12], four kinds of classification models, including Support Vector Machine (SVM), Bayesian Net, Nave Bayes and Decision Tree were selected to infer the type of each interaction. Brdiczka et al. [14] adopted speech features for the automatic detection of interaction group configurations.

A hierarchical approach of segmenting and recognizing human interactions dependent on who was speaking was adopted in [15]. In conversation analysis, turn patterns are usually referred to as the speaker turn patterns. In [1], speaker diarization is a term used to cluster all the speaker specific homogeneous sections of an individual speaker. Speaker diarization is a basic problem and use of speaker diarization gives a temporal structure of conversation to determine who was speaking at at that instant of time. An overview of the speaker diarization systems are provided in [16]. More recently, new types of multi-modal [17] features such as prosody, speaker turns etc. were added as features to classify meeting scenarios using a multi-stream Dynamic Bayesian Model (DBN) technique, which is adopted in [18, 19]. In [19], two DBN architectures were investigated: a two-level hidden Markov model (HMM) in which the acoustic observations were concatenated; and a multi-stream DBN in which two separate observation sequences were modeled. The first level of the DBN in multi-level DBNs decomposed the interaction patterns as sequences of sub-activities without labeling them or extracting any meaning from these patterns. The sub-activities are just parameters that are learned during training. This approach showed improvement, but the complex structure of the models made it hard to interpret. The second DBN processed other features and integrated them in a higher level.

Speaker turns [20, 21, 2, 22, 23] are mostly used to describe answers to questions like, Who?, When?, Where?. The speaker turn is used for speaker diarization, localization, floor control analysis [24]. There have been efforts similar to this to combine gesture, gaze and speech together. In [20], the audio features were extracted and segmented using the Bayesian

Information Criterion (BIC). The segmentation indicated speaker change points. This was the first time that a model selection framework such as BIC was used. A lot of work has been done on speaker diarization, which is based on this type of segmentation. The multi-model aspect was also introduced. In [2], the audio and video was combined to form a joint-feature vector. These features were then segmented using the BIC and then clustered. The corresponding video model was also built for these clusters in order to localize the speaker. The speaker turn patterns usually do not involve overlap speech. There have been some developments in incorporating overlap speech and classify the speech into activities, but it is still limited [25, 26, 27]. Labeling activities based the any feature set is a challenging problem. The activities allow us to describe a higher-level description of the conversation and action. The labeling of different types of conversation should include conversation dynamics of groups of speakers. One such effort is in describing a higher abstraction of conversation called the group turn. This group turn also includes overlap speech, as it is no longer only a signal-level description any more.

Group turn pattern is a conversational turn pattern that involves overlap speech, a single speaker and a group of speakers speaking simultaneously. Previously, in language analysis, this group turn pattern has been described [28]. Some other works such as in [29, 30, 25, 26, 27], have been partially successful in determining the overlap speech segments in a conversation. These are again used to describe actions or analyze the amount of overlap speech. In [31], overlap detection is done by an HMM-based technique. In our work, the focus is to create a multi-temporal scale-space for conversation to detect group turn patterns and distinguish this work from previous works of action recognition and speaker turn patterns.

CHAPTER 3

TURN PATTERNS

Turn patterns are a communication ordering of all the speakers in the conversation. This pattern shows us whose turn, at a particular time instant, it was to speak. The pattern shows the combination of information about the speakers such as, who was speaking, when, how many times a speaker spoke. These patterns also show the overlap and/or simultaneous speech involved in the conversation. Having the overlap speech in conversation can reveal the nature of the conversation. Longer overlaps is an indication of disagreement or quarrel. More overlap speech in conversations lead to more arguments than discussions.

In meeting rooms, a group of people conversing can be seen as proceeding through diverse conversational phases. These phases can be detected through the temporal structure of the conversation, according "who spoke when". These various phases are called as the conversation changes. Turn patterns describe the temporal structure of the meeting, which in turn offers an intermediate- level description of the meeting. The turn pattern does not describe what was being spoken, as the lexical features are not used, but it can describe how the conversation changed or progressed in time. We define a speaker having his or her turn to speak by "have the floor" [29]. We have defined conversation change to be in terms of speaker turns and group turns [29] [30], and a scale represents a temporal window that is used to find a conversation change in terms of the two turn patterns described below.

3.1 Speaker Turn Patterns

A turn consists of the sequence of talk spurts and pauses by a speaker who "has the floor". A speaker loses the floor when there is a pause by the speaker or when a speaker has fallen silent. A speaker gains the floor back when the speaker starts to speak and is not interrupted by a pause for at least 3 seconds [28]. Without this criterion, even the shortest unilateral sound would be designated as a turn. The time slice of 1.5 seconds was chosen because it is estimated to be the mean duration of a phonemic clause, and evidence shows that the phonemic clause is a basic unit in the encoding and decoding of speech. A speaker turn occurs when a speaker loses the floor and another speaker gains it.

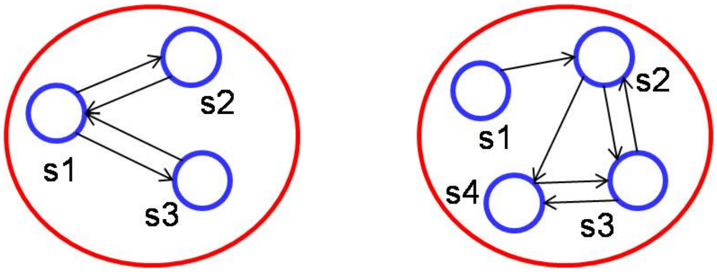


Figure 3.1: Conceptual sketch of speaker turn patterns. There are two figures, a sketch (left) of speaker turn pattern involving three speakers which is indicative of speaker student turn pattern, and the sketch (right) of a speaker turn pattern involving four speakers is indicative of a discussion or an argument.

In Figure 3.1, two conceptual speaker turn patterns are shown. A speaker turn pattern involving three speakers, S1, S2, and S3, are shown (left). This type of speaker turn pattern is typical in presentations where one speaker (S1) is dominant and has the floor during most of the conversation, and the other two speakers (S2 and S3) gain and lose the floor at sparse intervals, and therefore communication is always headed by the dominant speaker. Speakers S2 and S3 do not have any arrows between them, indicating that there is no speaker turn from speaker S3 to Speaker S2 and vice-versa. Another speaker turn pattern in shown (right), where there are 4 speakers involved in the conversation. In this speaker turn pattern, more interaction occurs between all the speakers. This indicates that all the speakers have gained or lost the floor more frequently, which in turn indicates discussion or argument.

Figure 3.2 shows the audio ground truth of an excerpt of a small conversation in the meeting-room corpus used. There are three speakers and all of them have gained the floor and lost it.

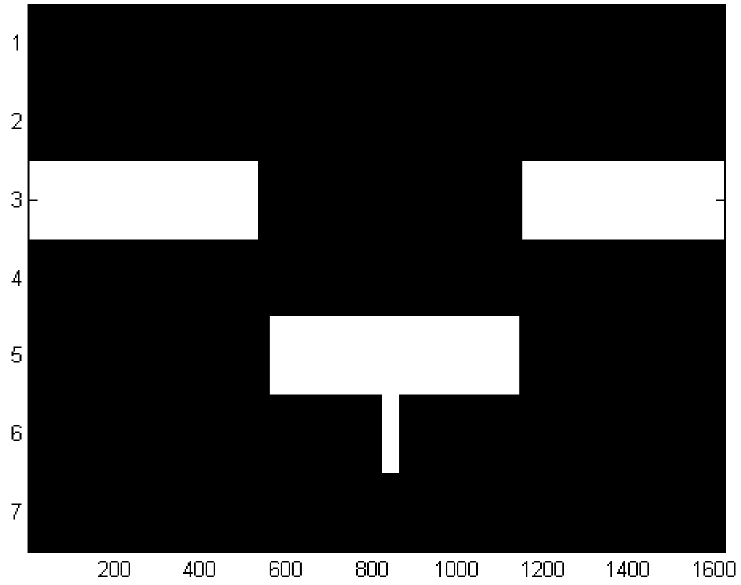


Figure 3.2: Sample speaker turn pattern. The vertical axis is the speaker ID and the horizontal axis is the time. The white patches indicate that a particular speaker is speaking.

3.2 Group Turn Patterns

A group turn occurs when a speaker or a group of speakers loose the floor, and a speaker or a group of speakers gain it. The definition of group is described in terms of simultaneous speech. Simultaneous speech is a type of speech where one or more speakers who do not have the floor are overlapping a speaker or a group of speakers who have the floor. An instance in which a group turn occurs from a speaker who has the floor, to a group of speakers, with the same speaker keeping the floor is called overlap speech. Overlaps are synonymous with interruptive simultaneous speech. Group turn is described to cover instances where individual turn takers who have the floor are effectively "drowned out" by the group.

Figure 3.4 shows the different group turn patterns as nodes G1, G2, G3...Gn. Each node can be considered as a super node consisting of speaker turn patterns. Each group turn pattern consists of a different number of speakers. The dynamics of each group turn pattern are also different. The first group turn pattern, G1, shows a type of speaker turn pattern. Each different speaker turn pattern can be segmented as the group turn patterns. Node G3, where there is

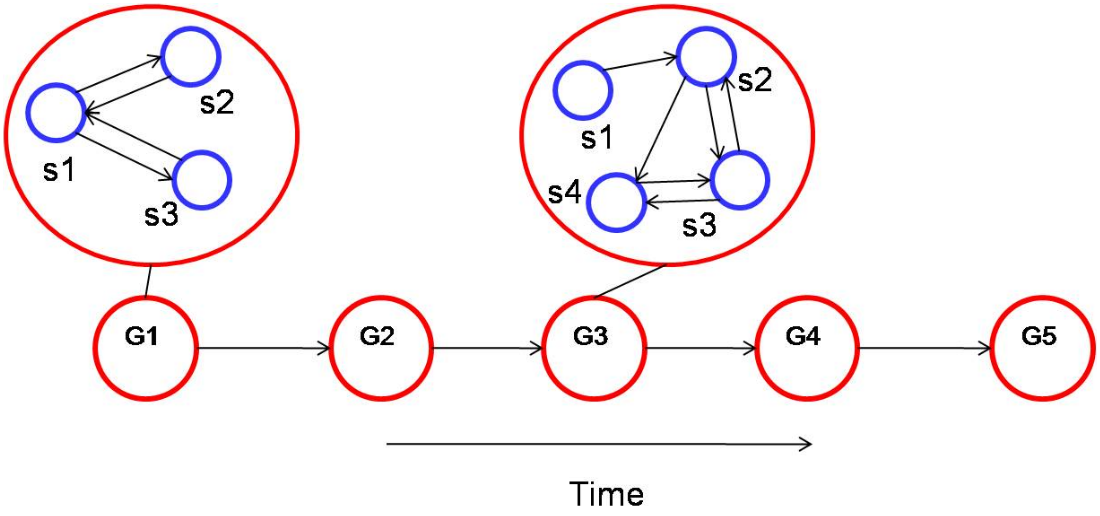


Figure 3.3: Conceptual sketch of five group turns. Each turn is different from each other.

more interaction between the speakers, has a different number of speakers involved in the conversation. Figure 3.5 shows the audio ground truth of an excerpt of a small conversation in the meeting-room corpus used. There are five speakers and all of them have gained the floor.

A change scale-space and its division will provide the necessary information to determine whether the change point at speaker turn or a group.

3.3 Commonly Occurring Group Turn Patterns

A polylogue contains a sequence of group turn patterns and these patterns describe speaker and simultaneous speech turn patterns. According to these turn patterns, we define conversational models that reflect the most commonly occurring group turn patterns and describe the turn changes with respect to a scale-space. There are four types of conversational models described:

1. Long Conversation w/ Short Overlaps

2. Short Exchange

3. Long Conversational Overlaps

4. Short Speech Exchanges

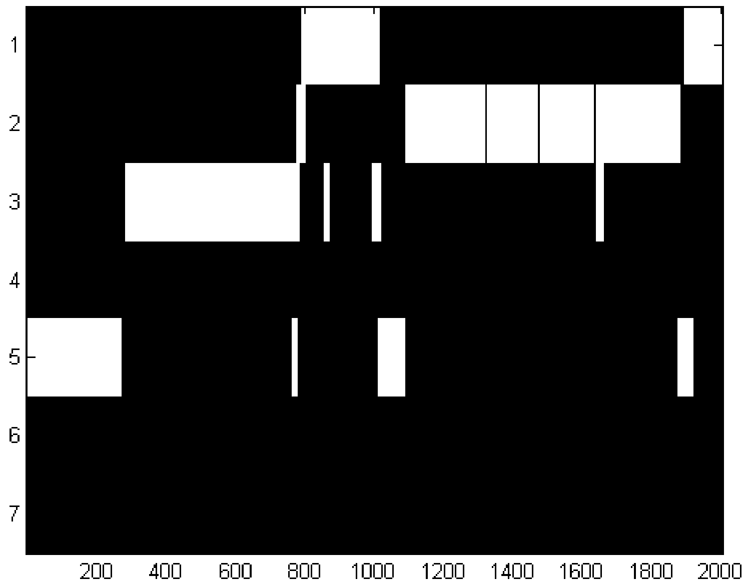


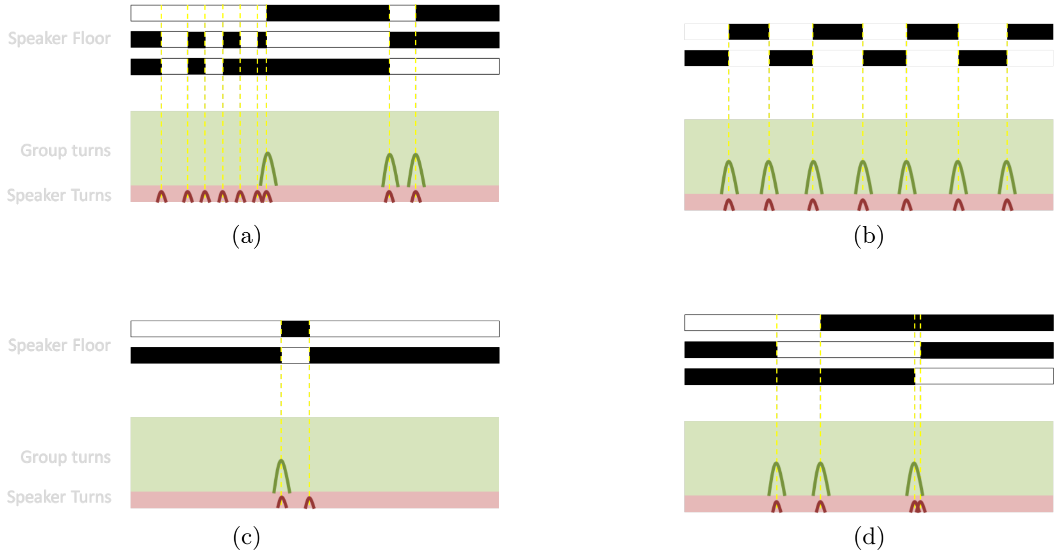
Figure 3.4: Sample group turn pattern. The vertical axis is the speaker ID and the horizontal axis is the time. The white patches indicate that a particular speaker is speaking. Different combinations of speakers speak at different time instances.

A conversation that suggests a dispute, quarrel and/or disagreement is an argument. Looking for these types of conversational turn patterns will lead to the detection of simultaneous speech. Constant small overlap exchanges indicate that there is disagreement.

3.3.1 Long Conversation with Short Overlaps

Long conversational change points will be detected at the group turn scale range and simultaneous speech with short length will be detected at the speaker turn scale range.

Figure 3.5[a] is a schematic corresponding to long, non-overlapping conversations, indicating a discussion with short overlaps, which is considered to be a responder(s) agreeing or acknowledging the communicator with words such as “Hmm”, “Uh”, and some laughter (giggles) by another person.



(a) (b)

(c) (d)

Figure 3.5: Schematics of different conversations. White horizontal bars show who has the floor. Black bars show that the speaker has fallen silent. For each speaker floor change, corresponding regions (group and speaker) in which peaks are shown.

3.3.2 Short Exchange

Short Exchange speech is another category of speech where there is a break in a monologue and there is very short speech in which the ratio of the monologue and this short speech is very high, which suggests a short exchange of speech. Short exchanges occur when a person is looking for approval or disapproval of the speech he is currently making. Words of approval only may not be considered short exchange; some spoken words that will not decrease the monologue to short speech ratio can only be considered short exchange.

Figure 3.5[d] depicts a conversation win which a speaker tends to wait for an approval or an acknowledgment or even a small sentence indicating the responder(s)'s view. Lower temporal scales will detect all these individual speaker changes or a turn.

3.3.3 Long Conversational Overlaps

Figure 3.5[c] shows the schematic of a conversation starting with long overlaps and continuing into a discussion. Long simultaneous speech will have change points at the group turn scale range. Different group patterns will have different temporal scales of conversations. The conversation that represents the switch into another state, is an indication that arguments have stopped and speakers have resumed discussions.

3.3.4 Short Speech Exchanges

Short speech exchanges happen when the speaker turns are rapid. The rapid speech might be due to some disagreement or small talk between the speakers, or a discussion that is in progress, in which, every speaker has very little to say. The ratio between the the speakers is less compared to short exchange speech.

In Figure 3.5[b], short speech exchange scenarios are shown, which are prominently found across the conversations that tend to be quarrels or disputes. Short exchanges would be detected at shorter scales as the conversation length of an individual speaker or a group of speakers would be short. Hence, the peaks will show up at speaker turn scale range, and if the conversation length falls in the scale range of group turns, then peaks would also be detected at that scale.

3.4 Synopsis

In this chapter, we describe two types of conversational turn patterns. The focus of this work is to detect group turns. The group turns are a turn-taking behavior in a group having a conversation. The changes to the conversations in the group turn will yield information on how the conversation was going, i.e, in terms of states like discussion, argument and presentation. Even though the detection states are still a problem, detecting group turn patterns will provide valuable insight into the group conversational pattern. The most commonly occurring group turn patterns can be found in everyday conversations where there is overlap, intrusions into someone's speech, question asking and agreeing with another person's speech. The group turn pattern will provide an idea as to which groups were conversing the most, leading to the dominant group of the speakers in the conversation.

CHAPTER 4

AUDIO-VIDEO CHANGE SCALE-SPACE

4.1 Features Used

We use both audio and video features to build the an audio-video change scale-space. The audio signal is captured at 44 kHz and was processed to extract 23 Mel-Frequency Cepstral Coefficients (MFCC) at a 30 Hz rate. This was done to bring the temporal dimensionality of the audio to the same rate as the video. MFCCs are then projected into principal component analysis (PCA) space to obtain a 23 dimensional feature vector.

The video frame rate is 30 Hz at a 720 x 480 resolution. We use the gray-scale difference image, as it is computationally less expensive. and performs better than optic flow on our data, where the inter frame displacements are high [33]. We down sample the difference image by a factor of 20. The image dimensions are further reduced to 23 by PCA to correspond with the dimensionality of the audio features. We combine the audio video features to create a combined audio and video feature set. These combined features are then projected into another PCA space. We normalize the audio and video by the covariances in the respective spaces. Figure 4.1 shows a video broken into audio and video features. We have the MFCC features extracted for an audio stream and shown as an image (left of features). We have the image difference features shown as an image (right of features).

4.1.1 Mel-Frequency Cepstral Coefficients (MFCC)

The audio features used in the process of building audio-video change scale-space are MFCC features. MFCC features are the dominant features used in any speech recognition system [34]. The features have the ability to represent the speech amplitude spectrum in a compact form. This has been the reason for their success. They are derived from a type of cepstral representa-

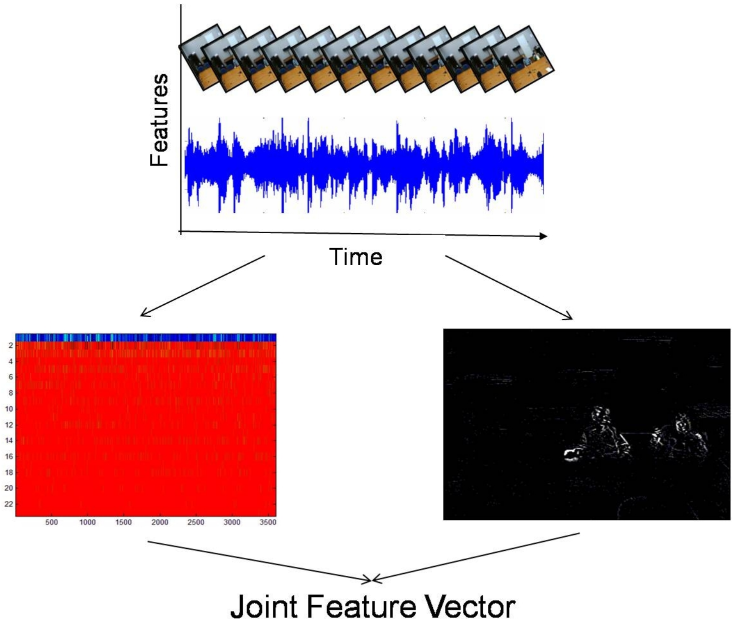


Figure 4.1: Features used in group turn segmentation. Audio features are Mel-frequency Cepstral Co-efficients (left) and video feature are the image differences (right). Using audio-video nonverbal cues, a multi-temporal scale conversation change is grouped into speaker and group turns.

tion of the audio stream. The cepstrum is formed by taking the Fourier transform of the audio spectrum. In mel-frequency spectrum, frequency bands are equally spaced on the mel-scale. This is a better approximation of speech signal than linear spaced frequecy bands [35]

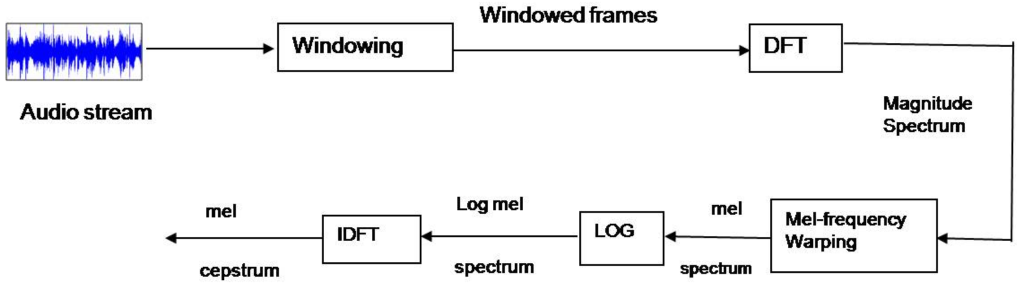


Figure 4.2: MFCC extraction flow diagram. The fourier transform of the input is taken, which is an audio signal. Mapped to the Mel scale and logs. A discrete cosine transform is performed to obtain the required number of MFCCs.

Figure 4.2 shows the process of extracting the MFCC features. The small speech signal sections that are statistically stationary are modeled. The window function is typically a

Hamming window. This removes the edge effect. DFT of the signal is taken and mel-frequency warping is done. A "mel" is a unit of special measure or scale of perceived pitch or frequency of a tone and is linear when freq is less than 1kHz. The frequency axis is warped according to the mel-scale.

Steps involved in calculating MFCC features for an audio stream include the following

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel-scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

The MFCC features are calculated at 30 frames per second. This is done to match the temporal dimensions of the video. Only the first 23 [2] MFCC features are considered. These features are then projected to a PCA space, but the dimensions are not reduced. Section 4.1.2 provides the necessary description of calculating principal component analysis.

4.2 Proposed Approach

The first step in the proposed method involves acquiring MFCC (audio) features and Image Difference (video) features. They are joined together to form the joint feature vector. the speech signal is then segmented into homogeneous segments - i.e, segments containing similar speech segments. In the first segmentation, the temporal window or scale used to calculate the segments is 3 seconds. The next step is to increase the scale at some interval and segment at each scale. This builds up a scale-space of segments and is called Audio-video change scale-space. This shows the change of different speakers having the floor at different scales. This scale-space is then broken up into two different turn scale ranges. The next step is to detect the group turns. The change values in the range of group turn patterns is summed to get the group turn BIC curve.

The audio-video change scale-space (ACSS) is a two-dimensional function over time (t) and scale (σ) whose value is given by

$$\mathrm{ACSS}(t,\sigma) = \Delta\mathrm{BIC}(t-\sigma, t+\sigma) \tag{4.1}$$

where, $\Delta\mathrm{BIC}(t-\sigma, t+\sigma)$ is the change in Bayesian Information Criterion (BIC) value between considering a single multivariate Gaussian model for the MFCC coefficients, X, over the time window $t-\sigma$ to $t+\sigma$ versus separate Gaussian models over $t-\sigma$ to t and over t to $t+\sigma$. We used the single Gaussian BIC representation as described in [20].

$$\begin{aligned} \Delta BIC(t-\sigma, t+\sigma) &= \frac{\sigma}{2}\left(\log|\Sigma_{X_{t-\sigma,t}}| + \log|\Sigma_{X_{t,t+\sigma}}|\right) \\ &\text{-}\sigma\log|\Sigma_{X_{t-\sigma,t+\sigma}}| - \sigma\lambda(d + \tfrac{d(d+1)}{2})\log N \end{aligned} \tag{4.2}$$

where λ is a penalty term (typically chosen as 1.0), d is the number of MFCC coefficients, and $|\Sigma_X|$ is the determinant of the covariance of the sequence of vectors, X. A large value of the ΔBIC indicates change in statistics.

We build a multiple scale BIC curve, $U(t)$, using:

$$U(t) = \sum_{\sigma \in groupturn} ACSS(t,\sigma) \tag{4.3}$$

The summed BIC curve was then low pass filtered and salient peaks were extracted. A peak is considered salient if it exceeds both its neighboring minima by $\alpha\tau$ [36], where τ is the standard deviation of the BIC curve and α is a constant. Higher the peaks, the higher the scale at which there was a change point. Low peaks are considered to be a speaker turn scale range change point. Higher peaks are considered to be at group turn scale range. Two close group turn peaks indicate the detected simultaneous speech.

Taxonomy of conversation change in Figure 4.3 illustrates the proposed method. The Bayesian Information criterion (BIC) is used to segment the joint feature vectors into segments. These segments are durations of the clip in which the audio and video information are similar. This segmentation is for the lowest scale (a temporal window) possible in the speech segmentation i.e, 3 seconds. Build a multi scale representation of the joint-feature vectors by

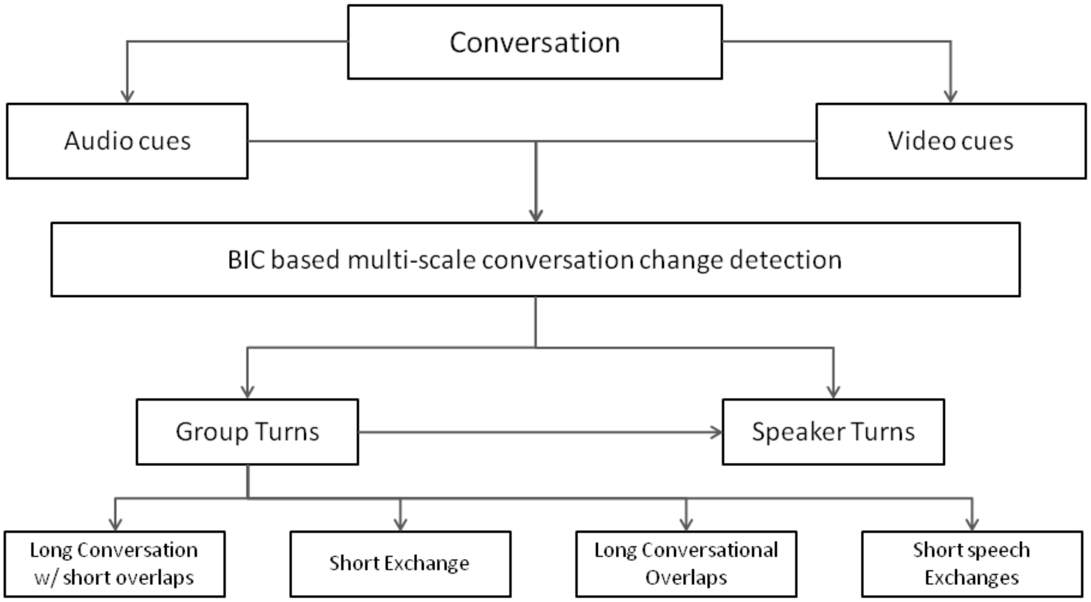

Conversation
Audio cues
Video cues
BIC based multi-scale conversation change detection
Group Turns
Speaker Turns
Long Conversation w/ short overlaps
Short Exchange
Long Conversational Overlaps
Short speech Exchanges

Figure 4.3: Taxonomy of conversation change. Using audio-video nonverbal cues, a multi temporal scale conversation change is grouped into speaker and group turns, and commonly occurring conversations for group turns are shown.

segmenting at each scale. The multi-scale representation of the features is broken up (with respect to scale) into two different regions – group turn and speaker turns.

Section 2.1 describes the segmentation procedure for a single temporal window (scale). The next two steps of building a scale-space and breaking the scales into different turns is described in sections 2.2 and 2.3 respectively.

4.2.1 Single-Scale Representation

The first step in building the multi-scale representation is the description of the model selection criterion to detect changes in statistics, the Bayesian Information criterion (BIC) [37]. Estimating the dimension of a model is a model selection criterion widely used in statistics. Using this model selection criterion, the changes to the joint audio-visual stream are detected. The audio signal is sampled at 16kHz, and the Mel-Frequency Cepstral Coefficients (MFCCs) are extracted using 32 filters with the bandwidth ranging from 166 Hz to 4000 Hz. These settings are chosen from a study [38] that found them to produce good results in speaker recognition and verification tasks. The MFCCs were extracted at 30 Hz, to match the temporal dimensionality of the video. The video features, which intend to capture motion, are obtained by image differences (three frames apart). The difference images are thresholded to suppress jitter noise, dilated by a 3 x 3 circular mask, down sampled from the original size of 480 x 720 to 48 x 72 and are vectorized. These features are then projected onto PCA space, to reduce dimensionality. The audio and video coefficients are joined together by multiplying features by a scaling factor, which is used to make the variances of the audio features equal to those of the video features.

The Bayesian Information Criterion (BIC) was introduced for speaker change detection in [20]. Consider a speaker loses the floor to another speaker or another group of speakers at a time instant t. To determine whether time instant t corresponds to a change-point, a time window (σ) preceding t is compared to a time window (σ) following t. The frames in the two windows are modeled parametrically, and if two sets of the frames in the corresponding window are deemed to be generating different models, the time instant t represents a change-point. The BIC, given a set of data $X = x_1, ..., x_N$, selects the model that maximizes the likelihood of the data. Since the likelihood increases with the number of model parameters, a penalty term proportional to the number of parameters d is introduced to favor simpler models. The BIC for a model M with parameters θ is defined as

$$BIC(M) = \log p(X|\theta) - \frac{\lambda}{2} d \log N \tag{4.4}$$

where λ is the penalty term (ideally equal to 1) and N is the number of feature vectors. The problem of determining the change point, which indicates a speaker change at the lowest scale, can be converted to a model selection problem. As given in Equation 4.2, the Gaussian BIC represents the models. Accordingly, there are two possible hypotheses. If we assume a unimodal Gaussian model for a speaker or a group of speakers having the floor, then the null hypothesis would be

$$H_0 : (x_{t-\sigma}, ..x_t.., x_{t+\sigma}) \sim N(\mu_o, \Sigma_o) \tag{4.5}$$

The alternate hypothesis is that two different models are needed to illustrate the data in each window.

$$H_1 : (x_{t-\sigma}, ..., x_t) \sim N(\mu_1, \Sigma_1)\ and\ (x_t, ..., x_{t+\sigma}) \sim N(\mu_2, \Sigma_2) \tag{4.6}$$

A positive value for the BIC justifies the alternative hypothesis and suggests that the time instant σ is a change-point.

The next step in speaker turn is to detect the peaks that are actually the individual speaker change point. In previous works, speaker turn is identified by removing all the overlap speech and running a silence detector to delete all the silence frames. This causes every change point to be an individual speaker change point.

This single-scale segmentation detects only speaker turn patterns. When a speaker turn is switched to a group of speakers, the single-scale BIC detects that change as a speaker change and not a group change. At a particular instant of time, if more than one speaker is speaking, there is overlap speech. This overlap speech is incorrectly segmented as a speaker change by the speaker diarization algorithms. Although it outperforms methods based on symmetric Kullback-Leibler (KL2) and generalized likelihood, the single-scale BIC method still fails in the case of overlap speech. In order to address this overlap speech problem, a multi scale representation is built. Even though this representation does not segment who and how many speakers were speaking in a group turn segment, it separates group turns from speaker turns.

4.2.2 Build Scale-Space: A Bottom-Up Approach

In section 4.2.1, a single-scale speaker change is explained. Detecting speaker turn using a single scale works only when the overlap speech and the silence frames are removed. But this is not ideally the case in many speech exchange systems. There is bound to be some amount of overlap speech involved in the conversations. In order to solve this problem, overlap speech sections has to be detected. Usually, very small overlaps do not matter as they indicate a

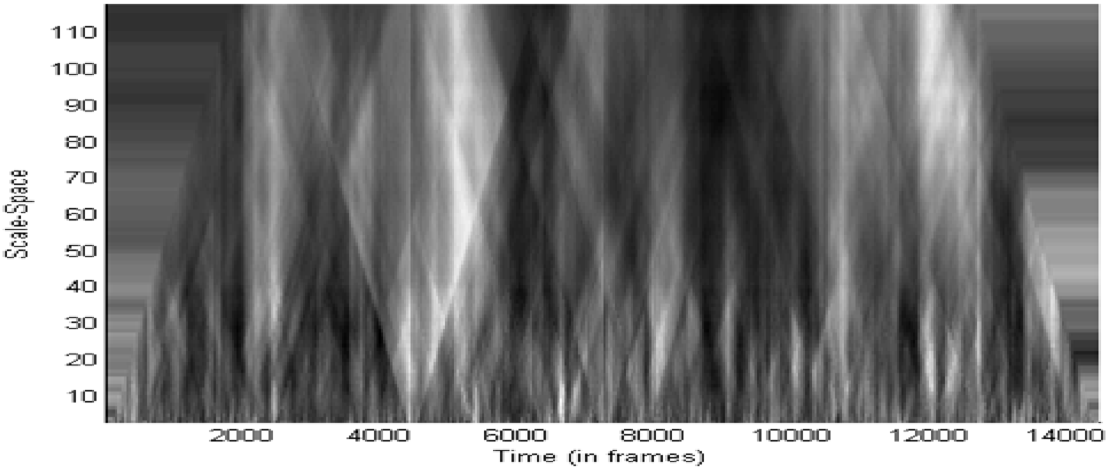
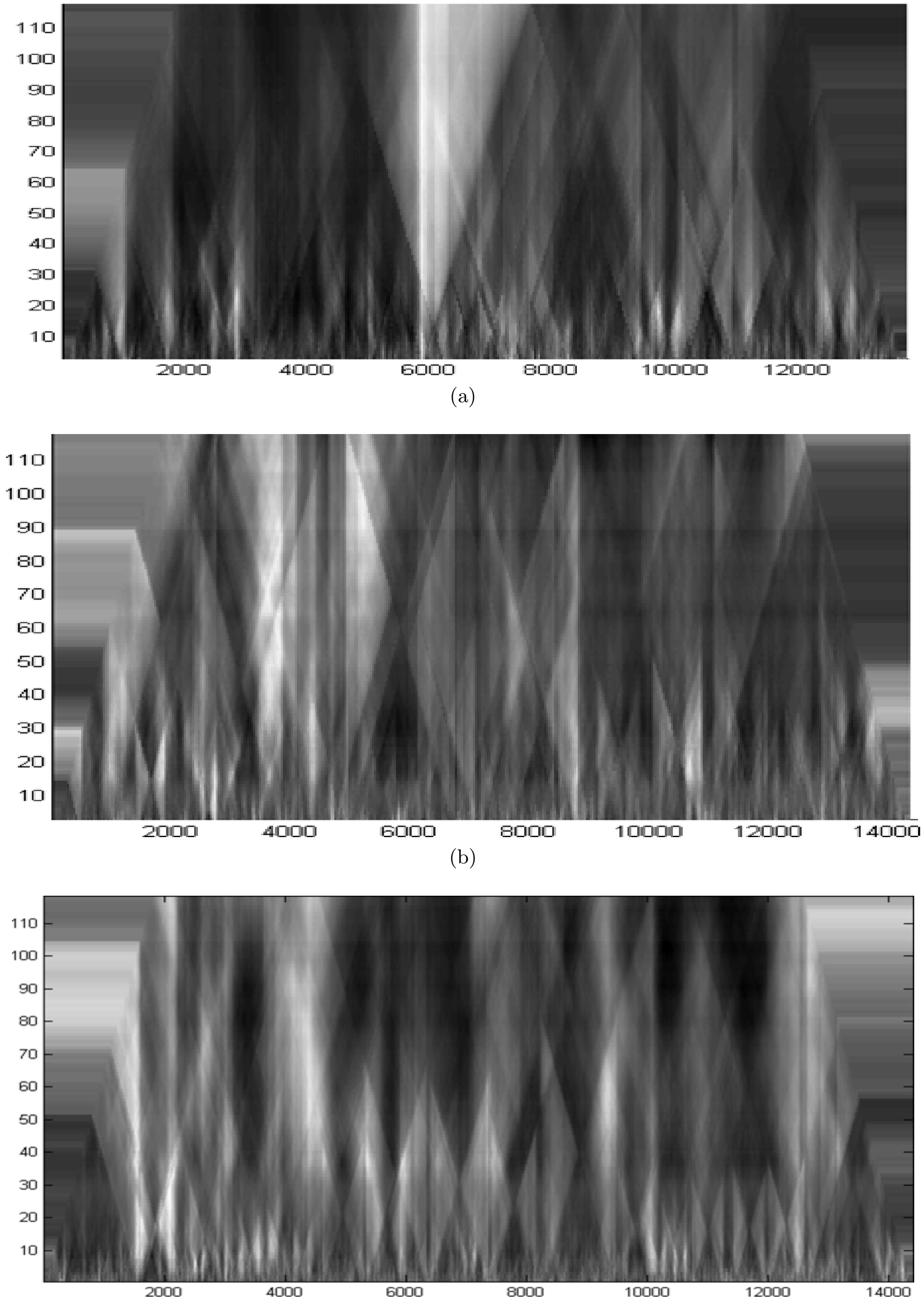


Figure 4.4: An example Audio-Video change scale-space. Time - Horizontal axis and scale - Vertical axis.

responder's agreement or disagreement to the communicator. These type of overlap speeches are collaborated with words such as "Hmmm", "Yes", "No". The goal is to detect overlap speech sections that have a group of speakers speaking together without having the floor. In section 4.2.2, an explanation on maximum scales used for each type of turns is presented.

In most recent works, there has been an effort to detect multi speaker speech activity [39]. Unsupervised learning of overlapped speech from multiple channels is one of the techniques used. But the basic problem of who is speaking at what time during overlap speech is still a problem. As there is a call for more robust and accurate speech activity detection systems, multi-party conversation in general, is currently the focus of much attention. One such approach is building a conversation change scale-space, which provides a model selection framework for detecting speaker change points at multiple temporal scales, using nonverbal features from audio and video. In order to build an audio-video change scale-space, a purely bottom-up approach is used.

The audio-video change scale-space (ACSS) is defined as a two dimensional function over time (t) and scale (σ), whose value is given in Equation 4.1. This scale space is visualized in Figure 4.4. It is a two dimensional plot of change against scale and time. Brightness indicates the likelihood of change for that time frame, at a particular scale. The brighter the value, the more likely is the change.



(a)

(b)

(c)

Figure 4.5: Three different Audio-Video change scale-space. The change scale-space is shown as intensity changes. Bright intensity corresponds to the likelihood of conversation changes at a particular scale (vertical axis) and time instant (horizontal axis) . This change scale-space is for an 8-minute conversation.

The entire conversation scale-space can be considered as capturing the texture of conversation. There are many short conversation patterns inside this texture. In Figure 4.5, examples of audio-video change scale-space are shown. There are three scale-space patterns shown. Each of the video, for which the scale-space is shown, has its own texture of conversation. To understand this scale space visualization, a good approach would be to look at all the brightest spots in the images. These indicate statistical changes in the joint-feature vector. As the scale goes longer, there are less time instants where the image is bright. This is because all the small speaker changes become statistically insignificant at coarser scales. In the finer scales, or the shorter scales, there tends to be more bright spots because of speaker changes happening.

In Figure 4.6, two different group turn patterns are shown. Each group turn and the corresponding window where the group has the floor is shown. The first group of speakers (red) has more speaker turns, which indicates more exchanges. When this type of conversation is involved, there are a lot of bright spots on the scale-space. The scale of the bright spots for this turn is dependent on the length of change of the speaker holding the floor. In the second group of speakers (blue), there are only three changes of speakers holding the floor. This is indicative of a single dominant speaker involved in the conversation.

4.2.3 Scale-Space Break-Up

In chapter 3, a discussion of two types of turn patterns is provided. These turn patterns can be identified in the audio-video change scale-space as scale ranges. For each turn pattern, a specific number of scales is allocated. This is a break of the scale in terms of speaker and group turns. In Figure 4.7, a conceptual break-up of the entire scale range is shown. The conceptual division of the scale-space into regions where speaker and group turn patterns can be detected.

Speaker turns requires scales in the finer lengths. Finer scales capture small changes in statistics and provide accurate change points for speaker change detection. The group turn scale range is decided depending on the length of the video. The speaker turn pattern range is small and is equal to 6 seconds [2]. In the case of scale-space break-up, the speaker turn range is increased from 3 seconds to 6 seconds. This is done because, even though to recognize a speaker turn requires only 3 seconds, an additional 3 seconds is alloted to deliberately miss all

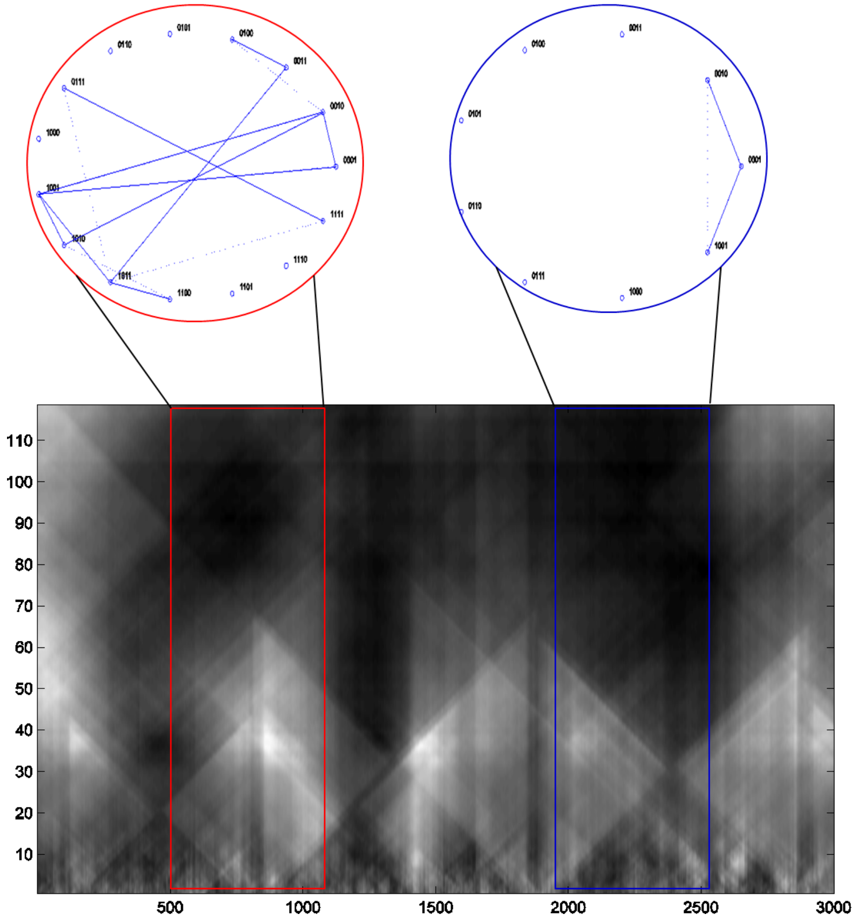


Figure 4.6: Two different group turn patterns with ground truth. The brightest spots indicate change. The first group turn (red) involves a different group of speakers loosing and gaining the floor. The second group turn (blue) involves a completely different set of speakers in the conversation. Scale - vertical axis, Time - horizontal axis.

the overlap speech or silence within the range of those extra 3 seconds. This results in coarser scales being assigned to the next level of turn patterns.

The next scale range is assigned to group turn patterns. In this type of turn patterns, coarser length change in a single speaker or group of speakers who have the floor is detected. The group turn scale range starts as the speaker turn scale range ends. In the dataset being used, all the videos are of the same length. As a result, the maximum range of this turn pattern is fixed. Ideally this range increases as the length of the video increases. Higher the scale range coarser is the detection of turns.

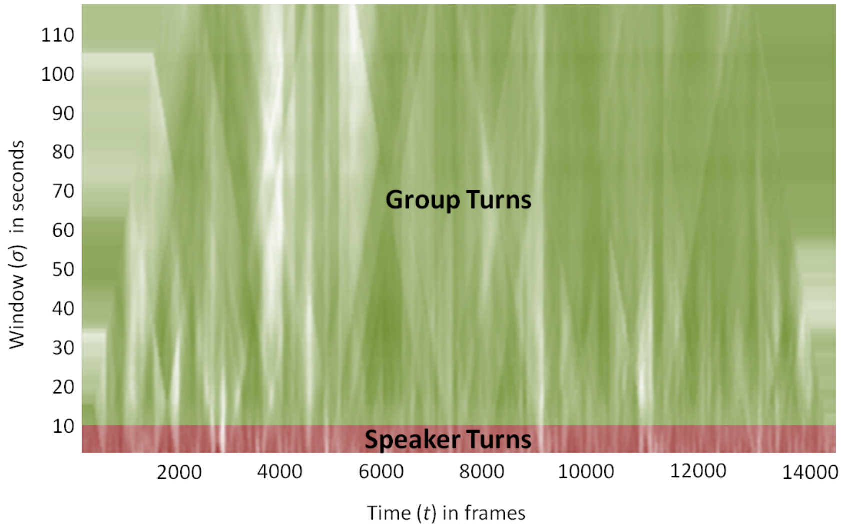


Figure 4.7: Conceptual break-up of a Audio-Video change scale-space. The change scale-space is shown as intensity changes. Bright intensity correspond to likelihood of conversation changes at a particular scale (vertical axis) an time instant (horizontal axis) . This change scale-space is for an 8 minute conversation.

Once the group turn scale range is decided, the next step is to detect the group turn patterns. To detect group turn pattern, at each instant of time, the BIC values of the entire group turn scale range is summed using the Equation 4.3. This gives a summed BIC curve that represent scales of only group turns. Change points are detected in this summed BIC curve in order to get the group turn. The summed BIC curve is then low pass filtered and salient peaks were extracted. A peak is considered salient if it exceeds both its neighboring minima by $\alpha\tau$ [36], where τ is the standard deviation of the BIC curve and α is a constant.

Figure 4.7 illustrates the conceptual break-up of this technique on an excerpt where five people conversing. The black areas indicate the time instant in which a single speaker or a group of speakers were speaking.

4.3 Synopsis

In this chapter, segmentation of conversation according to group turns is discussed. In order to segment, we build a scale-space representation of the entire conversation. This gives the temporal structure of the conversation and the changes taking place at various lengths. The segmentation, at each scale, is done using the Bayesian Information Criterion. The length of the temporal window, which is the scale, is varied to build the scale-space. The scale-space is then broken up into speaker and group turn ranges. This representation uses audio-visual joint features to build up a scale-space. A discussion of MFCC features is also added in this chapter.

CHAPTER 5

DATASET AND RESULTS

5.1 Dataset

The proposed approach is tested on a subset of the NIST pilot meeting-room corpus [40]. The corpus contains 19 meetings recorded in a room rigged with five cameras and four table mounted microphones. Of the five cameras, four are fixed cameras mounted on each wall of the room facing the central table, and the fifth is a floating camera that zooms onto salient activities in the scene such as the speaker, the white board, or the projection screen. Of the four table microphones, three are omni-directional microphones, and one is a 4-channel directional microphone. The meeting room setup is shown in Figure 5.1. Each participant in the room also wears a head microphone a directional lapel microphone.

The database available to us contains the five camera feeds for each meeting and the videos have a spatial resolution of 720 x 480 sampled at 29.97 Hz. There are two audio channels packaged with each video; one is a gain-normalized mix of all the head microphones, and the second is a gain-normalized mix of all the distant microphones. The audio streams are sampled at 44 kHz and has a resolution of 16 bits per sample. Of the 19 meetings, three meetings were excluded from the experiments because two of them did not have associated ground truth and the third consisted entirely of a presentation by one person. Audio-visual pairings are considered for each meeting by pairing each of the fixed cameras with one of the audio channels, resulting in 15 meeting clips. From each video (5-15), the first 90 seconds are discarded, and the next 8 minutes are chosen resulting in approximately 7 segments in each video. Videos (1-4) in Table 2.1, are randomly chosen, from each of the videos. We have $7x15 = 105$ segments.

In all of the meetings, participants are seated around a central table and interact casually. Depending on the type of the meeting, the participants discuss a given topic, plan events, play

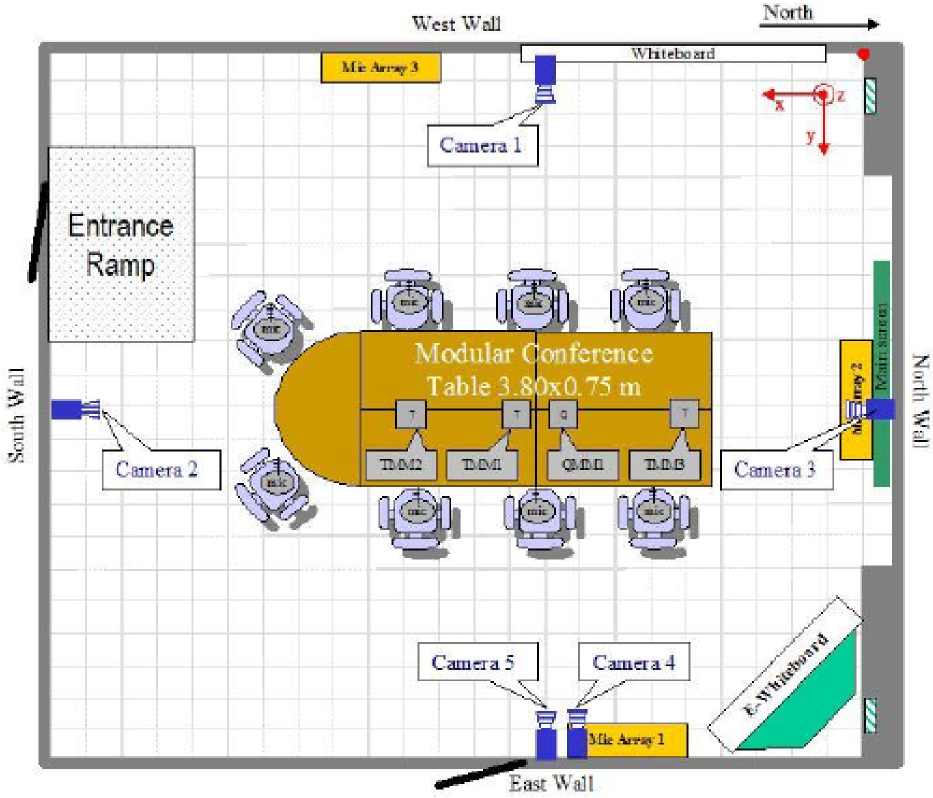


Figure 5.1: The NIST meeting room setup. Meetings are recorded using four fixed cameras, one on each wall and a floating camera on the east wall [40]. The audio is recorded using four table microphones and three wall mounted microphone arrays in addition to a lapel and a microphone for each participant.

games or attend presentations. From time to time, participants may take notes, stretch, and sip drinks. In some of the meetings, participants leave their chairs to use the white board or distribute materials. The audio and video signals from these meetings are quite complex because the meetings are unscripted and of long durations. Since only a single camera view is considered at a time, most faces are non frontal and sometimes participants are only partially visible. In some meetings, a participant may not be visible at all in a particular camera view. Even when the faces are frontal, they are often occluded by the person's own hand. Similarly, the audio signal is complex, consisting of short utterances, frequent overlaps in speech, and non-speech sounds such as wheezing, laughing, coughing, etc. Sample images of all clips from the dataset are shown in Figures 5.2. In order to quantitatively characterize the meetings in the dataset we use the following variables:

Table 5.1: Analysis of dataset. Analysis with respect to number of speakers and speaker entropy (speaker dominance) given by speaker speaking per time instant.

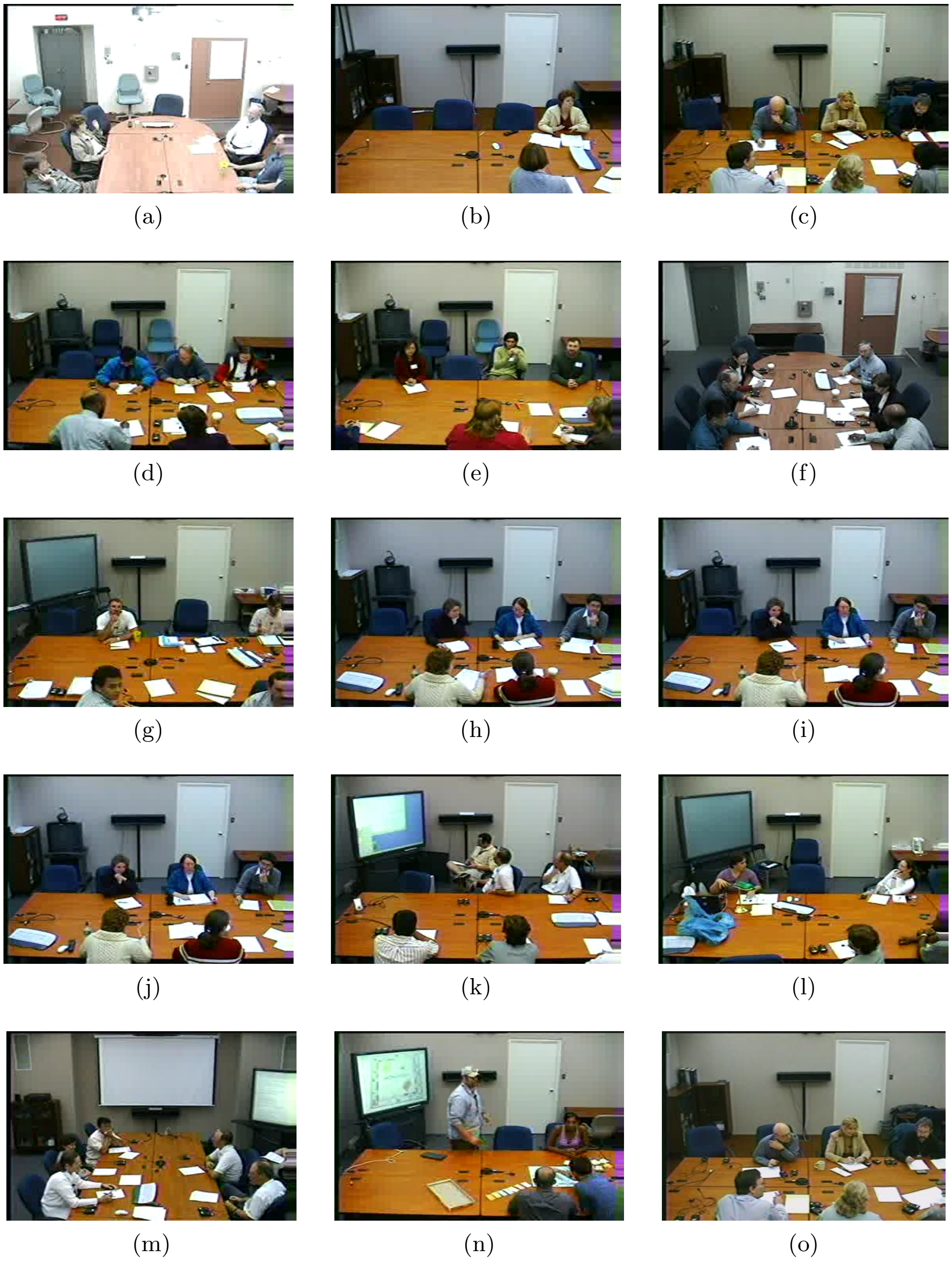
No.	Video Name	Type	No. of Segments	Speakers	Entropy
1	NIST_20020627-1010	Staff Meeting	7	6	2.13
2	NIST_20020305-1007	Planning	7	7	1.62
3	NIST_20020815-1316	Problem solving	7	4	1.63
4	NIST_20020213-1012	Planning	7	6	1.91
5	NIST_20011115-1050	Focus Group	7	4	1.25
6	NIST_20011211-1054	Planning	7	3	1.28
7	NIST_20020111-1012	Planning	7	6	1.37
8	NIST_20020213-1012	Staff Meeting	7	6	1.87
9	NIST_20020214-1148	Interaction w/ expert	7	6	1.85
10	NIST_20020304-1352	Game playing	7	4	1.59
11	NIST_20020305-1007	Planning	7	7	1.79
12	NIST_20020627-1010	Staff meeting	7	6	2.26
13	NIST_20020731-1409	Game playing	7	4	2.03
14	NIST_20020815-1316	Problem solving	7	4	1.61
15	NIST_20020904-1322	Interaction w/ expert	7	4	1.93

- Number of participants in the room (Speakers): We assume this metric will correlate with the video performance - the greater the number of speakers, the closer together they will be seated, leading to occlusions. Secondly, the overall amount of distracting motion may also be proportional to the number of speakers. Also, although the number of speakers may not directly correlate with audio performance, larger number of people results in increased background noise due to the various sounds that listeners make when they swivel in their chairs, tap the table or flip pages.

- Speaker Entropy (Entropy): This is a measure of speaker domination in a meeting. A low entropy indicates that only a few speakers speak for most of the time, whereas a high entropy indicates that the participants involved spoke more or less for about the same duration. The entropy is computed as

$$H_{conversation} = -\sum_{i}^{n} P(S_i)\log(P(S_i)) \tag{5.1}$$

$$P(S_i) = \frac{d(S_i)}{\sum_{i}^{n} d(S_i)} \tag{5.2}$$



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Figure 5.2: A frame from each clip from the dataset. The frames are from the same camera for all the dataset. This camera view (b-o) is from the west wall [40].

where N is the number of speakers involved in the meeting, d(S_i) is the total time duration for which person S_i speaks and P(S_i) is the percentage of time (i.e, probability) that person S_i speaks during the meeting.

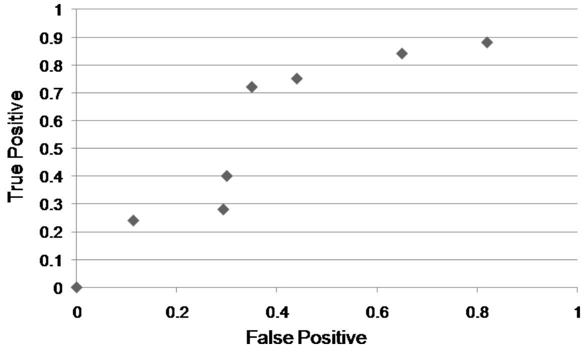
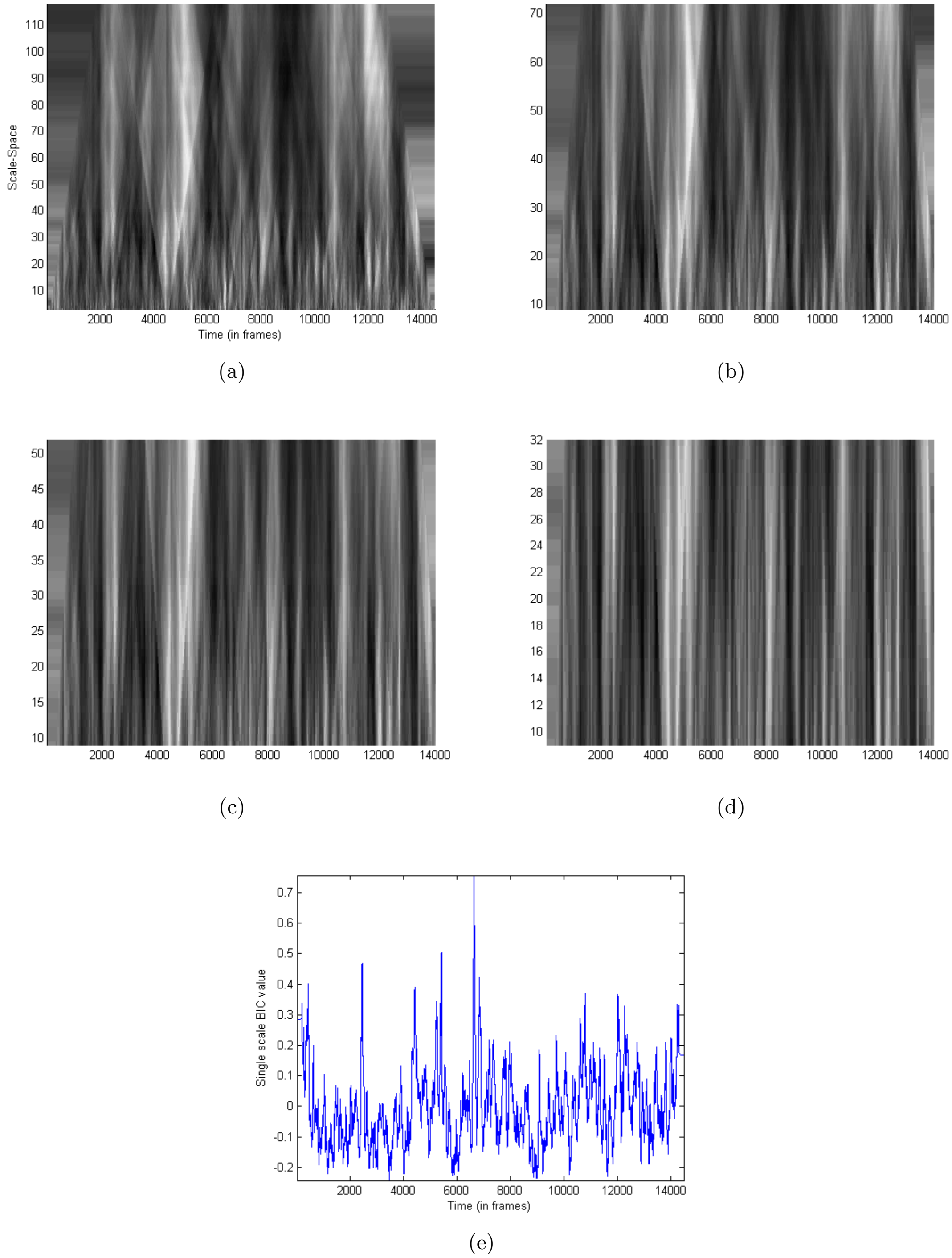


Figure 5.3: True and false positive plot for variable threshold. The threshold points correspond to peaks selection done for values ranging from max(U) to $max(U)/16$. The highest value with low false positive is for a threshold of $max(U)/3$. Their respective true positives and false positives are shown.

5.2 Results

5.2.1 Group Turn Scale Range

One of the first results in the group turn pattern detection is the visual comparison of changes in the entire scale-space with only the group turn region. The speaker turn range is eliminated and the group turn range is chosen. Figure 5.4 shows the comparison of the entire scale-space with different scale ranges of group turn. The result is that all the speaker turn patterns are eliminated and only the group turn pattern is considered. As shown in Figure 5.4, the changes are less dense than the original scale-space. This shows a higher abstraction of the conversation. The changes indicate a change in group turn rather than a single turn. The Figure 5.4 also shows that when the range of the group turn is reduced (b-d), distinguishing between group turn and speaker turn becomes harder. This is due to the scale range reducing to a single-scale BIC curve (e). The similar result is quantitatively analyzed later in the chapter.



(a)

(b)

(c)

(d)

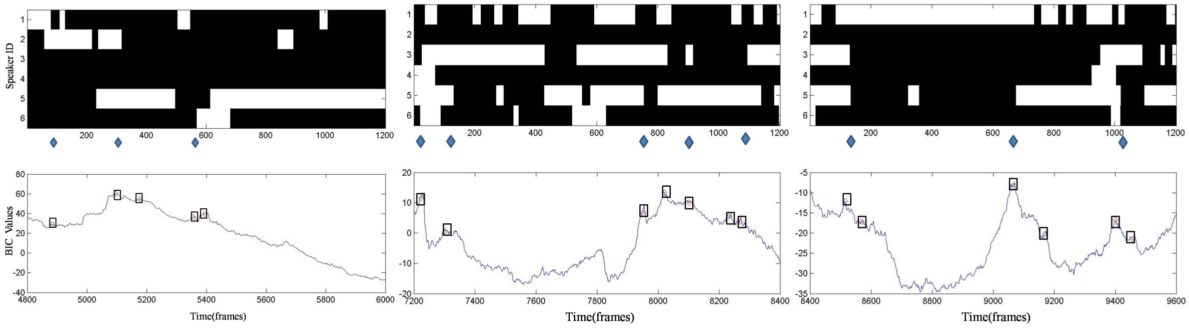
(e)

Figure 5.4: Reducing range of group turn scales. (a) is the entire scale space, (b-d) is the group turn scale-space reduced by 20 from 72 scales at (b) to 32 scales at (d) respectively. (e) is the single-scale BIC curve.

5.2.2 Peak Cut-Off in Group Turn

Cut-off value is in terms of maximum value of U, obtained from Equation 4.3, for each segment. The cut-off value greater than max(U)/2 has the lowest true positive and false positive rates. The cut-off value greater than max(U)/3 has the highest percentage of true positives identified for all the segments. In Figure 5.3, a true positive false positive plot for varying threshold values with a fixed group turn scale range is shown for the entire dataset. The plot shown in Figure.5.3, is the average values for the entire dataset.

Before group turn patterns can be detected, the entire scale space must be broken down into small equal segments, which are then marked group turn ground truth. In Figure 5.5, there are 3 different types of segments in which group turn patterns detected are shown. Each of these three results is a segment and consists of audio ground truth (top row), change scale-space and multi scale change points detected on the summed BIC curve. For each segment, the group turns were manually marked on the ground truth. They are marked by diamond shaped markers under the ground truth of the segments. Every change point detected as the best group turn, in the segment, is 1 second behind or ahead of the ground truth marked group turn. There are other change points detected in the same segment, which is the result of the change point from the previous segment. The cumulative rates of different cut-off values for peak selection are used for each clip.



(a) (b) (c)

Figure 5.5: Three different segments showing group turn patterns. Peaks (bottom) and the corresponding ground truth (top) is shown with hand marked group turn patterns (diamonds). In the audio ground truth (top), a black patch represents a speaker having the floor.

5.2.3 Group Turn Scale Range Analysis

In this section, we discuss the changes to the detection of group turns with respect to the scale chosen. The scale range of a group turn starts from 6 seconds in the scale range. This starting scale is chosen because the changes to the speaker turn can be identified at this scale [28]. For all the segments in the dataset, the maximum scale range for group turn is chosen at 60 seconds. This scale is reduced continuously and the group turn range detection is calculated. The highest scale (sigmaMax) chosen was 40. The peak detection rates are calculated at scale 40 and then the range is increased by one scale on either side of sigmaMax. The maximum range of the group turn is sigmaMax+sigmaMax/2 to sigmaMax-sigmaMax/2. The detection rate remains the same for a few ranges of scales, then changes. This is due to the nature of the conversation and the length of the conversation. As shown in Figure 5.6, the true positive and false positive plot for the scales at which there was a change in the detection rate is shown. This plot is for the entire dataset whose accuracies are averaged over all the videos. The lower the scale change, the fewer are the false positives and true positives. When the range is finally reduced to a single scale, the reduction is almost zero, as none of the group turn change points occurring in the conversation are detected. This also suggests that every group turn pattern is a speaker turn pattern and not vice-versa.

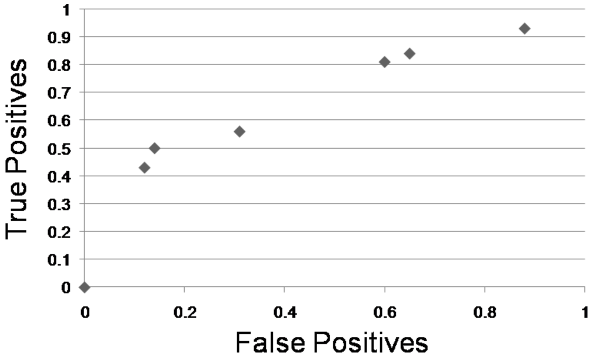
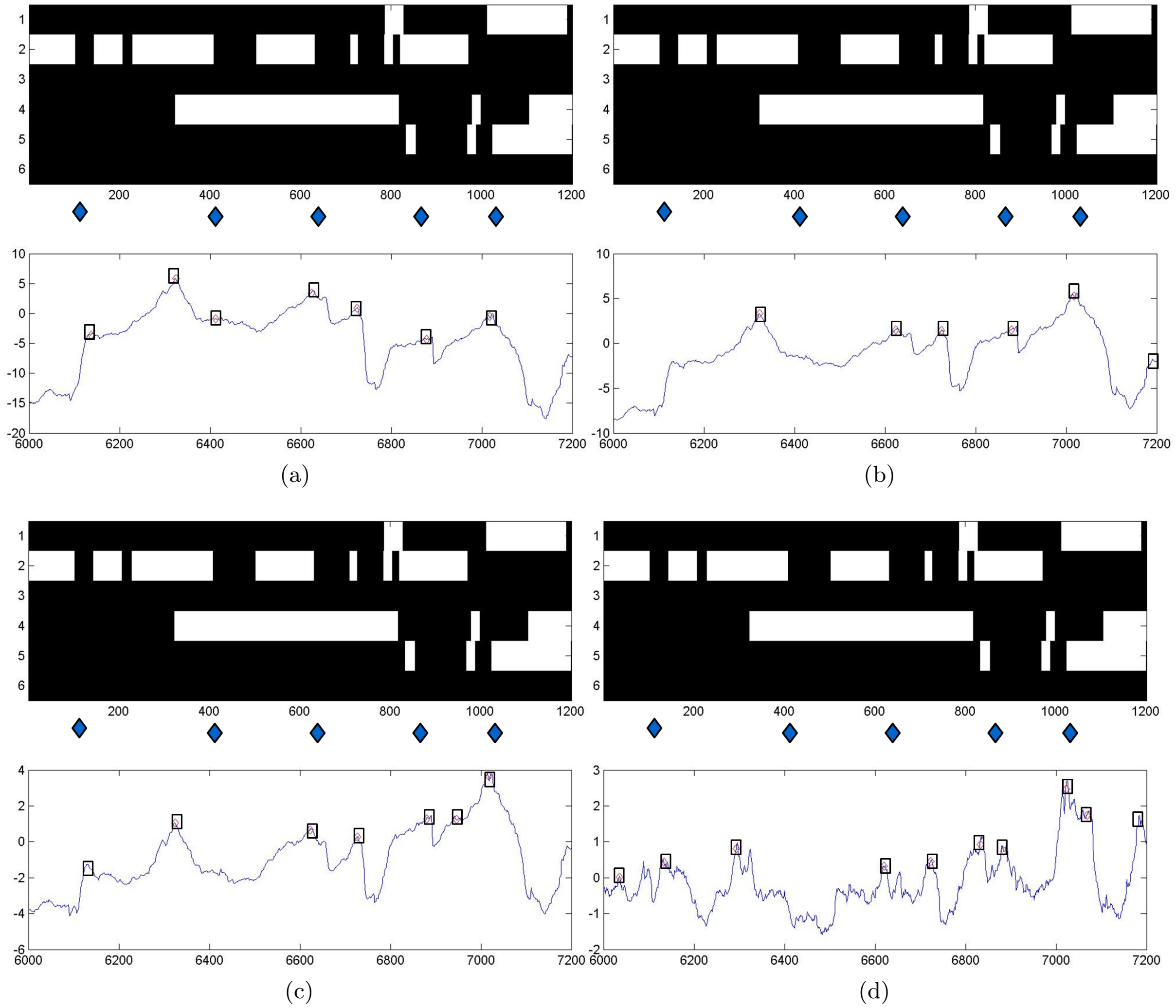


Figure 5.6: True and false positive plot for different group scale ranges. The changes are at the above points, which indicate the scale at which there was a change. The lowest scale will have fewer false positives.



(a)

(b)

(c)

(d)

Figure 5.7: Different segments of group turn with different scale ranges. Peaks (bottom) and the corresponding ground truth (top) is shown with hand marked group turn patterns (diamonds). In the audio ground truth (top), a black patch represents a speaker having the floor.

CHAPTER 6

CONCLUSIONS

6.1 Conclusions

The separation of speaker turns from group turns provides a higher-level abstraction of meeting room conversations. The speaker turn involves the detection of persons speaking individually, but the group turn provides detection of groups of speakers speaking together, indicating overlap speech. The group turn is an intermediate-level feature that can be used in the video classification and retrieval. It can also be used as a playback feature, in order to go to that section of the conversation in which a particular group of speakers is speaking.

Conversations are a set of complex turn patterns. Recognizing conversation changes using speaker and group turn patterns provides a rich description of meeting room conversations. We presented a novel representation, an audio-video change scale-space, that provides a snapshot of the conversations at multiple temporal scales, built in a purely bottom-up fashion. We demonstrated how this representation is used in automated group turn detection in meeting segments. There are two types of turn patterns described in this work, speaker and group turn patterns. We need two types of turn patterns to effectively describe a conversation change. The focus of this thesis is in distinguishing the speaker turn and the group turn.

6.2 Future Work

Future uses of this representation could include automatically generating rich descriptions of meetings by capturing multiple temporal scales. It is a intermediate-level feature to describe the conversations. The detection of another level of hierarchy to group turn is one of the main focuses in the future. Another focus is toward classification of different group turn patterns, in order to classify the conversation in different states such as presentation, discussion, argument, break and even silence.

REFERENCES

[1] S. Eggins and D. Slade. *Analysing casual conversation.* Equinox, 1997.

[2] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi. Audio segmentation and speaker localization in meeting videos. *In Proc. of ICPR*, 2:1150–1153, 2006.

[3] R. Krishnan, S. Sarkar. Detecting group turn patterns in conversations using audio-video change scale-space. *To be presented in ICPR*, 2010.

[4] M.R. Naphade and T.S. Huang. Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. *IEEE Transactions on In Neural Networks*, 13:793–810, 2002.

[5] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathouda, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *In Proc. of the IEEE Int. Conf. on Multimedia (ICME)*, 2006.

[6] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, P.Wellner D. Moore, and H. Bourlard. Modeling human interactions in meetings. *In Proc. of ICASSP*, 3, 2003.

[7] Dong Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings: A two-layer hmm framework. *In Proc. of the IEEE CVPR. Workshop on Event Mining*, 2004.

[8] S. Reiter, B. Schuller, and G. Rigoll. Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. *IEEE Transactions on In Neural Networks*, 13:793–810, 2002.

[9] M.R. Naphade and T.S. Huang. Segmentation and recognition of meeting events using a two-layered hmm and a combined mlp-hmm approach. *IEEE Transactions on In Neural Networks*, 13:793–810, 2002.

[10] T. Teixeira, J. Deokwoo, G. Dublon, and A. Savvides. Recognizing activities from context and arm pose using finite state machines. *In Proc. of third ACM/IEEE International Conference on Distributed Smart Cameras*, 1–8, 2009.

[11] Z. Yu, H. Aoyama, M. Ozeki, and Y. Nakamura. Collaborative capturing and detection of human interactions in meetings. *In Proc. of of PERVASIVE*, 65–69, 2008.

[12] Z. Yu, Z. Yu, Y. Ko, X. Zhou, and Y. Nakamura. Inferring human interactions in meetings: A multimodal approach. *In Proceedings of the 6th International Conference on Ubiquitous Intelligence and Computing*, 14–24, 2009.

[13] A. Nijholt, R.J. Rienks, J. Zwiers, and D. Reidsma. Online and off-line visualization of meeting information and meeting support. *The Visual Computer*, 22:695–976, 2006.

[14] O. Brdiczka, J. Maisonnasse, P. Reignier, and J.L. Crowley. Detecting small group activities from multimodal observations. *Applied Intelligence*, 30:47–57, 2009.

[15] Peng Dai, Linmi Tao, and Guangyou Xu. Audio-visual fused online context analysis toward smart meeting room. *IEEE Transactions on In Neural Networks*, 13:793–810, 2002.

[16] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1557–1565, 2006.

[17] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: ”who responds to whom, when, and how?” from gaze, head gestures, and utterances. *In Proc. of the ninth international conference on Multimodal interfaces*, 255–262, 2007.

[18] A. Dielmann and S. Renals. Dynamic bayesian networks for meeting structuring. *In Proc. of ICASSP*, 5:629–632, 2004.

[19] A. Dielmann and S. Renals. Dynamic bayesian networks for meeting structuring. *IEEE Transactions in Multimedia*, 9:25–36, 2007.

[20] S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. *In Proc. DARPA Speech Recognition Workshop*, 127–132, 1998.

[21] S. Know and S. Narayanan. Speaker change detection using a new weighted distance measure. *In Proc. of the ICSL*, 4:2537–2540, 2002.

[22] M. Kotti, E. Benetos, and C. Kotropoulos. Automatic speaker change detection with the bayesian information criterion using mpeg-7 features and a fusion scheme. *In Proc. of the IEEE International Symposium on Circuits and Systems*, 2006.

[23] M. Kotti, L.G.P.M. Martins, E. Benetos, J.S. Cardoso, and C. Kotropoulos. Automatic speaker segmentation using multiple features and distance measures: a comparison of three approaches. *In Proc. of the IEEE International Conference on Multimedia and Expo*, 1101–1104, 2006.

[24] S. Duncan. Some signals and rules for taking speaker turns in conversations. *Journal of Personality and Social Psychology*, 2:283–292, 1972.

[25] N. Campbell, T. Sadanobu, M. Imura, N. Iwahashi, S. Noriko, and D. Douxchamps. A multimedia database of meeting and informal interactions for tracking participant involvement and discourse flow. *In Proc. of the Language Resources and Evaluation Conference (LREC)*, 2006.

[26] N. Campbell and D. Douxchamps. Processing image and audio information for recognizing discourse participation status through features of face and voice. *In Proc. of the INTERSPEECH*, 2007.

[27] C.M. Adda-Decker, P.P.G. Adda, B.M. Philippe, and H. Benoit. Annotation and analysis of overlapping speech in political interviews. *In Proc. of the Sixth International Language Resources and Evaluation (LREC)*, 2008.

[28] J. Jaffe and S. Feldstein. *Rhythms of dialogue.* New York Academic Press, 1970.

[29] L.M. Dabbs, Jr., and R.B. Ruback. Vocal patterns in male and female groups. personality and social psychology bulletin. *Personality and Social Psychology Bulletin*, 518–525, 1984.

[30] A.J. Sellen. Speech patterns in video-mediated conversations. *In Proc. of SIGCHI*, 49–59, 1992.

[31] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. *In Proc. of ICASSP*, 4353–4356, 2008.

[32] S. Renals and D. Ellis. Audio information access from meeting rooms. *In Proc. of ICASSP*, 4:744, 2003.

[33] F. Quek, D. McNeill, R. Ansari, X.F. Ma, R. Bryll, S. Duncan, and K.E. McCullough. Gesture cues for conversational interaction in monocular video. *In RATFG-RTS*, 1999.

[34] L. Rabiner and B. Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc., 1993.

[35] S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Readings in speech recognition*, 65–74, 1990.

[36] S.S. Cheng and H.M. Wang. A sequential metric-based audio segmentation method via the bayesian information criterion. *In Proc. of Eurospeech*, 945-948, 2003.

[37] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[38] T. Ganchev, N. Fakotakis, and G. Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. *In Proceedings of the 10th International Conference on Speech and Computer, SPECOM*, 1:191–194, 2005.

[39] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at icsi. *In Proc. of the Human Language Technology Conference*, 1–7, 2001.

[40] M. Michel, J. Ajotand, and J. Fiscus. The nist meeting room phase II corpus. *In Proc. of MLMI*, 3:1–3, 2006.