Florian Iglisch, BSc

Multichannel Voice Activity Detection for ASR

Master's Thesis

to achieve the university degree of

Diplom-Ingenieur

Masters's degree programme: Electrical Engineering and Audio Engineering

submitted to

Graz University of Technology

Supervisor

Dipl.-Ing. Dr.techn. Martin Hagmüller

Signal Processing and Speech Communications Laboratory

Dr. Juan Andrés Morales Cordovilla

Graz, November 18, 2015

Abstract

The present master's thesis extends a voice activity detection (VAD) algorithm based on deep belief network classification to perform multi-channel VAD and room localization in a smart home environment. The simulation data used for evaluating the multi-channel VAD approaches is based on real life data recorded in an apartment where several microphone arrays have been applied to the walls and ceiling in each room. The database is a result of the European research project DIRHA which has been active until 2014. The extended VAD uses at least one signal per room. The signals are compared to each other by computing a spectral cross-correlation function on short segments between each channel. The maxima of the resulting functions are related to each other. This reveals the time difference of arrival of the signal in each channel. Given the time differences, the corresponding room where the signal originates from can be detected. The spectral correlation for detection is used before performing VAD where it locates acoustic events which then will be classified which proved to be more stable under some conditions. It is shown that in addition, channel selection plays an important role to improve the performance of the multi-channel VAD. While using only one channel per room increases computation speed, an algorithm being adaptive to different speaker positions is necessary. Several approaches are presented and compared to stationary channel sets. Overall, addressing recall as an important metric for VAD, it reaches 85% correctly detected and localized speech segments in best cases with an estimated standard deviation of less than 10%.

Zusammenfassung

In der vorliegenden Masterarbeit wird ein Deep Belief Network Klassifikator zur Stimmaktivitätserkennung (VAD) um einen Algorithmus zur Raumlokalisierung erweitert. Der Mehrkanal-VAD soll in einer Smart-Home Umgebung zum Einsatz kommen und eine robuste Lokalisierung des Raums in dem gesprochen wird durchführen. Die Leistungsfähigkeit des VAD wird anhand von Aufnahmen ermittelt, welche in einem realen Appartment ausgestattet mit Mikrofonarrays and Wänden und Decken im Rahmen des europäischen Forschungsprojekts DIRHA aufgenommen wurden. Der Mehrkanal-VAD benutzt mindestens ein Mikrofonsignal pro Raum. Die Signale werden untereinander mittels Kreuzkorrelation kurzer Abschnitte ihres Spektrogramms verglichen. Aus den zeitliche Unterschiede der einzelnen Maxima der Korrelationsfunktionen lässt sich der Kanal und der zugehörige Raum ermitteln, in die Signalquelle vermutet wird. Wird dieser Algorithmus vor der VAD berechnet dient es der allgemeinen Detektion und Lokalisierung von akustischen Ereignissen. Diese werden anschließend nur noch als Sprache oder Störgeräusch klassifiziert was sich unter bestimmten Bedingungen als vorteilhaft erweist. Ein zusätzlicher Faktor, der nachweislich Einfluss auf die Performance der Mehrkanal-VAD hat ist die Kanalauswahl. Im Idealfall ist ein Kanal pro Zimmer ausreichend, allerdings muss adaptiv dafür gesorgt werden, dass sich die Auswahl an unterschiedliche Positionen eines Sprechers anpasst. Diesbezüglich werden mehrere Ansätze präsentiert. Die Performance wird hauptsächlich mit einem Recall-Wert beschrieben, welcher korrekt erkannte und lokalisierte Sprachsegmente in Relation zu ihrer eigentlichen Gesamtzahl setzt. Ein Recall von 85% gemittelt über alle Versuche mit einer Standardabweichung von weniger als 10% konnten erzielt werden.

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all ma- terial which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis dissertation.

Date	Signature

Contents

Acronyms

AMS	Amplitude Modulation Spectrograms.
ASR	Automatic Speech Recognition.
DBN	Deep Belief Network.
DFT	Discrete Fourier Transform.
DIRHA	Distant-speech Interaction for Robust Home Applications.
LinDA	Linear Discriminant Analysis.
LPC	Linear Predictive Coding.
MFCC	Mel Frequency Cepstral Coefficients.
RASTA-PLP	Relative-Spectral Perceptual Linear Predictive Analysis.
RBM	Restricted Boltzmann Machine.
RMS	Root Mean Square.
SNR	Signal to Noise Ratio.
SPP	Speech Presence Probability.
SS	Spectral Subtraction.
STFT	Short Time Fourier Transform.
SVM	Support Vector Machine.
TDOA	Time Difference Of Arrival.
VAD	Voice Activity Detection.

Nomenclature

acc	Accuracy - measure of correct classified samples.
AMS	Amplitude Modulation Spectrograms feature.
C	cepstrum.
c	correlation index.
d	binary decision function.
F	total number of frequency bands.
f	frequency index.
f_0	fundamental frequency.
F_1	F-Score - combining RE and PR in a single value.
FN	number of False Negatives.
FP	number of False Positives.
FR	number of False Rooms.
K	total number of frames.
k	frame index.
k_{len}	frame length.
k_r	frame rate.
l	smoothing length.
MDFT_l	Mel Discrete Fourier Transformation feature with smoothing l.
MFCC_l	Mel Frequency Cepstrum Coefficients feature with smoothing l.
n	sample index.
N_{ch}	number of channels used.
N_{sp}	total number of samples containing speech.
PR	Precision - measure of amount of noise detected as speech.
q	quefrency index.
R	correlation.
RE	Recall - measure of actual detected speech.
s_r	sample rate.
TN	number of True Negatives.
TP	number of True Positives.
$v(k)$	frame dependent variance.
X_{est}	estimation of clean speech signal in frequency domain using noise suppression.
x_{est}	estimation of clean speech signal in time domain using noise suppression.
x_n	noise signal in time domain.

$X_{n,est}$	estimation of noise signal in frequency domain using noise suppression.
$x_{n,est}$	noise signal in time domain.
X_s	clean speech signal in frequency domain.
x_s	clean speech signal in time domain.
Y	noisy speech signal in frequency domain.
y	noisy speech signal in time domain.
$\boldsymbol{y}[n]$,$\boldsymbol{y}_A$,$\boldsymbol{y}_B$	vector of all time dependent speech signals, A and B denote sub groups of all available signals.

1 Introduction and Topic Review

The mobile life we know today and in particular the kind of interaction with our environment is a permanently changing process. New technologies and inventions allow us to combine more and more interactions into one single device such that a mobile phone has become a remote control for a lot of applications. When driving a car or staying at home, avoiding the need of a carried device by somehow monitoring the scenery can be seen as the next step of interacting with our environment like we interact with other people.

People use the term *Smart-Home* when referring to a monitoring system which is installed in a living environment such that certain situations cause a reaction by the system. Using microphones, the monitoring focusses on acoustic events like falling objects or people talking. Depending on that, the system automatically could call for help in an emergency or trying to understand spoken phrases extracting commands like switching the light on or off. It is obvious, that Smart-Home environments have a huge potential to support elderly or disabled people at home as well as improving the lifestyle and comfort in general.

Research in the field of acoustic monitoring aiming for Automatic Speech Recognition (ASR) in particular is still at its roots but develops rapidly as algorithms becomes more efficient and computational power increases. The European research project Distant-speech Interaction for Robust Home Applications (DIRHA) was one attempt to completely establish a smart-home environment by working in a real apartment with multiple channels per room.

In this thesis, the classical single channel Voice Activity Detection (VAD) approach will be extended to a multichannel VAD which is able to localise speech at room level without loosing time efficiency. To achieve this, the interaction of event localisation and VAD will be studied by comparing different approaches. A database which has been generated as part of DIRHA will be used for evaluation.

1.1 DIRHA - Distant-speech Interaction for Robust Home Applications

DIRHA was an European research project[1] with partners in Austria, Greece, Italy and Portugal aiming to build a smart-home environment being able to perform ASR and react when a known command has been detected. For each room one or more arrays of two to six microphones have

[1] http://dirha.fbk.eu

been applied to the walls or the ceiling yielding in a 40 channel signal in total for the apartment. The final goal of the projects was to implement a prototype with the fully working system.

To control complexity, the project was split in several research fields like VAD, which will be shortly reviewed in Sec. 1.2, ASR, source separation or localisation, which in the end should work together. The main focus lies on its robustness but also on its capability of working in real-time.

As people from four different countries work on this project, the system should be capable of all these languages. As a fifth language English is used for comparison purposes. To accomplish this, a basic database has been generated which contains a variety of environmental noise sources to simulate different sceneries. This database then has been extended with language dependent phrases or commands by using impulse responses measured at several discrete positions for all channels to include the spatial information. Hence, research can focus on different languages while keeping the data comparable [1].

Figure 1.1 shows the floor-plan of the apartment where the recordings have been produced. The black dots indicate where microphones have been placed while the coloured rectangles stand for the possible positions of a speaker during a recording. The four or eight arrows pointing away from the rectangles give information about speech direction.

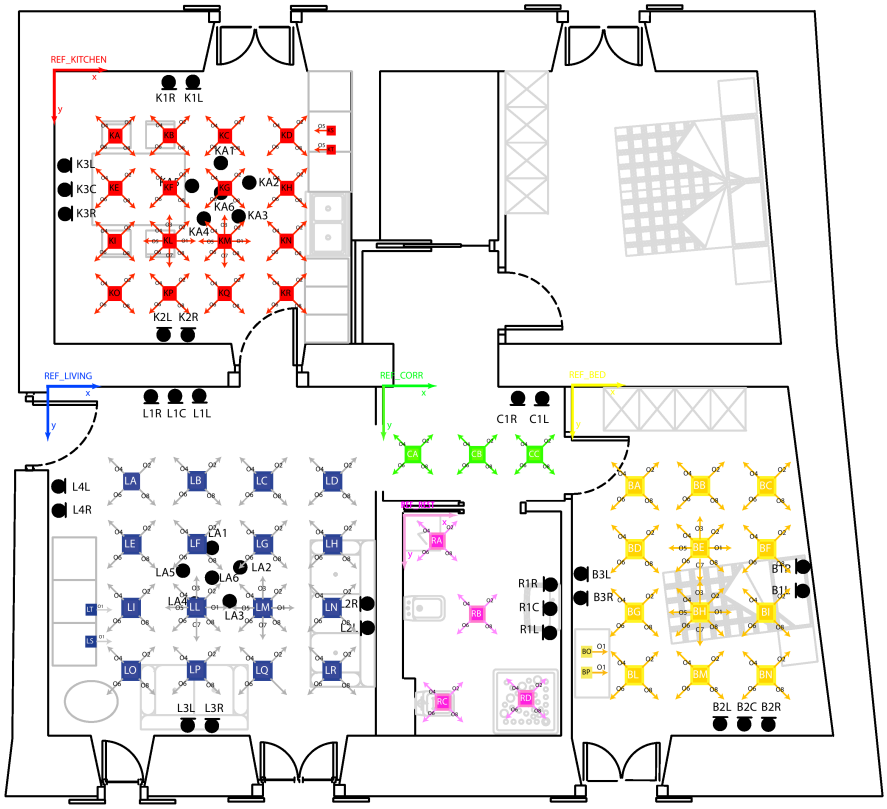


Figure 1.1: Floor-plan of the apartment where the recordings of the databases have been made. [1]

1.2 VAD - Voice Activity Detection

Today, VAD is widely used in many applications like telecommunications where different compressing schemes lead to a higher reduction in bandwidth, acoustical computer interfaces to interact with computers without the need of external devices or hearing aids to improve the signal quality. Depending on the requirements in terms of noise-type and Signal to Noise Ratio

(SNR) a VAD algorithm should be capable to cope with, the complexity of feature extraction and decision making varies a lot. For example with high SNR, Mel-Freq-Energy features and a simple threshold might be sufficient for good results whereas VAD in case of a reverberant environment with different noise sources like radio or tv, the selection of a robust feature set and classification algorithm combined with prior signal enhancement becomes a crucial task. In [2] a detailed literature review of popular VAD approaches of the two past decades can be found. In this review, algorithms are distinguished by the different types of feature categories they use as well as by differentiating between decision making approaches, i.e. whether they are based on thresholding, statistical modelling or machine learning techniques.

While these VAD algorithms have been constructed to perform on single channel signals, multi channel VAD is of an increasing research interest following various purposes. On the one hand the use of multiple microphones enables new kinds of signal enhancement or extend known approaches making them more robust. For example in [3], dereverberation is performed to increase the SNR while the authors in [4] use the additional information to extend a psychoacoustics based algorithm for better performance.

On the other hand considering meetings, multiple microphones extend the binary VAD task between speech and noise by adding the need for reducing the effect of crosstalk as well, as has been shown in [5].

When using arrays of microphones, beamforming techniques based on differences in time of arrival or signal power can be adapted to perform automatic source localisation. In addition to signal enhancement, exploiting this information makes it possible to track where speech comes from as well as increasing discrimination between multiple sources [6–8].

There exist few attempts to integrate VAD and ASR in a smart-home environment. In the CompanionAble European Project a monitoring system for interaction with a robot was implemented which showed promising results assuming certain constraints [9]. The study in [10] addresses elderly people as target group trying to automatically detect distress calls. As constraints only one microphone and a limited set of sentences has been used but evaluation is based on realistic played scenarios. First results showed a promising call detection rate of roughly 75%. As an example for controlling applications like TV or radio by speech can be found in [11]. Here, a recognition rate of about 93% could be achieved using 26 command words being recorded by two microphones in front of the user and classified using Gaussian Mixture Models.

Compared to the preceeding examples the preconditions for VAD in the DIRHA project are unique in terms of multiple signals from different rooms which have to be taken into account. Work referring to room localisation has been done in **??** In addition, while background noise establishes a realistic noise-floor, additional noise sources like hovering or working in the kitchen temporarily affects SNR in a realistic way. A VAD algorithm therefore should be capable both in discriminating between speech and noise by classification as well as robustly locating sound events in general to exploit spatial information provided by multiple microphones.

1.3 Outline

The multichannel approach developed in this thesis can be separated into localisation of speech or events and discrimination between speech and noise. As first attempts started with classification followed by localisation it will be shown that by reversing this order results improve.

For localisation two algorithms exist based on spectral correlation and onset detection respectively. In this case the latter can be seen as a baseline being outperformed by the first approach. Classification mainly is done using Deep Belief Network (DBN) but will be compared to different approaches for validation.

In addition, because of the large number of channels available, attempts have been made as

well to find an appropriate set of channels the algorithms work on. Stationary as well as adaptive approaches will be compared and their influence on the overall performance discussed.

The results obtained will be evaluated leading to a recommended multichannel VAD which performed best.

The thesis is organized as follows. Chapter 2 provides background information. It is meant as a review and leaves deeper discussion of the theory to literature. In Chapter 3, the main steps for performing multichannel VAD are discussed in detail. Here, the focus lies on developing the algorithms for room dependent VAD which is preceded by the basic description of the experimental settings and steps being used for signal enhancement and signal preparation. As several approaches to reach the goal have been made, a short summary about the algorithms used in the experiments is done at the end of the chapter. In Chapter 4 results will be presented which have been obtained when using the different sets of algorithms on the DIRHA-GRID corpus. In addition results of experiments with partial knowledge of the database like speech occurrence or speech origin are presented as well to support understanding of the performance of the algorithms. Chapter 5 contains the evaluation of the algorithms results depending on the setting. As the approaches for multichannel VAD separate the detection of voice and its origin, it can be shown that the performance depends on the order of these two steps in terms of classification performance of the DBN classifier. At the end of the chapter conclusion regarding the best approach leads to an outline of possibilities where further research could improve the performance of the VAD.

2 Background

The following chapter aims to give a rough overview concerning the theory behind the algorithms developed.

Section 2.1 addresses speech recognition mainly focussing on the characteristics of speech, how it can be described and eventually be detected. While this is a huge topic which has been covered a lot in literature this section is meant to be an introduction into the topics being used to develop the algorithms. For a detailed discussion of this complex research area the reader may be referred to standard literature.

Section 2.2 will give an overview about the relevant mathematics used in the developed algorithms for Voice Activity Detection (VAD). It is thought as a review stating important formulas and explaining some properties where needed. For derivations and deeper understanding the reader may be referred to standard literature.

2.1 Speech Recognition

Automatic Speech Recognition (ASR) in a complex acoustical environment of the real world faces the problem of finding a trigger which robustly activates further processing if speech exists and stops it otherwise. While one could do this by pressing a button for activation, it is of interest to automatically detect voice activity. For this, knowledge about speech is necessary how it can be described and how its characteristics can be exploited. Usually this is done by transforming the recorded signal, trying to focus on certain properties while suppressing everything else to extract features which together increase distinction between speech and noise. Thresholding or classifying this new signal eventually results in a binary signal which triggers the post processing of the recording.

The organisation of this section follows the three main steps mentioned: Sec. 2.1.1 introduces the speech model describing the basic characteristics of speech. Section 2.1.2 covers different domains the time signal may be represented in for extracting different types of features. Finally, Sec. 2.1.3 outlines classification. The latter basically refers to the work of [12] which has been taken as a basis for this thesis. For a general examination of the topics the reader may be referred to [13] or [14].

2.1.1 Characteristics of Speech

Speech basically is a mixture of noise-like sounds as plosives or fricatives as well as harmonic sounds like vowels or nasals. Here, when air passes through the vocal tract the vocal chords periodically open and close leading to a periodic train of short air pulses which result in a harmonic sound with fundamental frequency f_0. The spectrum of this sound is further formed by resonances in the vocal tract which change over time producing different characteristics of voiced speech. Several resonances, also called formants, in combination give each vowel its own characteristic.

For unvoiced speech, the vocal chords don't vibrate but are perturbed by turbulence in the vocal tract. Unvoiced phonemes like fricatives or sibilants as well as whispering therefore are just characterised by changes in the vocal tract.

Separating excitation in the glottis as source and the vocal tract as a filter, voiced and unvoiced speech or its mixture can be modelled as shown in Fig. 2.1. The source either is formed by a pulse train with a pulse frequency of f_0 or random noise which may be added to the pulses as well. The corresponding source signal is then filtered with the time varying filter $H(z)$ applying gain and formants for the final speech signal $s[n]$.

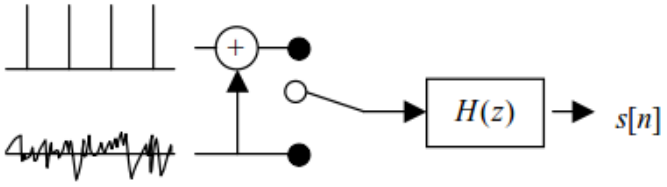


Figure 2.1: Source filter model of voiced or unvoiced speech [13].

The range of f_0 depends on the gender and lies for male and female together between 60 Hz and 500 Hz.

2.1.2 Transformation

A speech signal x_s usually is available as a sample train in time domain. While analysing x_s during overlapping timeslots using simple features like energy or zero crossing rate is easily possible, the information gain decreases rapidly if the time signal contains additional acoustic sources or if reverberation increases. On the other side, spectral analysis of x_s may reveal dominant periodicities and signal characteristics but at the loss of timing information.

The compromise to overcome these problems is to analyse short windowed sections of x_s with some overlapping. For each time slot the spectrum is calculated using Discrete Fourier Transform (DFT) in general, resulting in a spectrogram X_s. Computing the spectrogram of a time signal is also referred to as Short Time Fourier Transform (STFT) (see Sec. 2.2.1). X_s of x_s contains timing information as well as spectral distribution coming with the drawback of loosing both time and frequency resolution. For VAD time resolution is not that critical.

Taking the logarithm of X_s followed by the inverse Fourier transform leads to the cepstral domain whose coefficients have been shown to produce characteristic features for speech signals. Especially the preceding spectral compression by summing up certain frequency areas using a Mel filter-bank reduces bandwidth while retaining speech characteristics. For the Mel filter-bank psychoacoustical knowledge about spectral resolution of the ear is used which is high at low frequencies and decreases if frequencies become higher. For the mathematical definition of the resulting Mel Frequency Cepstral Coefficients (MFCC) see Sec. 2.2.1.

Computing features in time, spectral or cepstral domain aims to find a good representation of the time signal x_s while increasing discrimination between speech and noise and reducing bandwidth at the same time. Usually, features highly depend on the given task which makes

it very difficult, to define an optimal set. For this thesis, the feature sets used are based on a paper by Zhang et al [12]. The definitions of the set and its features can be read in Sec. 2.2.2.

2.1.3 Classification

Classification is the final step in the binary task of a VAD deciding if a time frame contains speech or not. In an ideal case features are good enough to define a simple threshold taking values above or below and rejecting everything else. But in reality different features work different well depending on the situation. Therefore more complex classification schemes are used to decrease the error rate.

As a preprocessing step features have to be unbiased and normalized to the same range. Without this step the numerical range of different features could vary a lot leading to dominating and disappearing features.

Zhang et al [12] did an extensive comparison with different classifiers on a huge dataset. The most efficient classifier used was a Deep Belief Network (DBN) which showed to produce the best results in most cases. While it needs a long training phase, classification by using a DBN can be done in real-time which is a major constraint for this thesis. Besides DBN the G.729B VAD [15] and Sohn VAD [16] are used as well for a better comparison of the results. The following outlines the basic principle behind the three classification schemes.

The G.729 is a speech coding standard which is used in telecommunication working at a low bandwidth of 8 kbit/s. With Annex B a VAD has been added to detect frames without speech to adapt encoding and compression ratio. The algorithm uses full- and low-band energy and zero-crossing rate per frame as well as linear prediction spectral coefficients as basis features. Computing the difference between the instant features and their corresponding average over a certain time period leads to four parameters on which the final decision can be made. A final smoothing reduces transient decisions taking past speech frames into account.

The Sohn VAD is based on statistical models of noise as well as speech plus noise in combination. Depending on the geometric mean of the estimated likelihood ratios per frequency band decision about speech in the current frame is made. A HMM based hangover scheme where each state depends on previous and the current observation prevents miss-detections during weak speech signals.

The DBN based VAD uses a pre-trained neural network for classification. DBN is a special case of a neural network consisting of a visible layer (bottom) whose units are represented by the computed features, some hidden layers with hidden units and the top layer representing the output labels. A DBN can be thought of as a stack of Restricted Boltzmann Machines (RBMs) [17] being trained from bottom to top. Each RBM consists of a visible and a hidden layer where the hidden units of one RBM form the visible units of the following. While each training of the RBMs is unsupervised, a final supervised fine-tuning of the whole stack using a back-propagation algorithm further improves the model. This two-phase training procedure helps to prevent over-fitting which usually is a problem using supervised training methods.

2.2 Mathematical Background

As mentioned in Sec. 2.1 the representation of an acoustic signal using special features helps to extract its characteristics while suppressing noise and reducing redundancy. For this the signal has to be analysed in different domains using standard transformations. Enhancing the signal by subtracting estimated noise further improves the results.

This section shortly reviews well known equations in Sec. 2.2.1 and introduces the definitions of the features used in this thesis in Sec. 2.2.2. Section 2.2.3 gives the mathematical background corresponding to noise suppression. For evaluation, performance measures are defined

in Sec. 2.2.4.

2.2.1 Signal Transformation

This section states the definitions of the signal transforms which in particular form the basis to compute the features used in the following chapters. They are well known and commonly used and therefore only introduced briefly.

The DFT is used to transform a discrete time signal y into its discrete spectral representation Y. It is defined as

$$Y\left[f\right] = \sum_{n=0}^{N} y\left[n\right] e^{-j2\pi fn/N} \tag{2.1}$$

If only short windowed sections of y are used it is referred to as STFT and Eq. (2.1) changes to

$$Y\left[f,k\right] = \sum_{n=0}^{N} y\left[n,k\right] e^{-j2\pi fn/N} \tag{2.2}$$

Shorter sections lead to a higher time resolution but a lower frequency resolution and vice versa.

The correlation of two signals is defined as

$$R_{i,j}[c] = \sum_{m=-\infty}^{\infty} y_i[m] y_j[m+c] \tag{2.3}$$

Using the convolution property of the DFT where the convolution of two time signals correspond to the product of their spectral counterparts correlation can be computed as follows

$$S_{i,j}[f] = Y_i[f] Y_j^*[f] \tag{2.4}$$

The complex conjugate of one signal ensures not to flip one signal against the other in time domain as it would be done for convolving two signals. In case of $y_i = y_j$, it is usually referred to as autocorrelation where the signal is correlated with itself. Performing autocorrelation makes it easy to detect periodicities inside the signal which might not be obvious due to noise.

The inverse transform of the logarithm of the spectrum is defined as cepstrum

$$C[q] = \frac{1}{N} \sum_{f=0}^{N-1} \ln(\left|Y\left[f\right]\right|) e^{+j2\pi fq/N} \tag{2.5}$$

Coming from the properties of the logarithm the product of two spectra equates to the sum of the corresponding cepstra.

2.2.2 Features

The following section introduces the features used for classification by defining them and explaining relevant properties. The features are selected following [12] which has been used when

demonstrating the advantages of the DBN classifier.

The MDFT_l feature represents a compressed version of the energy Fourier spectrum. A mel-filter-bank consisting of triangular filters is used to summarize regions of the spectrum in a non-linear manner which follows the perception of the ear. Mel frequencies are linear below 1 kHz whereas higher frequencies can be computed as [13]:

$$M(f) = 1125\ln(1 + f/700) \tag{2.6}$$

There are smoothed versions of MDFT_l used as well where the window length is indicated by the index l. Using a normalized rectangular window of l frames corresponds to a simple averaging of each frequency band over time. The averaged versions of the MDFT_l feature focusses on the long-term evolution of the speech signal suppressing rapid changes which may occur due to noise or short breaks which may be less important concerning VAD.

The MFCC feature is the cepstral representative of MDFT_l. It is computed as

$$\text{MFCC}_l(q) = \sum_{m=0}^{M-1} \ln(\text{MDFT}_l(f_{mel})) \cos\left(\pi q(m + 1/2)/M\right) \tag{2.7}$$

where f_{mel} stands for the compressed frequency bands. As for MDFT_l averaged versions of MFCC_l are used as well to gain information of different time properties.

Linear Predictive Coding (LPC) is a well known method in speech coding where each sample is linearly predicted by its p preceding samples [13]:

$$y[n] = \sum_{k=1}^{p} a_k y[n-k] + e[n] \tag{2.8}$$

a_k stands for the LPC coefficients and e denotes the prediction error or the difference between the original and the predicted signal which is also called the residual. Remembering the source-filter-model in Sec. 2.1.1 the a_k represent the filter coefficients of $H(z)$ i.e. the vocal tract while e stands for the excitation. Depending on the filter length the characteristic spectrum of voiced phonemes with its major resonances can be represented by the LPC coefficients while spectral noise between adjacent frequency bands is suppressed.

Relative-Spectral Perceptual Linear Predictive Analysis (RASTA-PLP) features predict the spectrogram values of MDFT_l the same way as for LPC. The difference besides the use of the perceptively compressed spectrogram lies in an additional prore-filtering of $\ln(\text{MDFT}_l)$ called RASTA-filtering which is used as a IIR bandpass filter with the transfer function [18]. On the one hand, the filter is meant to reduce reverberation effects on the other hand it smooths rapid changes between adjacent frames [13].

Amplitude Modulation Spectrograms (AMS) [19] are derived from the spectral information of magnitude progression in each frequency band of a mel-spectrogram. For each mel-channel small Hanning-windowed time-sections are Fourier transformed resulting in a modulation spectrogram of the channels amplitude. The final AMS features per frame are computed by grouping frequency bands below 1 kHz into 15 evenly spaced areas being weighted by a triangular shaped window and summarized. In addition to the initial features temporal and spectral delta features computed by the differences of adjacent frames and frequency bands respectively are appended to use information about the feature variation as well.

Feature selection is used to decrease the number of features keeping only those which contain the most information concerning a robust classification. During the work on this thesis a *First-Best* algorithm was used to recursively find the best working features to form a new smaller

set. The algorithm starts by performing classification for each single feature and determines the one producing the best result. Classification is performed again for all remaining features in combination with the selected one and again the best result identifies the next best feature to keep. Iterations are performed until results are converging or until a predefined number of features has been selected.

2.2.3 Noise Reduction

In this section Spectral Subtraction (SS) used for signal enhancement with respect to noise reduction will be explained. The goal is to increase the Signal to Noise Ratio (SNR) by reducing background noise which can be considered stationary over a certain time period.

The principle of SS is based on the assumption that noise is additive i.e. that the noisy signal y can be written as

$$y = x_s + x_n \tag{2.9}$$

where x_s denotes the clean signal and x_n stands for noise. If noise is known, subtracting it from the noisy signal leads to the original signal. In reality noise can only be estimated which after subtracting it from y leads to an estimation of the original signal:

$$x_{est} = y - x_{n,est} \tag{2.10}$$

Estimating the magnitude spectrum of the noise $X_{n,est}$ and subtracting it from the spectrum of the noisy signal Y is referred to as SS. Assuming time-changes of x_n being sufficient slow compared to x_s an adaptive estimation algorithm is considered to work best. The approach used in this thesis is based on minimum statistics using an implementation provided by Voicebox[2] based on the work of Gerkman et al. [20]. Here the algorithm consecutively estimates the less changing part of the speech signal by computing the a posteriori Speech Presence Probability (SPP) to estimate the noise power spectrum. The likelihood for speech presence $\mathcal{H}_1$ and speech absence $\mathcal{H}_0$ is assumed to be Gaussian distributed while the a priori probability concerning speech presence or absence is considered equally likely i.e. $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$. The main steps performed by the algorithm will be outlined in the following for a detailed derivation see [20].

For each frame, a posteriori SPP is computed by

$$P(\mathcal{H}_1|Y) = \left(1 + \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}(1+\xi_{\mathcal{H}_1})e^{-\frac{|Y|^2}{|X_{n,est}|^2}\frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)^{-1} \tag{2.11}$$

where $\xi_{\mathcal{H}_1}$ denotes the a priori SNR which in the default setting of the algorithm has been found to work optimal at 15 dB. After computing $P(\mathcal{H}_1|Y)$ it has to be ensured that the algorithm doesn't stagnate due to underestimation of noise. This is done by taking earlier a posteriori SPPs into account using a recursion and establish a limiter if these probabilities have been close to 1. The time smoothing recursion is computed by

$$\bar{P}(k) = 0.9\bar{P}(k-1) + 0.1P(\mathcal{H}_1|Y(k)) \tag{2.12}$$

To avoid stagnation a limiter is defined which resets $P(\mathcal{H}_1|Y(k))$ every time it reaches values

[2] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

above 0.99.

$$P(\mathcal{H}_1|Y(k)) \leftarrow \begin{cases} \min(0.99, P(\mathcal{H}_1|Y(k))), & \text{if } \bar{P}(k) > 0.99 \\ P(\mathcal{H}_1|Y(k)), & \text{else} \end{cases} \tag{2.13}$$

Considering $P(\mathcal{H}_0|Y) = 1 - P(\mathcal{H}_1|Y)$ the MMSE estimator can be obtained by computing the conditional expectation:

$$E\left(|N|^2|Y\right) = P(\mathcal{H}_0|Y)|Y(k)|^2 + P(\mathcal{H}_1|Y)|X_{n,est}(k-1)|^2 \tag{2.14}$$

Depending on the a posteriori SPP $X_{n,est}$ remains the same if the current frame is more likely to contain speech or it will be adapted to the current frame if it is more likely to contain noise. To suppress the influence of transients a final recursion is applied when computing the power noise estimate:

$$|X_{n,est}(k)|^2 = \alpha|X_{n,est}(k-1)|^2 + (1-\alpha)E\left(|N|^2|Y\right) \tag{2.15}$$

where $\alpha = 0.8$.

Figure 2.2 shows an example of enhancing the signal by subtracting the estimated noise from the original signal by comparing the different spectra of Y, $X_{n,est}$and X_{est}. The noise characteristic is smooth and especially sharp transients like onsets of syllables don't appear, which improves the difference between the more stationary background and impulsive sounds after subtraction.

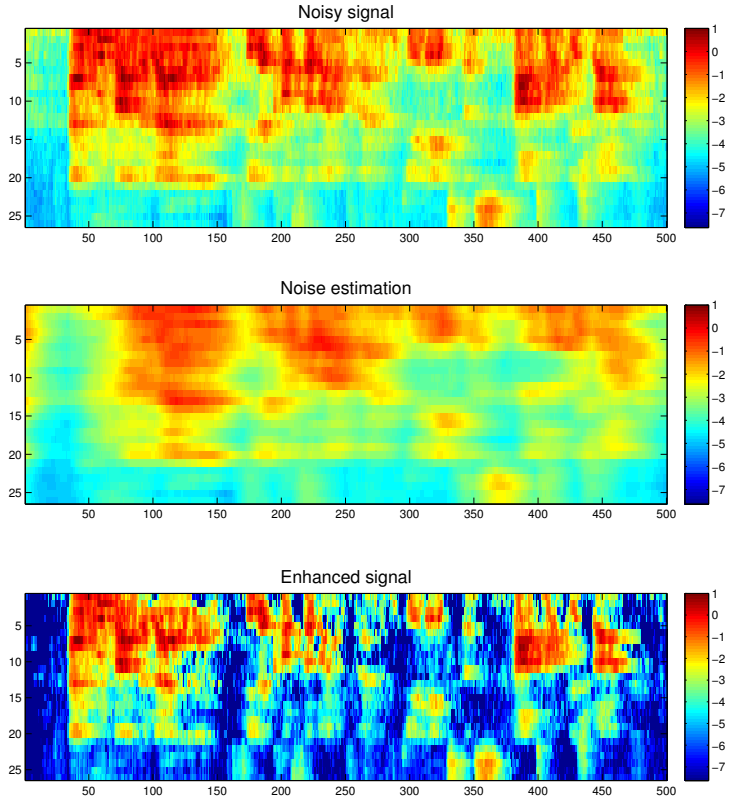


Figure 2.2: Noise subtraction using a noise estimation based on a minimum statistics approach. The top figure shows the spectrum of the noisy signal y. The middle figure shows its noise estimation. In the bottom plot, the spectrum after noise subtraction is shown. It is easy to see a much higher contrast between speech and background sources.

2.2.4 Performance Evaluation

The following outlines standard metrics used in literature to gain information about the classification performance and to allow comparison between different algorithms or settings.

For binary classification i.e. the decision between speech x_s or non-speech x_n made for each frame either results in a correct choice or not. In a testing scenario each decision can be compared to the ground truth and the result labelled with one of the following metrics. Assuming a frame containing speech, the number of true positives (TP) or hits corresponds to the number of frames where x_s has been detected, whereas the number of false negatives (FN) counts how often speech has been missed. In case of frames without speech the number of correct rejections corresponds to the number of true negatives (TN) whereas noise detected as speech or false alarms lead to the number of false positives (FP).

Usually, these statistics are used in relation to the total number of frames to describe the performance. In case of VAD the most important ratio describes the percentage of detected speech also known as recall (RE) by relating TP with the subset $N_{sp} = TP + FN$ of all frames containing speech:

$$RE = \frac{TP}{N_{sp}} \cdot 100\% \tag{2.16}$$

$RE = 0\%$ would correspond to the whole subset N_{sp} being classified as noise. In the best case when every speech frame has been detected i.e. $RE = 100\%$ a second parameter is needed containing information about additional false alarms which describes the precision (PR):

$$PR = \frac{TP}{TP + FP} \cdot 100\% \tag{2.17}$$

The characteristics of PR mainly depends on FP while its minimum value is bounded by TP. In the worst case of labelling everything as noise the term precision becomes meaningless and PR is therefore undefined.

Taking recall and precision a further parameter often is used as a single value performance descriptor called F-score (F_1) to gain a first feeling how the algorithm is working without looking at the details:

$$F_1 = 2 \cdot \frac{RE \cdot PR}{RE + PR} \tag{2.18}$$

For the current thesis, recall and precision have to be adapted as a correct decision depends on both, VAD and room detection. This means that a TP only occurs when detected speech is localised correctly which will be denoted by TP_{room}. Equation (2.16) changes to

$$RE = \frac{TP_{room}}{N_{sp}} \cdot 100\% \tag{2.19}$$

For precision, the FP term has to be adapted as it not only contains falsely detected voice activity, represented by FP but also any wrong localisation in case of speech presence which is denoted as FR. From this, Eq. (2.17) changes to

$$PR = \frac{TP_{room}}{TP_{room} + FP + FR} \cdot 100\% \tag{2.20}$$

3
Voice Activity Detector

The proposed room-aware Voice Activity Detection (VAD) algorithm is based on a single-channel VAD introduced by Zhang et al. [12]. This VAD uses a set of common features like Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) or Amplitude Modulation Spectrograms (AMS) which together form the feature vector for a Deep Belief Network (DBN). Zhang et al. compared this algorithm against common standards like G.729B [15], statistical based VADs like Sohn VAD [16] or machine-learning based VADs using Support Vector Machine (SVM). The DBN based VAD proved to outperform the competitors in terms of accuracy in most cases while still being capable of real-time computation.

The multi-channel VAD proposed in this thesis has been developed by separating the process of detection and localisation into the following steps: signal-enhancement, channel-selection, VAD and room-selection. These steps will be ordered according to Fig. 3.1 which represents the main signal flow between the recorded raw signals and VAD room decision respectively.

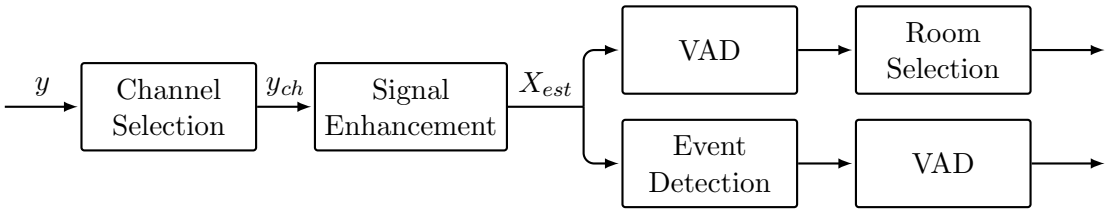


Figure 3.1: Basic signal-flow summarizing the main steps being performed for VAD.

Channel-selection mainly follows the intention of reducing the amount of computation power needed when using the whole set of 40 channels by finding a small subset of l channels. In the simplest case selecting only one signal per microphone group reduces redundancy as similarity between signals within a channel group is high compared to the similarity between signals from different groups. In a second attempt an adaptive channel selection tries to find an optimal signal for each room separately without any knowledge about the type of the current signal focussing on the cross-correlation between signals of each microphone group.

Signal enhancement is based on Spectral Subtraction (SS) where an estimation of the noise power spectrum $X_{n,est}$ is subtracted from the spectrum of the noisy signal Y. It aims to reduce long-term noise which changes slowly over time to increase the Signal to Noise Ratio (SNR). It is a pre-processing step where each signal of the 40 available channels is enhanced separately. The description of the basic principle and the algorithm used can be reviewed in Sec. 2.2.3.

VAD is done by classification of a feature vector using a DBN as proposed in [12]. Additionally, a simple VAD based on autocorrelation is introduced. For comparison reasons the G.729B VAD and Sohn VAD will be used as well.

Room-selection exploits the Time Difference Of Arrival (TDOA) of an acoustic event between different channels resulting from the limited sound velocity in air. While onset detection of single words or phrases would be a simple approach it heavily depends on the precision of the preceding VAD which has to be high right at the beginning of speech. This dependency can be minimised by taking the whole block of speech into account as the influence of noise corruption decreases and VAD can be considered more stable during the phrase. This can be exploited by computing cross-correlation of the spectrogram between two channels over time. Comparing the relative position of the correlation maxima reveals information about the channel, where the signal has been recorded first.

The split-up after the signal enhancement block in Fig. 3.1, indicates, that room detection not necessarily has to follow VAD. The performance when room localisation is done before VAD will be evaluated as well. To make discrimination of the two possibilities easier room detection being first will be referred to as event detection as in this case, the algorithm not only performs room localisation but also a certain kind of event detection of acoustic events in general being independent of speech presence or absence.

In this chapter the detailed discussion of the algorithms proposed is organized in the following sections.

Section 3.1 contains background information about the databases used, how they are built up and how signals will be used. In Sec. 3.2 different approaches of channel selection will be introduced. Section 3.3 covers feature sets and classification. In Sec. 3.4 room detection and the developed algorithms will be explained in detail.

3.1 Database

The database used for evaluating the proposed algorithms is part of the simulated corpus of the Distant-speech Interaction for Robust Home Applications (DIRHA) project[3]. It consist of a certain number of records of one minute duration at a sample rate of 48 kHz and 16 bit depth. The German version, called DIRHA_DE , has been used for evaluation. It contains German spoken phrases and commands which are allowed to overlap. This is an important requirement the localiser has to deal with but makes the simulation more realistic.

Figure 3.2 shows the floor-plan of the apartment where the recordings have been produced. There exist five rooms: the kitchen, labelled with **K**, the living room, labelled with **L**, the rest-room, labelled with **R**, the corridor, labelled with **C** and the bedroom, which is labelled with **B**. 40 microphones marked by the black dots are distributed by forming up to five arrays of different size in each room. The grouping depends on the size as well as the purpose of each room to ensure a good discrimination for localisation where it might be needed. For example, only the living room and the kitchen contain a group of six microphones at the ceiling. The coloured rectangles indicate possible positions of a person during a recording. The speech direction is marked by the four or eight arrows pointing away from the rectangles respectively. The quantization of space of positions results from the addition of speech to the database via measured room transfer functions.

The structure of one simulation contains the following components [1]:

- A keyword followed by a command;
- A spontaneous command (without the keyword);

[3] Detailed information can be found at http://dirha.fbk.eu/simcorpora

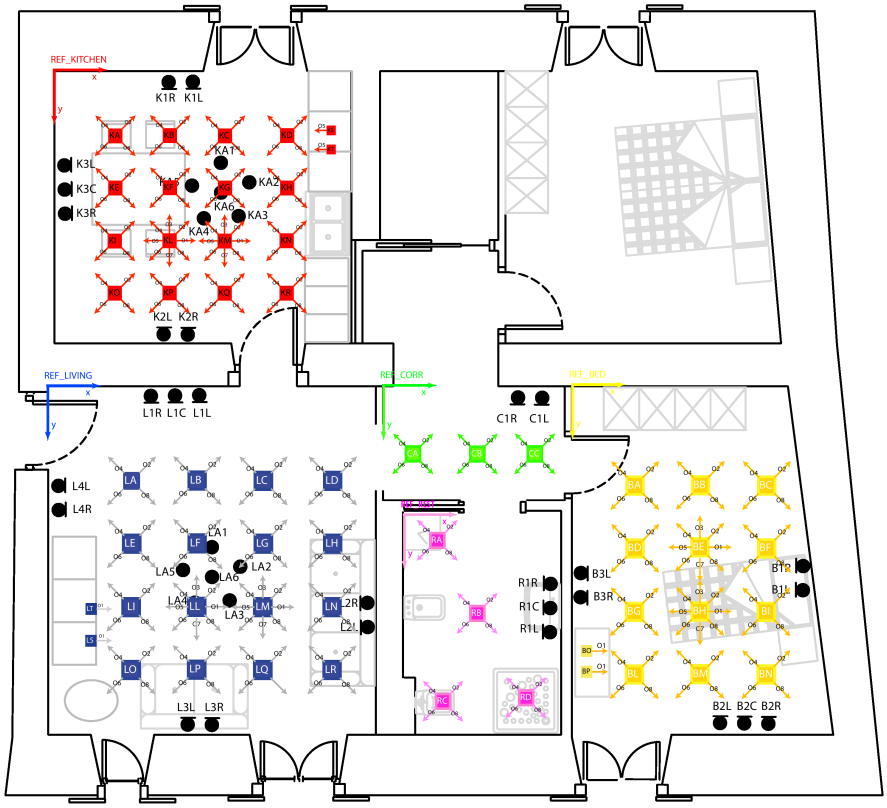


Figure 3.2: Floor-plan of the apartment where the recordings of the databases were made. The black dots indicate the position of the microphones. The coloured rectangles show possible speaker positions. The arrows at the stand for the direction the speaker gives the command. [1]

- A phonetically rich sentence;
- A segment of conversational speech;
- A variable number of localised non-speech sources (e.g., radio, TV, appliances, knocking, ringing, creaking and many others).

The commands and phrases can be located at every position marked by the coloured squares being spoken in any direction indicated by the arrows.

3.2 Channel Selection

Channel selection is the first step in the multi-channel VAD algorithm and comes even before signal enhancement. It is performed with the intention of reducing the number of signals for further processing while focussing on channels being close to possible noise sources.

Research concerning channel selection addresses different approaches. In [21] several state-of the art algorithms are compared and a new algorithm is proposed based on the variance of the speech intensity envelope.

Considering Fig. 3.2 there exist 40 microphones applied to the walls and the ceiling in each room. The number of available channels per room depends on its size and importance. For example the living room or the kitchen contain up to four microphone arrays on the walls whereas in the corridor or in the bathroom only one group of microphones is installed. For localising the correct room in case of speech several channel sets are constructed by using different combinations of channels and tested against each other in terms of distinction and efficiency.

The *maximum possible set* contains all available channels. Using this set for room localisation is very inefficient in terms of room localisation, i.e. using every channel per microphone array just increases redundancy. Therefore, the set is reduced by using only one channel per array. In this case the set is stationary using the whole spatial grid of microphones in the apartment with a minimum of 35% of the recorded signals. Recall from Fig. 3.2 that all rooms together contain 14 microphone arrays, the reduced channel set will be referred to as *Stat14Ch*.

The *minimum set* ensuring room discrimination can be achieved by selecting the most relevant channel per room. This increases efficiency in terms of computation power by considering the remaining signals per room being redundant. Defining the best signal per room is a difficult task. While in case of corridor and bathroom each signal can be considered equally important the best signal in the remaining rooms strongly depends on the speaker position and direction. Therefore, the simplest way of choosing a stationary set of five reference channels will only be considered as a baseline. The selected channels used for this set are in case of kitchen and living room the centre microphones KA6 and LA6 of the ceiling clusters while for the corridor the C1R channel as the more centred microphone and in the bathroom R1C is taken. In case of the bedroom B3R is selected assuming best distinctness between signals originating in bedroom and signals coming from outside. In the following, it will be referred to as *Stat5Ch*.

The time-varying positions and directions of different speakers make it necessary to adaptively select the best channel. Hence, for each time-block a new set of channels will be selected by finding the microphone which can be assumed to be at a minimum distance to the speaker. If chosen correctly the accuracy of both classification and localisation can be improved. As no further analysis or classification is involved, none of the channel selection approaches proposed distinguish between a speaker or a noise source. This is not a problem if speech and noise are coming from different rooms, it more likely supports the following room detection. If both sources are located in the same room, it will depend on the intensity of both sources which channel is being selected at the corresponding time-blocks.

The simplest approach of adaptive channel selection just computes the energy of each signal and compares the results room wise taking the signal containing the highest energy for each room respectively. In the following this will be referred as *Maximum-Energy-Channel-Selection*.

A more enhanced way of finding the best channel per room is to exploit the microphone-arrays using cross-correlation. This is done by selecting two channels per array which results in two time dependent vectors $\boldsymbol{y}_A$ and $\boldsymbol{y}_B$ whose elements $y_{Ai}[n]$ and $y_{Bi}[n]$ referring to the recorded signals are pairwise related. $n = 1 \dots N$ denotes time index in samples and $i = 1 \dots I$ stand for the channel index of all available channels. For each pair of elements, a cross-correlation will be performed

$$R_i[c] = \frac{1}{M-m} \begin{cases} \sum_{m=0}^{M-c-1} |y_{Ai}[m]||y_{Bi}[m+c]|, & \text{if } c \geq 0 \\ \sum_{m=0}^{M-c-1} |y_{Ai}[m-c]||y_{Bi}[m]|, & \text{if } c \leq 0 \end{cases} \tag{3.1}$$

where $-M < c < M$ denotes the time shift index and where each correlation function R_i denotes an element of matrix $\boldsymbol{R}$. Taking the absolute value of the time signals has empirically shown to provide more stable results.

For each pair $y_{Ai}[n]$ and $y_{Bi}[n]$ signals can be seen similar because of the small distance between microphones. Especially in case of events happening in a closer distance and hence are less destructed by reverberation or noise this similarity increases whereas a higher reverberation leads to a higher independence between signals. Therefore focussing on the maximum of the correlation function might reveal some useful information in terms of finding the best microphone group in each room.

The first approach addresses the maximum variance. Assuming background noise sources being corrupted by a lot of reverberation cross-correlation will result in broader peaks with less amplitude. In case of an acoustic source being close to the microphones cross-correlation between the two signals will be much stronger i.e. $R_i[c]$ for signal pair i at time shift c will result in a high but small peak and therefore in a higher variance. Finding the maximum variance for each room leads to five microphone pairs which can be assumed to be closest to an acoustic event. For one specific room this can be expressed in the following equation

$$j_{\max} = \arg\max_{j} \left(\frac{1}{2M-2} \sum_{c=-M+1}^{M-1} R_j[c] \right) \tag{3.2}$$

where $j = 1 \dots J$ denote the channel index of all channel-pairs of one room.

Aligning the corresponding signals $y_{Aj_{\min}}[n]$ and $y_{Bj_{\min}}[n]$ and taking the mean increases the SNR in case of acoustic events in this room (Eq. (3.3)). This is also known as *Delay and Sum Beamformer* and holds if the events direct signal is less reverberated and strong enough compared to background noise in order to be able to assign the correlation peak to this event. In this case background noise can be reduced, while the signal itself stays the same.

$$y_{ref} = \frac{y_{Aj_{\min}}[n] + y_{Aj_{\min}}[n - c_{\min}]}{2} \tag{3.3}$$

According to its basic principle this algorithm will be referred to as *Maximum-Variance-Channel-Selection* in the following sections.

The second approach addresses the peak position of each $R_{i,\max}$ which depends on the speaker position. If the speaker stands right in front of two microphones, the correlation maximum would be at $R_{i,\max}[c = 0]$. If the person moves to the left or right, the location of the correlation maximum becomes positive or negative. Finding the peak $R_j[c_{\Delta,\min}]$ with minimum deviation from zero comparing all correlation peak positions per room corresponds to the channel group with a minimum event to microphone path difference.

$$j_{\min} = \arg\min_{j}(|c_{j,\max}|) \tag{3.4}$$

where $c_{j,\max}$ denotes the correlation index of $R_{j,\max}$. Again, delay and sum beamforming is used to suppress noise.

In the following sections this algorithm will be referred to as *Minimum-Difference-Channel-Selection.*

3.3 Voice Activity Detection

The classification approach used in this thesis to label each time-frame as speech or noise are on the one hand the DBN which is a state-of-the-art classifier for speech [12]. On the other hand, a simple VAD algorithm based on variance and low frequency energy thresholding is used for comparison purposes (see Sec. 3.3.1). In addition, the G.729 VAD and Sohn VAD are used as a baseline.

3.3.1 Spectral Focussing VAD

Recall from the speech model of Sec. 2.1.1 the major part of a signal of spoken language is characterized by its fundamental frequency and strong first harmonics. Computing the spectrogram of a speech signal a striped pattern marks the corresponding time regions where each stripe corresponds to either the fundamental frequency or one of its harmonics. The more noisy the signal becomes the less visible the pattern occurs. The following algorithm, which will be referred to as *Spectral Focussing VAD*, aims to focus on the harmonic parts of speech to perform VAD. For this, the assumption has to be made that all relevant commands or words are clearly spoken but not whispered.

The first step addresses the extraction of spectral periodicities by applying a frequency dependent autocorrelation of each time frame of the spectrogram X_{est}.

Each frame of X_{R} can be seen as a symmetric function with a maximum at its centre corresponding to the frames energy. Assuming random noise with infinite length, any value except of the centre value would be zero while in the other extreme for a periodic source with infinite length, the same maximum occurs after every fundamental period in the simplest case. For shorter signals this behaviour approximately is the same, hence, computing X_{R} preserves the regions of vowels with its harmonic structure while flattening any non-periodic frame. As the focus lies in detecting periodic parts of speech, frequency bands of the spectrogram below 60 Hz and above 1 kHz are discarded when computing X_{R} to reduce the influence of possible periodic noise sources. Recalling the symmetric property of autocorrelation one half of X_{R} can be omitted.

The remaining part of the X_{R} consists of the frame energy at $c = 0$ which influences the first n adjacent values at $c = 1 \dots n$ and the remaining correlation values. As the energy value for noise or periodic sources can't be distinguished the first n correlation values will be omitted as well. From this, the function to compute X_{R} can be written as

$$X_{\mathrm{R}}(k, c) = \frac{F}{F - c} \sum_{m=1}^{F-c} X_{est}(k, m) X_{est}(k, m + c) \tag{3.5}$$

where $k = 1 \dots K$ denotes the frame index of K frames, $c = c_n \dots F - 1$ denotes the correlation index of interesting correlation bins and F stands for the number of frequency bands of X_{est}. The normalization factor compensates the effect of decreasing results with increasing c when correlating signals of finite length.

Figure 3.3 shows an example of X_{R}. The striped pattern in the second half correspond regions of voiced speech.

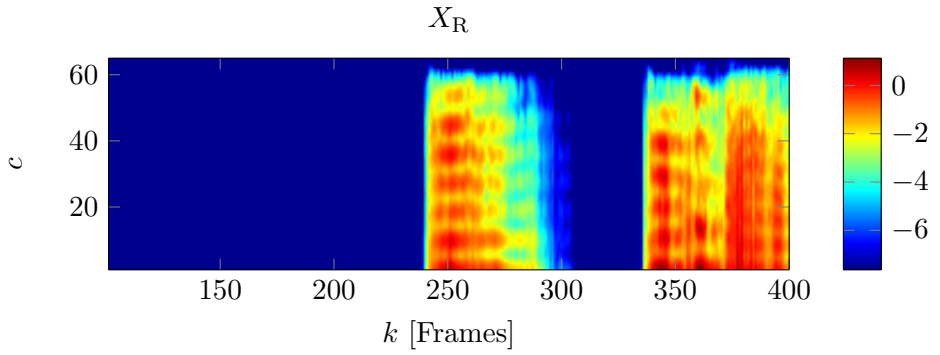


Figure 3.3: The pictures show an example of the spectral correlation. It contains the second block from the 'sim1.wav' taken from DIRHA_DE using channel LA6. It is the beginning of a spoken phrase which is indicated by the striped regions in the second half of X_{R}.

The decision whether the current frame contains noise or a periodic source is done by computing two simple features. Considering the flat versus peaky structure of X_{R} depending on signal characteristic computing the frame dependent variance will reflect this behaviour. Hence, using a threshold $v(k)_{thr}$, a decision function d_{var} can be computed as

$$d_{var}(k) = \begin{cases} 1, & \text{if } v(k) > v(k)_{thr} \\ 0, & \text{else} \end{cases} \tag{3.6}$$

where the frame dependent variance is computed as

$$v(k) = \left(\frac{1}{(2M-2)X_{\mathrm{R}}(k,0)} \sum_{c=-M+1}^{M-1} X_{\mathrm{R}}(k,c) \right) \tag{3.7}$$

which is normalized by the corresponding frame energy at $c = 0$. $c = c_n \dots F-1$ denotes the correlation index of interesting correlation bins.

The second feature is derived by computing the frame-maximum of a band-limited version of X_{est}. The band-limits correspond to the frequency range of f_0 which is $f_{min} = 60$ Hz and $f_{max} = 500$ Hz. Again a threshold is introduced resulting in a second decision function d_{max}

$$d_{max}(k) = \begin{cases} 1, & \text{if } |X_{est}|_{f_{fF},max}(k) > |X_{est}|_{f_{fF},max,thr} \\ 0, & \text{else} \end{cases} \tag{3.8}$$

where $|X_{est}|_{f_{fF},max}(k) = \max(X_{est}(k, f_{fF}))$ denote the frame dependent maximum signal and $f_{fF} = f_{min} \dots f_{max}$ stands for reduced number of frequency bands in f.

If both decision functions equal one, the frame contains a harmonic source with sufficient energy likely originating from voice. Taking this result short gaps of zeros will be removed originating from non-harmonic phonemes to gain continuous blocks in speech decision.

Figure 3.4 shows the second block of channel LA6 of *sim1.wav* from DIRHA_DE before (a) and after (b) thresholding. It can be seen, that the threshold used removes frames with less harmonic sounds resulting in a distinct onset. Although desired, it produces an unwanted fragmentation as the amount of harmonics in speech varies a lot. This problem can be reduced when using multiple channels exploiting sound propagation. While a couple of frames being set to zero (i.e. no speech), the same frames of an adjacent channel still contain enough energy to be labelled as speech. Hence, as long as a break between harmonic blocks is not detected in all channels it is assumed, that speech still exists. In case of Fig. 3.4 the second gap might be small enough to be labelled as speech due to the late arrival of the signal at distant microphones. Following this rule the number of onsets can be reduced which again leads to a lower false positive rate.

A major drawback of this algorithm is its lack in distinguishing speech from harmonic noise sources that meet the above threshold conditions. While the algorithm aims to perform as a fast but still robust VAD by focussing on the harmonic part of speech it should be taken as a simple approach which will be mainly used for comparison purposes.

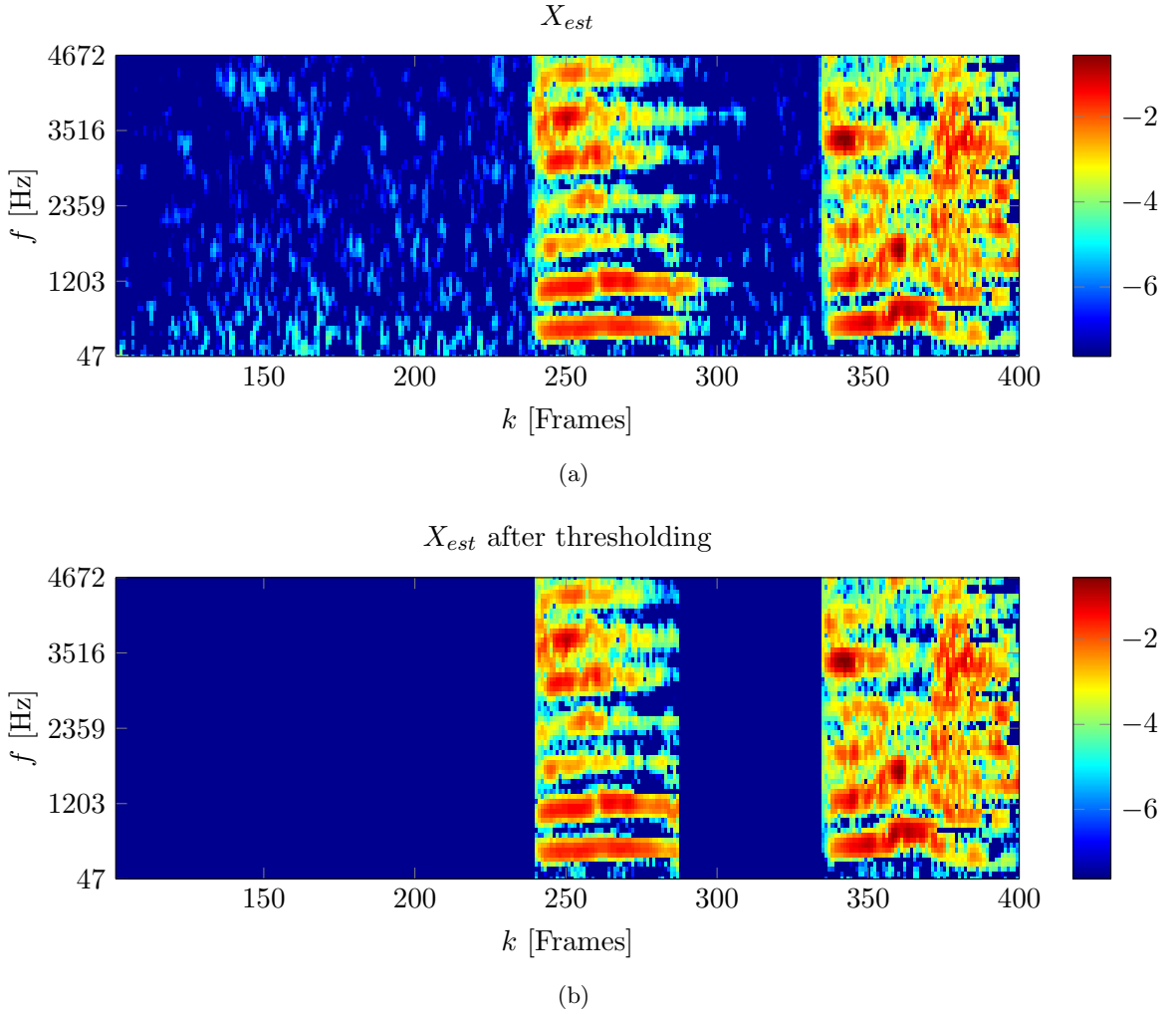


Figure 3.4: The pictures show an example of the thresholding process of speech likely sounds. It contains the second block from the 'sim1.wav' taken from DIRHA_DE using channel LA6. It is the beginning of a spoken phrase which is indicated by the strong harmonic speech regions in the second half of the spectrogram (a). These regions become sharp edged when thresholding is performed as can be seen in (b).

3.4 Room Detection

Detecting the correct room where speech originates is a crucial task for further signal processing. To achieve this, two algorithms have been developed based on the TDOA between different signals of each event.

The first approach follows VAD and can be seen as an onset-detection algorithm. Comparing the time of the beginning of each word or phrase between each recording enables room detection by taking the channel where the signal occurred first and assign it to its corresponding room. The algorithm will be explained in detail in Sec. 3.4.1.

The second approach is more complex to overcome the dependency of the VAD precision needed for detecting the correct beginning in each channel. It is based on a time-dependent cross-correlation of the spectrogram X_{est} which results in a correlogram per frequency bin followed by the localisation of each correlation maximum. The relative position between these maxima reveals information about the TDOA making it possible to select the appropriate room. As this algorithm mainly depends on the similarity between the channel-signals the spectrum is pre-enhanced in terms of contrast and transient. This makes it independent of its position in

the signal chain, i.e. whether it is preceding or following the VAD where it is either locating acoustic events in general or in the latter case determines the position of already detected speech blocks. This algorithm is discussed in detail in Sec. 3.4.2.

3.4.1 Room Detection based on Onset Detection

The spectrogram X_{est} analysed by a VAD can be written as

$$X_{est,vad}(k,f) = \begin{cases} X_{est}(k,f), & \text{if speech detected} \\ 0, & \text{else} \end{cases} \tag{3.9}$$

Taking this, word onsets can be defined at a frame k with $X_{est,vad}(k,f) \neq 0$ which is preceded by a frame being zero at all frequencies, i.e. $X_{est,vad}(k-1,f) = 0, \ \forall f$. To define an onset of a new word or phrase, this must be valid in all channels to avoid the detection of single onsets at a time, where the speech signal reaches more distant microphones like for example during short gaps between words. As these gaps are not necessarily detected in all channels, estimating the onset during this time could lead to wrong localisations when a gap occurs in a different room.

After finding an onset, the channel where the onset is detected first is considered to be closest to the speaker and hence the corresponding room is detected as origin. There might be cases, where the first onset of two channels from different rooms might occur at the same time frame k because of similar speaker to microphone distances or due to inaccuracies in VAD. If this is the case, the energy at frame k between these channels is compared by looking for the maximum. The frame with the higher energy is assumed to be closer at the source which of course only holds if the signal analysed is not disturbed by other noise sources.

The usual length of phrases is larger than the block size used, hence, assuming the case of correctly determine the onset in one block, the following might contain the same phrase but without its beginning. For this, the preceding decision is remembered and the corresponding room labelled containing speech, as long as enough signal is being detected in the interesting area between the first block-frame and its last one or rather the next onset if one exists. This procedure is important as it is far more difficult to determine the precise end of a word at the end of a phrase. For example, the end of a word often is spoken with less energy compared to its beginning which again might be missed by the VAD if distance increases.

Depending on the channel set used, it has been shown to be useful to omit onsets in cases when speech only has been detected in up to 30% of the channels used. Assuming speech-like noise falsely being classified as speech holds, if speech in general can be assumed to be detected in most of the adjacent rooms.

While the algorithm is based on very simple steps, it strongly depends on a robust VAD implementation being as precisely as possible at the beginning of words especially where some phrase or command starts. This dependency additionally increases with the low time resolution of the spectrogram. Furthermore it is not possible to locate two speakers at the same time in different rooms, due to the condition of a global onset which is necessary to relate local onsets to each other. As the DIRHA_DE database consists of overlapping speech segments onset detection cannot be used without further constraints.

3.4.2 Room Detection based on Correlation

As x_s contains more than one channel, it can be seen as a two dimensional matrix of size $\#(samples) \times \#(channels)$ and hence X_{est} containing a spectrogram for each channel it can be seen as a three dimensional matrix of size $\#(frames) \times \#(channels) \times \#(frequencies)$. X_{est} therefore forms the basis information for all of the following algorithms.

This algorithm segments X_{est} over time focussing on acoustic events. For each segment cross correlation functions R_{ij} between each channel are computed from which information about the events location can be extracted. Figure 3.5 shows the diagram of the steps being performed to achieve room detection.

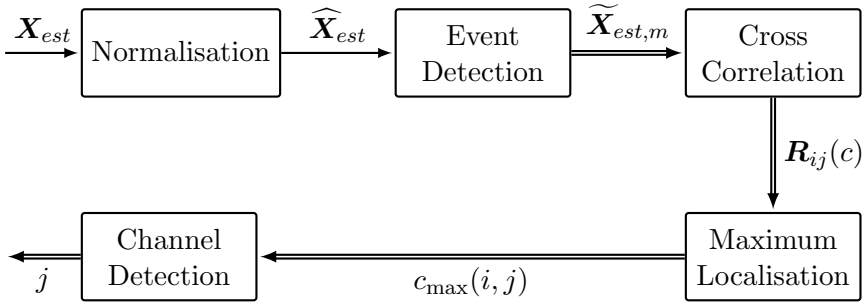


Figure 3.5: Flow-chart of the room detection using correlation. Input is a vector containing spectrograms of all channels the output is a binary mask corresponding to the room number. Note that after event detection multiple events may exist where the remaining steps have to be performed for each event which is marked by the double lined arrows.

Normalisation of $\boldsymbol{X}_{est}$ is performed in a first step to improve correlation results. it is divided into two steps. As the focus lies on transient signals normalization aims to extract regions of acoustic events and sharpen their spectral pattern. In the first step the Root Mean Square (RMS) value in each frequency band is subtracted

$$\widehat{\boldsymbol{X}}_{est}(k,f) = \boldsymbol{X}_{est}(k,f) - \sqrt{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{X}_{est}^2(k,f)} \tag{3.10}$$

Low values as well as the stationary part in each frequency signal of $\boldsymbol{X}_{est}$ become negative in $\widehat{\boldsymbol{X}}_{est}$. Setting these values to zero remains the spectral characteristic of transients and increases sparsity of the spectrum

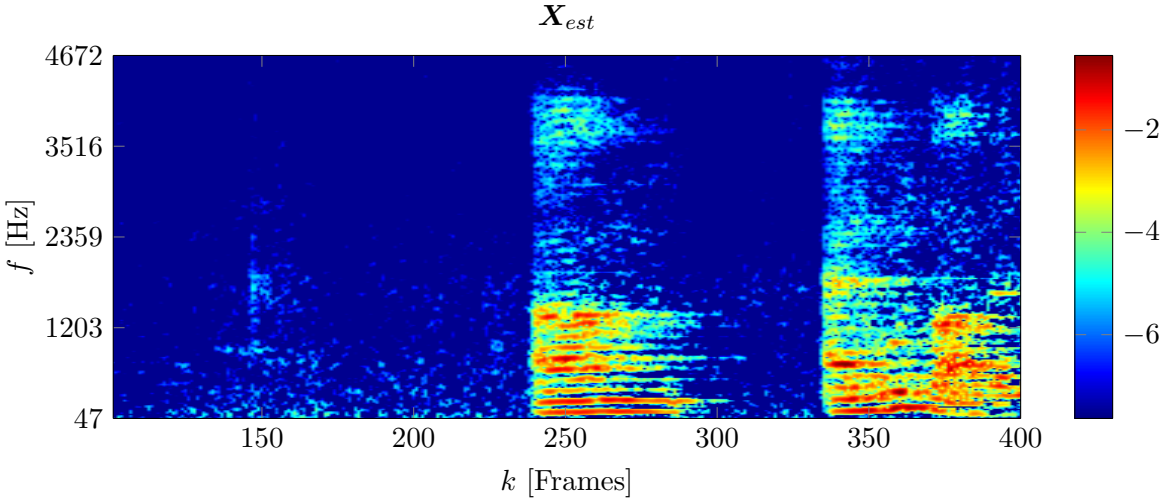
$$\widehat{\boldsymbol{X}}_0(k,f) = \begin{cases} \widehat{\boldsymbol{X}}_{est}(k,f), & \text{if } \widehat{\boldsymbol{X}}_{est}(k,f) > 0 \\ 0, & \text{else} \end{cases} \tag{3.11}$$

The second step doing normalization addresses equal amplitude scales per channel in each frequency band to ensure that the focus lies only on time information without being corrupted by varying dynamics between channels. To achieve this, every signal is divided by its maximum value over time:

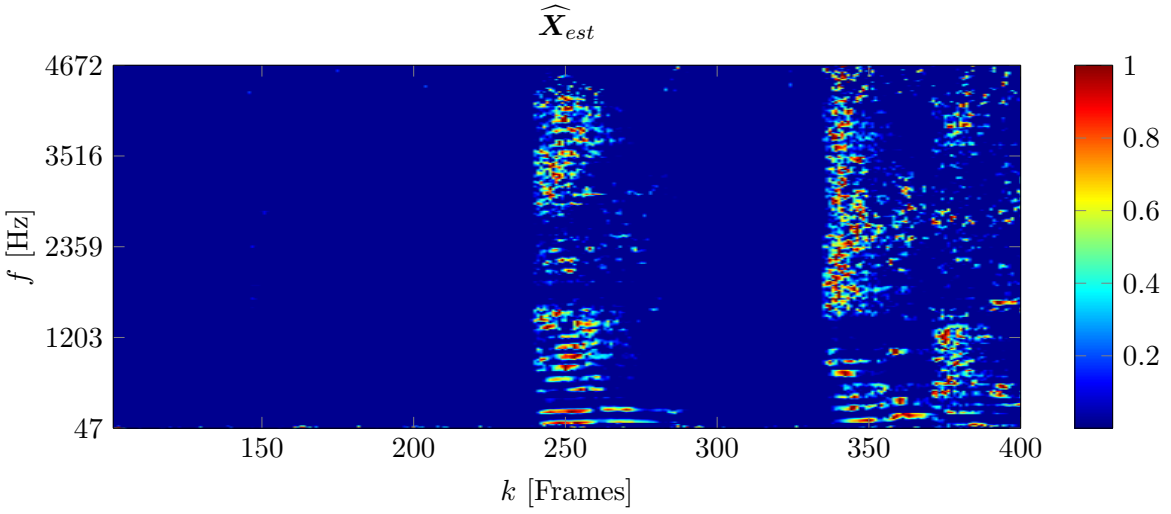
$$\widehat{\boldsymbol{X}}_{est}(k,f) = \frac{\widehat{X}_0(k,f)}{\max_k(\widehat{\boldsymbol{X}}_0(k,f))} \tag{3.12}$$

Figure 3.6 demonstrates the normalization process of the spectrogram. Compared to X_{est} in (a) the bottom plot reveals the characteristic pattern over a broad spectrum while suppressing weak information which is considered to be noise. In addition the signal is less stationary such that cross-correlation leads to a more robust positioning of the maxima between each other.

After normalisation, **Event Detection** is performed. Assuming a speech signal combined with some rhythmic background noise the latter may dominate $\widehat{\boldsymbol{X}}_{est}$ and lead to a false decision in localisation. To avoid this, the spectrogram is divided into smaller blocks around each detected



(a)



(b)

Figure 3.6: This figure demonstrates the normalization performed before performing spectral correlation. In this example, the spectrogram corresponds to the second block of the file 'sim1.wav' from DIRHA_DE of channel LA6. Note, that the normalized spectrum is shown with normal amplitude, while the top spectrum is plotted taking the logarithm. Especially in the lower frequency region the signal is less stationary while pattern at higher frequencies become more important. From this, cross-correlation maxima can be expected to become smaller and therefore time separation between each maximum improves.

acoustic event. For this, the mean value per frame in each channel is computed where higher values indicate signal activity.

$$\widetilde{X}_{est,i}(k) = \frac{1}{F}\sum_{f=1}^{F} X_{est,i}(k,f) \tag{3.13}$$

where $i = 1 \dots N_{ch}$ denotes the channel index.

As each acoustic event spreads through the apartment the average of $\widetilde{\boldsymbol{X}}_{est}$ of all channels is computed. On the one hand, this broadens the area being considered for correlation as different time occurrences are being overlaid. In addition, this reduces small gaps between words or

syllables, resulting after normalization.

$$\overline{X}_{est}(k) = \frac{1}{N_{ch}} \sum_{\imath=1}^{N_{ch}} \widetilde{X}_{est,i}(k) \tag{3.14}$$

$\overline{X}_{est}(k)$ is now thresholded by its mean which results in a binary mask indicating whether a frame belongs to the acoustic event or not.

$$m_{area}(k) = \begin{cases} 1, & \text{if } \overline{X}_{est}(k) > \frac{1}{K} \sum_{k=1}^{K} \overline{X}_{est}(k) \\ 0, & \text{else} \end{cases} \tag{3.15}$$

This will result in clusters of frames being labelled with a one around each acoustic event. To put them together into the desired area a median like rule is applied on m_{area}

$$\overline{m}_{area}(k) = \begin{cases} 1, & \text{if } \sum_{o=-k_a/2}^{k_a/2} m_{area}(k+o) > \frac{k_a}{2} \\ 0, & \text{else} \end{cases} \tag{3.16}$$

where k_a denotes a minimum duration an event is assumed to last.

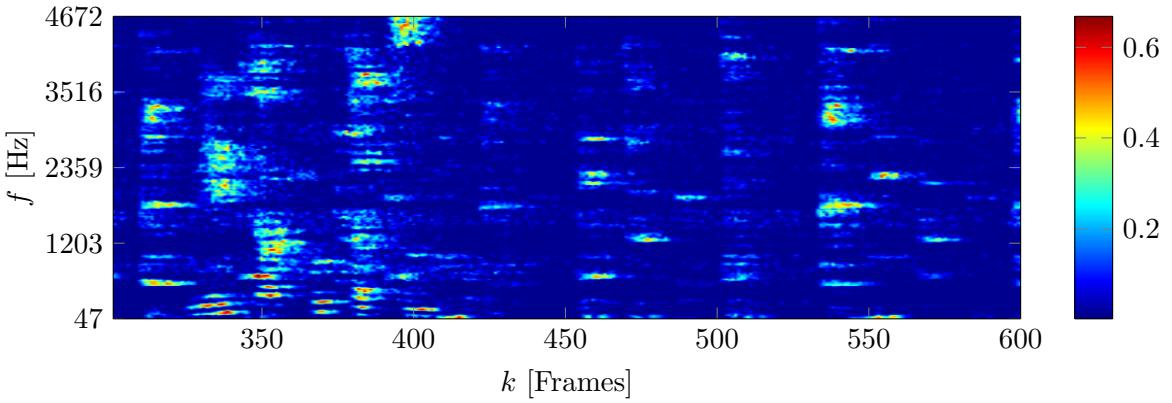
The event extraction process can be viewed in Fig. 3.7 based on a spectrogram containing a speech signal as well as music in the background. Taking the *sim15.wav* from DIRHA_DE a rhythmic music is played by a radio or TV station. When correlating signals of the whole three second block against each other, this noise type most likely dominates correlation results of single words. Separating the block in single short events can help to avoid this dominance, as in most cases, spoken words contain more energy compared to the noise source. In addition, the likelihood of strong patterns coming from the noise source is reduced when using shorter blocks for correlation. On the other hand, it has to be ensured to keep a minimum block-size to ensure sufficient resolution for maximum discrimination.

Figure 3.7 demonstrates the masking steps to extract interesting areas. In (a) the normalized spectrogram has been averaged over all used channels revealing important areas where signal spreads through the apartment. Summarizing this spectrogram leads to the bottom plot (b). Thresholding the mean signal results in short areas of interest represented by the dashed curve. Simple median filtering leads to the bold curve which in this example extracts to areas of interest. The first and broader one corresponds to the end of a speech signal, while the shorter one resulted from the rhythmic background-music.

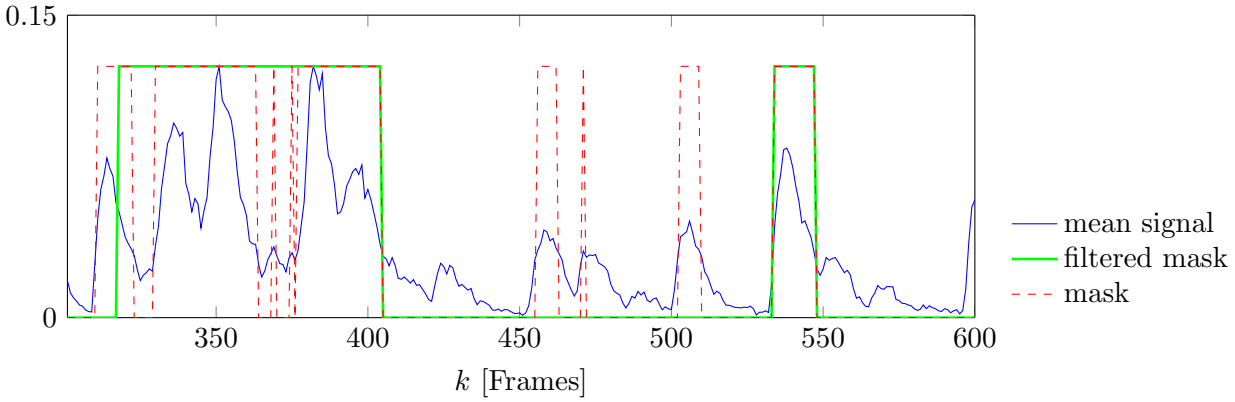
The **Cross-Correlation** can be a good indicator when looking for the same pattern in different signals. While the maximum of the resulting function not only denotes a high similarity but also the relative position between the original signals it also can be used to find the earliest appearance of the pattern. Using this for room localisation early tests have shown to be more stable when correlating spectrogram instead of the time-signals although they come with the lack of a low time-resolution.

The correlation function for spectrograms of infinite length can be written as

$$R_{ij}(c,f) = \sum_{k=-\infty}^{\infty} X_i(k,f) X_j(k+c,f), \qquad \forall i,j = 1 \dots N_{ch} \tag{3.17}$$



(a) $\tilde{\boldsymbol{X}}_{est}$ averaged over all channels



(b) Average and masking signals

Figure 3.7: This figure demonstrates the area extraction using the 'sim15.wav' file from DIRHA_DE which contains a disturbing rhythmic music as noise in the background. The top plot corresponds to the averaged normalized spectrogram of the fourth analysed block which further summarized leads to the bottom plot. Here, the dashed curve represent the masking after simple thresholding. Using a median-filter, the mask concatenates the smaller blocks leading to two areas shown by the bold curve which will be correlated.

where X_i and X_j are the spectrograms of two channels and $i, j = 1 \dots N_{ch}$ denote the channel number with N_{ch} as the number of channels. Note that in case of $i = j$ autocorrelation is performed. As the signal length is finite Eq. (3.17) has to be adapted referring to indexing and including a weight factor for bias compensation:

$$R_{ij}(c,f) = \frac{K}{K-|c|}\left(\left[\sum_{k=1}^{K+c} X_i(k-c,f)X_j(k,f)\right]_{c<0} + \left[\sum_{k=1}^{K-c} X_i(k,f)X_j(k+c,f)\right]_{c\geqslant 0}\right) \quad (3.18)$$

where $c = -K \dots 0 \dots K$ denotes the shifting of X_i against X_j being the index of each correlation function. K stands for the length of X_i and X_j.

At this point, R_{ij} contains a correlation function for every frequency band of X_{est}. Without further enhancement or frequency selection rules these functions are summed up to form an

overall correlation function per channel pair, i.e.

$$R_{ij}(c) = \sum_{f=1}^{F} R_{ij}(c, f) \tag{3.19}$$

where $f = 1 \dots F$ correspond to the number of frequency bands. $R_{ij}(c)$ can be seen in Fig. 3.8.

The next step addresses **Maximum Localisation**. In case of no transient signals i.e. without any short spoken commands or short noises like door smashes, water flush or using household aids we can assume stationary noise which is less correlated between rooms (see also Sec. 2.2). The structure of $R_{ij}(c)$ will be flat without distinct maxima except for the case of $i = j$ as autocorrelation contains a maximum at $c = 0$.

In the other case assuming an acoustic event A_e with sufficient energy to spread through most adjacent rooms. Even with the low time resolution of X_{est} TDOA of A_e between channels is detectable for signals in adjacent rooms. When the signal has been recorded at each channel, cross-correlation results in more or less distinct peaks for each R_{ij}. The peak location depends on the difference d_{ij} between l_i and l_j which correspond to the distance between the source of A_e and microphone i and j respectively. d_{ij} is proportional to $|c|$ i.e. a lower difference in source-microphone distances will lead to a correlation peak closer to $c = 0$. The worst case happens with $l_i = l_j$ and $i \neq j$ i.e. if the source position happens to be in the middle between two microphones. If reverberation conditions are the same it is impossible to perform room detection just by using correlation. The sign of c at the correlation peaks depends on the position of A_e related to the microphone i and j. For $l_i < l_j$ the signal arrives later at microphone j and $c > 0$. If $l_i > l_j$ the signal will arrive earlier at microphone j hence $c < 0$.

According to these properties, extracting the position of the maxima of R_{ij} is the next step of the algorithm, i.e.

$$c_{\max}(i, j) = \arg\max_{c}(R_{ij}(c)) \tag{3.20}$$

which results in a $N_{ch} \times N_{ch}$ matrix of correlation indices referring to each corresponding maximum. Each row contains peak positions of $R_{i=\text{const},j}$ which represent the perspective of one channel i being correlated with all channels j.

For **Channel Detection** Assigning the lowest position to its corresponding channel reveals where the event most likely originates from.

$$j_{\min}(i) = \arg\min_{j}(c_{max}(i, j)) \tag{3.21}$$

Whether the determined channels correspond to the shortest source-microphone distance or if they just accidentally occurred first due to the low precision in time or noise disturbances is controlled by a final rule: a room r is found to contain the acoustic source if $j_{\min}(i_r)$ correspond to channels inside this room. In case of using only one channel per room this would mean that the peak of the autocorrelation of room r has to be the first maximum. This means that if the first peak of the correlation functions of some rooms corresponds to r but inside this room $j_{\min}(i_r)$ indicates A_e is coming from outside this room it would be illogically to assign A_e to r. Using this rule, the worst case result would be if the distance between a person and the microphone in its room is similar or equal compared to the distance between the person and the microphone of the next room. This might lead to the location of the person in the other room respectively making a room detection impossible. The better case of this situation locates the person each time in the same room so that both rooms will be selected.

Figure 3.8 shows the result when performing spectral correlation using five channels. The title of each subfigure indicates which channel is correlated against all channels. Each centre peak corresponds to an autocorrelation function and therefore is not only the highest but also the most narrow peak compared to the cross-correlation maxima. As the latter tend to broaden a lot, the maximum of each function is marked by a circle to compare relative distances between the signals.

To present the underlying principle described above, the signal which has been used for this plot is a clean speech signal to avoid influence from noise sources. Therefore, the origin of speech can be assigned to the living room as the first maximum occurs in every case when the corresponding signal is correlated.

The correlation process of the algorithm described where room localisation is based on the results of the correlation in each frequency-band of the spectrogram mainly influences its robustness. In addition with its acoustic event based division of the whole spectrogram the algorithm becomes independent of a preceding VAD i.e. it could perform event-based room localisation first followed by the classification of each event. Furthermore, the algorithm is capable of locating more than one event at the same time. For this it is important that the interference between both events is at a sufficient low level to ensure not to miss one being dominated by the other. In addition the channel set used influences the success as well especially in terms of TDOA.

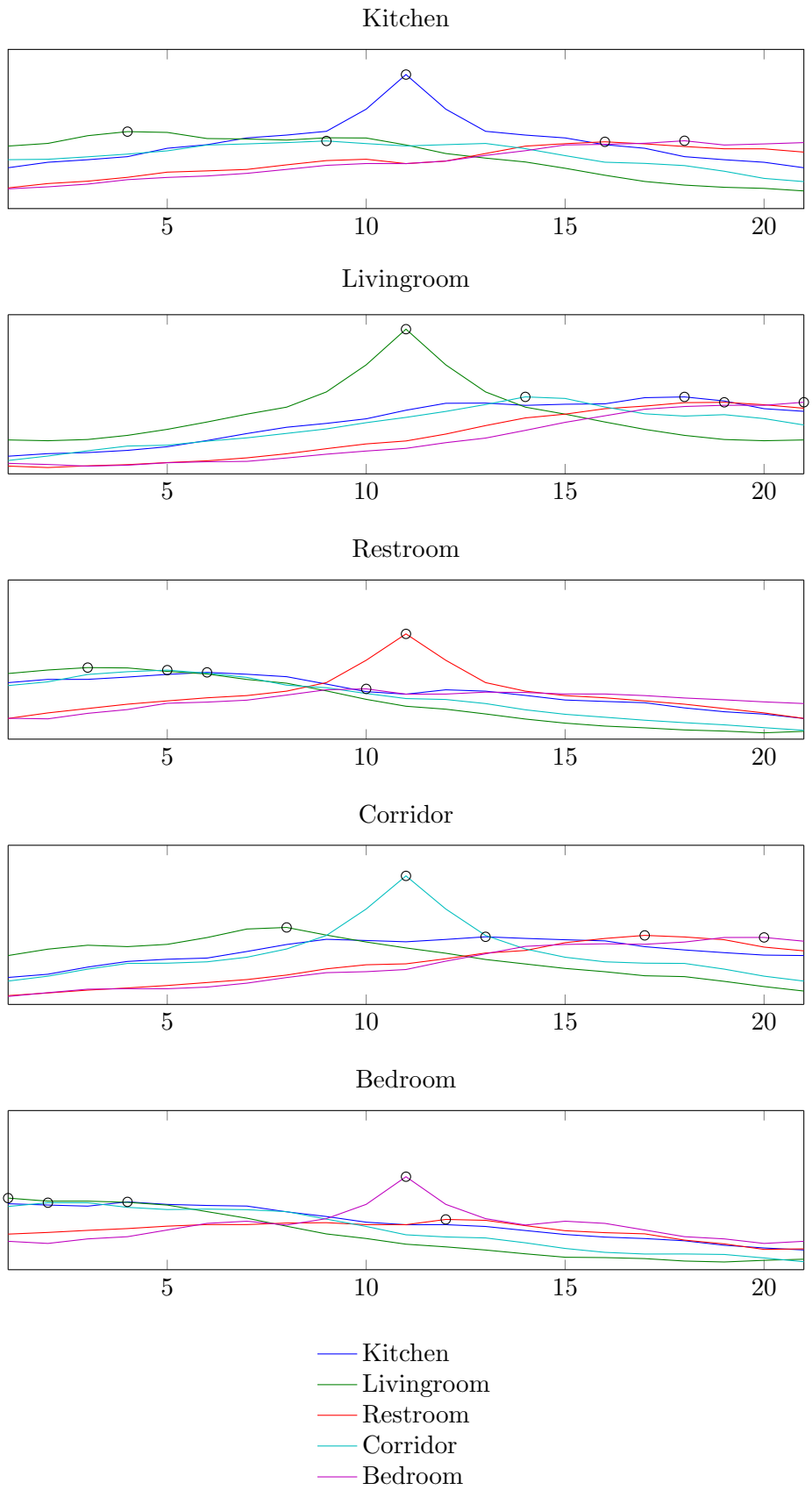


Figure 3.8: This figure contains a delay example with cross correlation functions. A clean speech version of 'sim1.wav' from DIRHA_DE has been used and the channel set has been selected using the $Maximum - Energy - Channel - Selection$ algorithm. Each maximum is marked by circle to increase visual distinction as cross-correlation tends to result in broad peaks. In all five cases, the correlation maximum of the living-room signal comes first compared to the others which indicates that speech originated in this room.

3.5 Conclusion

Using the different classification or VAD schemes specified with the developed room detection algorithms in the previous sections, a whole set of combinations of different multi-channel-VADs can be constructed to perform the desired tasks. This section is a short summary of the combinations used.

In addition to the DBN-classification VAD, a VAD focussing on certain spectral parts using thresholds has been proposed which contains sharp onsets and depends on the tonal part of speech.

As for localisation, two different approaches have been made. The simpler one, onset detection, depends on a preceding VAD by comparing the TDOA of the beginning of a word or phrase between different channels. Because of this, it cannot detect two different speakers at the same time.

The second room detection algorithm is more complex and based on spectral correlation. It extracts transient parts of the spectrogram which are cross-correlated to find the temporal appearance of the signal by comparing the correlation-maxima. The algorithm is capable of localising speech in different rooms at the same time. In addition, it can be used as event detector and localiser which can be followed by a VAD. In this case, classification only depends on signals where events have been localised which are more likely to be less destructed by reverberation and attenuation.

4 Results

In this chapter results are going to be presented using the algorithms discussed in Chapter 3. In all cases the German corpus DIRHA_DE has been used.

All computations are performed on segments $y_b[n]$ of the recorded signal y. Each segment is three seconds long and the hop size corresponds to one second. The block-length empirically has been shown to provide stable results without too much interference produced by background signals. The overlap is necessary to ensure robust Voice Activity Detection (VAD) where only a couple of frames at the beginning or at the end of the centre block contain speech. Each block will be labelled as containing speech activity or not. If voice has been detected a number marks the room where the source was localised. For better readability, no index b will be used for block indication.

When training a classifier overlapping is rejected to avoid redundancy and computation time by using each frame three times.

As the sampling rate of each signal equals $s_r = 16$ kHz, computing the spectrogram Y by using a frame length $k_{len} = 512$ samples and a hop-size $k_{hop} = 160$ samples yields in a frame rate $k_r = 100$ Hz and therefore 300 frames/block.

Depending on the experiment a different number of channels i forms the basis signal set for VAD and localisation which may change its sources every single block. In all cases block-wise noise reduction will be performed first to get reliable results. The adaptive algorithm used adapts noise estimation based on noise information from the preceding block following the minimum statistics approach in [20] (see Sec. 2.2.3). In the experiments, an implementation prvided by Voicebox[4] has been used.

Table 4.1 shows the features used when performing classification with the Deep Belief Network (DBN). Although based on [12] different dimensions have been used for the DFT- and MFCC-Features according to [6].

From this two other sets are derived which will be used for classification using the DBN. The smallest set representing a kind of reference for improvement simply contains the MDFT_8-Features but as energy mel spectrum. It is therefore referred to as *Mel-Freq-Energy Features* in the following sections. The second set consist of the 15 best performing features out of the full feature set after performing a Best-First search on all features and using gaussian mixture models for classification. In Tab. 4.1 the *Element* column contains these features where the number refers to the element inside the corresponding feature. Here, number one basically corresponds to the lowest frequency or its feature derivative in the first element of a multi dimensional feature.

[4] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Feature	Dimension	Element	Feature	Dimension	Element
Pitch	1	1	$MFCC_8$	13	2,4
MDFT	26		$MFCC_{16}$	13	12
$MDFT_8$	26		LPC	12	2,5,7,10,11
$MDFT_{16}$	26	21	RASTA-PLP	17	15
MFCC	13	1	AMS	135	47,91,95

Table 4.1: Feature set proposed by Zhang et al. with adapted dimensions according to [6]

For all experiments the sets of channels defined in Sec. 3.2 will be used for comparing efficiency in terms of performance vs number of channels and the influence concerning different adaptive approaches.

Tests are performed on each recording of the database leading to representation metrics like recall and precision for each scenario respectively. The recall represents the relation of all detected and correctly localised speech fragments, i.e. TP_{room} over all possible fragments of voice activity. The precision relates TP_{room} and the sum of correctly detected and localised speech fragments, falsely detected voice activity and any wrong localisation in case of speech (see Sec. 2.2.4). The results presented in the following section combine the corresponding metrics by computing the average and standard deviation of the single values.

The results are ordered in subsections as follows: Section 4.1 will present the results of noise subtraction between different algorithms as well as the performance of VAD and room detection which are extracted by using background knowledge. Section 4.2 presents the results concerning VAD and room localisation in terms of recall, precision and F-score (see Eqs. (2.18) to (2.20)).

4.1 Single Performance Analysis

In the following section, the focus lies on single modules of the algorithm. While in Sec. 4.1.1, results concerning noise subtraction algorithms are shown Secs. 4.1.2 and 4.1.3 presents the performance of VAD or room detection by exclude the influence between each other. To achieve this, background knowledge is used in terms of the interpretation of the results. On the one hand, when performing VAD localisation will be neglected, i.e. decision is made whether speech exists or not. On the other hand for room detection, only blocks containing speech are taken into consideration to provide a perfect VAD. Doing this, the overall results should be easier to interpret.

4.1.1 Spectral Noise Subtraction

Figure 4.1 plots the increase of accuracy when performing spectral subtraction before classification. Accuracy is defined as

$$acc = \frac{TP + TN}{N_{sp}} \tag{4.1}$$

The DBN VAD is used and room location is provided. As feature set are the Mel-Freq-Energy features at a varying smoothing-length l used. The baseline indicates the best result without spectral subtraction.

Using Voicebox, MMSE-implementations based on the work of R. Martin [14] and T. Gerkmann [20] is applied on the $MDFT_l$ representation of the spectrogram as well as on the clean uncompressed representation Y. The latter clearly shows the best improvement of up to 4%,

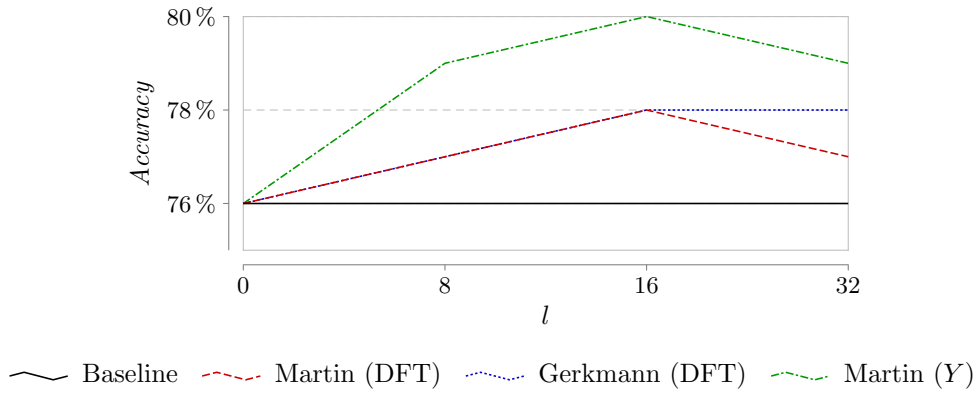


Figure 4.1: Accuracy increase of classification after performing spectral substraction.

while using a basic feature for classification. The performance between the both implementations is very similar. Therefore the choice to use the implementation of R. Martin was an empirical one.

4.1.2 Classification

Figure 4.2 presents the recall and precision values of the different channel sets, when neglecting the condition of finding the correct room. Hence, if a speaker has been detected it will count as hit even if the wrong room has been found. For DBN-classification, the full Zhang feature-set has been used.

The results are plotted for all three configurations containing DBN-classification, i.e. classification before room detection via onset detection or spectral correlation as well as classification following event detection. This is, because results differ dependent on the room detection scheme used.

When using spectral correlation the results occur to be quite similar. Taken the recal, both seem to be less dependent on the channel set used. When performing classification first, recall increases up to 2% but on the other hand precision drops in a similar matter. Interestingly, using 14 channels seem to have a stronger influence on precision when using classification first.

The result of VAD followed by onset detection is quite different. Especially recall drops highly which is not surprising as the onset algorithm partially performs classification as well when no clear onset exist. Since precision is around 95% on average and recall strongly improves when using a classification scheme like spectral focussing where thresholding produces sharp onsets, this dependency becomes more obvious. In addition, the results become dependent on the channel set used. For spectral focus, using the Maximum-Energy-Channel-Selectionset results in the highest recall while precision is best when using the 14 channels.

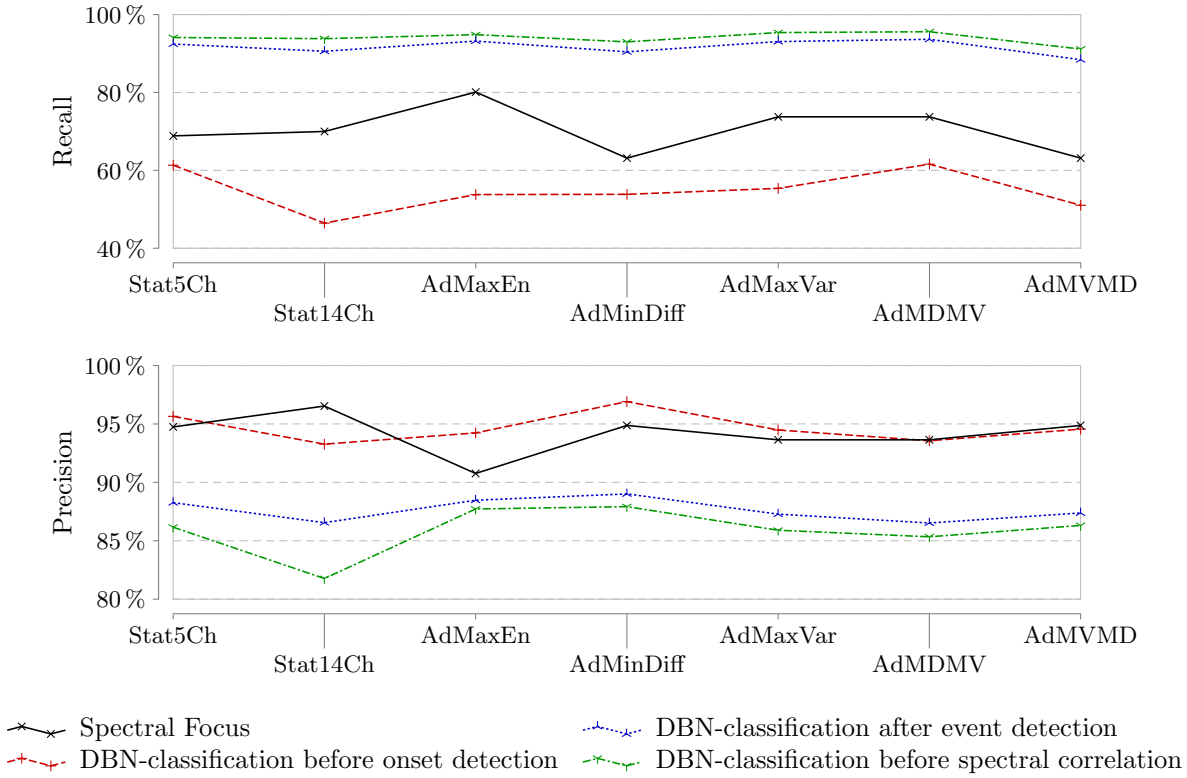


Figure 4.2: Recall and precision focussing on VAD performance. When DBN-classification and spectral correlation are used, the results are quite similar and less dependent on the different channel sets. When onset detection is used, its influence on classification can be seen as the detection algorithm drops signals in case of non-existent onsets. In this case, the threshold procedure of spectral focus outperforms DBN-classification. In addition, channel dependency increases where $Maximum - Energy - Channel - Selection$ results in the highest recall when using spectral focussing.

4.1.3 Room Detection

The results in Fig. 4.3 correspond to room detection performance. To focus on that, recall and precision are computed on the result of room localisation taking only signal blocks, where speech successfully has been detected. As in the preceding subsection, the Zhang feature set has been used for classification.

Taking recall, spectral correlation outperforms onset detection of about 5%. Considering the order between VAD and room detection, performing the latter first improves recall as well as precision. Onset detection after DBN-classification in contrast performs poorly where recall is about 20% lower in all cases. As only successfully detected speech blocks have been used for this result, the remaining dependency to VAD only refers to the characteristic of the spectrogram after performing classification which results in frequency dependent smooth and less predictable starts of a word.

The overall precision is around 70% in general. Except of the case of spectral correlation preceded by classification the use of 14 channels clearly improves precision of more than 5%.

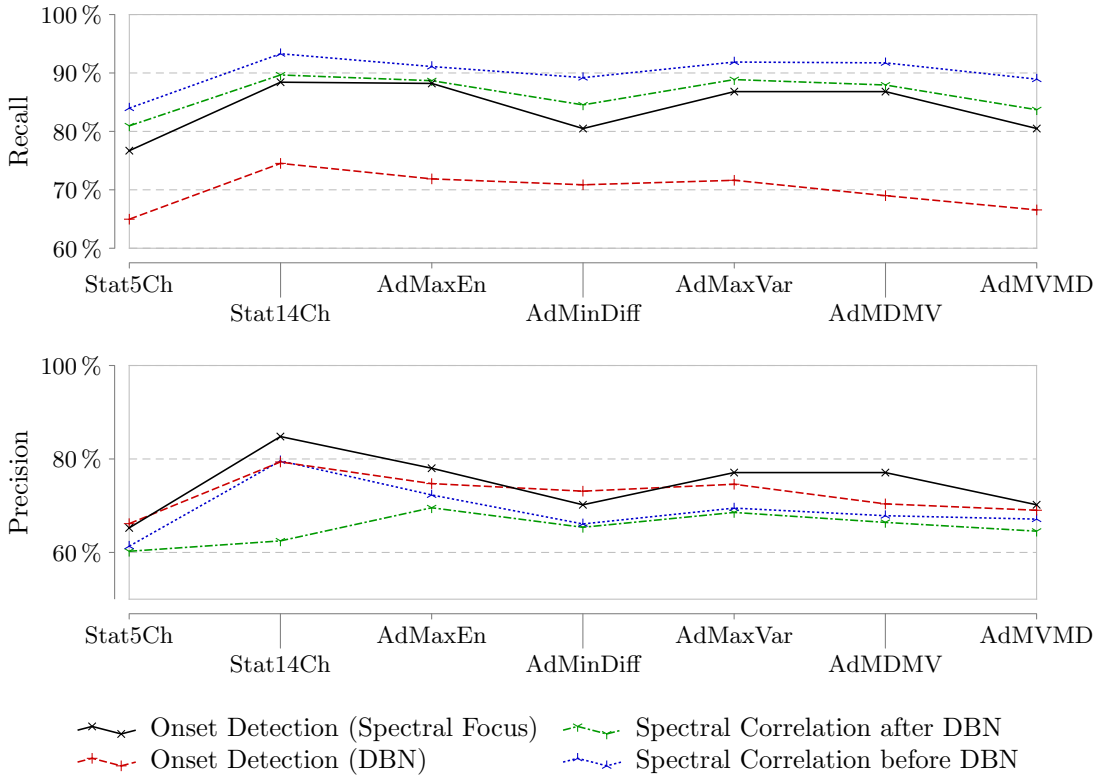


Figure 4.3: Recall and precision when focussing on room detection. Using spectral correlation outperforms onset detection in terms of recall while on the other hand precision is lower in all cases. Onset detection used on the DBN classified spectrogram performs poorly where recall is about 20 percent lower in general.

4.2 VAD and Room Detection

The following experiments are ordered according to the localisation algorithms proposed. Taking onset detection it will be used as room detector preceded by a VAD as it uses the beginning of detected voice activity for localisation. The VADs used will be the spectral focussing algorithm which will be compared to the G729B-VAD and Sohn-VAD as well as the DBN classifier, where different feature sets have been used.

Using the spectral correlation based localiser only DBN will be used assuming its better performance and to focus more on differences caused by varying channel- and/or feature-sets and because of its independence of acting as a event detector and localiser followed by a VAD or simply localising detected voice activity. The number of channel combinations used is increased by two variations of the Maximum-Variance-Channel-Selectionand Minimum-Difference-Channel-Selection. They will be used in combination by taking one when performing VAD and the other for localisation and vice versa. To increase readability when labelling plots, abbreviations are used in figures which are explained in Tab. 4.2.

Performance in terms of recall and precision has been computed on block level, i.e. blocks are labelled as speech or non-speech. A number indicates where speech originates from or has been estimated respectively. Ground truth for comparison is determined by extracting time information from the database. It is provided in samples and is transformed into block information for comparison.

The following sections give a short description of the experiments made and its characteristics and present the results.

Abbreviation	Long form
Stat5Ch	*5 stationary channels*
Stat14Ch	*14 stationary channels*
Stat40Ch	*All 40 channels*
AdMaxEn	*Adaptive channel set using Maximum-Energy-Channel-Selection*
AdMinDiff	***Ad**aptive channel set using Minimum-Difference-Channel-Selection*
AdMaxVar	*Adaptive channel set using Maximum-Variance-Channel-Selection*
AdMDMV	***Ad**aptive channel set using Minimum-Difference-Channel-Selectionfor classification and Maximum-Variance-Channel-Selectionfor room detection*
AdMVMD	*Adaptive channel set using Maximum-Variance-Channel-Selectionfor classification and Minimum-Difference-Channel-Selectionfor room detection*

Table 4.2: Meaning of the abbreviations used in the figures presenting the results.

4.2.1 Room Detection based on Onset Detection

The onset detection based room detection follows the simple rule, that parts of speech arrive at different times at different microphones. Selecting the channel with the earliest onset therefore leads to the room where the speech originated. Hence, VAD is a crucial part especially when it comes to detect the very first frames of a phrase. Depending on the first syllable in combination with background noise it can be very tricky to decide wether one frame is speech or not. But to find the earliest onset robust decision making is necessary to ensure proper time relations.

In the following two experiments two different VADs are used to perform this task. The first one is described in Sec. 3.3.1 which focusses on the tonal part of speech in a narrow frequency band. While it doesn't need a training stage by simply using thresholds for decision making, it can't distinguish between speech or background noise with strong harmonics in this frequency band. Similarly, it cannot be used in case of whispering. But for the energy rich voiced parts of speech finding the correct onset work well.

The second VAD is based on DBN-classification where the right feature set used is crucial for success. In the ideal case using the perfect features and being able to train the classifier with an large amount of data classification would work perfectly. As in reality the database usually is small and the feature set is not perfect, the algorithm will face problems in case of loud background noise with strong harmonics, soft-spoken sentences or noise being similar to speech. These problems especially occur at the beginning or the end of phrases which is one drawback in using this VAD as basis for a robust onset detection.

Experiment 1 - Spectral Focussing and Onset Detection

Figure 4.4 shows the results of the simplest algorithm combination: the spectral focussing VAD preceded by the onset detection algorithm as localiser. For comparison, results of the combination with the reference VADs and onset detection are given as well. These combinations have been made in both orders such that onset detection is acting as simple localiser (labelled with *-loc* in legend) as well as localising events when coming before VAD (labelled with *-ev*).

Results are plotted in terms of F-score, recall and precision over the different channel sets

used. Figures on the left side show the mean results over all tested signals while figures on the right side contain the corresponding standard deviation. The database used was DIRHA_DE and only test signals have been used for comparison reasons with the VAD using classification.

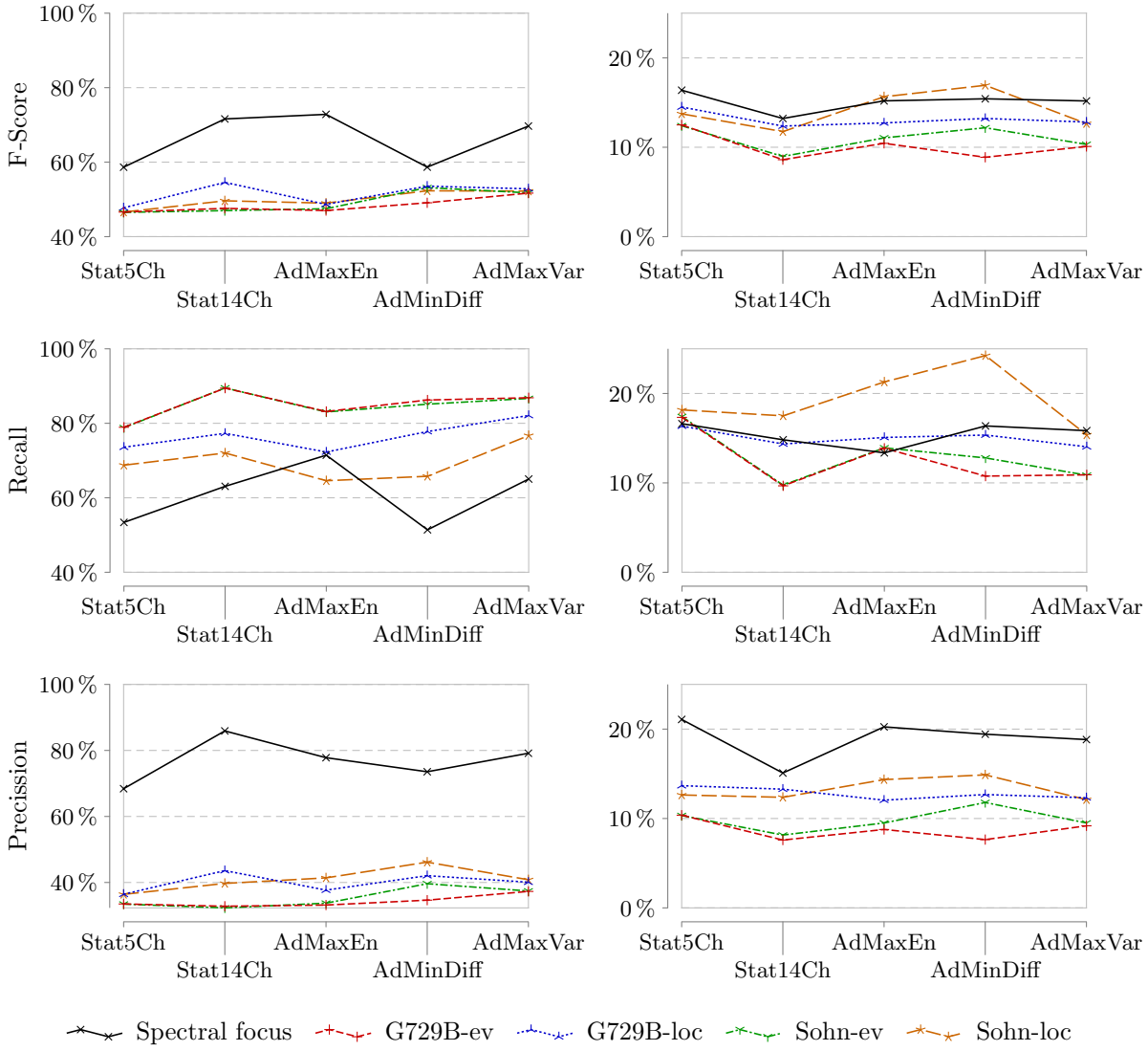


Figure 4.4: Results of Experiment 1 depending on the channel sets used: classification is used for VAD and room detection done by onset detection. The left three figures present the mean results over all tested signals in terms of F-score, recall and precision containing different feature sets. The right three figures contain the corresponding standard deviation. The used database was DIRHA_DE .

Just by looking at the F-score, differences between the known VADs (dotted and dashed lines) and the spectral focussing VAD are significant. While each combination of onset detection and the G729B and Sohn VAD respectively produce similar results, the spectral focussing VAD is depending on the channel-set more than 20% higher. It's strength lies obviously in its kind of focussing on the spectrum making it much more precise than the other algorithms. But its strength also becomes a disadvantage in terms of Recall. For most channel combinations it is outperformed by all other VADs.

This characteristic is a crucial one, as VAD isn't the only task but the very beginning of the Distant-speech Interaction for Robust Home Applications (DIRHA) project. For automatic speech recognition it must be ensured to detect not only the centre part of words but beginning and end as well. Therefore the result has to be taken with care.

As for the other VADs the result is exactly the opposite. Especially where onset detection is

used as event localiser, recall is very high while precision has dropped below 40%. So, both VADs take most events as voice activity resulting in a lot of false positives. Both implementations are tested with standard settings but one can assume, that they aren't adapted well enough for the task. In case of G729B it's origin has to be taken into account. Developed for mobile devices it works under totally different conditions. Furthermore its VAD had been added to increase compression during pauses in speech. Hence, it doesn't cost the algorithm if too much noise is misclassified.

Taking a look at standard deviation. Here, the result is very high in most cases making robust expectations about the classifiers outcome impossible. Only in case of event localisation values drop below 10%.

The more interesting part is the comparison between the different channel-sets. Again there exist distinct differences between the two kind of groups. The spectral focussing VAD gets its best results with the adaptive maximum energy channel set especially in terms of recall.

Experiment 2 - DBN-Classification and Onset Detection

This experiment uses DBN VAD for classifying the signal while onset detection is meant to perform room detection. The results can be seen in Fig. 4.5. As in the preceding experiment the DIRHA_DE database is used and the results are presented the same way only comparison now is performed between different feature-sets.

Comparing the mean values depending on the feature sets best case results of recall reach more than 70% in case of precision even more than 80%. But the overall performance using this classifier isn't sufficient. The problem might be the lack of reliably finding the start point of words. The classifier might detect single values per frame in one channel as being important and onset detection won't work any more.

The values become problematic when considering the high standard deviation values. None of them is lower than 15% hence the range of values to be expected around the mean value is still more than 30% in best case. A mean value of 70% in this case would indicate that at best performance classification results have to be expected in the range of 55% and 85%. This reduces the effective differences of the average results between different features or channel sets used as their range is highly overlapping.

Comparing the feature sets used, the reduced Zhang features tends to work better in terms of recall. In the best case it outperforms the full set up to 15% while standard deviation is about 20%. Compared with the high reduction of the number of features to little more than 5% the reduced feature set obviously provides better distinction between noise and speech at word beginnings increasing the robustness of onset detection.

Looking on the different set of channels differences mainly depend on the feature set used. The recall of the full Zhang feature-set is almost independent of channels used whereas for the remaining sets the adaptive sets found by maximum variance or energy seem to give better results. Taking standard deviation into account, it often increases with better performance, hence, decreasing the influence of the channel set used.

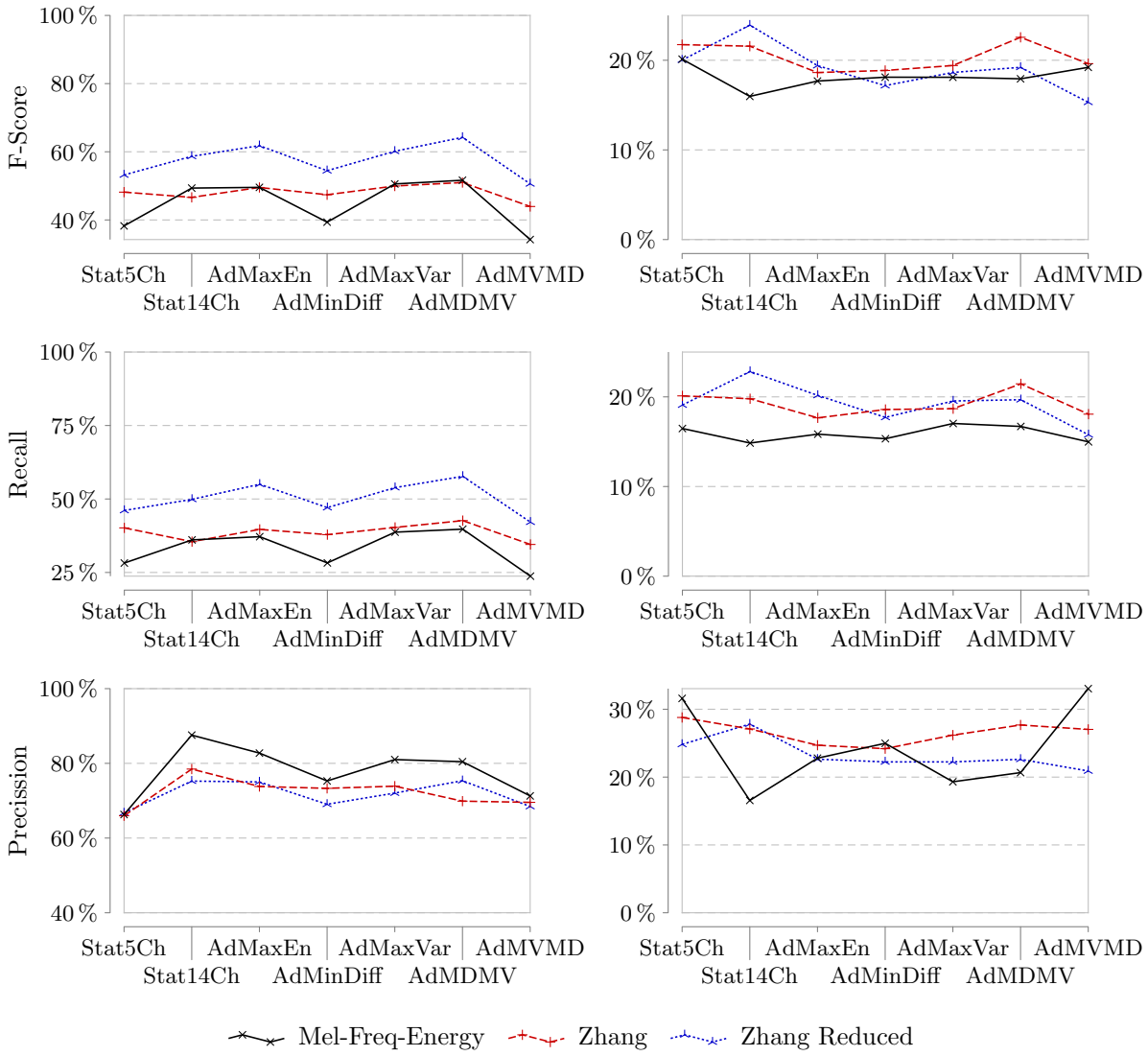


Figure 4.5: Results of Experiment 2 depending on the channel sets used: classification is used for VAD and room detection done by onset detection. The left three figures present the mean results over all tested signals in terms of F-score, recall and precision containing different feature sets. The right three figures contain the corresponding standard deviation. The used database was DIRHA_DE .

4.2.2 Room Detection based on Correlation

Room detection based on correlation is described in detail in Sec. 3.4.2. VAD is done by using a DBN as classifier. As mentioned in Sec. 4.2.1 the result increases with better features and with an increasing dataset. But the big difference compared to room detection done by finding the first onset that correlation includes many frames together looking for the same pattern. So in case classification at the edges of a phrase is not perfectly done, the phrase itself will be detected and the pattern matching process dominates as long as noise corruption doesn't dominate. The latter should be avoided by classification.

The following experiments use both algorithms but in different time order. In Experiment 1 room detection is done before classification which means that for each longer or group of shorter events decision is made in which room they occurred. Suppressing everything else, the classifier only has to detect voice activity. Experiment 2 does VAD in a first step suppressing everything not being marked as voice. Correlation now only uses this information for room detection. In both cases the German database has been used.

Experiment 1 - Event Detection and Localisation and DBN-Classification

In this experiment event detection and room localisation is performed first, followed by classification of the remaining signals in terms of speech or non-speech. When done perfectly the classifier only would get less reverberated signals which should be easier to distinguish. The classifier has been trained without room localisation first which highly improves results in terms of recall. The overall performance can be seen in Fig. 4.6.

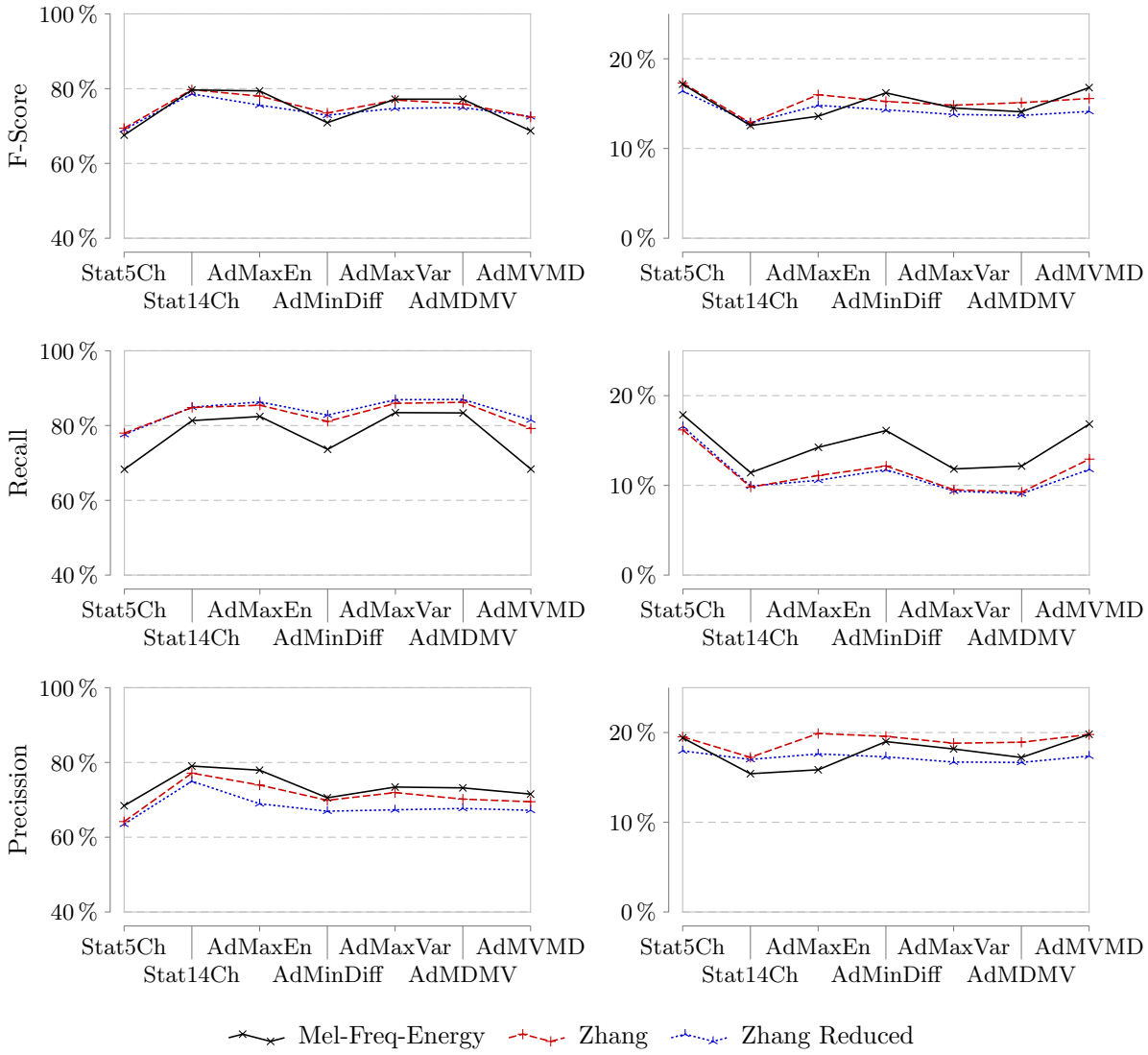


Figure 4.6: Results of Experiment 1 depending on the channel sets used: first room detection is done by correlation followed by classification to detect voice activity. The left three figures present the mean results over all tested signals in terms of F-score, recall and precision containing different feature sets. The right three figures contain the corresponding standard deviation. The used database was DIRHA_DE .

Looking at the mean values of the result two major conclusions can be made. First, the dependency between the feature sets is negligible although the Zhang features and the Mel-Freq-Energy features behave differently. So dependency increasing, when looking at recall and precision. Second, the overall performance is high especially in case of recall where high mean values and a low standard deviation make the result more reliable.

Taking the focus on recall which is important for further processing, the Zhang features outperform Mel-Freq-Energy. This is not surprising but looking at precision, the full set seem

to win against the reduced one. On the other hand similarly to the preceding section, standard deviation changes directly proportional which removes the advantage instantly.

Even for different channel sets, only the use of one signal per group seem to stabilize the results a little more. This could indicate, that the detection and localisation is the crucial part in this order which could be improved when using more channels per room in case adaptive selection is not robust enough.

The same holds for the little dependency of feature combinations. Again, if work has been done clearly, the training data reduces to signals with a far higher Signal to Noise Ratio (SNR) compared to the requirement of detecting VAD in every room in case voice energy is sufficiently high. Here, even the signal in the adjacent room lacks a lot of information and becomes more difficult to distinguish from similar sources.

Experiment 2 - DBN-Classification and Room Localisation

Compared with the preceding experiment the setting now is switched starting with classifying each signal for speech and use the result for room localisation. Figure 4.7 presents the results.

Focussing on recall the results look very promising reaching almost 90% in the best case. It looks quite similar to the preceding experiment although slightly worse in terms of precision and overall standard deviation.

Very interestingls is the almost identical result of the full and reduced Zhang feature sets. This seems very surprising at first as the results of using onset detection instead of spectral correlation showed to contain a significant dependency. On the other side it is worth to note that the correlation process uses the spectral information just being told in which area one should expect speech and where not. This makes tiny errors at the beginning or during a phrase less important as the correlation acts as a pattern search over a longer time period. Onset detection instead is very sensitive to tiny errors and therefore much more dependent on classification performance.

Again, the difference in using several channel combinations seems not that important at first but still reaches differences above 5%. In this case the results seem to making the adaptive set based on the maximum variance between the signals of a microphone pair the most reliable choice of channels. Lower standard deviation overall and higher mean values for precision lead to this result even if the difference is not that high.

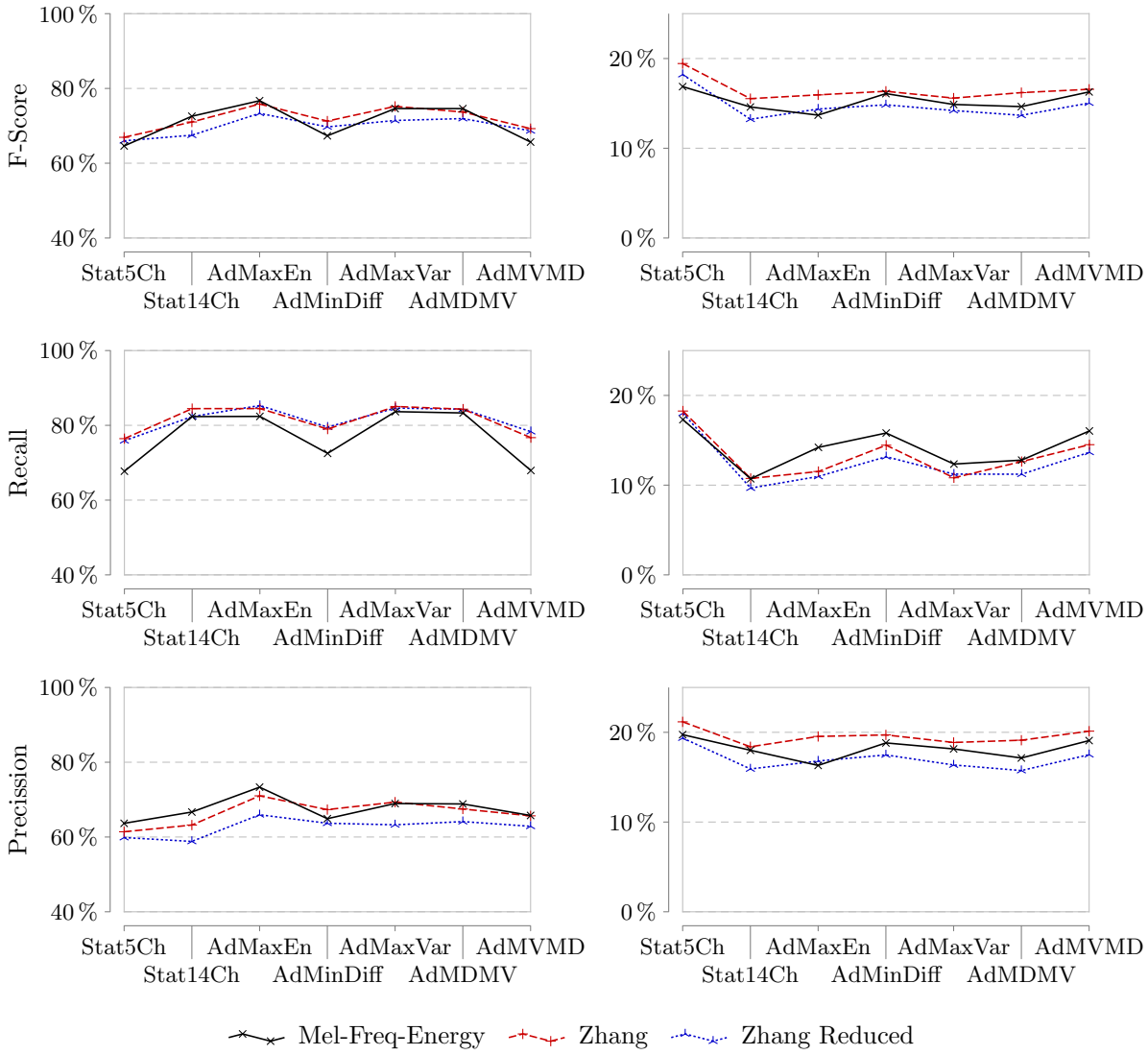


Figure 4.7: Results of Experiment 2 depending on the channel sets used: classification is used for VAD and room detection is done using correlation. The left three figures present the mean results over all tested signals in terms of F-score, recall and precision containing different feature sets. The right three figures contain the corresponding standard deviation. The used database was DIRHA_DE .

5 Conclusion

This thesis presents work on the task of using existing algorithms for Voice Activity Detection (VAD) extending them with a reliable and efficient source localiser to improve performance in multi-room environments. The mainly used VAD uses a Deep Belief Network (DBN) classifier which is -once trained - very efficient in both accuracy and speed.

Being part of the Distant-speech Interaction for Robust Home Applications (DIRHA) project the constraint not only addressed a reliable detection of voice activity but also detecting most of it. In terms of recall and precision this means, that a higher rate of false positives, i.e. a lower precision is more acceptable than a higher rate of speech being missclassified as noise, i.e. a lower recall.

During this work, it quickly showed to be a task containing lots of parameters to be tuned. While developing an algorithm for room detection, noise enhancement and feature set comparison have been equally important as finding a reliable way in reducing and optimising the number of signals used to reach that goal. In the end, room localisation algorithms have been developed as well as a simple VAD which could be used as a feature indicating voiced speech included in a larger feature set. Similarly several attempts in finding an adaptive way of combining the best signals per room making selection more robust have lead to more parameters.

It has been shown that starting with the work of [12] the VAD successfully could be extended by an algorithm being able not only to perform room detection but also to allow overlapping speech of speakers in different rooms.

The results to focus on the single components like VAD and room detection as well as on the overall performance of the complete algorithm. In the following sections the results will be further discussed. In Sec. 5.2 a comparison is made between the performance of the different settings. Section 5.3 provides a conclusion and Sec. 5.4 will give an outline about future work.

5.1 Overall Performance

The overall performance of the different configurations has been shown to be quite diverse. While for one setting average F-score values hardly exceeded 60% the next algorithm almost reached an F-score of 80%. Looking at the underlying metrics of F-score the combination of VAD and localisation algorithm either results in a high recall at the cost of being less precise or at a high precision but with an enormous drop-out rate. As for VAD, recall is the more important metric the comparison will focus on it. At best, an average recall of almost 87% has been reached.

Beside the average performance, a second important key aspect addresses standard deviation of recall and precision between the results of each test. In general the resulting standard deviation is high with values between 15% and 20% where recall usually performs better compared to precision. One reason for this addresses differences in signal quality such as noise sources or speech intensity.In addition the segmentation into 60 blocks per recording needs to be considered as well as a block is labelled as speech without care how much frames actually contain voice activity. On the other hand, RE and PR depend on two factors i.e. VAD and room detection which in a similar manner influences the variation. An evaluation about these factors however didn't have been performed which means, that the interpretation of the standard variation of the metrics cannot completely be projected onto the algorithm.

Looking at VAD and room detection separately gives insight whether one part of the detection and localisation process works better. In both cases the recall is above 85% in case of room detection it even exceeds 90% when using event detection and DBN-classification. For DBN-classification and spectral correlation, the difference can be seen when looking at the precision. Here, VAD is more stable mostly being above 85%, whereas the precision of room detection lies below 70% in most cases.

Using the G729B- and Sohn-VAD for comparison shows that most configurations outperform the results obtained with these VADs mostly due to low precision. As no special optimisation of the algorithms has been performed, the results have to been taken with care. Comparing the performance of the DBN classifier to additional classification schemes like Support Vector Machine (SVM) or Linear Discriminant Analysis (LinDA) had to be stopped due to limitations of the implementation used.

5.2 Comparison

When comparing the results of the different combinations of the algorithms used one combination doesn't seem to work together. When performing the classification of the spectrogram the influence of signal strength and background noise as well as the training data available lead to a less distinct decision at the start and end of a phrase. For the onset detection algorithm used, a clear word onset is essential especially when comparing the temporal occurrence between multiple signals. If this condition can't be met, the algorithm tends to loose certain speech fragments.

The spectral focussing approach aims to prevent this loss by thresholding the signal in such a way, that word beginnings become easy to detect for the algorithm as well as comparable between each channel. Results indeed improved compared to DBN-classification but still contain low values of the recall. Although precision is high, it won't be useful for VAD, as recall is about 70% in the best case which easily is outperformed by the remaining algorithm combinations.

Using DBN for classification and spectral correlation for room detection resulted in the best performance which could be observed. Therefore the following comparisons made address this combination. Testing the database by performing classification followed by room detection as well as by starting with localisation in terms of event detection followed by classification showed to be different in detail. While not very much apart, starting with localisation yields the highest recall on average.

Looking closer on the performance of the localisation-classification scheme in terms of VAD and room detection, the event detection approach is outperforming in case of room detection (see Sec. 4.1.3). The reason for this could be the normalization and segmentation process of the spectral correlation algorithm. Here, the complete signal ensures, that all transient details in all frequencies potentially being important will be enhanced such that correlation performs more robust. In case of classification being first, the transient characteristic of the spectrogram is altered in terms of where the beginning of speech has been detected in different channels and

how much frames are taken into consideration respectively.

After all, the probably largest difference lies in the training data. In case of classification being first, training has to be performed on all channels. This means, that the classifier has to distinguish between the clean speech signal and the reverberated and attenuated versions recorded in the adjacent rooms. When performing event detection, classification only needs to be performed where events have been localised which reduces computational cost as well as provides a more distinct way to distinguish between speech and noise data.

Recall from the results of the classification and correlation algorithms both combinations perform quite similar in terms of the different feature and channel sets used. Interestingly, the choice of the feature set doesn't seem to be very important as all three sets result in a similar F-score. When looking at the recall, differences between the log-spectrum features and the Zhang feature sets appear. Here, the latter outperform the simple feature set in both, the average and standard deviation results while being less precise at the same time. The differences between the full and the reduced versions of the Zhang feature set can be neglected. Compared to the huge difference concerning the number of features used the reduced feature set can be considered outperforming as it works much more efficient.

As for the features, the results in terms of different channel sets is very similar between the two combinations examined. While there is no single channel set being clearly superior to the remaining, some observations can be made. In case of starting with event detection, the constant 14 channel set as well as the adaptive 5 channel maximum energy set perform better in terms of precision although in this case being dependent on the underlying feature set. In case of classification being first, only the channel selection based on maximum energy seem to contain this advantage. For both combinations the dependency on the channel set of the recall clearly is stronger for the log-spectrum features with differences up to 15% while the variation between values for the remaining feature set won't exceed 10%.

5.3 Conclusion

Taking all together the question arises which algorithms in which combination produce the best results. As for VAD the highest possible recall is assumed to be the most important compared to precision or f-score, the combination of event detection followed by classification wins. Here, the highest recall on average with the lowest variation has been achieved.

Comparing the features, the reduced Zhang feature set holds the advantage of a low computational cost and a slightly better recall on the cost of a noticeable lower precision compared to the full feature set. Even with this drawback, it might be the better choice in the competing final algorithm.

At last, the choice of a channel set depends on several factors. The highest recall is reached by three adaptive approaches as well as the constant 14 channel set. The latter outperforms the remaining possibilities in terms of the overall performance b reaching the highest f-score as well as the lowest standard deviation. The major disadvantage in using this set is the number of channels which is almost three times the number of channels of the remaining sets. From these, the Maximum-Variance-Channel-Selectionwins the competition by focussing on the recall.

Deciding about the feature and channel set in this case is quite subtle as differences are minimal. Depending on additional approaches to improve the results, the need of reconsidering the decision about what to use might occur.

5.4 Future Work

The results being achieved can be seen as a good starting point for further investigation. Taking the setting which came closest to the constraints finding the major weaknesses of each component should improve the robustness of the algorithm.

As already mentioned, studying the influences of block-size or - more precisely - the influence of missing starts/stops of a command or phrase on standard deviation of all results can increase robustness for example when applying a hangover scheme. Depending on its characteristic, recall can be almost arbitrarily increased but on the cost of precision.

In addition attempts have been made in finding a more precise interpretation when miss-detections occur by simply dropping parts of the simulations where certain noise types occurred. While the result was expected to improve when suppressing the influence of certain noise types, it quickly proved to be much more complex.

Bibliography

[1] M. Matassoni, R. F. Astudillo, A. Katsamanis, and M. Ravanelli, "The dirha-grid corpus: baseline and tools for multi-room distant speech recognition using distributed microphones," in *Interspeech*, 2014.

[2] P. C. Khoa, "Noise Robust Voice Activity Detection," Master Thesis, Nanyang Technological University, 2012.

[3] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with multi-channel linear prediction and sparse priors for the desired signal," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, May 2014, pp. 23–26.

[4] J. Rosca, R. Balan, and C. Beaugeant, "Multi-channel psychoacoustically motivated speech enhancement," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 3, July 2003, pp. III–217–20 vol.3.

[5] S. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 1, pp. 84–91, Jan 2005.

[6] J. Morales-Cordovilla, M. Hagmuller, H. Pessentheiner, and G. Kubin, "Distant speech recognition in reverberant noisy conditions employing a microphone array," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, Sept 2014, pp. 2380–2384.

[7] M. Taghizadeh, P. Garner, H. Bourlard, H. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, May 2011, pp. 92–97.

[8] S. Stenzel and J. Freudenberger, "Time-frequency dependent multichannel voice activity detection," in *Speech Communication; 11. ITG Symposium; Proceedings of*, Sept 2014, pp. 1–4.

[9] P. Milhorat, D. Istrate, J. Boudy, and G. Chollet, "Hands-free speech-sound interactions at home," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 1678–1682.

[10] M. Vacher, B. Lecouteux, F. Aman, S. Rossato, and F. Portet, "Recognition of Distress Calls in Distant Speech Setting: a Preliminary Experiment in a Smart Home," in *6th Workshop on Speech and Language Processing for Assistive Technologies*, ser. 6th Workshop on Speech and Language Processing for Assistive Technologies. Dresden, Germany: SIG-SLPAT, Sep. 2015, pp. 1–7. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01179930

[11] N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," *Consumer Electronics, IEEE Transactions on*, vol. 57, no. 1, pp. 196–202, February 2011.

[12] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 697–710, April 2013.

[13] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[14] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission.* Wiley Publishing, 2008.

[15] A. Benyassine, E. Shlomot, H. yu 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," *Communications Magazine, IEEE*, vol. 35, no. 9, pp. 64–73, Sep 1997.

[16] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[17] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing [exploratory dsp]," *Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 145–154, Jan 2011.

[18] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, Mar 1992, pp. 121–124 vol.1.

[19] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J Acoust Soc Am*, vol. 126, no. 3, pp. 1486–1494, Sep 2009, 052909JAS[PII]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2757424/

[20] T. Gerkmann and R. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[21] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Commun.*, vol. 57, pp. 170–180, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2013.09.015