Dominant Speaker Identification for Multipoint Videoconferencing

Ilana Volfin

Dominant Speaker Identification for Multipoint Videoconferencing

Research Thesis

As Partial Fulfillment of the Requirements for
the Degree Master of Science

Ilana Volfin

Submitted to the Senate of the Technion—Israel Institute of Technology

Elul 5771 Haifa September 2011

Acknowledgement

The Research Thesis Was Done Under The Supervision of Professor Israel Cohen in the Faculty of Electrical Engineering.

I would like to express my gratitude to Prof. Israel Cohen for the supervision, guidance and support throughout this research.

I would also like to thank my family, Miryam, Gershon, Benny, Masha and Danniel, for your endless love and support.

The Generous Financial Help Of The Technion Is Gratefully Acknowledged.
This research was supported by the ISRAEL SCIENCE FOUNDATION (grant no.1130/11).

Contents

List of Figures

List of papers

- I.Volfin and I.Cohen, "Dominant Speaker Identification for Multipoint Videoconferencing," submitted to Computer Speech and Language

Abstract

A multipoint conference consists of N participants received through N distinct channels. Dominant speaker identification is the task of identifying which of the participants is the most dominant at a given time. The identification facilitates reducing the computational load on the communication network, on which the conference is conducted. In addition, it enables the conferees to focus their attention on the most active participant. The techniques applied to this problem so far, rely mostly on instantaneous measures of speech activity, with the underlining assumption that a switch in speaker is characterized by a rise in this activity. These methods neglect the prolonged properties of dominant speech, thus making them prone to false speaker switching. They also fail to discern between speech and non-speech transient audio occurrences.

In this thesis, we propose a dominant speaker identification method that reduces the number of false speaker switches and is also robust against transient audio occurrences. The proposed method is based on speech activity evaluation on time intervals of three different lengths. The speech activity for immediate, medium and long time intervals is evaluated and represented by three respective speech activity scores. In the process of score evaluation, we propose two models for the likelihood of the detected audio activity. One for the assumption that speech is present in the time interval and the other for speech absence. The unique set of scores acquired for each channel, facilitates the identification of the dominant speaker. The identification is based on a comparison of scores across all channels.

Objective evaluation of the proposed method is performed on a synthetic conference with and without the presence of transient audio occurrences. The performance is evaluated in terms of the time segments in which a false dominant speaker was identified. A segment of falsely detected dominant speaker usually follows a false speaker switch.

In this evaluation framework, the errors reflect the total number of false switches, the location of the falsely detected segment within the dominant speech burst and its duration. A qualitative evaluation on a segment of a real multipoint conference is conducted as well. The proposed method is compared with existing methods of dominant speaker identification. We achieve reduction in the number of false speaker switches and improved robustness against transient audio occurrences.

Abbreviations

AR	Auto Regressive
DFT	Discrete Fourier Transform
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
FCFS	First Come First Served
FEC	Front End Clipping
HMM	Hidden Markov Model
IID	Independent Identically Distributed
IP	Internet Protocol
ISDN	Integrated Services Digital Network
LPC	linear predictive coding
LR	Likelihood Ratio
LSA	Log-Spectral Amplitude
LT	Loudest Talker
MC	Multipoint Controller
MCU	Multipoint Control Unit
MFCC	Mel-frequency cepstral coefficients
MMSE	Minimum mean squared error
MP	Multipoint Processor
MSC	Mid Sentence Clipping
MS/I	Multi-Speaker/Interrupter
PCM	Pulse Code Modulation
PSD	Power Spectral Density
PSTN	Public Switched Telephone Network
SA	Spectral Amplitude

SE	Squared Error
SNR	Signal to Noise Ratio
SP	Spectral Power
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TFB	Tandem Free Bridge
TFO	Tandem Free Operation
VAD	Voice Activity Detection/Detector

Chapter 1

Introduction

1.1 Background

Videoconferencing was first presented at the 1964 World's Fair. This innovative technology required dedicated phone lines in a time when even having a phone at home was not taken for granted. In addition to the expensive system, line usage costs would reach as high as hundreds of dollars per hour. Today, almost five decades later, the technologies have immensely evolved. High speed processors at a relatively low cost and the migration of communication networks to the Internet, had brought these collaborative videoconferencing technologies not only into every house but even onto our palm-held smartphones.

Videoconferencing services are provided in different areas of life. Supporting the trend of globalization, it enables close cooperation between groups and individuals from around the world. In medicine, high level medical consultancy is provided to patients in remote or isolated areas. Videoconferencing also simplifies communication and collaboration in the academic community. Students get the opportunity to participate in lectures from leading researchers in their fields. Rare fields of knowledge may be conserved by allowing the existing research groups to collaborate. Specialists in very specific areas may be easily reached when needed for consultancy. Finally, it provides the services most commonly associated with videoconferencing. These are the business meetings and personal communications.

Videoconferencing may be conducted on the telephone, Internet or a hybrid type of network that combines between the two. Different architecture of connections were sug-

gested for a conference conducted over a network. These include the centralized and decentralized. In centralized architecture, the end-points are connected through a central control unit. This unit receives signals from all end-points, conducts signal processing operations and mixes the output signals that are sent back to the end-points. The decentralized architecture is further divided into *full mesh* and *multicast*. In full mesh, all end-points are connected and each end-point sends its output to all other end-points. In multicast, each end-point sends its output to a conference multicast address, from which it is forwarded to all other end-points. It is clear from the description above, that the conduction of a multiple participant conference, would have high resource demands from the system. This, for any chosen architecture. Hereon, we discuss the centralized architecture. However the discussion is valid for other architecture types as well.

The amount of information produced by N participants of the conference is large. All this information needs to be routed through the network and processed by the central control unit. The videoconferencing demands, of even a modest size conference, are in fact so high, that the advanced capabilities of the modern networks reach its limit. In order to provide the multipoint videoconferencing abilities, a solution is required for discarding unnecessary information. Such solution is provided by *speaker selection*.

In speaker selection, M most active speakers out of the total N participants are selected. The private case, where $M = 1$ is the *dominant speaker identification* problem. After selecting the M most active participants, part or all the information originating in the non-active participants may be discarded. The conference benefits from discarding information in several aspects:

1. Reducing the load on the network and central processing unit

2. Removing noise that originates from open idle microphones

3. Concentrating the conferees attention on the most dominant participants

There are two typical approaches to dominant speaker identification. The first follows the intuitive interpretation of dominance. This group of algorithms is often referred to as *loudest speaker* algorithms. The second approach is related to human conversational habits.

Dominance is often associated with loudness. Thus, many simple dominant speaker identification methods compare a measure of the signal in each channel, where the 'loudest' speaker is selected as the dominant one. The common measures are signal amplitude or power [1–3]. This approach might work in a case where the SNR in each channel is very high and there are absolutely no irrelevant disturbances, such as coughing, sneezing etc. In more realistic conditions, the instantaneous power of a noise signal on one of the channels can be higher than the power of a speech signal. This happens mostly during breaks between words, causing these methods to frequently switch away from the dominant speaker to the noisy channels. One of the solutions that was proposed for frequent switching is only allowing a speaker switch once in a time window of one or two seconds. In fluent speech, this time is equivalent to several words or even a short sentence. Thus, it may cause a wrongful switch to last at least one or two seconds or entirely miss a correct speaker switch.

Although we raise our voice, signaling our interest to engage in the conversation, naturally we anticipate when the current speaker is about to finish before barging in. This suggests taking a different approach to dominant speaker identification. It concentrates on detecting the points in the conversation where a switch in dominant speakers occurs. One of these methods is the Multi-Speaker/Interrupter (MS/I) algorithm [4]. In the first step, a voice activity detector (VAD) is employed and the participants are ranked according to the order of becoming active speakers. The VAD decision in this method is derived from either the signal power or the arrival of silence insertion descriptor (SID) frames. Next, a speaker is only allowed to be promoted in ranking if the ratio of its smoothed signal power in respect to a higher ranked speaker, passes a *barge-in* threshold. The barge-in mechanism reduces the number of false switching by introducing the demand of elevated *relative* activity into the switching process. This method however, does not address false switching due to transient audio occurrences. In fact, the barge-in mechanism would fail exactly in the case of such occurrences.

In this thesis we discuss the centralized multipoint videoconferencing. It consists of N participants received through N distinct channels, where each participant is using a single microphone and camera. In the discussed arrangement, only one speaker appears on the video display of each participant. The conference is administrated through a central

control unit, such as the multipoint control unit (MCU).

Voice Activity Detection and Dominant Speaker Identification

The VAD determines for each signal frame, whether it contains speech or not. It is a fundamental component in many speech signal processing applications as well as in communications. In speech enhancement, VAD is used for estimation of noise statistics. This information is then used to reduce the noise in following frames, where speech is present. In communications, the knowledge regarding speech presence may be used in order to only transmit signal frames that contain speech, thus conserving bandwidth.

VAD is in base a binary classification problem, where its two classes are *speech is present* and *speech is absent*. A general VAD consists of two main components, signal frame representation and a classification method. The representation is a set of temporal or frequency-domain properties of the signal frame, their combinations or functions of them. When choosing a representation, the aim is to maximize the separation abilities with the available system resources. When the whole frame representation is brought down to a single value, simple thresholding may serve as the classification method. If the frame is represented by a large set of features, the classification is often obtained using machine learning techniques, such as the support vector machine (SVM).

Voice activity detection is a closely related task to dominant speaker identification. Many VAD methods are using speech specific features. These methods are specifically oriented towards detecting speech, rather than assuming that each 'loud' frame is speech, as done by the simple methods. This is both an advantage and a disadvantage for using speech specific VADs in the dominant speaker identification problem. On the one hand, speech specific activity detection is necessary for robust dominant speaker identification. On the other hand, simultaneous instantaneous audio activity may exist on more than one end-point. In this case, the instantaneous information is insufficient for determining which of the end-points transmits the dominant speech. Although VAD methods cannot be used as-is, the conceptual closeness of the two can contribute to creating dominant speaker identification methods that are voice activity detection in a broader sense.

1.2 Motivation and Goals

Speaker selection is a vital component in a videoconferencing system. Common approaches for speaker selection, rely on a simple measure of signal level in the channel, as measured by its power or amplitude [1–3]. Dominant speech activity is characterized by prolonged speech activity during which these measures vary considerably, including near zero levels during breaks between words. In fact, the non-dominant transient disturbances, are typically characterized by high energy, which leads to frequent false speaker switching. The difference between the dominant and non-dominant audio activity, which is often related to temporal duration, is not captured by such instantaneous measures. More advanced methods [4, 5], make use of a *barge-in* mechanism, in order to reduce the frequent false switching. In these methods, the speakers are ranked according to their instantaneous audio activity. A promotion in ranking is allowed only if the instantaneous audio activity, or the temporally smoothed version of it, on a certain channel passes a certain barge-in threshold. These methods provide a means for a speaker to barge into a conversation by raising his voice, in the same time providing a loophole for a false switch due to transient audio occurrence that is characterized by concentrated burst of energy.

The goal of this thesis, is to develop a dominant speaker identification method that is robust against false speaker switching. Another goal is to design an algorithm that is robust against non-dominant transient audio occurrences. We exploit the natural properties associated with dominant speech. These are related to speech structures of different typical lengths. Our goal is to represent each signal frame in terms of presence or absence of these structures in its surrounding. This information about each frame on a distinct channel is then passed on for comparison across all channels. Based on this representation, the comparison stage may better discriminate between isolate transient audio occurrence on one channel from fluent speech activity on a dominant channel. The dominant speaker identification results of the proposed method, outperform methods that are based on a comparison of an instantaneous audio activity measure, both in the number of false speaker switches and in robustness against transient audio occurrences.

1.3 Overview of the Thesis

In this thesis, we propose a novel method for dominant speaker identification. The proposed method assigns speech activity scores to each signal frame, according to speech activity evaluation on three time intervals of different length. Each time interval of a certain length, is assumed to consist of a sequence of smaller sub-units. The speech activity in each time interval is determined by the number of its active sub-units. The activity in a sub-unit is determined by a thresholding operation, where the sub-unit is considered active if its value is above the threshold. For each interval length, two likelihood models are assumed on the number of active sub-units under the assumption of speech presence or absence respectively. Speech activity scores for each of the three time intervals are obtained from the ratio between those two likelihood models. The scores refer to the level of speech activity in the immediate medium and long time intervals that precede and include the currently observed signal frame. The obtained set of scores, provides a representation of speech activity which enables the discrimination between isolate audio activity, that is typical of non-dominant activity, and fluent audio activity that is characteristic of dominant speech activity.

We show in the experimental results, that by using information from time intervals of different lengths, the number of false speaker switches is reduced, compared to existing methods. We also show that the proposed method is more robust to false switching due to transient audio occurrences.

1.4 Organization

This thesis is organized as follows. In Chapter 2, we briefly present the subject of videoconferencing. We discuss the various types of videoconferencing, types of communication networks on which videoconferencing is conducted and the accepted architectures of connections between the conference end-points. Finally, we focus on the multipoint videoconferencing as it is the particular interest of this thesis. In Chapter 3 we present two popular speech processing applications, namely voice activity detection (VAD) and speech enhancement. In that chapter, we focus on the elements that contribute to the

development of the proposed dominant speaker identification method, later in this thesis. In Chapter 4, we introduce the dominant speaker identification method that makes use of information derived on time intervals of different length. We describe in detail the components of the proposed method and we test its performance under different conditions. Finally, in Chapter 5 we conclude our work and propose directions for future research.

Chapter 2

Videoconferencing Systems

2.1 Introduction

Videoconferencing brings people together for a collaborative meeting in a variety of applications. It is currently being used for education and learning, healthcare and medicine, conference meetings and personal communication. The use of videoconferencing saves time and money for organizations by eliminating travel costs. It also allows people in remote regions or rural areas the access to professional services such as medical consultation or remote education that would not be available to them otherwise.

Since the first *Picturephone* had been introduced by Bell Labs at the 1964 World's Fair, videoconferencing has changed immensely. The picturephone required dedicated lines and the service was too expensive for the general use, hence it was discontinued soon after its release. The next videoconferencing system was offered to the public in 1982, by Compression Labs., the system was priced $250,000$ with a line cost of $1,000$ per hour. Over the years, the costs of videoconferencing systems and network usage, had dropped by about an order of magnitude every five years. Following the rapid technological evolution in communication networks, hardware and the widely available high speed Internet, today videoconferencing services are affordable and widely available. Commercial systems are available to be purchased or hired, primarily for organizations. The PC based videoconferencing systems were introduced as an extension to the traditional videoconferencing, but these gain more and more popularity as an independent form of collaborative communication. Some popular software solutions, such as Skype or MSN messenger, are

available for the home and small business users, and it is only a matter of time for a viable videoconferencing solution to be available for the smartphone market.

Videoconferencing technology has several advantages over a face to face meeting. One advantage of this technology is that it is green, in accordance with the growing initiative of many governments and organizations. Videoconferencing helps in reducing the air pollution and energy consumption, produced by the transportation means that are involved in delivering each participant to the meeting location. In addition, since participants are not required to travel to the meeting destination, they participate in the conversation with a higher energy level, which makes a more efficient use of the conference time. Another advantage is that in contrast to a physical conference meeting, setting up a videoconferencing session is much simpler. It requires much less preparation time, since there is no need to coordinate flights or accommodation for the participants. Finding a convenient time-slot to hold the conference is much easier when the meeting cost is low and several conference sessions can be held to resolve all open questions. It is also simpler to add a "last minute" participant to the videoconference.

Another use of videoconferencing is in services such as remote education and medical consultance, that are delivered to remote and rural areas. Remote education is practiced in countries such as Canada, Australia and the US, where individuals or small groups of remote students gain access to high level education. In the virtual classroom, the basic interaction between teacher and students is augmented by interactive presentation of the material, such as graphical demonstrations, provided by the resources of the conference environment [6]. In addition, the instantaneous access to resources on the web combined with communication with team members and experts through videoconferencing connections [7], creates a participatory learning experience which is known as increasing knowledge retention [8].

Videoconferencing attempts to provide a satisfactory replacement for an actual face to face meeting. As such, it has to take into account the human factors involved in a group interaction. It had been noted in researches, that knowledge and ideas are best transferred and assimilated when the technology provides a *participatory* environment. Participants, while dispersely located, should be able to see, hear and interact with other participants in as natural fashion as possible.

Videoconferencing provides an efficient and cost-effective solution, when the option of an actual meeting is unavailable. In some scenarios it might even be preferred over the physical meeting [9].

The chapter is organized as follows. In Section 2.2 different types of videoconferencing are presented, varying in the number of participants in each terminal, the network infrastructure it is operating on, and the type of connections between the end-points. Section 2.3 elaborates on multipoint videoconference and its most important components, the MCU and the speaker selection methods.

2.2 Types of Videoconferencing

In this section, we describe the differences between types of videoconferencing technologies in three aspects:

1. The number of participants at each conference end-point
2. Communication infrastructure used for the conference
3. Arrangement of the connections between the end-points

2.2.1 Group vs. Personal conferencing

In the aspect of the number of participants in each end-point of the conference, videoconferencing is classified into two types *group* and *personal* conferencing.

Group videoconferencing

In group videoconferencing, the activity of a group of conferees is transmitted to other participating end-points, the system used for this purpose can be further divided into *set-top* and *dedicated* systems [10].

Set-top videoconferencing systems These are complete conference systems that sit on top of a cart, or in a small boardroom. These systems are portable and are suitable for small groups. Set-top systems are used for business and administration meetings, small group educational sessions and for telehealth clinical sessions.

Dedicated systems These systems are built into a room and are not meant to be moved. These are usually the most expensive videoconferencing systems, also providing the highest quality service. These systems are ideal for large group conferencing, such as administrative meetings, distance education and medical conferences. Most of these systems support both ISDN and IP based end-points (discussed later in this chapter).

Personal videoconferencing

Personal videoconferencing is usually operated using a *desktop* system. These systems are typically used for a *one to one* communication, where every participant sits in a different location, also denoted as a *multi-point conference*. Personal videoconferencing systems are easy to use and do not require skilled maintenance personnel, as some group conference systems do. The hardware requirements, for a basic use, are satisfied by a common PC, these include a monitor, a single camera, microphone and speakers or headphones. A basic videoconference session can be set up using a PC and software, available for download from the Internet. The desktop videoconferencing systems are very popular for personal communications, provided by popular software applications such as Skype, MSN and Yahoo. More expensive solutions that provide a higher quality video and audio signals that are more fire-wall friendly, are available for medium to large size organizations, offered by companies like Polycom and Cisco.

2.2.2 Communication Infrastructure

In this section we discuss the two different types of communication infrastructure, typically used for conducting a videoconference: *ISDN* and *TCP/IP* based [10].

ISDN

The Integrated Services Digital Network (ISDN) is a communication standard for digital transmission of voice, video and data over the Public Switched Telephone Network (PSTN). Since the data is transmitted over telephone lines, there are several advantages for ISDN based videoconferencing. The telephone lines are not connected to the Internet, hence this is a more secure way of transmitting information. The bandwidth of an ISDN

line is guaranteed, which insures a continuous quality of service (QoS). The bandwidth of the connection can be increased by adding more lines. Hence, overall an ISDN connection provides a secure, high quality connection. There are some disadvantages in using ISDN. Adding an ISDN line in order to increase bandwidth, has an additional cost. In addition, the usage of the lines is charged per usage time. Overall, a videoconference, several hours long, with multiple participating cites, may be expensive. Another disadvantage is that establishing a multi-party ISDN call can be complicated and take up to 30 minutes, since each participant has to be called at a specific location.
Communication over the ISDN based networks are according to the ITU-T H.320 recommendation [11].

TCP/IP

Transmission Control Protocol (TCP)/ Internet Protocol (IP) is named after the two comprising protocols. The TCP/IP can be used over a heterogeneous network, connecting computers using different types of networks. TCP/IP network is cheaper to use, since once the connections are set up, it is used for no additional cost. The available high-speed IP connections include DSL, cable modem, satellite, wireless broadband and fiber optic. IP networks are also called packet-based, since the information is sent in *packages*, and the protocols that provide the audio-visual communication are defined in the ITU-T H.323 recommendation [12].

2.2.3 Conference Architecture

There are different architectures for videoconferencing, each design is evaluated in terms of perceived signal quality, scalability, controllability and compatibility with existing standards. We describe here three common conference architectures [13]. The *centralized*, where connections between end-points are conducted through a central processing unit, the *decentralized* architecture, in which data is transferred between end-points without a mediating bridge and the *tandem-free* architecture, which proposes a hybrid between the two aforementioned architectures.

Centralized

In this setup, the connection between end-points is provided by a *conference bridge* or *Multipoint Control Unit* (MCU). Each end-point connects to the conference bridge, and the bridge establishes a one-to-one data and signaling connection with each end-point. The bridge receives the audio data from each end-point, summing tailored output streams to return to each end-point.

In order to avoid the summation of background noise from non-active end-points, the conference bridge employs a *speaker selection* method to select M active speakers out of N participants. Thus, $M+1$ output sums are formed, one for each active speaker, excluding his own speech signal, and one more stream for the remaining $N-M$ unselected conferees. A selection algorithm usually chooses one to three signals for output, where the selection is based on a comparison of some signal properties between all end-points, specific methods are described below.

Conference bridges are available on two types of platforms.

1. Software based bridges, hosted on dedicated computers. These bridges are more appropriate for smaller conferences, scalability is achieved by sharing speech processing functions over multiple servers.

2. DSP based bridges, running on DSP based media-servers. These bridges are meant for large scale applications and scalability is achieved by adding DSP cards to the server.

The use of a centralized bridge causes reduction in signal quality due to the decoding and encoding operations, which are also referred to as *tandem* arrangement. Additional degradation results from the delay imposed by the signal processing. Trying to lessen the impact of the bridge on signal quality, several techniques were offered. These techniques, regarded as *Select and Forward* conference bridges, select the active speakers with a low level of signal decoding, and forward the signals to the destination end-points, avoiding or reducing the amount of signal processing in the mixing stage.

Forgie [14] selected a single speaker, using First come First served method (FCFS), thus only one signal was broadcasted to all participants.

Nahumi [15] achieved a voice activity detection (VAD) decision using partially decoded

signals. During single-talk, only one speech signal was transmitted to all participants, no decoding or encoding needed. During multi-talk, only the active speakers signals were decoded, mixed and encoded prior to redistribution. This reduces the number of processed signals from N to the M active ones.
Champion [16] selected and forwarded the signal of a single speaker during single-talk, and streams of primary and secondary speakers, selected by the FCFS approach, during multi-talk.

Decentralized

In a decentralized architecture, data is exchanged between end-points without using a centralized bridge. In this setup, speech quality is improved since data arrives directly from the origin without any mid-way processing. The presence of a single unit with processing abilities of a conventional bridge is not required, but the end-points are expected to be able to perform some signal processing operations on incoming signals. The decentralized conferencing further branches into the *Full Mesh*, and *Multicast* conferencing models.

Full Mesh In full mesh conference, every two end-points are connected. Each end-point transmits $N-1$ copies of its outgoing signals and receives $N-1$ signals, which translate, for the worst-case, to $N^2 - N$ streams through the network. The signaling control of the conference is usually centralized at a central server, to maintain conference control. Since each end-point has to at least be able to decode and mix $N-1$ streams, this arrangement suites a small scale conference, with a large bandwidth availability.

Multicast In a multicast conference, each end-point transmits a single copy of its stream to all other end-points, through conference multicast address, instead of sending out $N-1$ identical streams. On the input side, it receives $N-1$ streams, sent in the same way by the other end-points. This configuration conserves bandwidth by reducing the number of output streams.

Tandem Free Operation

The tandem free bridge is a hybrid between the signal degrading, *Centralized* and the bandwidth consuming *Decentralized* architectures. In this model, a tandem free bridge (TFB), is a multitalker select-and-forward conference bridge. Once the active talkers are identified by the bridge, M compressed signals are forwarded to the $N - M$ end-points where they are decoded and mixed. In this model, some of the parameters needed in the speaker selection process, can be encoded into the output stream of each end-point. Each end-point receives only up to M incoming packets, each arrival time interval, which reduces the workload on the end-points.

2.3 Multipoint Videoconferencing

Multipoint videoconferencing is a type of a *personal* videoconferencing with three or more participants, each in a separate location, communicating through ISDN or IP based network. The end-points are mediated by a conference bridge or an MCU. In this videoconferencing configuration, each conference participant is transmitting its audiovisual information to the MCU, the MCU decodes the incoming information from all end-points and performs conversion from different protocols, if needed. Audio and video processing is carried out, followed by re-encoding and transmitting an input signal back to every end-point.

2.3.1 MCU

The MCU provides the capability of three or more end-points to communicate in a multipoint conference. The MCU consists of two parts. A mandatory *Multipoint Controller* (MC), which controls the end-points. It is responsible for the signaling between end-points, and it controls the conference resources. The optional part of an MCU are the *Multipoint Processors* (MP), these units provide the centralized processing of video audio and data, in the multipoint conference [12].

Audio Processing

The audio signal, is the most important feature of a videconference. If the audio is poor, the cooperation between the participants will be poor [8].
The MCU is required to be able to process audio data in a variety of formats, with various data rates [17], where the most common translation ability is between H.323 (IP) and H.320 (ISDN) standards. The MCU has to be able to decode, mix and appropriately encode each data format, depending on the receiving end-point. The typical issues that are encountered during conference audio processing and need to be resolved are:

- Echo Removal - the signal originating from an end-point, has to be removed from the mixed input audio stream prepared for that end-point by the MCU. In some conferencing architectures, this task is performed by the end-point itself.

- Overflow - may occur when a number of signals are mixed, and the signal sum is greater than the allocated dynamic range. Some typical solutions for this problem are, for example [18], by *saturating* the sum result to the most positive or negative value that the system can represent, or *scaling* the incoming audio samples to prevent the overflow. If either of these methods is applied, it has to be taken into account in the echo-removal task.

- Background Noise - in a large conference, when the audio signals originating from open microphones of inactive end-points are mixed, they add up to distracting background noise. A solution to this problem can be by mixing only signals originating from end-points that pass a certain activity threshold.

In order to deal with the aforementioned issues, and for a more efficiently conducted conference, the MCU has to be able to perform the following tasks:

- Simple Mixing - combination of the audio signals from all participants for an output signal that is distributed to the participants.

- Selective Mixing - selective mixing is applied to prevent saturation, or to remove the background noise. The latter is achieved by speaker selection methods that are elaborated on later in this chapter. The selective mixing should be transparent to the participants.

- Private connections - upon request, the MCU provides a point-to-point connection between two end-points. This function is supported by the signaling abilities of the MCU.

Video Processing

The video stream, accompanying the audio, adds a natural dimension to the videoconferencing session. It allows the participants to meet face-to-face, it transmits facial expressions and body language. The processing of video typically requires more resources, compared to audio processing. In order to preserve the quality of the videoconference, it is also necessary to maintain the original synchronization between the audio and video signals.
The video streams are handled in one of two typical methods:

Selection Where a single video stream is selected to be forwarded to each participant. This does not require decoding and can be made relying on the audio signal.

Mixing Where each user can see more than one participant simultaneously. In this scenario, the MCU selects a subset of participants, by some criteria. The signals are decompressed and assembled into a *split-screen* image. The composed image is compressed and redistributed to the participants. This mode of operation, reduces the video quality due to decompressing-mixing-compressing operations an increases the delay that is caused by additional processing.

2.3.2 Active Speaker Selection

Processing the input from N channels can be a heavy and time consuming task for the conference bridge. In addition, the information passed through the network is partly redundant, and the conference may benefit from its removal. Thus, in a conference of N participants, M participants are chosen for output, usually with $1 \leq M \leq 3$.

Speaker selection algorithms must satisfy some requirements to ensure high speech quality (partly adopted from [19]):

- Audible Switching - the selection process should not introduce audio artifacts into the mixed output signal. Such artifacts may be caused by speech clipping, or different volume levels of differen speakers.

- Interactivity - the selection process should be transparent to the participants. It should preserve natural conversation events such as overlap between speakers or barging-in into the conversation.

- Resistance to Noise - the selection algorithm should reduce background noise and have some immunity against transient audio occurrences, such as paper handling, knocks, sneezing etc.

- Lack of Discrimination - loud and quiet participants should be treated equally, both should have the same opportunity being mixed into the output stream.

In the following, we present some popular methods for speaker selection. Although detecting speech presence by the signal level in the channel might also be considered a VAD, we differentiate between the two groups of methods. We refer to the first two methods as *Level Based* since the speakers are selected based solely on the signal level in the channel. We refer to the second set of methods as VAD based, since they employ the VAD as a first step, and the speaker selection is based on some additional criteria.

Level Based

The Level Based methods are a group of Loudest Talker (LT) algorithms, in which the speakers are selected according to speech energy, or loudness. In this approach, the speakers are ranked according to the level of the measured feature, once every certain time-interval. These algorithms are most popular for Pulse Code Modulation (PCM) systems, that are used in telephony based (PSTN) networks. Examples of such algorithms follow.

Speaker Selection based on Signal Power the method described in [3], ranks the speakers according to the power value. The speaker with the highest power is selected as current speaker and two additional participants are selected as active speakers. In order

to prevent frequent switching of the current speaker, the decision is made every $1-2$ seconds.

Speaker Selection based on Signal Amplitude in [2], the dominant speaker is identified as the speaker with the largest signal amplitude, only if the amplitude exceeds a preselected threshold level.

These algorithms are prone to frequent speaker switching, when two speakers have the same signal level simultaneously, or in the presence of transient non-speech audio occurrences.

Voice Activity Detection Based

The Voice Activity based methods, employ a VAD to detect speech activity in the current time-frame for each participant. The order in which the participants become active speakers, ranks the participants, where the M top ranked participants are mixed for output.

First Come First Served (FCFS) ranks the speakers according to the FCFS criterion. These methods employ a VAD to detect speech activity in a current time-frame. An active participant list keeps track of M most active speakers. When one of the active speakers stops speaking, he is removed from the list while the rest of the participants are promoted in ranking.

Multi-Speaker/Interrupter (MS/I) This algorithm was designed for tandem free operation (TFO) conferencing and is described in [4]. The MS/I algorithm assigns speaker privileges according to order of activity, assigned by a VAD, power signal envelope in time-frame i, $\hat{E}_i$ and a *Barge In threshold*, B_{th}. The spectral envelope is calculated as:

$$\hat{E}_{i+1} = \max(\hat{E}_i, \beta\hat{E}_i + (1-\beta)E_i) \tag{2.1}$$

where E_i is the instantaneous signal power, in frame i and β is a weighting factor. A barge-in threshold is used to prevent spurious switching between speakers with close $\hat{E}_i$'s.

A participant with priority m can be promoted to priority $m-k$ if

$$10\log\left(\frac{\hat{E}^m}{\hat{E}^{m-l}}\right) > B_{th}, \forall l = 1, ..., k, \quad k \leq m \tag{2.2}$$

After an active conferee stops talking, $\hat{E}_i$ is decayed exponentially, for a time period of T_h seconds, after which $\hat{E}_i = 0$. The spectral envelope, integrates long-term information, making a more robust decision regarding a speaker switch.

2.4 Summary

This chapter has reviewed the topic of videoconferencing. The different types of videoconferencing systems in three aspects were introduced, group vs. personal conferencing, circuit switched (ISDN) vs. packet based (IP), and the aspect relating to the architecture of connections between the end-points. The final videoconferencing system that is used by the consumer depends on his needs and the financial and technological resources at his disposal.

The multipoint videoconferencing topic was elaborated on, in section 2.3 in order to place it in the broader context of videoconferencing for future discussion in this thesis.

Chapter 3

Speech Processing Methods

3.1 Introduction

In this chapter we elaborate on two widely used speech processing applications, voice activity detection (VAD) and speech enhancement. These two problems had been attracting research efforts for several decades. We use some of the elements and general approaches from the discussed works in the course of this thesis. Hence, we saw fit to elaborate on them here.

Voice activity detection is a key component in speech processing and communication applications. The VAD decides, for each signal frame, whether it contains speech or not. It is known that during a telephone call, each speaker is only active for about 35% of the time [20]. Hence, detecting only the time periods in which the speaker is active allows turning off the transmission, or reducing the coding rate, during these periods [21, 22]. This may help in reducing the average coding-rate, co-channel interference [23], and extend battery life. In speech enhancement, the VAD decision enables estimation of noise statistics, during the non-speech periods [24]. Nosie statistics are used in the following frames, where speech is present, for noise reduction.

A VAD algorithm is formulated as a binary classification problem. The objective of a VAD is to determine for each signal frame if it contains speech or not. The first step in VAD algorithms is the representation of a signal frame. This may be by simple signal attributes such as power, amplitude or a frequency representation. Other representations are more complicated and include speech specific features. Some popular features are

the linear predictive coding (LPC) coefficients and mel-frequency cepstral coefficients (MFCC), zero crossing rate, power content in specific frequency bands etc. In statistical methods [25–27], in addition to a certain signal representation, a likelihood model is also assumed on the distribution of the representative coefficients given speech is present or absent. The second step in VAD is the classification. The two generally used methods for classification are thresholding or by applying machine learning techniques.

The objective of a speech enhancement algorithm is to reconstruct the spectral coefficients of the clean speech signal from the coefficients of the signal contaminated by noise. Traditional methods deal with additive noise, that is assumed to be uncorrelated with the speech signal. Speech enhancement methods usually concentrate on the estimation of the spectral amplitude of speech, while the phase is taken as the phase of the noisy signal, having relatively small perceptual importance [28]. One of the earlier methods for spectral amplitude estimation was the *spectral subtraction*, where the short term power spectral density (PSD) of the noise is estimated, and is subtracted from the PSD of the observed signal. The squared root of the estimated PSD of speech is taken as the estimator of the spectral amplitude [29]. Another method is the Wiener filter approach [29], which proposes to filter the noisy signal with a linear filter, that minimizes the mean squared error (MMSE). The result of this minimization is a non-causal short time Wiener filter. More recent approaches to speech enhancement, are the *spectral enhancement* methods. In these methods, the spectral magnitude estimation is based on a minimization of a certain distortion measure, $d\big(X_l(k), \hat{X}_l(k)\big)$, between the clean spectral coefficient, $X_l(k)$ and its estimate, $\hat{X}_l(k)$. These methods also assume a certain statistical model on the distribution of the spectral coefficients of speech [30–32]. Spectral enhancement are the methods we discuss in more detail later in this chapter.

The chapter is organized as follows. In Section 3.2 we introduce the topic of VAD. In Section 3.3 we introduce spectral speech enhancement methods. The chapter is summarized in Section 3.4.

3.2 Voice Activity Detection

Voice activity detection is an important component in speech processing and communication applications. Let $x(n)$ and $d(n)$ denote speech and uncorrelated noise signals respectively, where n is a discrete time index. The observed signal is given by $y(n) = x(n)+d(n)$, or in the time-frequency domain, by $Y_l(k) = X_l(k) + D_l(k)$ where l and k are time and frequency indices respectively. The objective of a VAD is to determine, for each signal frame, whether it belongs to a speech or noise segment. Speech presence or absence is typically represented by two hypotheses, which assign each frame to one of the two classes, as follows.

$$\begin{aligned} H_0: &\quad \text{speech is absent in frame } l: \quad Y_l(k) = D_l(k) \\ H_1: &\quad \text{speech is present in frame } l: \quad Y_l(k) = X_l(k) + D_l(k) \end{aligned}$$

The basic components in each VAD method are representation and classification. In the following we present some of the popular approaches to VAD. For each approach we shortly describe the motivation and mode of operation. We then present some examples for each approach.

3.2.1 Machine learning based voice activity detection

In this approach, firstly, a set of features is assigned to each time frame. Let $\Psi(l) \in \mathcal{R}^n$, be the set of features representing frame l in the feature space. Since this is a binary classification problem, each observation belongs to one of two classes. The labels of the classes are denoted by $y \in \{-1, 1\}$. In a machine learning technique, a separating function is trained on a training set with known labels. A new observation may then be classified by evaluating the value of the separating function when applied to the new observation. Given a set of training vectors, $\{\Psi_i\}_{i=1}^N$, and their respective class labels $\{y_i\}_{i=1}^N$, $y_i \in \{-1, 1\}$, a separating function $f(x)$ is trained. A new observation, $\Psi(l')$, is classified according to $f(\Psi(l'))$. A popular technique used for this purpose is support vector machine (SVM) classification, in this case $y(\Psi(l)) = \text{sign}\{f(\Psi(l))\}$. Different features are used by different methods in the classification process, some examples are as follows. In [33], the feature vector comprises of features used by G.729 annex B, [34]: *spectral distortion, full band*

energy difference, low band energy difference and the *zero crossing difference*. In [35], the feature vector is composed of sub-band signal-to-noise-ratio. In [36] *a priori*, *a posteriori* and *predicted* SNR values are concatenated into a feature vector.

Classification by a Support Vector Machine

The SVM is a means for binary classification. It is a supervised classification method, providing a separating hyperplane, based on a set of training vectors. The optimal separating hyperplane is the solution to an optimization problem, providing the maximal margin separating between the two classes. The training vectors located on the edges of the margin are called *support vectors*. The parameters of the separating hyperplane are determined based on these vectors. The linear SVM is formulated as follows.
Let $\omega \in \mathcal{R}^n$ denote the normal vector to the separating hyperplane. Consider the function

$$f(x) =< \omega, x > +b \tag{3.1}$$

where $x \in \mathcal{R}^n$ is a general vector in the feature space, and b is the bias term. The separating hyperplane is obtained for $f(x) = 0$, the hyperplanes $f(x) = 1$ and $f(x) = -1$ represent the margin on both sides of the separating hyperplane, which is clear of training vectors. The distance between the hyperplanes that form the margin is $\frac{2}{\|\omega\|}$, this is what needs to be maximized, i.e $\|\omega\|$ needs to be minimized. All training vectors also fulfill $y_i(< \omega, x_i > +b) \geq 1$, since no training vectors are located inside the margin. In particular, the training vectors located on the edges of the margin are called *support vectors*, these vectors fulfill $y_i(< \omega, x_i > +b) = 1$, and they also determine the values of the parameters, ω and b, in (3.1). Thus the optimization problem to be solved is

$$\begin{aligned} \min_{\omega,b} . \quad & \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{N} \alpha_i \big[y_i(< \omega, x_i > +b) - 1\big] \\ \text{s.t} \quad & \alpha_i \geq 0 \ , \ i \in \{1, N\} \end{aligned} \tag{3.2}$$

With N the number of vectors in the training set, and $\{\alpha_i\}_{i=1}^{N}$, the Lagrange multipliers. By solving (3.2), ω is obtained as

$$\omega = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{3.3}$$

where α_i are non-zero only for the respective x_i, which are the *support vectors*.
The classification of a new vector, $\tilde{x}$ is obtained as

$$\tilde{y} = \text{sign}\{f(\tilde{x})\} \tag{3.4}$$

Substituting (3.3) into (3.1), we have

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i < x_i, x > +b \tag{3.5}$$

A nonlinear separating function may be obtained by replacing the linear inner product in (3.5) by a non-linear kernel function, $K(x_i, x)$.

3.2.2 Statistical Voice Activity Detection

In statistical voice activity detection, a statistical model is assumed on the distribution of representative signal coefficients given each of the hypotheses, H_0 or H_1. Let Y denote the representative coefficients of a signal frame. It is assumed that the conditional probability distribution of Y given speech and non-speech hypotheses, $p(Y|H_1)$ and $p(Y|H_0)$ respectively, are known. The test statistic, according to which the presence of voice activity is determined, is the likelihood ratio (LR)

$$\lambda = \frac{p(Y|H_1)}{p(Y|H_0)} \tag{3.6}$$

Typically, it is more convenient to use the log-likelihood ratio, that is acquired by taking the natural logarithm of both sides of (3.6).

$$\Lambda = \log p(Y|H_1) - \log p(Y|H_0) \tag{3.7}$$

The determination of voice activity is made according to comparison of Λ to a threshold, η

$$\Lambda \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.8}$$

if Λ is greater than the threshold, hypothesis H_1 is accepted, otherwise H_0 is accepted.

Examples of Statistical VAD Methods

We would like to present two existing methods for statistical VAD here. These methods are interesting for the different models assumed on speech coefficients. The different models result in a different formulation of the log-LR.

Shin et al. 2008 [37] This work assumes that the DFT coefficients of noise and noisy speech are independent and follow the Laplacian distribution. The likelihoods of a single frequency band are

$$\begin{aligned} p(Y_k|H_0) &= \frac{1}{\lambda_d(k)} \exp\{-2(|Y_{k,(R)}| + |Y_{k,(I)}|)/\sqrt{\lambda_d(k)}\} \\ p(Y_k|H_1) &= \frac{1}{\lambda_d(k)+\lambda_x(k)} \exp\{-2(|Y_{k,(R)}| + |Y_{k,(I)}|)/\sqrt{\lambda_d(k)+\lambda_x(k)}\} \end{aligned} \tag{3.9}$$

where $Y_{k,(R)}, Y_{k,(I)}$ denote the real and imaginary parts of the k^{th} DFT coefficient respectively. $\lambda_d(k)$ and $\lambda_x(k)$, are the variances of noise and clean speech respectively. Let $\mathbf{Y} = [Y_0, Y_1, ..., Y_{M-1}]$ denote frequency representation of the full time-frame, composed of M frequency bands. Then, assuming independence in the frequency domain, $p(\mathbf{Y}|H_i) = \prod_{k=0}^{M-1} p(Y_k|H_i), i \in \{0,1\}$, and the log-LR is

$$\Lambda = log\left(\frac{p(\mathbf{Y}|H_1)}{p(\mathbf{Y}|H_0)}\right) = \log\left(\prod_{k=0}^{M-1}\frac{p(Y_k|H_1)}{p(Y_k|H_0)}\right) = \sum_{k=0}^{M-1}\log\frac{p(Y_k|H_1)}{p(Y_k|H_0)} \tag{3.10}$$

Substituting the model assumptions (3.9) into (3.10) the log-LR is

$$\Lambda = \sum_{k=0}^{M-1}\left[\frac{1}{1+\xi_k} + 2\Big(|Y_{k,(R)}| + |Y_{k,(I)}|\Big)\left(\frac{|Y_k| - \sqrt{\lambda_d(k)}}{|Y_k|\sqrt{\lambda_d(k)}}\right)\right] \tag{3.11}$$

where $\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}$ is the *a-priori* SNR in frequency bin k, and $|Y_k| = \sqrt{|Y_{k,(R)}|^2 + |Y_{k,(I)}|^2}$. Two separate values are suggested for the threshold, conditioned on the VAD decision from the previous frame.

$$\exp\{\eta_i\} \propto \frac{P(H_n = H_0|H_{n-1} = H_i)}{P(H_n = H_1|H_{n-1} = H_i)},\ i \in \{0,1\} \tag{3.12}$$

Mousazadeh and Cohen 2011 [27] In this paper, the speech signal is modeled by an autoregressive-generalized autoregressive conditional heteroscedasticity (AR-GARCH) process. The AR(p)-GARCH(1, 1) process is described by the following equations.

$$x_t = \sum_{i=1}^{p}\alpha_i x_{t-i} + \epsilon_t \tag{3.13}$$

$$\epsilon_t = \sigma_{t|t-1} v_t \tag{3.14}$$

$$\sigma^2_{t|t-1} = \beta_0 + \beta_1\epsilon^2_{t-1} + \beta_2\sigma^2_{t-1|t-2} \tag{3.15}$$

where (3.13) describes the auto-regressive (AR) model, according to which the clean speech signal evolves in time. In (3.14) v_t are zero mean independent and identically

distributed (IID) random variables with unit variance. The GARCH$(1,1)$ model according to which $\sigma^2_{t|t-1}$ evolves in time, is described by (3.15). Let y_t denote the speech signal corrupted by additive white Gaussian noise

$$y_t = x_t + d_t \tag{3.16}$$

Let $\theta = [\beta_0, \beta_1, \beta_2, \alpha_1, ..., \alpha_p]^T$, denote the vector of model parameters, and y_1^{t-1} all observations from $t = 1$ up to time $t - 1$.
Since $(x_t|y_1^{t-1}) \sim \mathcal{N}\Big(\sum_{i=1}^{p} \alpha_i \hat{x}_{t-i}, \hat{\sigma}^2_{t|t-1}\Big)$ and $(d_t|y_1^{t-1}) \sim \mathcal{N}(0, \sigma^2)$ are independent,

$$(y_t|y_1^{t-1}; \theta) \sim \mathcal{N}\Big(\sum_{i=1}^{p} \alpha_i \hat{x}_{t-i}, \hat{\sigma}^2_{t|t-1} + \sigma^2\Big) \tag{3.17}$$

The LR is formed by

$$\lambda_t = \frac{p(y_t|y_1^{t-1}; \theta, \mathbf{H_1})}{p(y_t|y_1^{t-1}; \theta, \mathbf{H_0})} \tag{3.18}$$

The correlation between consecutive speech samples is incorporated into the log-LR similarly to what is done in [38], to help prevent clipping of weak speech.

3.2.3 Voice Activity Detection based on long term information

These methods are especially interesting in the context of this thesis, because they refer to temporal dependency in a sequence of observations. In speech, adjacent time-frames have a higher probability of belonging to the same class. This property of speech was exploited for the design of VAD algorithms. In these algorithms, the VAD decision for time-frame n, is made on a sequence of observations denoted by $\mathcal{Y}_n$. We present two VAD methods that formulate the log-LR, based on long term information. Both methods employ the Gaussian model for speech coefficients, but they differ in the way temporal dependence between neighboring time frames is treated.

Sohn 1999 [25] The algorithm in this paper, assumes a Gaussian statistical model on the conditional distribution of DFT coefficients of the signal. The M-long vector $\mathbf{Y} = [Y_0, Y_1, ..., Y_{M-1}]$, denotes the vector of DFT coefficients of a single time-frame in the

observed signal. The DFT coefficients are assumed IID, hence:

$$\begin{aligned} p(\mathbf{Y}|H_0) &= \prod_{k=0}^{M-1} \frac{1}{\pi\lambda_d(k)} \exp\Big\{ - \frac{|Y_k|^2}{\lambda_d(k)} \Big\} \\ p(\mathbf{Y}|H_1) &= \prod_{k=0}^{M-1} \frac{1}{\pi[\lambda_d(k)+\lambda_x(k)]} \exp\Big\{ - \frac{|Y_k|^2}{\lambda_d(k)+\lambda_x(k)} \Big\} \end{aligned} \tag{3.19}$$

The LR of a single frequency band is

$$\lambda_k = \frac{p(Y_k|H_1)}{p(Y_k|H_0)} = \frac{1}{1+\xi_k} \exp \frac{\gamma_k \xi_k}{1+\xi_k} \tag{3.20}$$

where

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \text{ and } \gamma_k \triangleq \frac{|Y_k|^2}{\lambda_d(k)}$$

a-priori and *a-posteriori* SNRs respectively. The log-LR of a single time frame is given by the geometric mean of the LR of the constituting frequency bands,

$$\Lambda = \frac{1}{M} \sum_{k=0}^{M-1} \log \lambda_k \tag{3.21}$$

The temporal dependency between consecutive frames is introduced here through a Hidden Markov Model (HMM) based hang-over scheme. The process is assumed to have two states, where the state at time n is represented by q_n and it is either H_0 or H_1. The Markov process is assumed to be time invariant and the transition probabilities are represented by $a_{ij} = P(q_n = H_j | q_{n-1} = H_i)$. The process is assumed to be stationary, and it is assumed to have $P(q_n = H_i) = pH_i$, where $pH_0 = \dfrac{a_{10}}{a_{10}+a_{01}}$ and $pH_1 = \dfrac{a_{01}}{a_{10}+a_{01}}$ are steady state probabilities. The LR here is based on the vector of all observations $\mathcal{Y}_n = \{\mathbf{Y}(n), \mathbf{Y}(n-1), ..., \mathbf{Y}(1)\}$, such that

$$\mathcal{L}(n) \triangleq \frac{p(\mathcal{Y}_n|q_n = H_1)}{p(\mathcal{Y}_n|q_n = H_0)} = \frac{P(\mathcal{Y}_n, q_n = H_1)}{P(\mathcal{Y}_n, q_n = H_0)} \frac{pH_0}{pH_1} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.22}$$

Define $\alpha_n(i) \triangleq p(\mathcal{Y}_n, q_n = H_i)$, using the forward procedure [39], $\alpha_n(i)$ is given by

$$\alpha_n(i) = \begin{cases} pH_i \; p(\mathbf{Y}(1)|q_1 = H_i) & \text{if } n = 1 \\ (\alpha_{n-1}(0)a_{0j} + \alpha_{n-1}(1)a_{1j}) \cdot p(\mathbf{Y}(n)|q_n = H_i) & \text{if } n \geq 2 \end{cases}$$

Define the recursive term $\Gamma(n)$

$$\Gamma(n) = \frac{\alpha_n(1)}{\alpha_n(0)} = \frac{a_{01} + a_{11}\Gamma_{n-1}}{a_{00} + a_{10}\Gamma_{n-1}} \Lambda(n) \tag{3.23}$$

The log-LR is given by

$$\mathcal{L}(n) = \frac{pH_0}{pH_1} \Gamma(n) \tag{3.24}$$

Ramirez et al. 2005 [40] In this paper, a similar model to [25] is used for the conditional distribution of the DFT coefficients of the signal, and the content of different sub-bands is also assumed IID, (3.19). The approach proposed in this paper, differs in the assumption regarding the dependency between adjacent time-frames, and in the observation vector which represents the current frame. In time, the multiple observation log likelihood ratio test (MO-LRT) is employed, assuming the temporal observations are independent. The observation vector is non-causal and symmetric in respect to the current observation, $\mathcal{Y}_l = \{\mathbf{Y}_{l-m}, ..., \mathbf{Y}_{l-1}, \mathbf{Y}_l, \mathbf{Y}_{l+1}, ..., \mathbf{Y}_{l+m}\}$.
Define $\Phi(l) = \log \frac{p(\mathbf{Y}_l|H_1)}{p(\mathbf{Y}_l|H_0)} = \sum_{k=0}^{M-1} \log \frac{p(Y_l(k)|H_1)}{p(Y_l(k)|H_0)}$ as the single-frame LR.
The log-LR of the non-causal observation vector is

$$\mathcal{L}_{l,m} = \sum_{q=l-m}^{l+m} \Phi(q) \tag{3.25}$$

or, the log-LR in (3.25), can be recursively computed,

$$\mathcal{L}_{l+1,m} = \mathcal{L}_{l,m} - \Phi(l-m) + \Phi(l+m+1) \tag{3.26}$$

3.3 Speech Enhancement

The objective of speech enhancement is to obtain an estimator for the clean speech signal, based on observations of the signal corrupted by noise. This problem may be formulated as estimation of spectral components of speech, from a signal degraded by statistically independent additive noise. Spectral speech enhancement methods, which we discuss here, estimate the magnitude of clean speech by minimizing the expected value of a certain distortion measure.

3.3.1 Spectral Enhancement

Let $x(n)$ and $d(n)$ denote speech and uncorrelated noise signals respectively, where n is a discrete time index. The observed signal is $y(n) = x(n) + d(n)$, or in the STFT domain, we have

$$Y_l(k) = X_l(k) + D_l(k) \tag{3.27}$$

where l and k are time and frequency indices respectively. The clean speech signal in the time domain, $x(n)$, is obtained by the inverse STFT transform. The signal $X_l(k)$, in (3.27) is complex, hence $X_l(k) = A_l(k)e^{j\phi_l(k)}$, with $\phi_l(k)$ representing the phase and $A_l(k) = |X_l(k)|$ the magnitude. The conditional variance of the speech spectral amplitude, is denoted by $\lambda_l(k) = E\{|X_l(k)|^2|H_1\}$.

The enhanced speech signal is estimated separately for each sub-band, based on the observed (noisy) signal $Y_l(k)$. In spectral speech enhancement, an estimate of the clean speech is obtained by minimizing the expected value of a certain distortion measure, $d\big(X_l(k), \hat{X}_l(k)\big)$

$$\hat{X}_l(k) = \min_{\hat{X}} E\Big\{ d\big(X_l(k), \hat{X}_l(k)\big) \Big| \mathcal{Y} \Big\} \tag{3.28}$$

where $\mathcal{Y}$ is a set of observations. Some popular choices for the distortion measure are:

1. Squared Error (SE) -

$$d_{SE}\big(X_l(k), \hat{X}_l(k)\big) \triangleq |X_l(k) - \hat{X}_l(k)|^2 \tag{3.29}$$

2. Spectral Amplitude (SA) -

$$d_{SA}\big(X_l(k), \hat{X}_l(k)\big) \triangleq |A_l(k) - \hat{A}_l(k)|^2 \tag{3.30}$$

3. Log-Spectral Amplitude (LSA) -

$$d_{LSA}\big(X_l(k), \hat{X}_l(k)\big) \triangleq |\log A_l(k) - \log \hat{A}_l(k)|^2 \tag{3.31}$$

4. Spectral Power (SP) -

$$d_{SP}\big(X_l(k), \hat{X}_l(k)\big) \triangleq |A_l^2(k) - \hat{A}_l^2(k)|^2 \tag{3.32}$$

Speech Modeling

The DFT speech coefficients, given $\lambda_l(k)$, and the speech presence or absence hypothesis in frame l, are assumed to be generated by

$$X_l(k) = \sqrt{\lambda_l(k)} V_l(k) \tag{3.33}$$

where $\{V_l(k)|H_0\}$ are identically zero, and $\{V_l(k)|H_1\}$ are statistically independent complex random variables with zero mean, unit variance and IID real and imaginary parts.

The distribution of speech coefficients, is defined by the statistical model assumed for $V_l(k)$, under H_1. Let $V_{\rho,l}(k)$, $\rho \in \{R, I\}$, refer to the real and imaginary parts of $V_l(k)$, respectively. Three commonly used models for the distribution of $V_{\rho,l}(k)$, given speech is present, H_1, are [32]: the Gaussian model,

$$p(V_{\rho,l}(k)|H_1) = \frac{1}{\sqrt{\pi}} \exp\big(-V_{\rho,l}^2(k)\big) \tag{3.34}$$

the Gamma model,

$$p(V_{\rho,l}(k)|H_1) = \frac{1}{2\sqrt{\pi}} \left(\frac{3}{2}\right)^{1/4} \big|V_{\rho,l}(k)\big|^{-1/2} \exp\left(-\sqrt{\frac{3}{2}}\big|V_{\rho,l}(k)\big|\right) \tag{3.35}$$

and the Laplacian model

$$p(V_{\rho,l}(k)|H_1) = \exp\big(-2\big|V_{\rho,l}(k)\big|\big) \tag{3.36}$$

Given one of the models (3.34)-(3.36), and the MMSE estimate for $\lambda_l(k)$ (discussed later in this Section),

$$\hat{\lambda}_l(k) = E\big\{A_l^2(k)\big|\mathcal{Y}\big\} \tag{3.37}$$

we have the distribution for the real and imaginary parts of the spectral speech coefficients, $p\big(X_{\rho,l}(k)\big|\lambda_l(k)\big)$.

Spectral Gain Function

Typically, speech enhancement algorithms propose methods for estimating the spectral amplitude, $A_l(k)$. As for the phase, it was shown in [30], that the MMSE estimator for the complex exponential $e^{j\phi_k}$, is obtained by solving the following constrained optimization problem,

$$\begin{aligned} \min_{e^{j\hat{\phi}_k}} \quad & E\big\{|e^{j\phi_k} - e^{j\hat{\phi}_k}|^2\big\} \\ \text{subject to} \quad & |e^{j\hat{\phi}_k}| = 1 \end{aligned} \tag{3.38}$$

and it is equal to the complex exponential of the noisy phase, which does not affect the magnitude estimation.

In order to estimate $A_l(k)$, we go back to the expression in (3.28), that may be written as

$$\hat{X}_l(k) = \min_{\hat{X}_l(k)} \int\int d(X_l, \hat{X}_l) p\Big(X_l\big|\mathcal{Y}, \lambda_l(k)\Big) p\Big(\lambda_l(k)\big|\mathcal{Y}\Big) dX_l(k) d\lambda_l(k) \tag{3.39}$$

Assuming we have $\lambda_l(k)$ (3.37), (3.39) reduces to

$$\begin{aligned}\hat{X}_l(k) &= \min_{\hat{X}_l(k)} \int d(X_l(k), \hat{X}_l) p\Big(X_l \Big| Y_l(k), \lambda_l(k)\Big) dX_l(k) \\ &= \min_{\hat{X}_l(k)} E\big\{ d(X_l, \hat{X}_l) \big| Y_l(k), \lambda_l(k) \big\} \end{aligned} \tag{3.40}$$

The expression in (3.40), can also be written as

$$\hat{X}_l(k) = G(\xi_{l|l-1}(k), \gamma_l(k))^2 Y_l(k) \tag{3.41}$$

where $G(\xi_{l|l-1}(k), \gamma_l(k))$ is the *spectral gain function*. $\xi_{l|l-1}(k) = \frac{\lambda_{l|l-1}(k)}{\lambda_{d_l}(k)}$ is the one-frame-ahead *a priori* SNR, and $\gamma_l = \frac{|Y|^2}{\lambda_{d_l}(k)}$ is the *a-posteriori* SNR. Denoted by $\lambda_{l|l-1}(k)$, is the conditional one-frame-ahead spectral variance of speech, its estimator is $\hat{\lambda}_{l|l-1}(k) = E\big\{ A_l^2(k) \big| \mathcal{Y}_0^{l-1}(k) \big\}$ and $\mathcal{Y}_0^{l-1}(k) = [Y_0(k), Y_1(k), ..., Y_{l-1}(k)]$. The expression for the estimator of the spectral magnitude, is hence

$$\hat{A}_l(k) = G(\xi_{l|l-1}(k), \gamma_l(k)) |Y_l(k)| \tag{3.42}$$

Thus, the estimator for the spectral magnitude of the clean signal, is the magnitude of the noisy signal adjusted by the spectral gain function, while the phase used for estimation, is the phase of the original noisy signal. The expression for the spectral gain function, $G(\xi_{l|l-1}, \gamma_l)$, in (3.42), is obtained by minimizing (3.40), after selecting a distortion measure from (3.29)-(3.32), and the statistical model for speech from (3.33)-(3.36).

The corresponding spectral gain functions to the distortion measures in (3.29)-(3.32) for the Gaussian speech model are, [41, 30, 42, 43]

$$G_{SE}(\xi_{l|l}) = \frac{\xi_{l|l}}{1 + \xi_{l|l}} \tag{3.43}$$

$$G_{SA}(\xi_{l|l}, \gamma_l) = \frac{\sqrt{\pi v_l}}{2\gamma_l} \Big[(1 + v_l) I_0\left(\frac{v_l}{2}\right) + v_l I_1\left(\frac{v_l}{2}\right) \Big] \exp\left(-\frac{v_l}{2}\right) \tag{3.44}$$

where $v_l \triangleq \frac{\xi_{l|l}}{1+\xi_{l|l}} \gamma_l$.

$$G_{LSA}(\xi_{l|l}, \gamma_l) = \frac{\xi_{l|l}}{1 + \xi_{l|l}} \exp\left(\frac{1}{2} \int_{v_l}^{\infty} \frac{e^- t}{t} dt \right) \tag{3.45}$$

$$G_{SP}(\xi_{l|l}, \gamma_l) = \sqrt{\frac{\xi_{l|l}}{1 + \xi_{l|l}} \Big(\frac{1}{\gamma_l} + \frac{\xi_{l|l}}{1 + \xi_{l|l}} \Big)} \tag{3.46}$$

3.3.2 A-priori SNR Estimation

A-priori SNR Estimation, is an important component in speech enhancement methods. The *a priori* SNR is estimated separately for each sub-band.

Decision Directed

This estimator was proposed by Ephraim and Malah [30]. Assuming speech is present, the estimator is given by

$$\hat{\xi}^{DD}_{l|l-1}(k) = \alpha \frac{\hat{A}^2_{l-1}(k)}{\lambda_{d_{l-1}}(k)} + (1-\alpha)\max(\gamma_l(k) - 1, 0) \tag{3.47}$$

where the term $\frac{\hat{A}^2_{l-1}(k)}{\lambda_{d_{l-1}}(k)}$ represents an SNR term from the previous frame, and the term $\gamma_l(k) - 1$ is the ML estimator of the *a priori* SNR, which depends on the current frame only. The weight parameter $0 < \alpha < 1$ determines the relative importance of each term in (3.47), in the calculation of $\hat{\xi}^{DD}_{l|l-1}(k)$, for the current frame. This parameter controls the trade off between noise reduction and signal distortion, as follows. The term $\gamma_l(k)-1$, that is based on the corrupted signal from the current frame, introduces *musical noise* [44] in low SNR conditions. A value of α, close to 1, reduces the musical noise, but on the other hand, the most recent available signal information is only weight in by $(1-\alpha)$, which introduces distortion. A value of $\alpha = 0.98$ was determined by simulations and informal listening tests [30].

Recursive *a priori* SNR Estimator

In [45,31,32] recursive estimators for *a priori* SNR are proposed. The derivation of these estimators follows the Kalman filtering rational, as they are obtained in two steps. The first step is the "propagation", where we are given an estimate $\hat{\lambda}_{l|l-1}$, conditioned on the noisy measurements up to frame $l-1$. In the "update" step, $\hat{\lambda}_{l|l-1}$ is updated to $\hat{\lambda}_{l|l}$, using the new noisy observation Y_l. Once $\hat{\lambda}_{l|l}$ is obtained, we have $\hat{\xi}_{l|l} = \frac{\hat{\lambda}_{l|l}}{\lambda_{D_l}}$ as the *a priori* SNR.

Propagation This step requires a model for the propagation of the speech variance, $\hat{\lambda}_{l|l-1}(k)$, in time. The model provides an estimator (or predictor) for the spectral ampli-

tude variance in frame l, based on the information obtained up to frame $l-1$. Among the models proposed for the propagation are

1. The DD approach [45]

$$\hat{\lambda}^{\mathrm{DD}}_{l|l-1}(k) = \max\Big\{(1-\alpha)\hat{\lambda}_{l-1|l-1}(k) + \alpha\hat{A}^2_{l-1}(k), \lambda_{min}\Big\}$$

2. The GARCH(1,1) model [31]

$$\hat{\lambda}^{\mathrm{GARCH}}_{l|l-1}(k) = \lambda_{min} + \mu\hat{A}^2_{l-1}(k) + \delta\big(\hat{\lambda}_{l-1|l-2}(k) + \lambda_{min}\big)$$

where

$$\lambda_{min} > 0,\ \mu \geq 0,\ \delta \geq 0,\ \mu + \delta < 1$$

Update The estimate for the speech spectral variance, in time-frame l is updated by the observation Y_l, as follows.

$$\hat{\lambda}_{l|l}(k) = E\Big\{A_l^2\big|\hat{\lambda}_{l|l-1}(k), Y_l(k)\Big\} = G\Big(\hat{\xi}_{l|l-1}(k), \gamma_l(k)\Big)^2\Big|Y_l(k)\Big|^2$$

The spectral gain function $G\big(\hat{\xi}_{l|l-1}(k), \gamma_l(k)\big)$, for a Gaussian speech model, can be chosen out of (3.43)-(3.46).

3.4 Summary

In this chapter, we have introduced VAD as a binary classification problem that is represented by two hypotheses. We have discussed the statistical and feature based approaches to VAD, and we presented several short and long-term based VAD approaches.

We have also discussed the speech enhancement problem in this chapter. We presented the spectral enhancement method, that is stated as an optimization problem, which aims at minimizing a certain distortion measure. Each distortion measure combined with a statistical model on the distribution of speech spectral coefficients, defines a spectral gain function. The estimate of the clean speech signal is then obtained by the application of the gain function to the spectral coefficients of the noisy signal.

Another topic we addressed in this chapter is the estimation of *a priori* SNR, which is an important component in speech enhancement applications. The decision directed, and the recursive methods for its estimation were presented.

Chapter 4

Dominant speaker identification for multipoint videoconferencing

4.1 Introduction

Multipoint videoconferencing technology has been existent since the early 1960s. Throughout this period, it had transformed from an expensive technology restricted for use in large organizations, to cheap and easy to use applications available in almost every home.

In multipoint videoconferencing, three or more dispersedly located participants connect for a meeting, over telephone or Internet based networks. Typically, the meeting is controlled by a central processing unit, which is in charge of routing signals between participants.

The incorporation of video into audioconferencing, had significantly raised the amount of information transmitted through the network. In addition to increased bandwidth consumption, it raises the amount of information that is processed by the central processing unit. An effort has been made to offer solutions for reducing the load on the network. Most of these solutions involve the identification of the most active participants, through a process referred to as *speaker selection*. Once the active speakers are selected, the remaining audiovisual information may be discarded, thus relieving the network.

Many works in the field of improving the efficiency of data traffic in videoconferencing, rely on speaker selection as a vital component [46, 47]. However, little research attention

has been devoted to the speaker selection task itself. Simple methods are based on indicators of the signal level in the channel, as measured by its amplitude or mean power [1–3]. Not only the signal level does not necessarily indicate the presence of speech, it is also a very instantaneous measure of audio activity. Thus, measuring the signal level in the channel provides very little information regarding a prolonged speech activity, that may indicate the presence of dominant speech activity.

More advanced speaker selection methods make use of speech specific activity indicators. This, for example, by incorporating a voice activity detector (VAD). In [4], a VAD is used to identify the active parties. The VAD decision in this method is derived from either the signal power or the arrival of silence insertion descriptor (SID) frames. The active speakers are then ranked by the order of becoming active. A speaker can be promoted in ranking, only if its smoothed signal power exceeds a certain *barge-in* threshold. The ranking list keeps a continuous record of the M most active participants. Another method is proposed in [5]. This method is based on a set of speech specific features and a machine learning technique that classifies each signal frame into either voice or noise. In order to prevent spurious speaker switching, the barge-in mechanism, originally proposed in [4], is used in this work as well. These two methods, although constituting an advancement over the level based methods, still concentrate on instantaneous measures for speech activity. No special attention is devoted to long-term properties of dominant speech in the speaker switch mechanism. The barge-in mechanism, that is proposed as the switching mechanism, increases the vulnerability of these algorithms to false switching due to transient noises.

In this chapter, we present a novel approach for dominant speaker identification, based on speech activity evaluation on time intervals of different lengths. The lengths of the time intervals we use, correspond to a single signal frame, a part of a word and few words to a sentence. This mode of operation allows the capturing of basic speech events such as words and sentences. Sequences and combinations of these events may indicate the presence of dominant speech activity (or lack of it). Another unique ability offered by the proposed method, is the distinction between transient audio occurrences that are isolated and those that are located inside a speech burst.

Integration of long-term speech information, had already been proven effective in VAD

applications [25, 48, 40]. Long term information was used in the aforementioned methods in order to determine whether speech is present in a currently observed time-frame. We find this approach well suited to our problem, since dominant speech activity in a given time-frame would be better inferred from a preceding time interval, than from any instantaneous signal property. Hence, we incorporate the approaches from these VAD works into the proposed method.

Objective evaluation of the proposed method is performed on a synthetic conference with and without the presence of transient audio occurrences. In addition, we test the proposed method on a segment of a real five channel audioconference. Results are compared with existing speaker selection algorithms. We show reduction in the number of false speaker switches, and improved robustness against transient audio occurrences.

The chapter is organized as follows. In Section 4.2, we present the problem statement of dominant speaker identification. In Section 4.3, we present the proposed method. In Section 4.4, we describe the proposed method for SNR estimation. We present the two proposed approaches for the speech activity score evaluation method in Section 4.5, where one is based on a single observation, and the other introduces temporal dependence between consecutive time-frames, basing the score on a sequence of observations. Experimental results are presented in Section 4.6. This work is concluded in Section 4.7.

4.2 Problem Statement

A multipoint conference consists of N participants received through N distinct channels. The objective of a dominant speaker identification algorithm, is to determine, at a given time, which one of the N participants is the dominant speaker. We discuss an arrangement, where each participant receives a video feed from only one other participant. In the proposed embodiment, the video stream of the dominant speaker is sent to all participants, while the dominant speaker himself receives the video stream from the previous dominant speaker. Throughout this chapter, we use the terms channel, participant, user and speaker interchangeably, as referring to a conference end-point.

We define a *speech burst* as a speech event composed of three sequential phases: initiation, steady state and termination. In the first phase, speech activity builds up. During

the second phase, speech activity is mostly high, but it may include breaks in activity due to pauses between words. Finally, in the third phase, speech activity declines and then stops. Typically, a dominant speech activity is composed of one or more consequent speech bursts. We refer to the point where a change in dominant speaker occurs, as a *speaker switch* event.

The challenges in the dominant speaker identification problem, arise from equipment, surrounding and personal characteristics of the different users. The type of equipment used, may introduce noises such as crosstalk (speakers) or reverberations (far talking microphone). The quality of the sensor affects the SNR of the incoming signal. The type and level of noise, in the surrounding of the speaker, influence the ability of the system to identify each speaker as dominant. The presence of transient noises, characterized by short duration and high energy, may distract the decision regarding the dominant speaker. Finally, personal characteristic of the speaker, such as loudness or quality of voice, may also affect the identification of the dominant speaker.

The desired behavior of a dominant speaker identification algorithm is as follows.

- No false switching should occur during a dominant speech burst. Both transient noise occurrences and single words that are said in response to or in agreement with the dominant speaker, are considered transient occurrences. These should not cause a speaker switch.

- A speaker switch event cannot occur during a break in speech between two dominant speakers. It has to be triggered by a beginning of a speech burst.

- A tolerable delay in transition from one speaker to another, in a speaker switch event, is up to one second.

- When simultaneous speech occurs on more than one channel, the dominant speaker is the one who began speaking first.

- The relative loudness of the voice of a speaker, should not influence his chance to be identified as the dominant speaker.

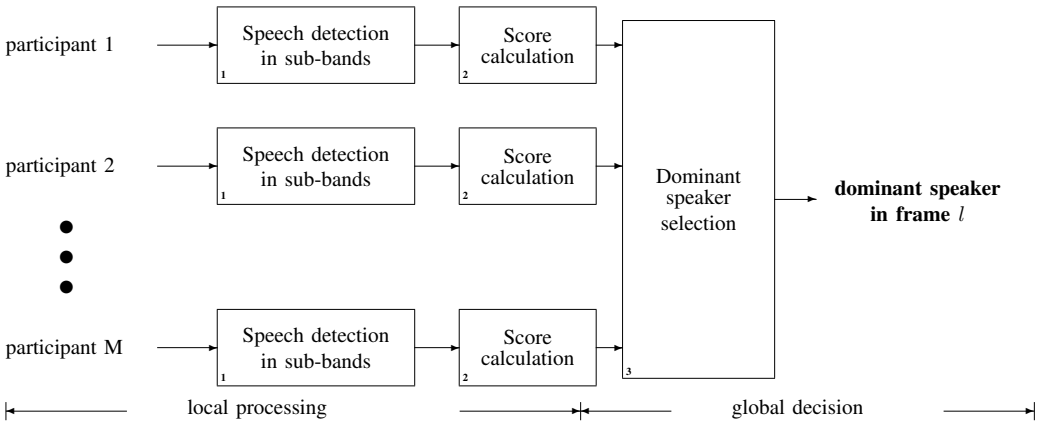


Figure 4.1: A flow-chart describing the proposed method; Block 2 is described in more detail in Figure 4.2

4.3 Dominant speaker identification based on time intervals of variable length

The proposed method for dominant speaker identification, consists of two stages, a *local processing* and a *global decision*, as depicted in Figure 4.1. In the first stage, the speech activity scores are evaluated for the immediate medium and long time-intervals. The lengths of the chosen time-intervals correspond with the lengths of: one signal frame, a part of a word and several words. In the second stage, the dominant speaker is identified based on the speech activity scores derived in the first stage (Figure 4.1 block 3). This stage is designed to detect speaker switch events. We assume that a speaker switch event can be inferred from a rise in the three speech activity scores on a certain channel, relatively to scores of the dominant channel. The rational of speaker switch event detection is presented in 4.3.2.

The implementation of the proposed algorithm is summarized in Figure 4.4.

4.3.1 Local Processing

In this stage, the signal in each channel is processed separately. The objective of our approach is to place each signal frame into a broader context than its instantaneous audio activity. This is accomplished by processing the currently observed frame, by itself, in

addition to a medium-length preceding time interval and in addition to a long time interval that precedes it. Thus, each time we move up to a longer time interval, the speech activity obtained in the previous step is analyzed again, in a broader context.

The motivation for this mode of processing is the nature of the signals we expect to receive through the channels. The expected types of signals during a multipoint conference are:

1. Silence or stationary noise at different power levels.

2. Transient audio occurrences such as knocks, caughing, sneezing etc.

3. Fluent and continuous speech, consisting of words and sentences

Each of these signal types would yield a typical combination of score values. This would allow us to discriminate between the different types of signals. Specifically, in the *global decision* stage, it would enable the discrimination of dominant speech activities on a certain channel, from non-dominant activity on other channels.

In the proposed approach, we relate to each time interval as composed of smaller sub-units. The speech activity in each time interval is determined according to the number of *active* sub-units, by attributing a *speech activity* score to this number. In the process of score evaluation, we propose two models for the likelihood of number of active sub-units. One for the hypothesis of *speech presence* and one for *speech absence*. The score is obtained from the ratio between these two likelihoods. The score evaluation method is fully described in Section 4.5. The speech activity evaluation process consists of three sequential steps, referred to as *immediate, medium* and *long*. The input into each step is a sequence of the number of active sub-units, acquired in the previous step.

For the step of immediate speech activity evaluation, we use a frequency representation of the frame, to test for speech activity in sub-bands. We operate on the frequency range that corresponds to the range of voiced speech. Let this range be denoted by $k \in [k_1, k_2]$ and the total number of sub-bands in this range, by N_1. As the frequency representation, we use the SNR value in each sub-band. Let it be denoted by $\underline{\xi}_l = \left\{\xi(k,l)\middle| k \in [k_1, k_2]\right\}$, where l is a discrete time index. The method for obtaining the time-frequency representation of SNR is presented in detail in Section 4.4. Next, we find

the number of *active* sub-bands in the frame. A sub-band is considered active, if its SNR is higher than a threshold ξ_{th}. The threshold value was obtained from a speech training set. The number of active sub-bands, in time frame l, is denoted by $a_1(l)$.

Since SNR measures the quality of the signal in each sub-band, it is a good indicator for speech presence. Noise may also exhibit locally high SNR values. This is partly dealt with in the score evaluation process, where high number of active sub-bands, given a non-speech frame has lower probability. It is further addressed in the processing for the medium and long time intervals, where isolated instantaneous spike of noise will be attributed a low score.

The thresholding approach serves several purposes in the discussed problem. For example, in the immediate time processing, a high amplitude noise in isolate number of sub-bands, would only result in low $a_1(l)$. Whereas, if we had used a measure that relies on the absolute value of SNR, such activity would have resulted in an indicator for high activity. Another advantage is the equalization between loud and quiet speakers. A sufficiently high value of SNR, in a similar number of sub-bands, results in a similar measure of activity for both loud and quiet users, thus eliminating the effect of loudness. We proceed with the thresholding approach in the next steps of speech activity detection, motivated by this rationale.

The number of active sub-bands is provided as an input into the *Score calculation* block, Figure 4.1 block 3. In this block, two additional thresholding steps are carried out, for time intervals of medium and long length. The input into the speech activity evaluation step for the medium length time interval, is a sequence of the number of active sub-bands in the last N_2 frames. We denote it by $\underline{\alpha}_l = \left\{a_1(l-m)\middle| m \in [0, N_2 - 1]\right\}$. Next, $\underline{\alpha}_l$ is thresholded by α_{th} and the number of active frames in the medium length time interval that precedes frame l is obtained. The number of active frames is denoted by $a_2(l)$, where $0 \leq a_2(l) \leq N_2$. This number indicates the amount of independent instantaneous activities, in a group of N_2 sequential frames. Let us assume that a certain frame, p, exhibited high number of active sub-bands, $a_1(p)$. The amount of activity in its neighbors, would determine if frame p contains an isolate noise spike, or that it is part of a longer audio activity. The amount of neighboring activity, in the medium time-term, would be indicated by $a_2(p)$.

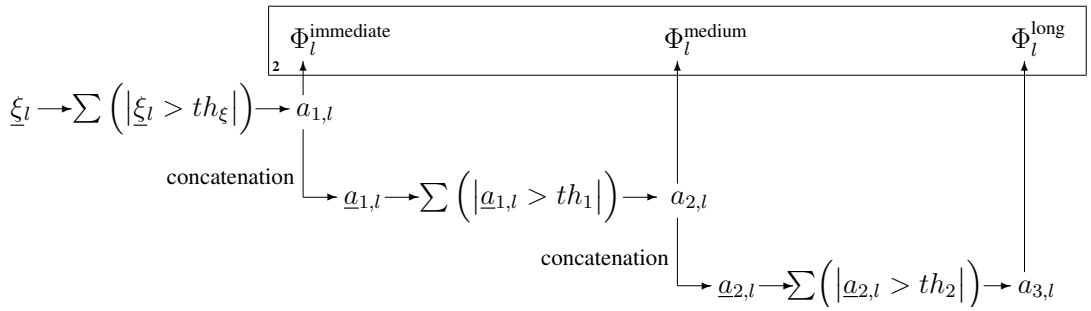


Figure 4.2: Speech activity scores evaluation process

Finally, a sequence of speech activity indicators for medium time-terms, is provided as input into the long term speech activity evaluation step. This sequence is denoted by $\underline{\beta}_l = \left\{a_2(l - mN_2)\middle| m \in [0, N_3 - 1]\right\}$, where N_3 is the number of medium length blocks constituting the long time interval. The number of active medium length blocks, is obtained by thresholding $\underline{\beta}_l$ by β_{th}. This number is denoted by $a_3(l)$. An active block of medium length, indicates a short transient occurrence. According to the rationale in the previous steps, low $a_3(l)$ would indicate that the transient is isolate, while high $a_3(l)$ is likely to indicate a part of a speech burst.

After obtaining $a_1(l), a_2(l)$ and $a_3(l)$, we have a good representation of the speech activity history of time-frame l. We also achieved here a *backward inference* by each of these values, on the shorter time intervals. A speech activity score is now attributed to the speech activity indicators, $a_1(l), a_2(l)$ and $a_3(l)$. The summary of the score evaluation process is illustrated in Figure 4.2. We denote the set of scores in frame l by $\Phi_l^{\text{immediate}}, \Phi_l^{\text{medium}}$ and Φ_l^{long}. The scores from the distinct channels are provided into the *Dominant speaker selection* block (Figure 4.1), where this information is translated into dominant speaker identification.

4.3.2 Global Decision

The objective of this stage is to return the number of the channel associated with the dominant speaker. This stage is activated once in a decision-interval. This both in order to reduce the processing load and also to serve as a smoothing factor. This stage is designed to utilize the properties highlighted by the scores that were obtained in the previous stage, for dominant speaker identification. The approach we take in this stage

is of detecting *speaker switch* events, rather than selecting a dominant speaker once in a decision-interval. Once a dominant speaker is identified, he remains dominant until the speech activity on one of the other channels justifies a speaker switch. In the following, we refer to the non-dominant channels as *competing* (for dominance).

The expected score behavior of the dominant speaker and a channel that justifies a speaker switch are described as follows. The type of channels (dominant or competing) referred to by the score, are indicated in brackets. During the dominance period of the dominant speaker, the score $\Phi_{\text{long}}(\text{dominant})$, is expected to be high. The scores for the immediate and medium time-terms, $\Phi_{\text{medium}}(\text{dominant})$ and $\Phi_{\text{immediate}}(\text{dominant})$, are allowed to be low for periods that correspond to breaks between words. As for the speech activity on the competing channels that justifies a speaker switch, all three scores are expected to be high. The long-term speech activity $\Phi_{\text{long}}(\text{competing})$ has to be high, to demonstrate prolonged speech activity. In addition, both $\Phi_{\text{medium}}(\text{competing})$ and $\Phi_{\text{immediate}}(\text{competing})$ are expected to be high, to indicate an offset of speech.

The speaker selection algorithm is illustrated in Figure 4.3. We propose the following realization for detecting speaker switch events, in accordance with aforementioned expected behavior of scores. Each decision-interval, three comparison terms, $\underline{c}_1$, $\underline{c}_2$ and $\underline{c}_3$, are defined. These comparison terms contain the information about the ratio between speech activity in the competing channels and speech activity in the dominant channel, for the long, medium and immediate time-terms respectively. In the next step, a set of channels demonstrating speech activity that might justify a speaker switch, is found. The set of indices representing these channels, is stored in j. In case j contains a reference to more than one channel, a dominant channel, denoted by j^*, is selected based on the highest speech activity score for the medium time-term.

4.4 *A priory* SNR Estimation

4.4.1 Estimator Derivation

We propose a method for *a-priori* SNR estimation, using the approach in [45], where the computation is composed of a *propagation* and *update* steps.

IF (l mod *decision interval* $== 0$) **DO**:

COMPUTE

$$\underline{c}_1 = \log\left(\frac{\Phi^l_{\text{long}}(\text{everyone})}{\Phi^l_{\text{long}}(\text{dominant})}\right)$$

$$\underline{c}_2 = \log\left(\frac{\Phi^l_{\text{medium}}(\text{everyone})}{\Phi^l_{\text{medium}}(\text{dominant})}\right)$$

$$\underline{c}_3 = \log\left(\frac{\Phi^l_{\text{immediate}}(\text{everyone})}{\Phi^l_{\text{immediate}}(\text{dominant})}\right)$$

IF exists $\{j : c_1(j) > 3 \;\&\; c_2(j) > 2 \;\&c_3(j) > 0\}$,

$$j^* = \max_j\{c_2(j) : c_2(j) > 2\}$$

$$\text{Dominant}(l) = j^*$$

ELSE

$$\text{Dominant}(l) = \text{Dominant}(l-1)$$

ELSE

$$\text{Dominant}(l) = \text{Dominant}(l-1)$$

Figure 4.3: The dominant speaker identification algorithm

Let $x(n)$ and $d(n)$ denote clean speech and uncorrelated noise signals, respectively. The observed signal is $y(n) = x(n)+d(n)$ or in the time-frequency domain, $Y_l = X_l+D_l$, where l is the time-frame index and $Y_l, X_l,\ D_l$ are vectors of the respective DFT coefficients. The vector of speech coefficients, X_l is a complex-valued vector, $X_l = A_l e^{j\phi_l}$, with ϕ_l representing the phase and $A_l = |X_l|$ the magnitude. Let $\mathcal{Y}^{l-1} = \{Y_t | t = 0, 1, ..., l-1\}$ represent the set of measurements up to frame $l-1$. The estimator for the conditional variance of X_l, conditioned on $\mathcal{Y}^{l-1}$, is denoted by $\hat{\lambda}_{l|l-1}$, and is referred to as the *one frame ahead conditional variance*. The variance of noise, D_l, is denoted by λ_{D_l}.

The *propagation* step aims to estimate $\hat{\lambda}_{l|l-1}$. It is obtained by assuming a model on the propagation of $\hat{\lambda}_{l|l-1}$ in time, conditioned on values estimated up to frame $l-1$. We propose to model the propagation of $\hat{\lambda}_{l|l-1}$ in time, as a random GARCH(1,1) process [31], so that the *propagation* step is given by

$$\begin{aligned}\hat{\lambda}_{l|l-1} &= E\left\{\lambda_{min} + \mu|X_{l-1}|^2 + \delta\left(\lambda_{l-1|l-2} - \lambda_{min}\right)\right. \\ &\quad \left.\left|\hat{A}_{l-1}, \hat{\lambda}_{l-1|l-2}\right.\right\} \\ &= \lambda_{min} + \mu\hat{A}^2_{l-1} + \delta\left(\hat{\lambda}_{l-1|l-2} - \lambda_{min}\right) \qquad (4.1)\end{aligned}$$

where

$$\lambda_{min} > 0,\ \mu \geq 0,\ \delta \geq 0,\ \mu + \delta < 1 \tag{4.2}$$

In the *update* step, we acquire $\hat{\lambda}_{l|l}$, given the current measurement, Y_l and $\hat{\lambda}_{l|l-1}$, as $\hat{\lambda}_{l|l} = E\left\{A_l^2 \left| \hat{\lambda}_{l|l-1}, Y_l \right.\right\}$. This is obtained by applying a spectral gain function, $G\left(\hat{\lambda}_{l|l-1}, \gamma_l\right)$, to Y_l [45]

$$\hat{\lambda}_{l|l} = G\left(\hat{\lambda}_{l|l-1}, \gamma_l\right)^2 |Y_l|^2 \tag{4.3}$$

Where $\gamma_l = |Y_l|^2/\lambda_{D_l}$ denotes the *a-posteriori* SNR.

In the proposed method, we chose the spectral gain function, $G_{SP}\left(\lambda_{l|l}, \gamma_l\right)$ that minimizes the expected spectral power distortion measure, $d_{SP}(X_l, \hat{X}_l) \triangleq \left[A_l^2 - \hat{A}_l^2\right]^2$. Assuming that the spectral coefficients of speech and noise are modeled by independent, zero mean, Gaussian random variables [49], G_{SP} is given by [43]

$$G_{SP}\left(\hat{\lambda}_{l|l-1}, \gamma_l\right) = \sqrt{\frac{\hat{\lambda}_{l|l-1}}{\lambda_{D_{l-1}} + \hat{\lambda}_{l|l-1}} \left(\frac{1}{\gamma_l} + \frac{\hat{\lambda}_{l|l-1}}{\lambda_{D_{l-1}} + \hat{\lambda}_{l|l-1}}\right)} \tag{4.4}$$

The estimator proposed in this work is obtained by substituting (4.4) into (4.3)

$$\hat{\lambda}_{l|l} = \frac{\hat{\lambda}_{l|l-1}}{\lambda_{D_{l-1}} + \hat{\lambda}_{l|l-1}} \left(\lambda_{D_{l-1}} + \frac{\hat{\lambda}_{l|l-1}|Y_l|^2}{\lambda_{D_{l-1}} + \hat{\lambda}_{l|l-1}}\right) \tag{4.5}$$

Finally, the *a-priori* SNR estimator is obtained by dividing both sides of (4.5), by λ_{D_l}

$$\hat{\xi}_{l|l} \triangleq \frac{\hat{\lambda}_{l|l}}{\lambda_{D_l}} \tag{4.6}$$

4.4.2 Relation to the Decision Directed Estimator

We would like to compare the proposed estimator to the decision-directed estimator (DD) of Ephraim and Malah [30] that may be written as

$$\hat{\lambda}_{l|l}^{DD} = \alpha \hat{A}_{l-1}^2 + (1-\alpha)\max\{|Y_l|^2 - \lambda_{D_l}, 0\} \tag{4.7}$$

where α $(0 \leq \alpha \leq 1)$ is a weighting factor that controls the trade off between noise reduction and signal distortion.

It was shown in [45], that $\hat{\lambda}_{l|l}$, in (4.5), can be rewritten as

$$\hat{\lambda}_{l|l} = \alpha_l \hat{\lambda}_{l|l-1} + (1-\alpha_l)(|Y_l|^2 - \lambda_{D_l}) \tag{4.8}$$

where the time dependent term α_l, is given by

$$\alpha_l \triangleq 1 - \frac{\hat{\lambda}^2_{l|l-1}}{(\lambda_{D_{l-1}} + \hat{\lambda}_{l|l-1})^2} \tag{4.9}$$

For the proposed estimator (4.5), assuming in (4.1) that $\delta = 0$ and taking $\mu \to 1$, (4.8) takes the form

$$\hat{\lambda}_{l|l} = \alpha_l(\lambda_{min} + \hat{A}^2_{l-1}) + (1 - \alpha_l)(|Y_l|^2 - \lambda_{D_l}) \tag{4.10}$$

and

$$\alpha_l = 1 - \frac{(\lambda_{min} + \hat{A}^2_{l-1})^2}{(\lambda_{D_{l-1}} + \lambda_{min} + \hat{A}^2_{l-1})^2} \tag{4.11}$$

In the DD method (4.7), α is set a-priori, and is usually close to 1. The time dependency of α_l introduces data driven flexibility, as follows. When SNR is low, α_l (4.9) is closer to 1, and the estimator in (4.8), relies mostly on the spectral amplitude estimator from the previous frame, while the term $(|Y_l|^2 - \lambda_{D_l})$, based on the current frame, is mostly discarded. This is a desirable behavior, since $|Y_l|$ is a random variable, and may significantly vary from the noise spectrum estimate, λ_{D_l}, creating spectral peaks which translate to *musical noise* in the time domain [44].

In [45], the estimator

$$\hat{\lambda}^{RS}_{l|l} = \alpha^{RS}_l \hat{A}^2_{l-1} + (1 - \alpha^{RS}_l)(|Y_l|^2 - \lambda_{D_l}) \tag{4.12}$$

with

$$\alpha^{RS}_l = 1 - \frac{\hat{A}^4_{l-1}}{(\lambda_{D_{l-1}} + \hat{A}^2_{l-1})^2} \tag{4.13}$$

that is a private case of the estimator in (4.10), with $\lambda_{min} = 0$, is discussed. It had been concluded that introducing the time dependent term α^{RS}_l, results in a smoother estimator, compared to the DD estimator. The reduced form of the GARCH based estimator, in (4.10), introduces additional regularization into the estimator in (4.12), by adding the coefficient λ_{min} to the term $\hat{A}^2_{l-1}$. The coefficient λ_{min} prevents from weighing in a zero estimate for the spectral amplitude, thus regularizing $\hat{\lambda}_{l|l}$ estimate by forcing it to be at least λ_{min}. Another influence of λ_{min} is observed when comparing the time dependent weighting factor between (4.11) and (4.13). In (4.11) λ_{min} imposes $0 < \alpha_l < 1$, which means that both the immediate time term $(|Y_l|^2 - \lambda_{D_l})$ and the estimate based on the past measurements, $\hat{A}^2_{l-1}$, are being weight into the calculation of $\hat{\lambda}_{l|l}$ in (4.10).

The full GARCH(1,1) model (4.1), has two additional parameters, μ and δ , which can also be adjusted, weighing in information computed up to time-frame $l-2$. The GARCH based method was chosen over the traditional DD approach for this work, due to the data-driven flexibility it offers. The validity of this choice was verified by comparing it to the performance of the DD estimator in the speech enhancement application described in [31], where the proposed estimator yielded comparable results.

4.5 speech activity score evaluation

In this section, we formulate the speech activity scores evaluation method. As discussed in 4.3, the speech activity score for a certain time-interval is determined by the number of active sub-units in a representative vector. We consider the log-likelihood ratio of the number of active sub-units, as the respective speech activity score. In order to determine the likelihood ratio of the number of active sub-units, we need to assume a likelihood model on this number, under the assumption that it originates in a speech or non-speech signal segment, denoted by H_1 and H_0 respectively. We propose two approaches for the score calculation method: the first approach uses the information from a single observation, and the second approach makes use of a sequence of observations in the likelihood-ratio formulation process. We also discuss the difference between the two approaches and the expected difference in performance.

4.5.1 Modeling the number of active sub-units

Let $\underline{\nu}_l = [\nu(l), \nu(l-1), ..., \nu(l-N_R+1)]$, denote an N_R long representative vector. The vector $\underline{\nu}_l$ is thresholded by the threshold value ν_{th}, resulting in a binary vector $\underline{\nu}_{l,\text{binary}}$. The vector $\underline{\nu}_{l,\text{binary}}$ is summed

$$v(l) = \sum_{m=1}^{N_R} \nu_{\text{binary}}(m) \tag{4.14}$$

The value, $v(l)$ is the number of active sub-units out of the total number of entries, N_R, in the original vector, $\underline{\nu}_l$. We propose to model this number as follows:

1. *Given H_1: speech is present*

   We regard every active sub-unit as a success in a Bernoulli trial, where $P(x) =$

$p^x(1-p)^{(1-x)}$, with $x \in \{0,1\}$, and p is the probability of success, equal for all vector entries. The vector length is N_R, thus we compute the probability of $v(l)$ successes out of N_R experiments. Hence, we assume this number follows the Binomial distribution

$$\begin{aligned} P(v(l)|H_1) &\sim \text{Bin}(N_R, p) \\ &= \binom{N_R}{v} p^{v(l)}(1-p)^{N_R - v(l)} \end{aligned} \tag{4.15}$$

2. *Given H_0: speech is absent*

   When speech is absent, we expect a lower probability for a higher number of active sub-units, hence we assume an Exponential distribution

$$P(v(l)|H_0) \sim \exp(\lambda) = \lambda e^{-\lambda v(l)} \tag{4.16}$$

The distribution constants, λ and p , are obtained from a training set.

4.5.2 Score Evaluation

Given an observation vector X_l, representing time-frame l, and two possible classes of its origin, H_0 and H_1. The likelihood of the observation to belong to each class, $i \in \{0,1\}$ is given by $p(X_l|H_i)$. Accordingly, the likelihood ratio is given by [50]

$$\Lambda_l = \frac{p(X_l|H_1)}{p(X_l|H_0)} \tag{4.17}$$

We define the speech activity score as the log-likelihood ratio of the observation vector. It is obtained by taking the natural logarithm of (4.17)

$$\Phi_l = \ln\left(\frac{p(X_l|H_1)}{p(X_l|H_0)}\right) \tag{4.18}$$

In the following sections, two approaches for score evaluation are proposed. In the first approach, X_l is the number of active sub-units in the current representative vector. In the second approach, X_l is a vector consisting of a sequence of the number of active sub-units, on a sequence of the respective representative vectors. The scores are hereafter referred to as *Binomial* and *Binomial-sequential* respectively.

Single observation approach

In this approach the score is based on the number of active sub-units, $v(l)$, in the representative vector, $\underline{\nu}_l$. Substituting the model assumptions, from (4.15) and (4.16), into (4.18), we have the speech activity score based on a single observation

$$\begin{aligned}\Phi_l^{\text{single}} &= \ln\left(\frac{p(X_l = v(l)|H_1)}{p(X_l = v(l)|H_0)}\right) \\ &= \ln\left(\frac{\binom{N_R}{v}p^{v(l)}(1-p)^{N_R - v(l)}}{\lambda e^{-\lambda v(l)}}\right) \\ &= \ln\binom{N_R}{v(l)} + v(l)\ln p + (N_R - v(l))\ln(1-p) \\ &\quad - \ln\lambda + \lambda v(l)\end{aligned} \tag{4.19}$$

Multiple observation approach

We follow the approach proposed in [25] for a VAD application. In this method, an HMM is used to recursively update the likelihood ratio of frame l, using all previous frames.

We base the score in this approach on a sequence of N sequential values of active sub-units $\underline{v}_l^N = [v(l-N+1), v(l-N+2), ..., v(l)]$, taken from the N preceding and including, the observed time-frame, l. The hidden Markov model here consists of two states:

$$\zeta_l = \begin{cases} H_{1,(l)} & \text{speech is present in frame } l \\ H_{0,(l)} & \text{speech is absent in frame } l \end{cases}$$

With the state dynamics described by
$a_{ij} = p(\zeta_l = j|\zeta_{l-1} = i)$. For speech signals, it is more likely that a frame of speech would be followed by a frame of speech rather than by a frame of silence. This notion is introduced by setting $a_{00}, a_{11} > a_{01}, a_{10}$.

The likelihood-ratio in this approach is

$$\begin{aligned}\Lambda_l^{\text{sequential}} &= \frac{p(X_l = \underline{v}_l^N|H_1)}{p(X_l = \underline{v}_l^N|H_0)} \\ &= \frac{p(\underline{v}_l^N, H_1)}{p(\underline{v}_l^N, H_0)} \cdot \frac{pH_0}{pH_1}\end{aligned} \tag{4.20}$$

Where pH_0 and pH_1, are steady state probabilities that are determined by $pH_0 = \frac{a_{10}}{a_{10}+a_{01}}$ and $pH_1 = \frac{a_{01}}{a_{10}+a_{01}}$.

Introducing the notations $\alpha_l(1) \triangleq p(\underline{v}_l^N, H_1)$ and $\alpha_l(0) \triangleq p(\underline{v}_l^N, H_0)$, $\alpha_l(j), j \in \{0,1\}$ is recursively computed using the forward procedure [51]:

$$\alpha_k(j) = \begin{cases} P[X_k = v(k)|H_{j,(l)}] \\ \cdot [\alpha_{k-1}(0)a_{0j} + \alpha_{k-1}(1)a_{1j}], l-N+1 < k \leq l \\ P[X_k = v(k)|H_{j,1}]p(H_{j,1}) \quad , k = l-N+1 \end{cases}$$

Denote the recursively computed term

$$\begin{aligned} L_l &= \frac{\alpha_l(1)}{\alpha_l(0)} \\ &= \frac{p(v(l)|H_1)}{p(v(l)|H_0)} \cdot \frac{\alpha_{l-1}(0)a_{01} + \alpha_{l-1}(1)a_{11}}{\alpha_{l-1}(0)a_{00} + \alpha_{l-1}(1)a_{10}} \\ &= \Lambda_l^{\text{single}} \cdot \frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}} \end{aligned} \tag{4.21}$$

It is important to note for future discussion that the term L_l is large when speech is present and it is small in the absence of speech.

Substituting (4.21) into (4.20), we have

$$\begin{aligned} \Lambda_l^{\text{sequential}} &= L_l \cdot \frac{pH_0}{pH_1} \\ &= \Lambda_l^{\text{single}} \cdot \frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}} \cdot \frac{pH_0}{pH_1} \end{aligned} \tag{4.22}$$

Where a_{10} and a_{01} are determined a priori.

Finally, we have the speech activity score formulation for the approach based on a sequence of observations

$$\begin{aligned} &\Phi_l^{\text{sequential}} \\ &= \Phi_l^{\text{single}} + \ln\left(\frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}}\right) \\ &\quad + \ln\left(\frac{pH_0}{pH_1}\right) \end{aligned} \tag{4.23}$$

The intuitive advantage of the sequential approach may be observed in the structure of the sequential score (4.23). The sequential score (4.23) differs from the single score, Φ_l^{single}, by the addition of the terms $\ln\left(\frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}}\right)$ and $\ln\left(\frac{pH_0}{pH_1}\right)$. The term $\ln\left(\frac{pH_0}{pH_1}\right)$ is a constant bias term added to the score in all cases. The term $\ln\left(\frac{a_{01} + L_{l-1}a_{11}}{a_{00} + L_{l-1}a_{10}}\right)$ on the other hand, influences the score differently, in the presence or absence of speech. With the constant values $a_{00}, a_{11} > a_{01}, a_{10}$ set a-priori:

1. In the presence of speech, L_{l-1} is large, so the term $\ln\left(\frac{a_{01}+L_{l-1}a_{11}}{a_{00}+L_{l-1}a_{10}}\right) \rightarrow \ln\left(\frac{a_{11}}{a_{10}}\right) > 0$, hence $\Phi_l^{\text{sequential}} > \Phi_l^{\text{single}}$,

2. When speech is absent, L_{l-1} is small, and the term $\ln\left(\frac{a_{01}+L_{l-1}a_{11}}{a_{00}+L_{l-1}a_{10}}\right) \rightarrow \ln\left(\frac{a_{01}}{a_{00}}\right) < 0$. So that in this case, $\Phi_l^{\text{sequential}} < \Phi_l^{\text{single}}$

Hence, the *Binomial-sequential* score achieves better separability between speech presence and absence, in comparison to the *Binomial* score. On the other hand, the integration of more observations from the past, introduces a delay into the speaker switching process. This happens because the new information regarding the *speaker switch* is masked by the old information of dominance. Both these differences are exhibited in the experimental results in section 4.6.

4.6 Experimental Results

In this Section, we compare the performance of the proposed method to other dominant speaker identification methods. Since a standard evaluation framework for the task of dominant speaker identification does not exist, we propose several experiments and objective error measures, to test the basic requirements for this type of system.

Throughout this section, we denote the proposed method when used with the single-observation approach for the score evaluation, 4.5.2, as the *Binomial* method and when used with the sequential-observation approach ,4.5.2, as the *Bin-Sequential* (*Bin-Seq* in figures 4.6 and 4.7) method. In case the distinct name of the method is not specified, the performance of the two methods was similar. The parameter set that was used in the experiments is provided in Figure 4.5.

In the first experiment, the dominant speaker identification algorithms are evaluated in a simple task of switching to the dominant speaker in the presence of stationary noise. For this purpose, a synthetic multipoint conference was simulated by concatenating speech segments taken from the TiMit database. Three speakers were randomly chosen from the database and several speech bursts were concatenated on a distinct channel, for each speaker. The speech bursts in each channel were spread along the conference length, such

For all channels i

For all time frames l

For frequency bins $k_1 < k < k_2$

Compute $\underline{\xi}_{i,l} = \{\xi_i(k,l) | k \in [k_1, k_2]\}$, as described in Section 4.4.

Count the number of active sub-bands in $\underline{\xi}_{i,l}$, using:

$$a_{i,1}(l) = \sum_{k=k_1}^{k_2} \text{sign}[\max(\xi_i(k,l) - \xi th, 0)]$$

where sign(x) is the Signum function.

Compute the score $\Phi_{i,l}^{\text{immediate}}$ using one of the methods described in Section 4.5.

Construct the vector $\underline{\alpha}_{i,l} = \{a_{i,1}(l-m) | m \in [0, N_2 - 1]\}$

Compute $a_{i,2}(l) = \sum_{m=0}^{N_2-1} \text{sign}[\max(\alpha_{i,l}(m) - \alpha_{th}, 0)]$

Compute score $\Phi_{i,l}^{\text{medium}}$ using one of the methods described in Section 4.5.

Construct the vector $\underline{\beta}_{i,l} = \{a_{i,2}(l - mN_2) | m \in [0, N_3 - 1]\}$

Compute $a_{i,3}(l) = \sum_{m=0}^{N_3-1} \text{sign}[\max(\beta_{i,l}(m) - \alpha_{th}, 0)]$

Compute score $\Phi_{i,l}^{\text{long}}$ using one of the methods described in Section 4.5.

Perform comparison of scores $\left\{\Phi_{i,l}^{\text{immediate}}, \Phi_{i,l}^{\text{medium}}, \Phi_{i,l}^{\text{long}}\right\}$ across channels, as described in Section 4.3.2 and determine which of the channels contains dominant speech.

Figure 4.4: Dominant speaker identification algorithm, based on speech activity information from time intervals of different lengths.

Fs = 16KHz	window length = 64 samples	overlap = 50%	
$k_1 = 2$, $k_2 = 12$	$N_2 = 33$	$N_3 = 16$	
$\xi_{th} = 3$	$\alpha_{th} = 5$	$\beta_{th} = 32$	
$p_{\text{immediate}} = 0.5$	$p_{\text{medium}} = 0.5$	$p_{\text{long}} = 0.5$	
$\lambda_{\text{immediate}} = 0.78$	$\lambda_{\text{medium}} = 24$	$\lambda_{\text{long}} = 47$	
	$pH_0 = pH_1 = 0.5$		
$a_{10} = 0.1$	$a_{11} = 0.9$	$a_{01} = 0.2$	$a_{00} = 0.8$

Figure 4.5: Algorithm parameters that were used in the experiment.

that each speech burst requires a switch in the dominant speaker. White noise in the range of -2 to 5 dB SNR was added to all signals. The algorithms were applied to the signals and the dominant speaker was identified once in a time-period denoted by *decision interval*.

Quantitative analysis was performed on the synthetic test set to examine whether an identification method fulfills the expectations, as stated in Section 4.2.
The analysis includes the following measures:

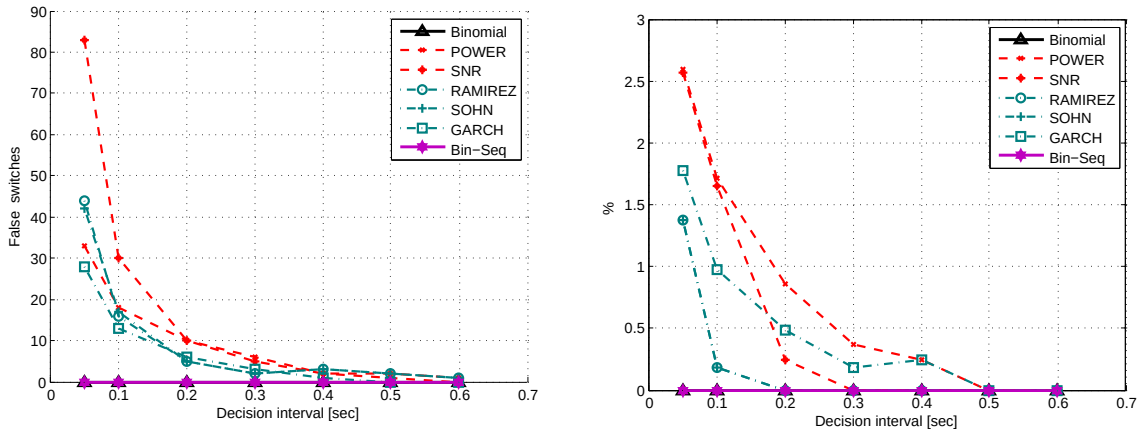
- *False speaker switches* - number of false switches to a non-dominant speaker.
- *Front End Clipping* (FEC) - error in detecting the beginning of a speech burst. The signal on each channel consists of several isolated speech bursts. The FEC error for each speech burst is obtained. Then, the mean FEC error for the whole conference, is computed as the mean value of individual FEC errors of all speech bursts in all channels. We assume that a tolerable delay in switching to the dominant speaker, in terms of mean FEC, is one second. All tested methods fulfilled this requirement.
- *Mid Sentence Clipping* (MSC) - clipping occurring in the middle of a speech burst. This is the most disturbing type of error since it causes a switch of speaker in the middle of a dominant speech burst. It is computed as the ratio between the undetected mid section of speech burst [samples] (UMSB) to the length of speech burst[samples] (LOSB) in percent.

$$\%\text{MSC} = 100 \cdot \frac{\text{UMSB}}{\text{LOSB}}$$

The performances of the two proposed methods were compared to the following methods:

- Three methods that identify the dominant speaker by applying a VAD to each channel, and identifying the speaker with the highest VAD score as dominant. The VAD methods used for the comparison are denoted by *Ramirez*, *Sohn*, and *GARCH*, and are described in [40], [25] and [27] respectively.
- A method that identifies the dominant speaker as the one with highest signal power. It is referred to as the *POWER* method throughout the comparison.
- A method that identifies the dominant speaker as the one with the highest SNR. It is referred to as the *SNR* method throughout the comparison.

Since the comparison is made once in a decision interval, the instantaneous values of the VAD score, SNR and POWER, are not necessarily representative values of the decision



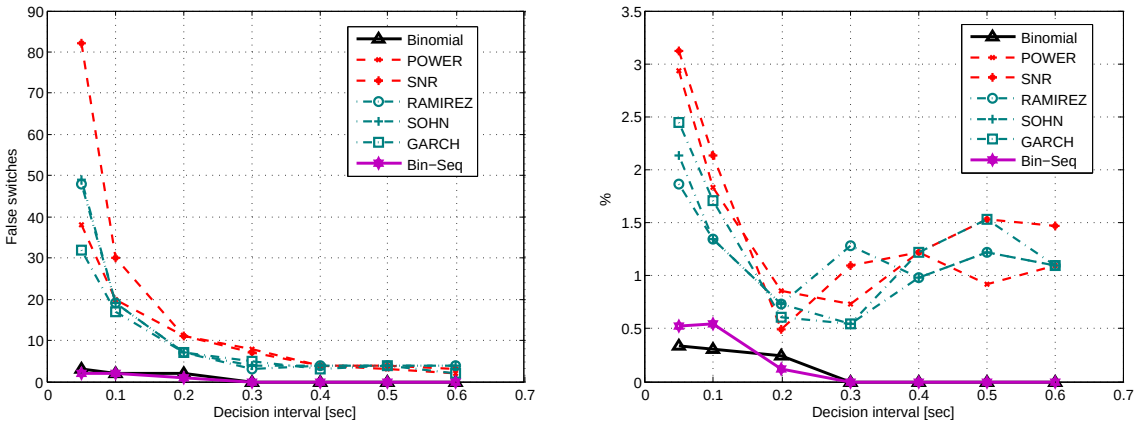
(a) (b)

Figure 4.6: (a) False speaker switches; (b) Mid Sentence clipping

interval. Hence, the maximal values of the VAD score, SNR and POWER, in the decision interval were used for the comparison. The results of this experiment are displayed in Figure 4.6, where the false switching and MSC errors are plotted as a function of the decision interval. In Figure 4.6(a), *POWER*, *SNR* and *VAD* based methods , all show frequent false speaker switching. For the proposed method, both the false switching and the MSC errors are zero.

In the second experiment, we test the robustness of the algorithms to transient noise. Transient noise occurrences of door knocks and sneezing, were added to the signals in the synthetic conference of the first experiment. The quantitative influence of the transient occurrences, is presented in Figure 4.7 and can be compared to the results in Figure 4.6. There is a rise both in the number of false switches and a respective rise of the MSC error, for all methods. The proposed method is affected by the transient occurrences, when a very short decision-interval ($0.05 - 0.2$ sec) is used (Figure 4.7(a)). The false switching that occurs with the proposed method is of shorter duration, in comparison to the other methods. This can be observed in the relative rise of the MSC errors in Figure 4.7(b) in comparison to Figure 4.6(b), for decision interval in the range $0.05 - 0.2$ sec.

The difference between the two proposed methods, as discussed in section 4.5.2, is shown in Figure 4.7, where for decision-intervals of $0.05 - 0.2$ sec, false switching occurs. The *Binomial-Sequential* method has a slight advantage over the *Binomial* method, in the



(a) (b)

Figure 4.7: Synthetic experiment with the presence of transient noise: (a) False speaker switches; (b) Mid Sentence clipping

number of false speaker switches, as displayed in Figure 4.7(a). This is expected because the *Binomial-Sequential* score is based on a sequence of observations. Thus, when more information from the past is integrated into the decision, it takes more time for the dominance of a speaker to dissipate or to build up. In case of a transient, its duration is too short for building up speech activity that would indicate dominance. That same sequential processing, causes a delay in switching back to the actual dominant speaker, as exhibited in Figure 4.7(b) by the higher value of the MSC error.

In Figure 4.8, we present a qualitative comparison, in the presence of transient noise. The proposed method is compared to the *POWER* and two VAD based identification methods, *Ramirez* and *GARCH*. The decision interval in this comparison is 0.3 seconds. In general, it is noticeable that the algorithms we are comparing to, are more responsive to a dominant speaker switch. This is due to the instantaneous nature of their scores. This responsiveness causes false switching to occur when there is a significant rise in energy on another channel, disregarding the cause for this rise. For the proposed method, the use of long-term information facilitates distinction between speech and transient audio occurrences.

The third experiment was performed on a segment of a real 5 channel multipoint conference, depicted in Figure 4.9. On this segment, only channels 2 and 4 contain

speech. Channel 1 contains a high level of stationary noise and some crosstalk from other channels. Channels 3 and 5 contain only crosstalk from other channels. The y axis in Figure 4.9 was scaled to the amplitude of the signals in the channel. Taking the maximal signal amplitude in channel 2 as a reference 1, the maximal signal amplitudes in channels 1, 3, 4 and 5 are $0.4, 0.5, 0.5$ and 0.1 respectively. In this experiment, the proposed method was compared to the *POWER* method.

The proposed method switches correctly from channel 2 to channel 4. It ignores a high energy transient that occurs during the dominant speech burst in channel 4. From channel 4, the proposed algorithm switches back to channel 2 and stays on this channel. It remains on the dominant speaker in spite of an utterance of the words "yes yes" and a noise transient on channel 3. Thus, the proposed method behaves according to the requirements stated in Section 4.2. The POWER method, on the other hand, switches frequently to channel 1, which contains stationary noise, during the first dominant speech burst in channel 2. It also switches to the noisy channels, during the dominant speech burst in channel 4 and throughout the second dominance period in channel 2.

4.7 Conclusion

We presented a novel dominant speaker identification method, for multipoint videoconferencing. The proposed method is based on evaluation of speech activity on time intervals of different length. The speech activity scores for the immediate, medium and long time-terms, are evaluated separately for each channel. Then, the scores are compared, and the dominant speaker in a given time-frame is identified, based on the comparison. We proposed two approaches for the score evaluation method. The single observation based approach, for which the scores enable a faster reaction of the identification algorithm to speaker switches. In the second approach, the score is based on a sequence of observations. This makes the algorithm more robust to transient audio occurrences, but is slower in responding to changes. The information from time intervals of different length, enables the proposed method to distinguish between speech and non-speech transient audio occurrences. Experimental results have shown the improved robustness of the proposed method against transient audio occurrences and frequent speaker switching, in comparison

to other speaker selection methods.

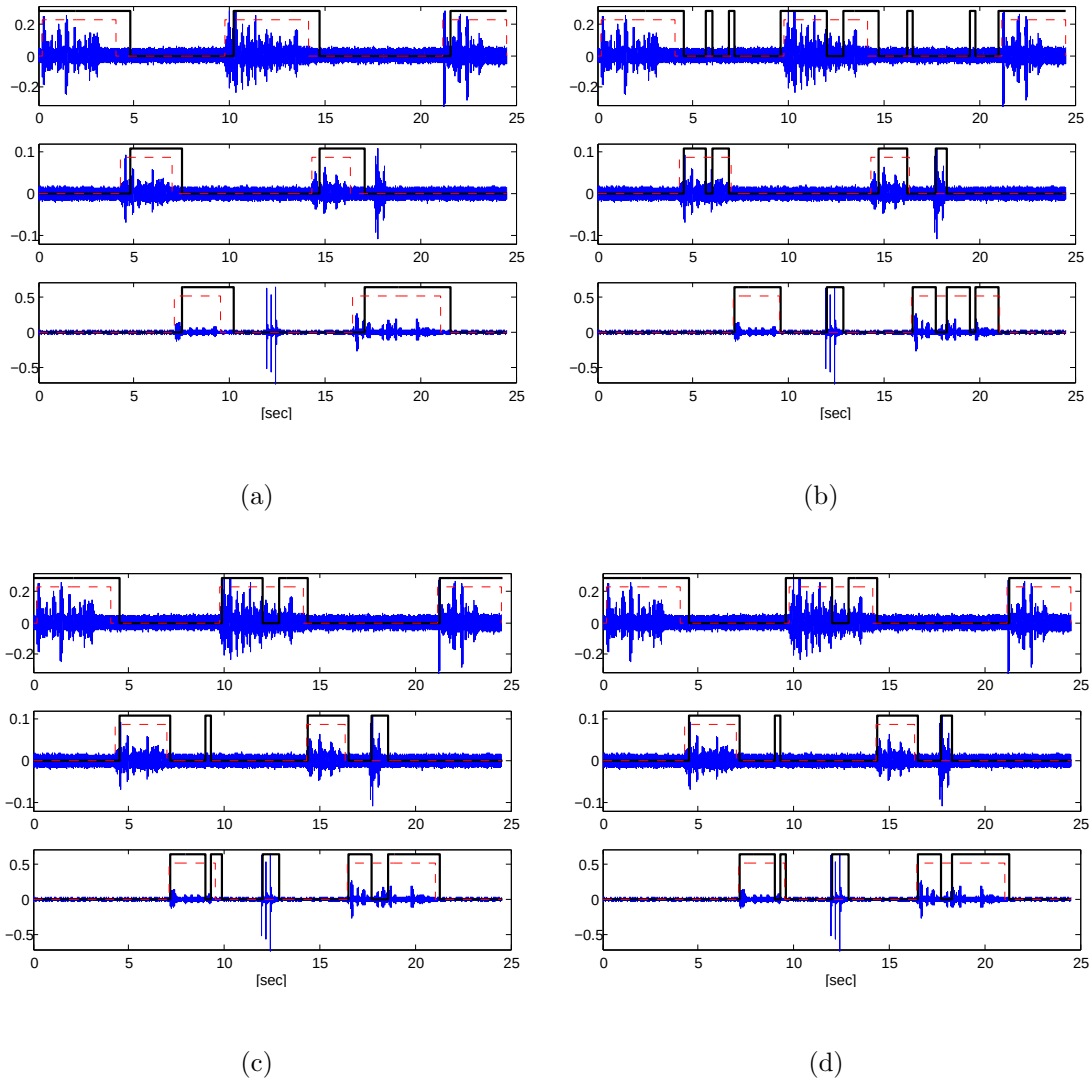
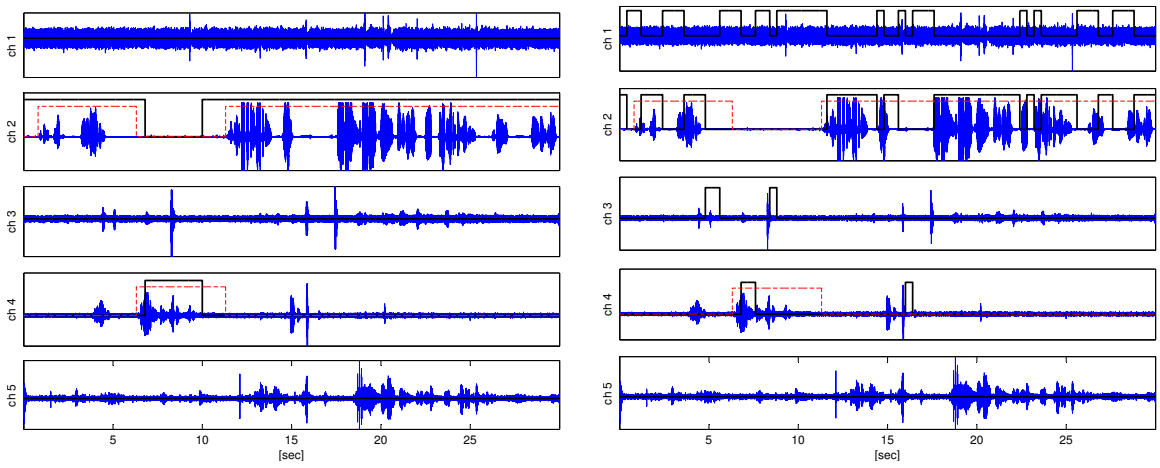


Figure 4.8: Results of dominant speaker identification, for decision-interval of 0.3 sec: (a) Dominant speaker identification by the proposed method; (b) dominant speaker identified by *POWER* method; (c) dominant speaker selected by the method based on *Ramirez* VAD; (d) dominant speaker selected by the method based on *GARCH* VAD ; the decision of the algorithm is marked by the higher *solid bold* line and the hand marked decision is marked by the low *dashed* line



(a) (b)

Figure 4.9: Experimental results on a real 5 channel multi-point conference for *decision interval* of 0.4sec; (a) Dominant speaker is selected by the proposed method; (b) dominant speaker is selected by the *POWER* method. The decision of the algorithm is marked by the high solid line and the hand marked decision is marked by the low dashed line

Chapter 5

Conclusion

5.1 Summary

In this thesis, we addressed the problem of dominant speaker identification in multipoint conferencing. We discussed the importance of dominant speaker identification in video-conferencing applications, and presented a novel method for the identification.

We introduced a novel method for dominant speaker identification. The proposed method consists of two stages. In the first stage, the signal in each channel is processed independently and the speech activity in each channel is evaluated. This method addresses the audio activity in the current frame as part of three time intervals, each of a different length. Thus, the speech activity is evaluated for the immediate, medium and long time intervals. In the second stage, the obtained speech activity information facilitates the selection of the channel that exhibits the most dominant speech activity. The dominant speaker identification is performed by comparing the unique set of speech activity scores obtained for each channel.

Two approaches are proposed for the speech activity score evaluation process. These approaches were inspired by works in the field of VAD. One approach evaluates the scores based on the speech activity information on the three time intervals, for the current frame alone. The second approach introduces dependency between sequential frames via an HMM. The two approaches represent a tradeoff between the algorithm responsiveness to changes and its robustness to false switching due to transient audio occurrences.

In the score evaluation process, each of the three time intervals is treated as composed

of smaller sub-units. The score is determined by the number of active sub-units in the time interval. In particular, we propose two likelihood models for the number of the active sub-bands, under the assumptions of speech presence or absence respectively. For the speech presence assumption, we propose a 'success' model, where each active sub-unit is considered a success in a Bernoulli trial. For the assumption of speech absence, a larger number of active sub-units has a smaller probability to occur. Hence, the exponential probability is chosen to model this likelihood.

5.2 Future Research

The method proposed in this thesis can be further explored in the following aspects:

Non causal processing We have discussed the use of causal time intervals of different lengths. We believe that non-causal processing may prove beneficial for the proposed method. Non causal processing provides a 'glimpse' into the future audio activity, hence the content of this future may infer on the currently observed frame. Since, this type of processing introduces a delay into the selection process, it may only be used in applications that allow it.

Temporally variable thresholds The thresholds used in the proposed method are set a-priori, disregarding the SNR of the incoming signal. However, the distribution of the number of active sub-units varies according to the SNR of the incoming signal. The higher the SNR threshold, less sub-units pass it, especially when the SNR of the signal is relatively low. Hence, introducing a mechanism that would determine the threshold according to the measured SNR in a previous frame or a sequence of them, would bring the distribution of the number of active units closer to the assumed likelihood models.

Preprocessing - Speech Enhancement In our method, the signals are analyzed without any preprocessing, to enable low computational complexity. Application of a speech enhancement algorithm may improve the performance of the proposed method. Speech enhancement may also eliminate the need for time-varying thresholds.

Preprocessing - Echo detection or cancellation The proposed method does not deal with echoes from other channels. Echo signals have the same properties of speech. The application of an echo cancellation or echo detection algorithm as a preprocessing step, may provide a solution to this problem. By saving an *echo presence* flag, the scores of a channel containing echoes can be reduced so that it will not be identified as dominant speech in the selection stage.

Bibliography

[1] W. Kwak, S. Gardell, and B. Mayne Kelly, "Speaker identifier for multi-party conference," US Patent No.: 6,457,043 B1, Sep. 2002.

[2] Y. F. Chang, "Multimedia conference call participant identification system and method," US Patent No.: 6,304,648 B1, Oct. 2001.

[3] Y. Kyeong Yeol, P. Jong Hoon, and L. Jong Hyeong, "Linear PCM signal processing for audio processing unit in multipoint video conferencing system," in *Proc. IEEE Third Symposium on Computers and Communications (ISCC98)*, Athens , Greece, Jun. 1998, pp. 549 –553.

[4] P. Smith, P. Kabal, and R. Rabipour, "Speaker selection for tandem-free operation VoIP conference bridges," in *Proc. IEEE Workshop on Speech Coding*, Tsukuba , Japan, Oct. 2002, pp. 120 – 122.

[5] X. Xu, L. wei He, D. Florencio, and Y. Rui, "Pass: Peer-aware silence suppression for internet voice conferences," in *Proc. IEEE International Conference on Multimedia and Expo*, Toronto , Canada, Jul. 2006, pp. 2149 –2152.

[6] R. Colvin Clark and A. Kwinn, *The New Virtual Classroom*, 1st ed. Pfeiffer, 2007.

[7] R. Petersen, ""Real World" connections through videoconferencing we're closer than you think!" *TechTrends*, vol. 44, pp. 5–11, 2000. [Online]. Available: http://dx.doi.org/10.1007/BF02763308

[8] H. Molyneaux, S. O'Donnell, H. Fournier, and K. Gibson, "Participatory videoconferencing for groups," in *Proc. IEEE International Symposium on Technology and Society (ISTAS 2008)*, Fredericton, New Brunswick, Canada, Jun. 2008, pp. 1 –8.

[9] J. Vinsonhaler, J. Johnson, L. Braunstein, D. Henderson, R. Boman, and R. Gilliland, "A comparison of collaborative problem solving using face to face versus desktop video conferencing," in *Proc. IEEE Thirty-First Annual Hawaii International Conference on System Sciences*, Kohala Coast, Hawaii, USA, Jan. 1998, pp. 127 –134.

[10] S. Liu, H. Molyneaux, and B. Matthews, "A technical implementation guide for multi-site videoconferencing," in *Proc. IEEE International Symposium on Technology and Society (ISTAS 2008)*, Jun. 2008, pp. 1 –8.

[11] "Narrow-band visual telephone systems and terminal equipment," ITU-T Rec. H.320, Mar. 2004.

[12] "Packet-based multimedia communication systems," ITU-T Rec. H.323, Nov. 2000.

[13] P. Smith, P. Kabal, M. Blostein, and R. Rabipour, "Tandem-free operation for VoIP conference bridges," *IEEE Commun. Mag.*, vol. 41, pp. 136–145, May 2003.

[14] J. Forgie, C. Feehrer, and W. P, "Voice conferencing technology final report," MIT Lincoln Lab., Tech. Rep. DDC ADA074498, Mar. 1997.

[15] D. Nahumi, "Conferencing arrangement for compressed information signals," US Patent 5,390,177, Feb. 1995.

[16] T. Champion, "Multi-speaker conferencing over narrowband channels," in *Proc. IEEE Military Communications Conference (MILICOM'91)*, McLean, VA , USA, Nov. 1991, pp. 1220 –1223.

[17] M. Willebeek-LeMair, D. Kandlur, and Z.-Y. Shae, "On multipoint control units for videoconferencing," in *Proc. IEEE 19th Conference on Local Computer Networks*, Minneapolis, MN , USA, Oct. 1994, pp. 356 –364.

[18] T. Chua and D. Pheanis, "Bandwidth-conserving real-time VoIP teleconference system," in *Proc. IEEE Third International Conference on Information Technology: New Generations (ITNG 2006)*, Las Vegas, NV , USA, Apr. 2006, pp. 535 –540.

[19] P. J. Smith, "Voice conferencing over IP networks," Master's thesis, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada, Jan. 2002.

[20] P. T. Brady, "A technique for investigating on-off patterns of speech," *Bell Systems Technical Journal*, vol. 44, pp. 1 – 22, 1965.

[21] S. Bruhn, E. Ekudden, and K. Hellwig, "Continuous and discontinuous power reduced transmission of speech inactivity for the GSM system," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM'98). The Bridge to Global Integration*, Sydney, Australia, Nov. 1998, pp. 2091 –2096.

[22] S. Chawla, H. Saran, and M. Singh, "QoS based scheduling for incorporating variable rate coded voice in bluetooth," in *Proc. IEEE International Conference on Communications(ICC 2001)*, Helsinki, Finland, Jun. 2001, pp. 1232 –1237.

[23] A. Gersho and E. Paksoy, "An overview of variable rate speech coding for cellular networks," in *Proc. IEEE International Conference on Selected Topics in Wireless Communications*, Vancouver, Canada, Jun. 1992, pp. 172 –175.

[24] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Québec, Canada, Oct. 1993, pp. 85 –86.

[25] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1 –3, Jan. 1999.

[26] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio Speech, and Language Process.*, vol. 14, no. 2, pp. 412 – 424, Mar. 2006.

[27] S. Mousazadeh and I. Cohen, "AR-GARCH in presence of noise: Parameter estimation and its application to voice activity detection," *IEEE Trans. Audio Speech, and Language Process.*, pp. 916–926, May 2011.

[28] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679 – 681, Aug. 1982.

[29] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. IEEE Press, 2000, ch. 8.

[30] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109 – 1121, Dec. 1984.

[31] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation," *Speech Communication*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[32] ——, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Processing*, vol. 86, no. 4, pp. 698 – 709, Apr. 2006.

[33] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Proc. IEEE 6th International Conference on Signal Processing*, Beijing, China, Aug. 2002, pp. 1124 – 1127 vol.2.

[34] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, pp. 64 –73, Sep. 1997.

[35] J. Ramírez, P. Yélamos, J. Górriz, C. Puntonet, and J. Segura, "SVM-enabled voice activity detection," in *Advances in Neural Networks (ISNN2006)*, Jun. 2006, pp. 676–681.

[36] Q. H. Jo, Y. S. Park, K. H. Lee, and J. H. Chang, "A support vector machine-based voice activity detection employing effective feature vectors," 2008. [Online]. Available: http://ietcom.oxfordjournals.org/cgi/content/short/E91-B/6/2090

[37] J. W. Shin, H. J. Kwon, S. H. Jin, and N. S. Kim, "Voice activity detection based on conditional MAP criterion," *IEEE Signal Process. Lett.*, vol. 15, pp. 257 –260, Feb. 2008.

[38] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498 – 505, Sep. 2003.

[39] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Pearson Education, 1993.

[40] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689 – 692, Oct. 2005.

[41] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586 – 1604, Dec. 1979.

[42] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725 –735, Apr. 1992.

[43] P. Wolfe and S. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. IEEE 11th Signal Processing Workshop on Statistical Signal Processing*, Aug. 2001, pp. 496 –499.

[44] O. Cáppe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345 –349, Apr. 1994.

[45] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870 – 881, Sep. 2005.

[46] S. Shaffer and W. Beyda, "Reducing multipoint conferencing bandwidth," US Patent No.: 6,775,247 B1, Aug. 2004.

[47] M. Howard, R. Burns, C. Lee, and M. Daily, "Teleconferencing system," US Patent No.: 6,775,247 B1, Aug. 2004.

[48] J. Ramírez, J. C. Segura, C. Benítez, Ángel de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, Apr. 2004.

[49] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443 – 445, Apr. 1985.

[50] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990.

[51] N. Shimkin, "Estimation and identification in dynamical systems," 048825 Lecture Notes, Fall 2009.