ADVANCEMENTS IN AUTOMATIC SPEAKER AND SPEECH PROCESSING IN CO-CHANNEL SPEECH

by

Navid Shokouhi

APPROVED BY SUPERVISORY COMMITTEE:

____
John H. L. Hansen, Chair

____
Carlos Busso

____
Issa Panahi

____
P. K. Rajasakeran

This thesis class file

is dedicated to ...,

who ...

ADVANCEMENTS IN AUTOMATIC SPEAKER AND SPEECH PROCESSING IN CO-CHANNEL SPEECH

by

NAVID SHOKOUHI, BS

DISSERTATION

Presented to the Faculty of
The University of Texas at Dallas
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY IN
ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

November 2016

ACKNOWLEDGMENTS

November 2016

PREFACE

This dissertation was produced in accordance with guidelines which permit the inclusion as part of the dissertation the text of an original paper or papers submitted for publication. The dissertation must still conform to all other requirements explained in the “Guide for the Preparation of Master’s Theses and Doctoral Dissertations at The University of Texas at Dallas.” It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts which provide logical bridges between different manuscripts are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the student’s contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the dissertation attest to the accuracy of this statement.

ADVANCEMENTS IN AUTOMATIC SPEAKER AND SPEECH PROCESSING IN CO-CHANNEL SPEECH

Publication No. ___________________

Navid Shokouhi, PhD
The University of Texas at Dallas, 2016

Supervising Professor: John H. L. Hansen

350 word Abstract.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Co-channel speech refers to single-channel audio signals that contain more than one speaker. A wide range of terms have been used to describe various aspects of co-channel speech, which we will clarify throughout this chapter. We consider both conversational speech and artificially mixed streams as co-channel. All signals treated in this study are single-channel recordings. Of all such data, a subset may have more than one "active" speaker at the same time, i.e. multi-speaker segments, which we label as "overlapped speech". Overlapped regions are segments of a co-channel signal where both speakers are simultaneously active. This categorization is summarized in Fig. **??**.

The specifics of recording conditions are overlooked in this study. For example, information that relates to each speakers distance from the microphone and room acoustics. This is intentional, since most of the difficulty in dealing with co-channel speech arises from it being single-channel, which implies limited access to spatial information as well as other sources of meta-data.

There are different ways of categorizing co-channel data in terms of how it is generated. This study considers a semantic classification that focuses on the speakers' interactions with one-another. We divide co-channel data into two subgroups:

- conversational co-channel speech
- co-channel speech with independent parties

Conversational co-channel speech refers to recordings in which speakers acknowledge other parties in the recording and engage in dialog. Alterations of speech production by each speaker are an important artifact of conversational co-channel speech, which are the result of conscious and unconscious reactions of the foreground speaker and interferer(s) during speech especially at or around overlaps. Examples of such alterations are raised pitch and energy level [2]. The ESPN show First Take is filled with arguments between the shows two regular sports commentators Stephen A. Smith and Skip Bayless. First Take is a perfect example of an exaggerated version of the above-mentioned changes in speech production. These changes are problematic in automatic speech applications and are considered a type of distortion. Consequently, the treatment will be directed towards applications that suffer the most from such alterations, predominantly speech recognition.

Co-channel data with independent parties, are examples of co-channel data where the speakers do not interact with each other. Cross-talk between separate channels is considered a source of such co-channel speech. The main characteristic of such data is that speakers are not aware of each other and therefore do not pertain to normal conversational mannerisms. That includes following turn-takings rules and limiting overlapped speech. Artificially generated co-channel data (mixing independent channels) is another example of co-channel speech with independent parties. A considerable portion of this study will focus on this type of artificially generated co-channel data to analyze performance of overlap detection and also speaker recognition. We rely on this data since it provides the flexibility of controlling the amount of overlapped speech. As we will show in this chapter, conversational co-channel speech does not necessarily contain sufficient overlapped data for some of our experiments.

1.1 Background

1.2 Approach

The goal of this thesis is to provide tangible solutions to problems caused by co-channel speech in automatic speech technology. We argue that part of these issues are caused by overlapped speech (direct speech interference), which plays a significant role in making co-channel speech a difficult problem. The presence of overlapped speech can be detrimental to speaker diarization and speech recognition systems. There is no clear and unique way of labeling or transcribing overlapped segments. In speaker diarization, it becomes difficult to assess system performance at overlaps. The same goes for speech recognition. Aside from determining which is the "primary" speaker, recognizing speech at overlaps is more difficult due to interference from other speakers. For this and other reasons detailed in the next chapter, the first portion of this study is devoted to overlapped speech detection. Our approach to overlap detection will be to focus on developing signal processing techniques to detect and separate overlap from single-speaker speech.

Although overlap is considered an important aspect of co-channel speech, in many conversational speaker recognition data, the amount of overlap in co-channel speech could be considered negligible for some tasks. For example, in the case of speaker recognition, we are usually interested in a specific speaker, but the audio file may contain other speakers as well. The standard approach in dealing with this problem is speaker diarization. Speaker diarization is considered a "signal-level" solution to this problem. In addition to devoting one chapter to speaker diarization, a novel approach is presented with the intention of bypassing the use of speaker diarization in the aforementioned scenario while preserving speaker recognition performance. This approach is to remove unwanted speaker-dependent information from latent variable subspaces generated from audio files. We refer to such solutions as "subspace level". We propose using two techniques to remove interfering speakers from co-channel data. This is performed in two different ways:

1. Remove interfering speakers in the feature subspace level: i-vector subspace factorization.

2. Remove interfering speakers in the signal level: speaker diarization.

Each of these methods will be described in a separate chapter. Speaker diarization, will attempt to recognize and group speech that belongs to the same speaker in a co-channel audio stream. While subspace factorization maps speaker-dependent models to a subspace that will only contain parameters identifying the speaker of interest (aka primary speaker).

CHAPTER 2

FRONT-END SIGNAL PROCESSING

2.1 Introduction

Overlapped speech constitutes a significant amount of research in co-channel speech. To such an extant that in many cases co-channel and overlap are considered synonymous. This chapter is focuses on proposing signal processing techniques to recognize and instances of co-channel data where speakers overlap; overlapped speech detection. Single-channel recordings from meetings or conversations are examples during which speakers may overlap. Separating the resulting mixture becomes especially difficult when one does not assume prior knowledge about speaker identities or speech content.

Most studies on overlapped speech have focused on separating the target or suppressing interfering speech (**?**). Often to de-noise and thereby improve the performance of automatic speech applications (**???**) (primarily speech recognition). However, over the past decade, due to vast developments in recognition systems such as speaker identification (SID) and diarization, a growing trend of detecting overlapped regions has been observed. In speaker identification, the presence of interfering speech in conversational speech styles not only reduces the effectiveness of trained speaker models but also increases the uncertainty in scoring test files with overlapped regions (**?**). Removing overlapped segments increases model reliabilities which consequently improves recognition (**?**). State-of-the-art speaker diarization systems are also currently at a stage where one of the main sources of error is the presence of overlapped speech (**??**). One of the main reasons overlaps become a source of confusion in speaker diarization systems is that there is no basis for selecting ground-truth in overlapped

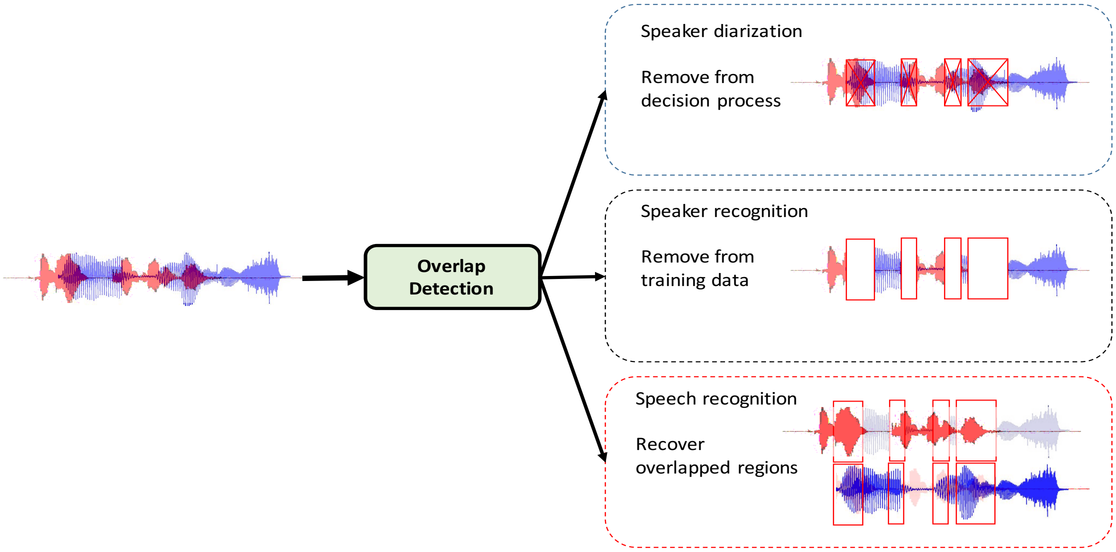

Overlap Detection
Speaker diarization
Remove from decision process
Speaker recognition
Remove from training data
Speech recognition
Recover overlapped regions

Figure 2.1. Applications of overlap detection. Top: In speaker diarization, removing ignoring overlapped regions provides a more fair assessment of diarization performance. Middle: Removing overlaps from in speaker recognition increases the reliability of training data. Bottom: Overlap detection results can be used as an initial step to recover overlapped regions.

regions. This makes evaluating speaker diarization systems more challenging[1]. Fortunately, for applications such as speaker identification and diarization it is rarely necessary to separate the target from interfering speaker in overlapped speech, since preserving speech content is not a priority. One can improve system performance by detecting and excluding overlapped segments for both SID and diarization.

Replacing interferer suppression and target separation with overlapped speech detection, is sometimes called "usable speech detection"[2] (?). An overlapped speech detection system can be used in any of the aforementioned tasks as a data purification step or a signal processing front-end.

[1] Future chapters will describe co-channel speech data in speaker diarization in more detail.

[2] In order to avoid any confusion between this study and the assumptions made in (?), we use the more general term overlapped speech detection.

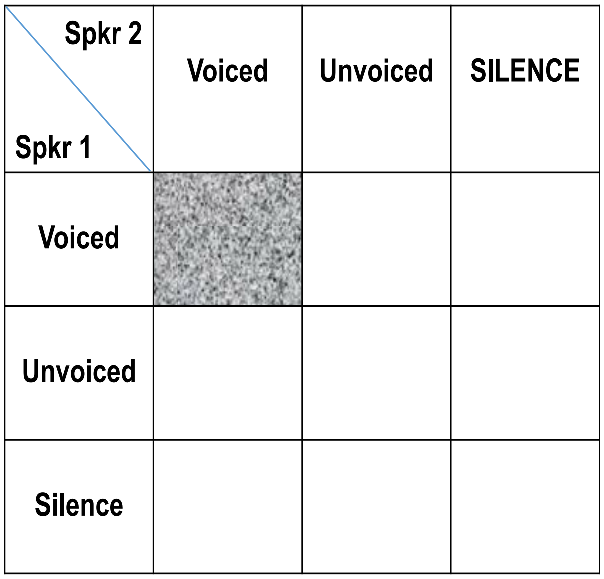
Signal processing front-end solutions proposed to dealing with co-channel speech in this study focus solely on overlapped speech detection. Figure 2.7 summarizes the utilizing overlap detection in speech processing technology.

Detecting overlapped segments has previously been considered in tasks such as speaker identification (SID) and speaker diarization (??). In such problems, the presence of a secondary speaker either decreases model reliability (in training), or introduces confusion in the decision-making process by distorting test files. In cases where speech is of contextual value, such as in speech recognition, the traditional approach is to somehow enhance the target speaker or weaken interfering speakers. Unfortunately, removing unwanted speech at overlaps is not straightforward and requires prior knowledge of one or both speakers. Such difficulties further motivate the use of overlapped speech detection. Detecting overlaps is computationally advantageous when one has the luxury of neglecting overlapped data (?). As is the case for speaker recognition and diarization (?). This study proposes methods for overlap detection in monophonic speech. By detecting overlapped speech, we are able to remove them from the training and decision-making process.

2.2 Background

Traditionally, studies have used spectral harmonicity as a key factor in detecting overlapped speech (??). This approach is motivated by the fact that two fundamental frequencies exist in most instances of overlapped speech which disarranges the harmonic structure observed in single-speaker speech. As a side-note here, we point out that most focus in overlapped speech has been at regions where both speakers produce "voiced" speech. In (?) a classification of different types of segments in co-channel speech is presented. Figure 2.2 is adopted from (?).

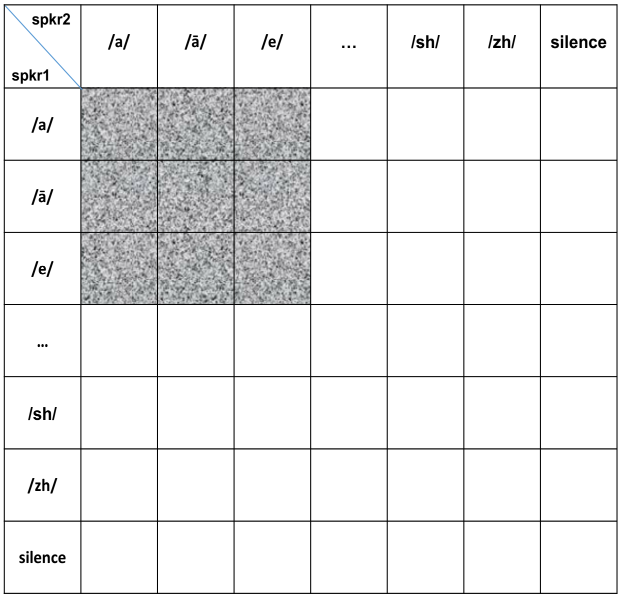
Most of our focus will be on the voiced-voiced cell. Merely because detecting other regions becomes far more difficult. A more detailed classification of overlapped regions is presented in (?), where a grid containing all phones is used to rank-order overlapped segments in terms



Spkr 2 Spkr 1	Voiced	Unvoiced	SILENCE
Voiced	▨		
Unvoiced			
Silence			

Figure 2.2. Classification of different segments in a co-channel file. In overlap detection we are interected in the voiced-voiced (shaded) region.

of difficulty. The analysis in (?) expands Fig. 2.2 as shown in Fig. 2.2. General consensus is to focus on detecting voiced-voiced overlap detection, which from now on we will refer to as overlap detection.



spkr2 spkr1	/a/	/ā/	/e/	...	/sh/	/zh/	silence
/a/	▨	▨	▨				
/ā/	▨	▨	▨				
/e/	▨	▨	▨				
...							
/sh/							
/zh/							
silence							

Figure 2.3. phone-based expansion of overlapped segments in Fig. 2.2.

Concentrating on voiced speech allows us to use more discriminating harmonic structures to detect overlaps. In (?), the peak-to-valley ratios in frame-based spectral autocorrelations are introduced as a discriminating feature for overlapped speech detection through the same assumption. Spectral flatness measure, the ratio of geometric to arithmetic means calculated from spectral bins in a speech frame, has also been used as a measure to capture harmonicity and has been used to detect the presence of overlapped speech (?). Another related characteristic is observed when monitoring fundamental frequencies along time. Adjacent pitch period comparison (APPC) presented in (?) uses the temporal variation of estimated "pitch" periods as a measure to detect "usable" speech with the assumption that temporal variations of adjacent pitch periods are significantly higher in overlap. A multi-pitch tracking algorithm proposed in (?) was used in (?) to estimate coexisting fundamental frequencies in the presence of multiple speakers. Regions where more than one fundamental frequency is estimated are labeled as overlap. The multi-pitch tracking technique described in (?), decomposes speech into sub-bands and pitch estimation is only performed on reliable sub-bands.

A slightly different, yet fundamentally similar, approach to distinguish overlapped speech is to use speech kurtosis which measures higher order moments of the signal statistics (?).

A number of studies have considered investigating spectral characteristics at formant frequency locations when dealing with overlapped speech. Giuliani et al. use a filter-based approach to improve speech recognition rates for different instances of meeting conditions by adding a detection step that separates double-speaker speech from single-speaker audio (?). This was accomplished by cascading two-layer sub-band filters to capture formant characteristics. Formant frequency information was obtained by filtering the signal at sub-bands with center frequencies and bandwidths corresponding to nominal F_1, F_2, and F_3 values for all English vowels. One of the reasons Formant-based overlapped speech analysis has received less attention is the difficulties in modeling pole interactions at overlapped regions, which is an issue for linear predictive modeling and other commonly used formant tracking techniques. Characterizing pole interactions using standard LP models easily becomes intractable in the presence of more than one source. Add to this complication, the unknown speaker locations with respect to each other and the microphone. As a result, we focus our attention to nonlinear speech models, some of what have proven more successful in the scenarios described above.

Nonlinear speech models, including the AM-FM speech model (**?**) have been used in previous studies to model speech resonances without any specific requirements for the source signal. These energy operators have also been used to deal with signals with more than one source (**?**), aka co-channels signals [3]. Maragos et al. use higher order energy operators to develop an algorithm that simultaneously demodulates the components of a co-channel mixture in AM-FM modulated signals (**?**). Litvina et al. separate speech from music using the Teager energy operator (TEO) separation algorithm (**?**) (**?**), where they used the extracted components to design a time-varying filter and suppress the interfering signal. Similar multicomponent signal decomposition techniques have been addressed using energy operators to separate narrow-band signals (**???**).

Our goal is to incorporate sub-band analysis to design a technique suitable for **overlapped speech detection**. Two algorithms are proposed that incorporate sub-band analysis for overlap detection.

- using TEO methods on narrow-band components to detect speech harmonics.
- apply cosine functions across sub-band outputs to magnify the presence of multiple harmonics.

[3] Co-channel is a more general terminology used to described multi-component signals. In the case of speech, co-channel speech may refer to any single-channel recording that contains speech from multiple speakers, regardless of whether there is overlap.

2.3 Pyknograms

We propose a novel approach for overlapped speech detection based on an enhanced spectrogram. These spectrograms, called Pyknograms, were first introduced by Potamianos and Maragos in (**??**) and are calculated by applying multi-band demodulation in the AM-FM speech model framework (**?**)[4]. Pyknograms provide a more prominent representation of harmonic trajectories, which we propose to use as a means to detect the presence of interfering speech.

2.3.1 Energy Operators and the AM-FM speech model

In Pyknograms (**?**), the harmonic structure of speech is enhanced by decomposing spectral sub-bands into amplitude and frequency components. This multi-band analysis uses the AM-FM speech model (**?**) to decompose sub-bands and thereby calculate corresponding instantaneous frequencies and bandwidths: (2.3), (2.2). Pyknogram extraction locates dominant peaks in the spectrogram from instantaneous frequencies. To extract Pyknograms, the speech signal is initially passed through a filter-bank (we have modified the algorithm to use logarithmically spaced Gamma-tone filters, while (**?**) uses linearly-spaced Gabor filters). Filter-bank outputs are then decomposed into amplitude and frequency components using the discrete energy separation algorithm (DESA-1) (**?**), where the frequency and amplitude components of a given sub-band, $x(n)$, are calculated using the discrete energy operator,

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1), \tag{2.1}$$

$\Psi[x(n)]$ is energy operator used to estimate amplitudes and instantaneous frequencies, as shown in Fig. **??**.

[4] The authors in (**?**) used the term "Pyknogram" which stems from the Greek word "pykno" meaning dense. Pyknograms represent highly resonating regions in time-frequency plots as populated scatter plots, hence the term density.

$$f(n) = \frac{1}{2\pi} \arccos\Big(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\Big), \tag{2.2}$$

$$|a(n)| = \sqrt{\frac{\Psi[x(n)]}{\sin^2(2\pi f(n))}}. \tag{2.3}$$

2.3.2 Pyknogram Extraction

Pyknograms are estimated from DESA-1 outputs. The weighted average of instantaneous frequency components (see (2.4)) is used to derive a short-time estimate of the dominant frequency in each sub-band over time-frame units (typically 25 msec) (**?**). Frequencies are weighted using the estimated signal power ($|a(n)|^2$). The average frequency computed for each frame/sub-band (time-frequency unit) can be viewed as the 1^{st}-order moment of instantaneous frequencies.

$$F_w(t) = \frac{\sum_t^{t+T} f(n)a^2(n)}{\sum_t^{t+T} a^2(n)}, \tag{2.4}$$

The algorithm also provides a means to estimate weighted bandwidths for each resonance, (2.5). What we refer to here as bandwidths are essentially 2^{nd}-order frequency moments.

$$B_w(t) = \sqrt{\frac{\sum_t^{t+T} (\dot{a}(n)/2\pi)^2 + (f(n) - F_w)^2 a^2(n)}{\sum_t^{t+T} a^2(n)}}, \tag{2.5}$$

where $f(n)$ and $a(n)$ are instantaneous frequency and amplitude values from (2.2) and (2.3). In (2.4), the instantaneous frequencies are averaged over the t^{th} frame using squared instantaneous amplitudes as weights. T in (2.2) is the number of samples per frame, from $n = t$ to $n = t + T$. $\dot{a}(n)$ is the first difference of $a(n)$ (i.e., $a(n) - a(n-1)$). The per-frame values of F_w provide initial estimates of spectrogram peaks. This results in a time-frequency t-f representation of the overall signal, where time units correspond to frames and frequency units to filter-bank sub-band indexes.

In (?), the bandwidth values defined in (2.5) are used for analysis purposes. Here, we use them in overlap detection systems to determine the reliabilitiy of $t-f$ units. Our assumption is that large Pyknogram bandwidths correspond to higher uncertainty in frequency estimates. We investigate this in following sections by adding an uncertainty term to our frequency estimate proportional to the estimated bandwidth:

$$\tilde{F}_w(t) = F_w(t) + \epsilon_t, \tag{2.6}$$

where

$$\epsilon_t \sim \mathcal{N}(0, B_w(t)). \tag{2.7}$$

As a final step, dominant harmonic peaks are selected by comparing the average frequency estimates with filter-bank center frequencies. According to (?), points at which filter-bank center frequencies coincide with the weighted frequency estimates from (2.4) are more reliable in estimating spectrogram peaks. The assumption being that frequency estimates are more accurate when aligned with a filter in the filter-bank. This defines the condition through which initial F_w values are tested to detect whether they correspond to prominent peaks. At frame t:

$$F_w(c) = c \quad \Longleftrightarrow \quad \{c \in peaks\} \tag{2.8}$$

where c are the filter-bank center frequencies. Note that center frequencies are distributed in a logarithmic scale. Another peak selection condition (as shown in Fig. 2.7) is to limit the relative variance of selected frequencies with respect to center frequencies.

$$\frac{\partial F_w(c)}{\partial c} < thr \tag{2.9}$$

This condition limits non-harmonic anomalies that break the patterns in regular speech trajectories. Since such patterns are frequently observed in overlapped data, we omit this restriction from the peak-picking step.

One of the advantages of the peak-picking constraint in (2.9) is the quantization of spectrograms onto filter-bank center frequencies. This allows the mapping of all signals onto a unified space defined by the filter-bank, which enables reliable comparison within the time-frequency space.

Using an energy operator based approach helps avoid assumptions on the number of speakers in the signal. AM-FM decomposition is suitable since it relies on signal resonances and does not restrict signals to a specific structure or number of speakers (as opposed to models such as linear prediction). The final time-frequency representation is called a Pyknogram and is denoted $S_{pyk}(t, f)$ as a function of time (t) and frequency (f). Using Pyknograms, we would like to investigate overlap detection methods.

Discontinuities in the Pyknogram layout is an indication of interfering speech. An analogy for speech harmonic patterns are skiing tracks left behind on a snowy surface. In the single-speaker case, the patterns leave parallel tracks that progress relatively slowly over time and correspond to fundamental frequency harmonic tracks. In the presence of an interfering speaker, these patterns are distorted by similar but intersecting tracks, which adds sudden jumps along the time axis (as shown in Fig. 2.9). Since the majority of speakers are only capable of producing one fundamental frequency at each time instance, it is expected that the harmonic tracks should be consistent across time. This keeps harmonics parallel over short time intervals. The presence of a second speaker creates harmonic tracks that in general do not follow the same patterns, hence discontinuities are observed along time in Pyknograms. We use variations across adjacent frames as our measure of overlapped speech.

2.3.3 Unsupervised overlap detection Pyknograms

The average Euclidean distance between consecutive frames across all frequencies can be used to detect sudden jumps in Pyknograms along time. Similar to the technique used for spectral flux estimation (**?**). The distance function, D_{ovl}, at frame t is computed as the 2-*norm* distance between consecutive Pyknogram frames, $S_{pyk}(t, f)$ and $S_{pyk}(t - 1, f)$.

$$D_{ovl}(t) = \sqrt{\sum_{f} \left(\left(S_{pyk}(t,f) - S_{pyk}(t-1,f) \right)^2 \right)} \tag{2.10}$$

where t and f respectively correspond to the frame index (time) and filterbank bin (frequency).

Overlapped segments are expected to have higher D_{ovl} values as compared to single-speaker speech. Figure 2.9 shows instances where sudden jumps are observed in the Pyknogram of an overlapped signal. The average value of these distances for all frames in a speech segment corresponds to the amount of overlapped regions (higher values are associated with greater overlap).

We evaluate the performance of our proposed detection metric on overlapped speech from the GRID database (**?**)(see Sect.2.3.4 for more details on GRID). A key factor that determines the difficulty of detecting the presence of overlapped speech is the signal to interference(SIR) value. Greater absolute SIR values correspond to regions where one of the speakers has lower impact on the signal energy. Therefore it is more difficult to detect the occurrence of overlap in signals as the SIR moves away from $0dB$. Notice we use absolute SIR, since in overlap *detection* there is no difference between target and interfering speakers.

Another important factor in detecting overlap is that the SIR value will change across different frames within a single file, which is due to the non-stationary nature of speech. This poses major restrictions on the effectiveness of overlap detection evaluation, since providing frame-based ground-truth becomes unrealistically difficult. One must therefore rely on ensemble measurements over complete speech files for which the average SIR is known. This notion is illustrated in Fig. 2.10, where D_{ovl} distributions (histograms) extracted on a per-frame basis are compared with ensemble D_{ovl} distributions associated with each file. The “scores” (D_{ovl} values) in Fig. 2.10 are pyknogram distances calculated using (2.10). The top figure (Fig. 2.10-a), shows the distribution of scores per *frame* (i.e. 25msec intervals) for overlapped (target) and clean (non-target/single-speaker) *files*. Figure 2.10-b shows the ensemble score distributions (average score over all frames in a file, which are typically 2 seconds long). The task in overlap detection is to separate the two classes in each plot (dark blue from light blue). As observed in these distributions, the per-frame classes are almost indistinguishable (Fig. 2.10-a), while in Fig. 2.10-b the classes show much better separation.

2.3.4 Evaluation

This section evaluates our proposed pyknogram-based overlap detection system in terms of *accuracy, robustness,* and *precision*. Evaluation tasks for each SIR category are in the form of standard binary classification problems, where target examples are from a collection of files with fixed SIR values and non-target files are clean (single-speaker) files. We measure system performance using detection equal error-rates (EER; where false-positive and false-negative errors are equal). EER values are presented in Fig. 2.12 for different SIRs. The expectation is that the detection algorithm should be consistent across a range of SIR values (i.e. robustness). As for precision, we are interested to know how short signals can be before overlap detection performance significantly drops (noting the observation in Fig. 2.10).

Bellow, a collection of overlap detection features are presented that have previously been used to detect overlapped regions (**???**). To the best of our knowledge, overlap detection results on this database have not been reported for any of the following features, therefore we rely on our own implementations.

Baseline features

- *Speech kurtosis*: Kurtosis has been reported as an effective measure to detect the presence of multiple speakers in overlapped signals by several studies (**???**). It has been shown that overlapped speech exhibits lower kurtosis compared to single-speaker speech (**?**). The kurtosis of a zero-mean random variable x is defined as:

$$k_x = \frac{E\{x^4\}}{(E\{x^2\})^2} \tag{2.11}$$

In this case x refers to speech samples in a given frame.

- *Spectral flatness measure (SFM)*: The ratio of geometric to arithmetic means of spectral magnitudes across frequency within each frame (?). For the i^{th} frame:

$$sfm_i = \frac{\frac{1}{N}\sum_{n=1}^{N} X(f_n)}{\sqrt[N]{\prod_{n=1}^{N} X(f_n)}} \tag{2.12}$$

where $X(f_n)$ corresponds to the magnitude spectrum at frequency f_n and N is the total number of frequency bins.

- *Spectral autocorrelation peak-valley ratio (SAPVR)*: described briefly in Sec. 2.1, this feature uses the dominance of peaks in the spectral autocorrelation in each frame as a measure to detect overlaps (?).

Data: Monaural Speech Separation Challenge

The data used in our controlled experiments is from the monaural speech separation and recognition challenge (a.k.a speech separation challenge (SSC)) (?). The objective there was to permit a large-scale comparison of techniques for the overlapped speech problem (?). Participants were asked to identify keywords in sentences spoken by a target talker when mixed into a single channel with a background talker speaking sentences of the same structure but with different content. The data used in SSC was obtained from the larger GRID corpus (?), which is a multi-talker audio-visual sentence corpus that supports computational-behavioral studies in speech perception. In our study, we only use the audio content which consists of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). The sentences are structured in the following format.

<command><color><preposition><letter><number><code>

For example, “lay white at X six now”.

The test and training set contain the same set of talkers. Seven overlapped sets are available, one clean and the rest composed of sentence pairs artificially summed at 6 signal-to-interference ratios (SIR) (+6, +3, 0, 3, 6, 9 dB). Since file durations are short (typically less than 5 seconds) and the utterances contain negligible pauses, it is reasonable to consider the average SIR values, provided for each file, a fair representation of the amount of overlap. This also allows the assumption that the entire signal is overlapped (see Fig. 2.11). We have down-sampled all files to 8kHz to match telephone recordings. Note that the experiments conducted in this study do not comply with the objectives of the speech separation challenge described in (**?**).

This corpus is isolated from variabilities other than overlapped speech, which makes it useful to study the effects of overlap. To the best of our knowledge, this dataset is the most organized publicly available corpus that contains large, as well controlled, amounts of overlapped speech (note that we are mostly interested in *overlapped speech* and not *co-channel speech* as defined and distinguished in the introduction). Among the corpus' other advantages is the fact that segments are short which makes the definition of a signal-to-interference ratio more appropriate. Had the signals been longer, say a few minutes long, the notion of a signal-to-interference ratio across the entire signal would have been less applicable, due to the non-stationary nature of speech.

number of speakers	18 (male) 16 (female)
average file duration	1.9 (sec)
noise	interfering speakers clean,+6, +3, 0, −3, −6, −9 dB
sampling rate	8 KHz

Table 2.1. Summary of data used for SID experiments

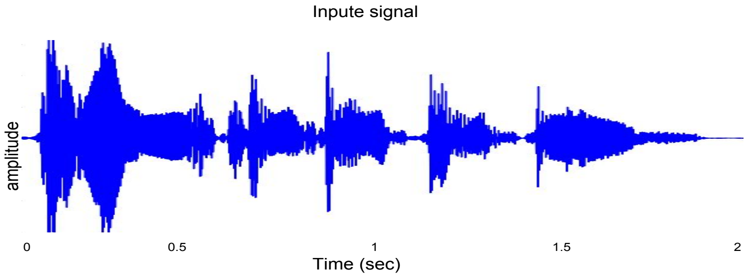
Overlapped speech detection vs. SIR (Robustness & Accuracy)

Here the performance of pyknogram-based overlap detection is compared with the three baseline algorithms across different SIR values. The goal is to monitor the chances in EER as SIR values increase. The target/non-target files used in this binary classification task are obtained from a pool of overlapped and clean files. In each task, overlapped files with the same SIR are used as target examples and the overlap detection score (or feature value) assigned to them is compared against the scores estimated for clean files to compute the binary classification EER. Figure 2.12 compares performances for the proposed and baseline systems across SIR values of $0, 3, 6$ and $9dB$.

Overlapped speech detection vs. segment length

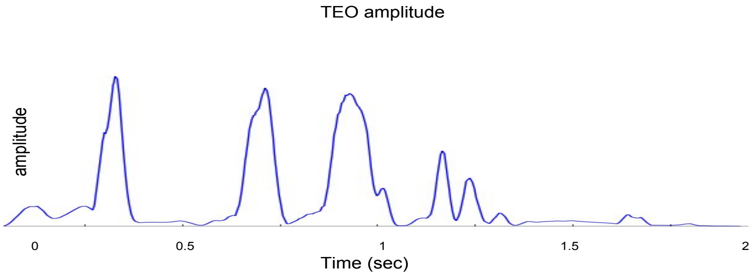
A main concern in dealing with overlapped regions is that overlap decisions are less reliable as segment lengths become shorter. This restricts algorithm precision in terms of the ability to detect overlap in a frame-based framework. Precision is most valuable in tasks such as speaker diarization in conversational speech, where overlap mostly occurs at speaker transitions in turn-takings. The goal of this phase is to evaluate system precision and compare pyknogram-based detection with baseline features. In other words, how short can overlap segments get before observing a significant drop in system performance. Once again, overlap detection performance is measured through the detection EER. Figure 2.13 shows the change in system performance as shorter duration segments are used to obtain overlap decisions.

2.4 Gammatone Sub-band Frequency Modulation Spectra



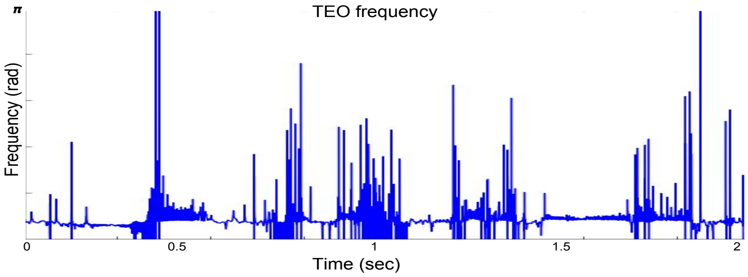
3.5in

Figure 2.4. Input signal.



3.5in

Figure 2.5. Outputs of DESA-1: Signal amplitude component estimated using TEO, Eq. (2.3).



3.5in

Figure 2.6. Outputs of DESA-1: Signal frequency component estimated using TEO, Eq. (2.2).

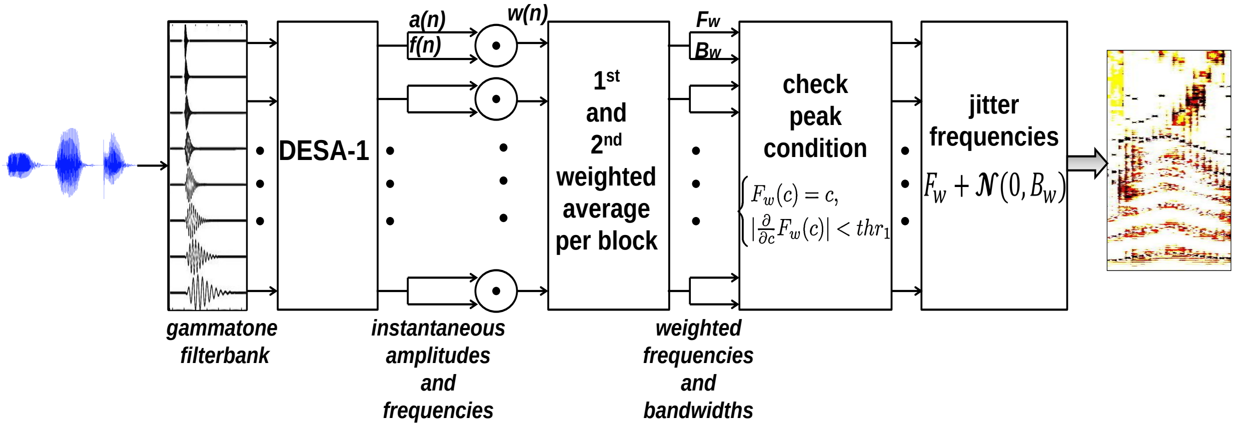

a(n)
f(n)
w(n)
DESA-1
1st and 2nd weighted average per block
Fw
Bw
check peak condition
$F_w(c) = c,$
$|\frac{\partial}{\partial c}F_w(c)| < thr_1$
jitter frequencies
$F_w + \mathcal{N}(0, B_w)$
gammatone filterbank
instantaneous amplitudes and frequencies
weighted frequencies and bandwidths

Figure 2.7. Pyknogram extraction block-diagram.

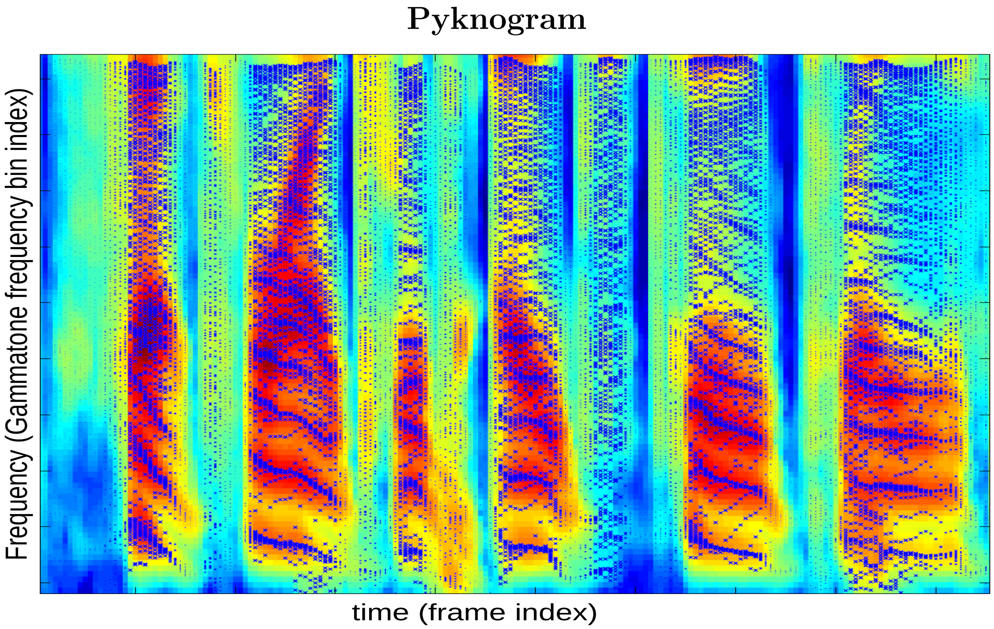


Figure 2.8. Pyknogram for a given speech signal. The spectrogram is plotted in the background for comparison. Pyknogram markers have been scaled by the amplitudes of corresponding t-f units. Frequencies are scaled to equivalent rectangular bandwidth (ERB) rate.

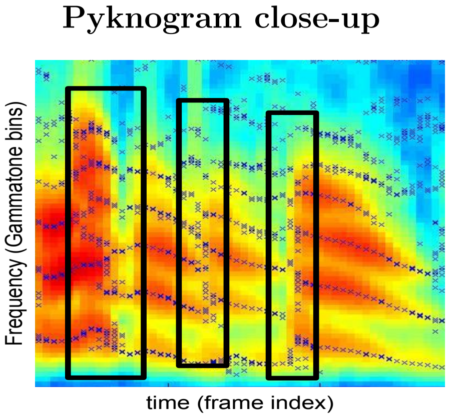


Figure 2.9. A closer look on Pyknograms for overlapped speech. The enclosed patches show discontinuities that occur in the presence of an interfering speaker.

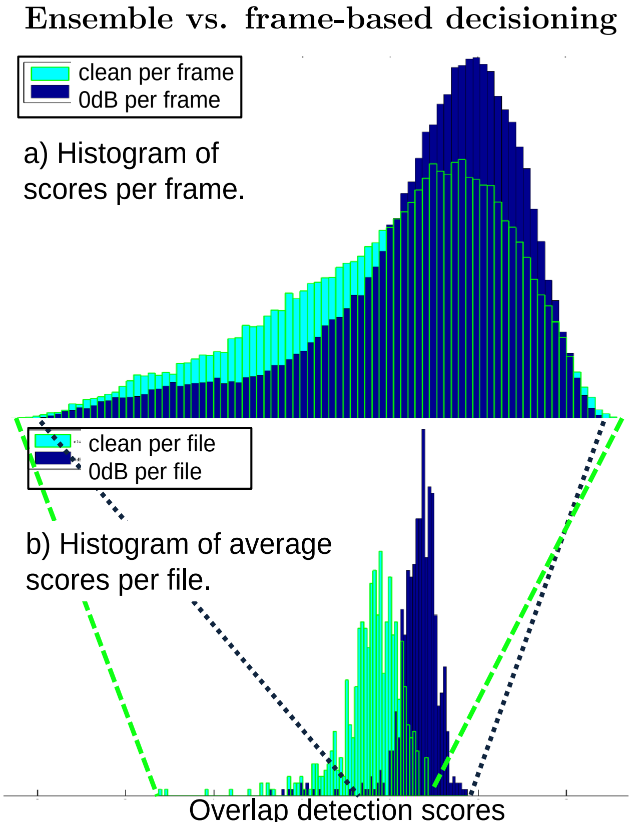


Figure 2.10. The effect of ensemble decisioning on distinguishability of overlapped regions. a) shows score per frame histograms and b) shows the histogram of ensemble scores. Using multiple frames to make a decision helps separate the distributions of clean and overlapped segments.

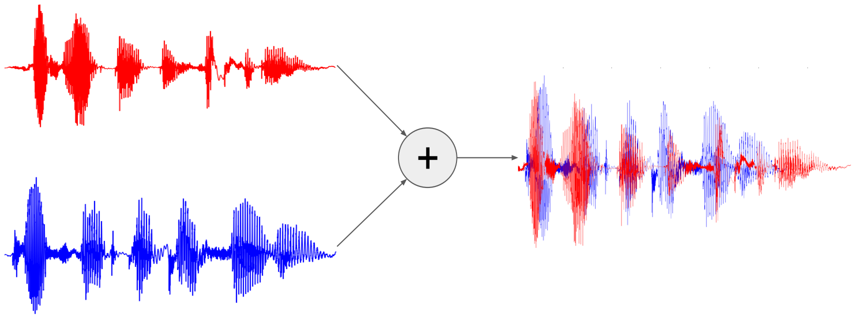


Figure 2.11. Example of the mixing process for a 0dB SIR overlapped signal. As shown on the right, it is fair to assume that overlap occurs throughout the signal.

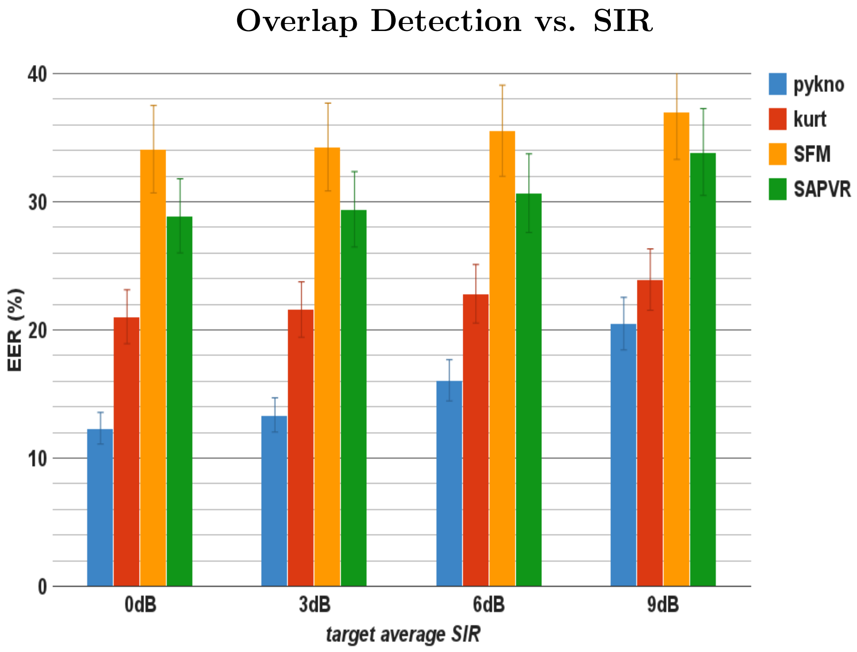


Figure 2.12. Overlap detection EER for different SIR values. The higher the SIR, the more difficult it is to detect the presence of interfering speakers.

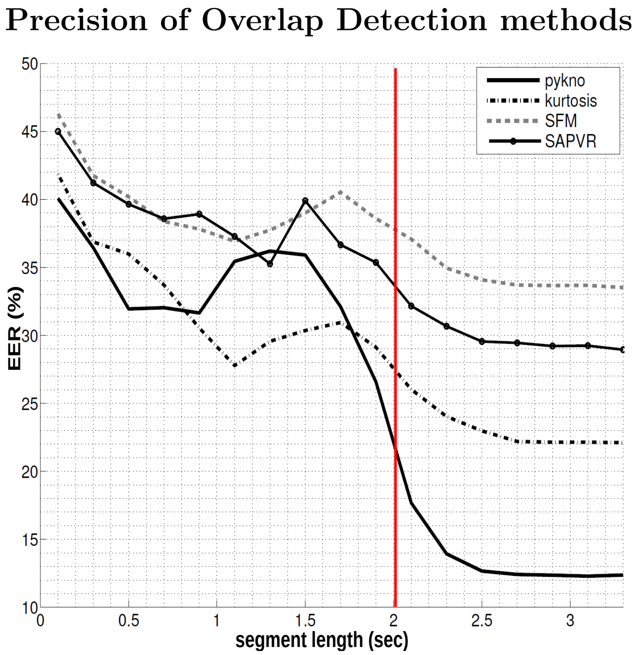


Figure 2.13. Overlap detection EER as a function of segment length. The plot shows that signal lengths should be at least 2 seconds for the algorithms to start reaching their best performance.

CHAPTER 3

CONCLUSION

APPENDIX

VITA

Navid Shokouhi