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

A wide range of terms have been used to describe various aspects of co-channel speech, which we will to 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, i.e. multi-speaker speech, which we label as "overlapped speech". Overlapped regions are segments of a co-channel signal where both speakers are active at the same time. 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 environment. This is intentional, since most of the difficulty in dealing with co-channel speech arises from it being single-channel, which implies our interest in limiting access to spatial information as well as other sources of meta-data.

Alterations in speech production are an important artifact of co-channel speech, which occur solely in conversations and are the result of conscious and/or unconscious reactions of the foreground speaker and interferer(s) during 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, our treatment will be directed towards applications that suffer the most from such alterations, predominantly speech recognition.

One can argue that co-channel speech has hardly received the attention it deserves compared to other speech related problems. An indication of this remark being the ambiguity in terminology used in the literature. Nevertheless, its presence becomes more noticeable with the increasing demand for automatic speech interfaces as well as the increased diversity where speech data is captured in naturalistic settings. Speech applications have pushed boundaries, forcing researchers to show more interest in tackling problems that have previously been only partially solved.

This dissertation aims to provide tangible solutions to the issue of co-channel speech for automatic speech applications. The first half focuses on speaker recognition/verification by developing an understanding of the effects of speaker interference on various stages of a given recognition system. Assessments are based on different categories of co-channel speech described in the previous section as well as the various standard realizations of a speaker recognition system, namely GMM-UBM (Gaussian mixture model - Universal background model) and i-vector/PLDA (classifying speaker identity vectors, aka i-vectors, using probabilistic linear discriminant analysis). Algorithms developed to improve speaker recognition in co-channel speech can also be used to improve diarization. Most of such approaches attempt to detect regions of overlapped speech. In current speaker diarization systems a significant amount of errors are caused by overlaps [3], hence the importance of overlap detection.

The second half will consider audio stream analysis for diarization with emphasis on speech recognition in co-channel speech. When it comes to speech recognition, aside from the direct effects of interfering speech, the foreground speaker is implicitly affected. This is projected in terms of changes in speech production and mannerisms. It is shown in [1] how an entire word is articulated differently enough that the acoustic model is no longer able to recognize individual phones. This is due to the presence of an interfering speaker, who may not necessarily be active during the time-lapse of the aforementioned word. These are instances in which co-channel speech affects speech production. Changes in speech

production have negligible effects on speaker recognition tasks, at least in standard tasks where there is access to considerably large amounts of data to model speaker identities and are therefore overlooked in the first portion of this study. One cannot say the same for speech recognition, since the acoustic models are meant to represent finer acoustic characteristics as compared to speaker recognition. This and many others are among the challenges of speech recognition in co-channel speech.

1.1 INTRODUCTION

Overlapped speech is referred to a monophonic audio recording in which at least two speakers are simultaneously active. 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 (**??**). Overlaps are a source of confusion in speaker diarization systems, since there is no basis for selecting ground-truth in overlapped regions. This makes evaluating speaker diarization systems more challenging. A reasonable work-around is to ignore overlapped regions when evaluating diarization performance. 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. This task, which replaces interferer suppression and target separation with overlapped speech detection, is sometimes called "usable speech detec-

tion"[1] (**?**). 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.

Traditionally, studies have used spectral harmonicity as a key component in detecting overlapped speech (**??**). This approach is motivated by the fact that two fundamental frequencies exist in many instances of overlapped speech which disarranges the harmonic structure observed in single-speaker speech. 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 conclusive summary of common features used to detect overlapped speech for improved speaker diarization is presented in (**?**).

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

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

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.

In this study, we use the AM-FM speech model along sub-bands(**?**) to model resonances. An energy operator based approach (**??**) is used to track harmonics in each sub-band (overlapped or single-speaker) and analyze the signal in those regions to determine whether speech is overlapped. Energy operators have previously been used to deal with signals with more than one source (**?**), aka co-channels signals [2]. 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*. The motivation for sub-band decomposition is to be able to use TEO methods on narrow-band components and detect speech harmonics.

---

[2]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.

1.2 OVERLAPPED SPEECH DETECTION

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 magnify the presence of a 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 a method for overlap detection in monophonic speech. By detecting overlapped speech, we are able to remove them from the training and decision-making process.

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 (**?**)[3]. Pyknograms provide a more prominent representation of harmonic trajectories, which we propose to use as a means to detect the presence of interfering speech.

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: (1.3), (1.2). Pyknogram extraction locates dominant peaks in the spectrogram from instantaneous frequencies. To extract Pyknograms, the

[3]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.

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,

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1), \tag{1.1}$$

$\Psi[x(n)]$ is energy operator used to estimate amplitudes and instantaneous frequencies, as shown in Fig. 1.2.

$$f(n) = \frac{1}{2\pi} \arccos\Big(1 - \frac{\Psi[x(n) - x(n-1)]}{2\Psi[x(n)]}\Big), \tag{1.2}$$

$$|a(n)| = \sqrt{\frac{\Psi[x(n)]}{\sin^2(2\pi f(n))}}. \tag{1.3}$$

The weighted average of instantaneous frequency components (see (1.4)) is used to derive a short-time estimate of the dominant frequency in each sub-band over time-frame (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.

$$F_w(t) = \frac{\sum_t^{t+T} f(n)a^2(n)}{\sum_t^{t+T} a^2(n)}, \tag{1.4}$$

The algorithm also provides a means to estimate weighted bandwidths for each resonance, (1.5). What we refer to here as bandwidths are essentially 2^{nd}-order frequency moments.

$$B_w(t) = \sqrt{\frac{\sum_t^{t+T} (\dot{a}(n)/2\pi)^2 + (f(n) - F_w)^2 a^2(n)}{\sum_t^{t+T} a^2(n)}}, \tag{1.5}$$

where $f(n)$ and $a(n)$ are instantaneous frequency and amplitude values from (1.2) and (1.3). In (1.4), the instantaneous frequencies are averaged over the t^{th} frame using squared instantaneous amplitudes as weights. T in (1.2) 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 (1.5) 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:

$$\tilde{F}_w(t) = F_w(t) + \epsilon_t, \tag{1.6}$$

where

$$\epsilon_t \sim \mathcal{N}(0, B_w(t)). \tag{1.7}$$

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 (1.4) 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:

$$F_w(c) = c \quad \Longleftrightarrow \quad \{c \in peaks\} \tag{1.8}$$

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. 1.3) is to limit the relative variance of selected frequencies with respect to center frequencies.

$$\frac{\partial F_w(c)}{\partial c} < thr \tag{1.9}$$

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 (1.9) 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. 1.5). 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.

1.2.1 Unsupervised overlap detection

The average Euclidean distance between consecutive frames across all frequencies can be used to detect sudden jumps in Pyknograms along time. Much like 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 $S_{pyk}(t-1, f)$.

$$D_{ovl}(t) = \sqrt{\sum_{f} \left(\left(S_{pyk}(t, f) - S_{pyk}(t-1, f) \right)^2 \right)} \tag{1.10}$$

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 1.5 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.1.3 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. 1.6, 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. 1.6 are pyknogram distances calculated using (1.10). The top figure (Fig. 1.6-a), shows the distribution of scores per *frame* (i.e. 25msec intervals) for overlapped (target) and clean (non-target/single-speaker) *files*. Figure 1.6-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. 1.6-a), while in Fig. 1.6-b the classes show much better separation.

1.3 Experiments

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. 1.8 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. 1.6).

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.

1.3.1 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:

$$k_x = \frac{E\{x^4\}}{(E\{x^2\})^2} \tag{1.11}$$

  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:

$$sfm_i = \frac{\frac{1}{N}\sum_{n=1}^{N} X(f_n)}{\sqrt[N]{\prod_{n=1}^{N} X(f_n)}} \tag{1.12}$$

  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. 1.1, this feature uses the dominance of peaks in the spectral autocorrelation in each frame as a measure to detect overlaps (**?**).

1.3.2 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. 1.7). 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 1.1. 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 1.8 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 1.9 shows the change in system performance as shorter duration segments are used to obtain overlap decisions.

1.4 CASE STUDY: SPEAKER VERIFICATION IN OVERLAPPED SPEECH SIGNALS

Overlapped speech is a common phenomenon in audio recordings that are used in speaker identification (SID) tasks. In this section, in order to show the detrimental effects of adding overlapped data to speaker verification, we present a case study of speaker recognition on data from the monaural speech separation challenge (**?**). Since most speaker verification applications are focused on spontaneous (as opposed to text-dependent) speech, a large portion of the data are recorded from telephone or face-to-face conversations, which are prone to overlap. Examples of overlapped speech vary from instances as short as back-channeling (such as filled pauses, "aha") in a regular conversation to intentional long duration overlaps used to hold the ground in arguments, which clearly has a more substantial impact on verification accuracy. In (**?**), Shriberg et al. provide an analysis of the amount of overlapped speech in Switchboard and other corpora comprised of conversational speech. Based on

the criteria used in their work (derived for automatic speech recognition purposes), 12% of words are considered overlapped in Switchboard, contributing to a large portion of the database. The frequency of overlap, however, is merely one of the factors contributing to speaker verification performance. For example, here we show that placing overlaps in train vs. test data also plays a significant role in determining system performance.

The SID experiments use 12 dimensional MFCC features (13 excluding the 0^{th} coefficient) plus Δ and $\Delta\Delta$, which adds to a total of 36 dimensional features. 512 mixtures were used to form the Universal background model (UBM). Each speaker's Gaussian mixture model (GMM) was obtained through MAP adaptation of the means.

1.4.1 Overlaps in test data

As a comparison benchmark, we first evaluate SID performance under clean train and test conditions on the SSC data. Gaussian mixture models (GMM) are adapted from a Universal back model (UBM) trained on TIMIT files (**?**). For each model speaker, there are 500 utterances in SSC, which are all used in the training process. Test files are available in all SIR conditions. As expected, lower SIR values correspond to higher equal error rates. The presence of a secondary speaker, clearly causes confusion in the score distribution, leading to less separability between target and imposter trials. SID performance under clean test files and those with average SIR ranging in $+6, +3, 0, -3, -6, -9dB$ are provided in Fig. 1.10.

It is worth mentioning that the authors were tempted to compare these results with stationary noise experiments. However, contrary to our expectations, we observed that performances were better in the overlapped condition when compared to white Gaussian noise and speech-shaped noise interference, even for negative SIR values. We find this to be a misunderstanding caused by comparing stationary and non-stationary noise through the same measurement procedure, which is the SIR (or SNR). For a given target speech file, adding a certain amount of stationary noise will affect all frames, whereas in the case of non-stationary noise (here speech) only a portion of the frames receive non-uniform interference.

This leads to incomparable results under presumably similar conditions which we decided to exclude from this study to avoid confusion.

1.4.2 Overlaps in train data

We also examine the effect of adding overlapped speech to train files. Figures 1.13 and 1.14 compares the effects of adding overlapped speech in train and test files.

An interesting observation is the higher rate with which the EER increases when the SIR drops for the test condition. We believe this is due to the fact that in train conditions, the training of Gaussian mixture models tends to cancel out the effect of the interfering speech. For each speaker, the GMM is trained on a set of features, some of which are influenced by the desired speaker and the rest influenced by the interfering speakers. Since multiple training files are used to model each speaker (different training files have different interfering speakers), the GMM tends to converge to a common locale in the feature space, which belongs to the speaker for whom the models are being trained. We call this effect averaging out (or cancelling out) of the interfering speakers. This to some extent slows the growth in EER as the data becomes noisier in train files. Such cancellation, however, does not exist across test files.

1.5 OVERLAP DETECTION SCORES AS META-DATA FOR SID

Using meta-data to yield more accurate decisions is a common practice in SID evaluations (**??**). Incorporating quality measures such as speech activity detection (SAD) and effective file durations can significantly improve SID performance (**??**) regardless of system architecture (be it i-vector, GMM-UBM, or any other system). Meta-data provides lower-level scores that help increase the distinguish-ability between target/imposter trials. In this study, part of the confusion in score distribution is caused by the presence of interfering

speakers. We, therefore, use the scores from overlap detection algorithm(s) as secondary information to improve overall speaker verification performance.

There are several approaches through which quality-measures can be applied in a binary classification scenario (**???**). Here, we use a stacking approach, called Q-stack, in which the quality measures (here overlap decisions) are concatenated ("stacked") with speaker verification decisions (**?**). The resulting vector is a high-dimensional score vector which allows more separability due to the additional information provided by the stacked dimensions. The stacked score vectors are then processed with a support vector machine (SVM) classifier. SVM parameters are trained using a development set extracted from a separate subset of the data. In our experiments, the development set consisted of $10,000+$ trials, a quarter of which were clean trials and the remaining $7,500+$ trials contained overlapped test files with $0, 3, 6dB$ SIR levels. An evaluation set of size $18,000$ trials with similar specifics and target-imposter ratio was used to test overall system performance.

Table 1.2 shows the improvements obtained by using the overlap detection scores individually and in combination groups. The other two features, kurtosis and SFM, show less correlation, however provide significant complementary information when combined and used alongside SAPVR and pyknogram features. The best result is obtained when all four features are concatenated, since each overlap detection system may yield better performance in certain scenarios.

The authors suggest that better individual performances from SAPVR and SFM is because of the nature of their definition which makes them superior in distinguishing harmonic structures. Since speaker identities are mostly influenced by voiced speech, this assists the speaker recognition task in quantifying the amount of voiced speech. Pyknogram-based detection is designed to locate harmonic discontinuities as opposed to the presence of harmonics.

Our experiments show that the best performance is obtained using an SVM with a radial basis function (RBF) kernel. The SVM parameter(s) (here γ) were determined through

SID	pykno	kurtosis	
[HTML]C0C0C0[HTML]343434 ✓			
[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓		
[HTML]C0C0C0 ✓		[HTML]C0C0C0 ✓	
[HTML]C0C0C0[HTML]343434 ✓			[HTI
[HTML]C0C0C0 ✓			
[HTML]C0C0C0[HTML]343434 ✓	[HTML]C0C0C0 ✓	[HTML]C0C0C0[HTML]343434 ✓	
[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓		
[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓		
[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓	
[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓	
[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓	[HTML]C0C0C0 ✓	

Table 1.2. SID performance (EER) with and without overlap detection scores as meta-data. Grey cells highlight the features used in each experiment. The relative change in EER is presented in the last column.

cross-validation on the development set. Class weights (i.e., target/imposter weights for the SVM classifier) and the cost (aka slack) parameter were selected according to the DCF parameters (C_{fa}, C_{miss}, and $prior$, (**?**)) used throughout the SID experiments.

We also conducted an experiment using ideal overlap labels (labels from ground-truth) in the Q-stack paradigm which resulted in an upper bound in performance of 8.74% EER (23% relative improvement). We note that for the Q-stack algorithm, the relative drop in EER from using all overlap features is approximately 20%, which is not far off from when ground-truth labels are used. This confirms the effectiveness of the selected overlap detection features/scores.

1.6 CONCLUSION

An overlap detection method based on enhanced spectrograms (pyknograms) was introduced which led to effective detection accuracy across different SIR levels. The proposed method was compared with existing overlap detection features in terms of accuracy, robustness across different signal-to-interference levels, and precision. We also investigated various properties of overlapped speech and its effect on speaker identification (SID); including

signal-to-interference ratios, signal duration, and the difference when overlapped speech is introduced to test vs. train files. Our experiments on a specialized database for overlapped data showed that the presence of an interfering speaker is more visible when introduced to test files as the SIR increases. An additional finding was the improved performance in overlapped speech detection when using ensemble decisions, instead of decisions based on individual frames. The final study considered using overlapped detection results as meta-data for a given speaker verification task. The meta-data was incorporated using the Q-stack algorithm and a support vector machine (SVM) classifier to improve verification performance by taking advantage of a high-dimensional score-space. We established a lower bound for the achievable EER for the Q-stack paradigm by calculating the results using ground-truth overlapped labels, which yields a 23% relative improvement. Using the proposed overlap detection system and other existing features the relative improvement was 20%, a mere 3% off the best achievable performance given by the lower bound.

Mail All Correspondence To:

Prof. John H.L. Hansen
Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Dept. of Electrical Engineering, University of Texas at Dallas
2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, U.S.A

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

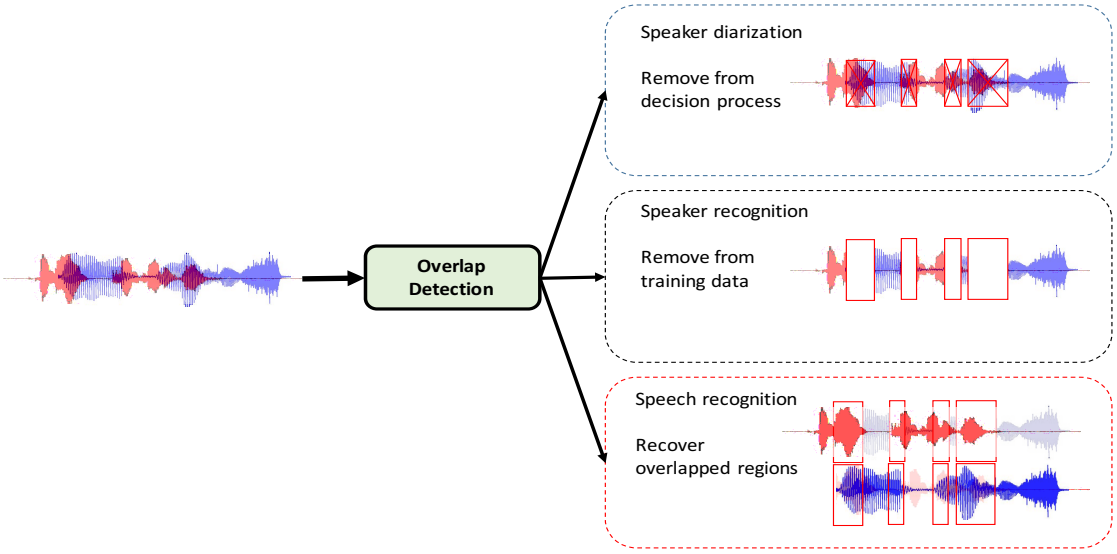
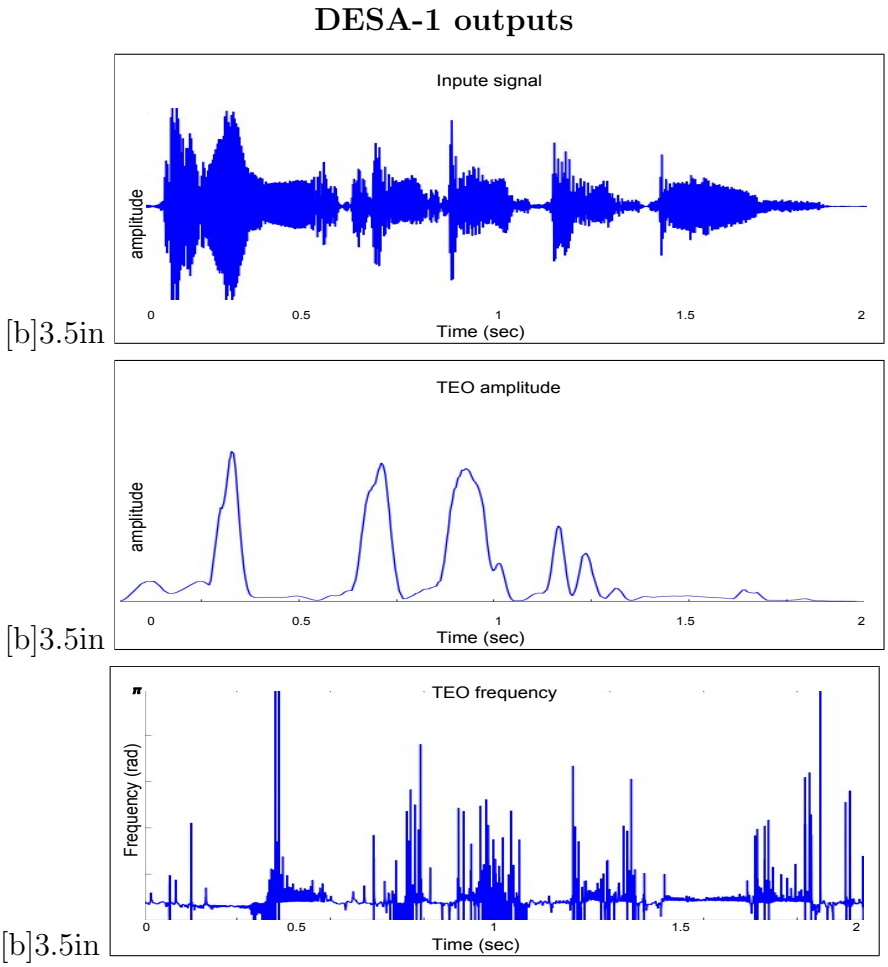


Figure 1.1. Applications of overlap detection. Top: In speaker diarization, removing ignoring overlapped regions provides a more fair assessment of diarization performance. Middle: Removing overlaps from in speaker recognition increases the reliability of training data. Bottom: Overlap detection results can be used as an initial step to recover overlapped regions.



[b]3.5in

[b]3.5in

[b]3.5in

Figure 1.2. Instantaneous amplitude and frequency component. These are the outputs of DESA-1. Top: Input signal. Middle: Signal amplitude component estimated using TEO, Eq. (1.3). Bottom: Signal frequency component estimated using TEO, Eq. (1.2).

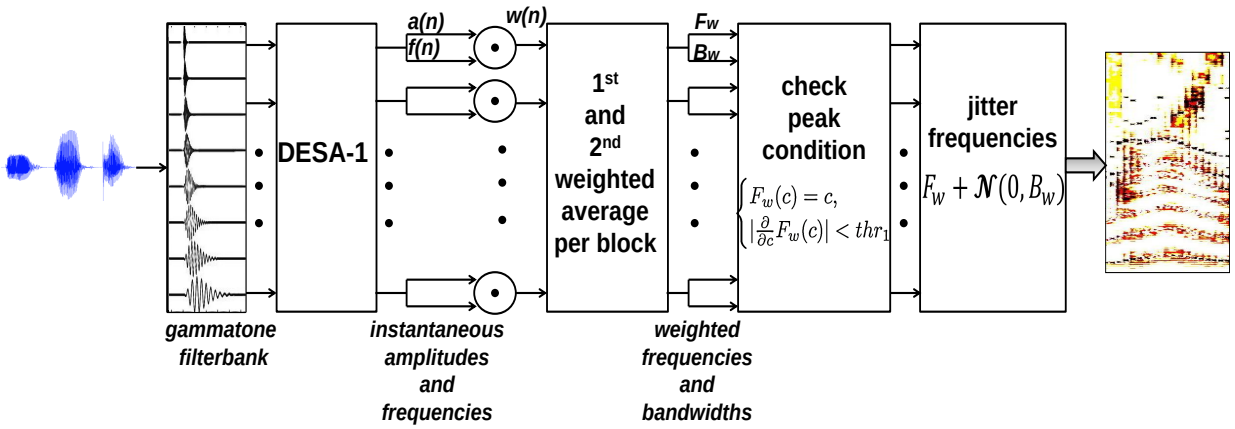


Figure 1.3. Pyknogram extraction block-diagram.

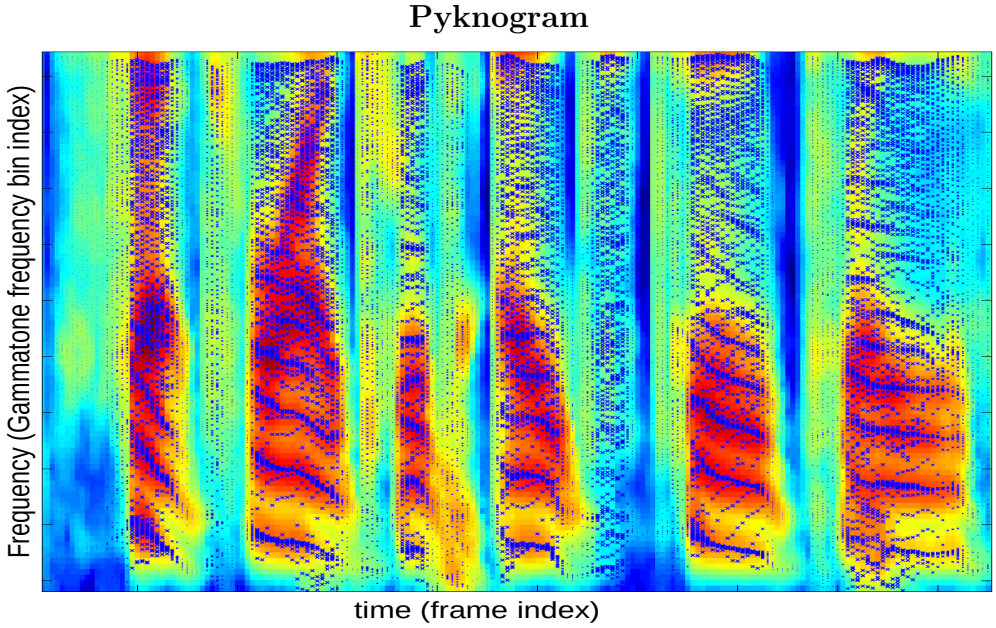


Figure 1.4. Pyknogram for a given speech signal. The spectrogram is plotted in the background for comparison. Pyknogram markers have been scaled by the amplitudes of corresponding t-f units. Frequencies are scaled to equivalent rectangular bandwidth (ERB) rate.

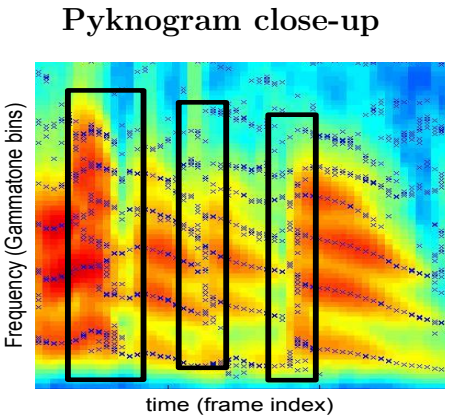


Figure 1.5. A closer look on Pyknograms for overlapped speech. The enclosed patches show discontinuities that occur in the presence of an interfering speaker.

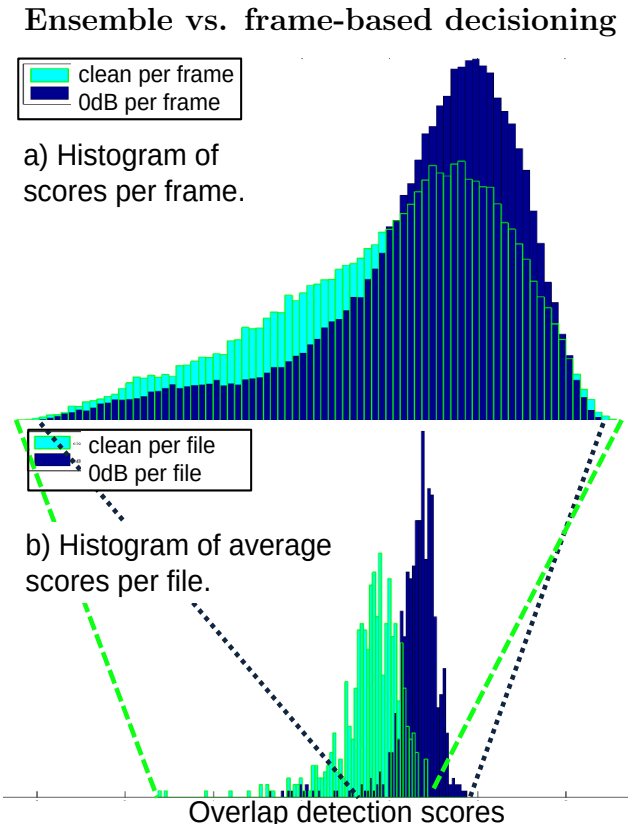


Figure 1.6. The effect of ensemble decisioning on distinguishability of overlapped regions. a) shows score per frame histograms and b) shows the histogram of ensemble scores. Using multiple frames to make a decision helps separate the distributions of clean and overlapped segments.

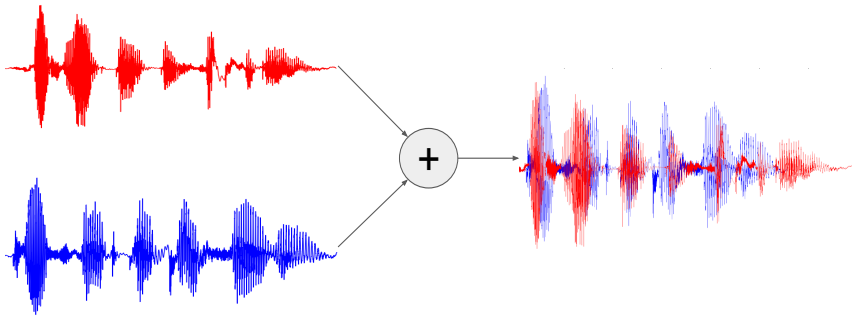


Figure 1.7. Example of the mixing process for a 0dB SIR overlapped signal. As shown on the right, it is fair to assume that overlap occurs throughout the signal.

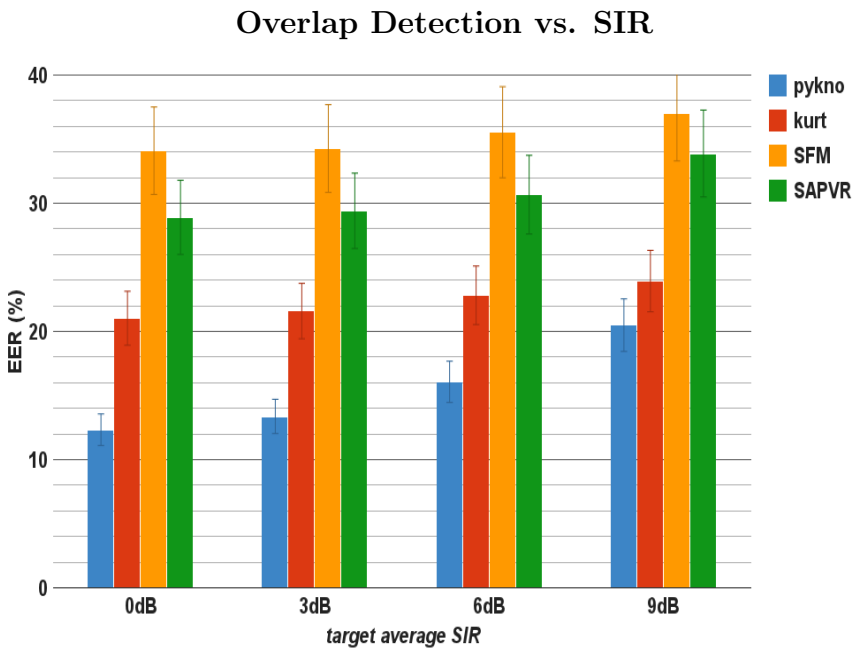


Figure 1.8. Overlap detection EER for different SIR values. The higher the SIR, the more difficult it is to detect the presence of interfering speakers.

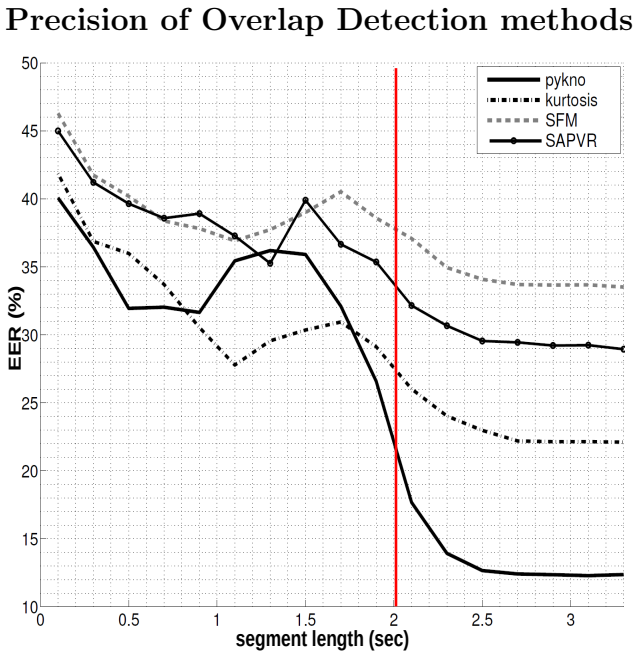


Figure 1.9. Overlap detection EER as a function of segment length. The plot shows that signal lengths should be at least 2 seconds for the algorithms to start reaching their best performance.

[t]0.5

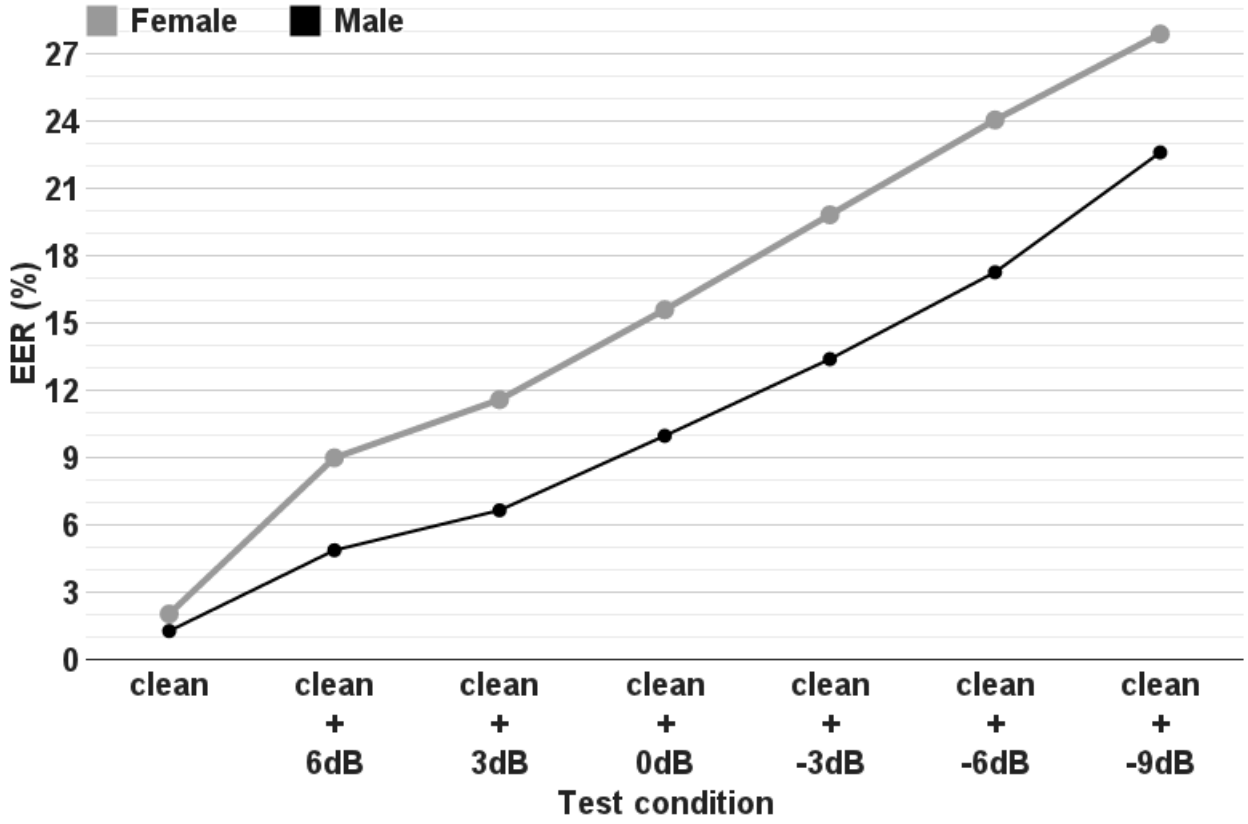
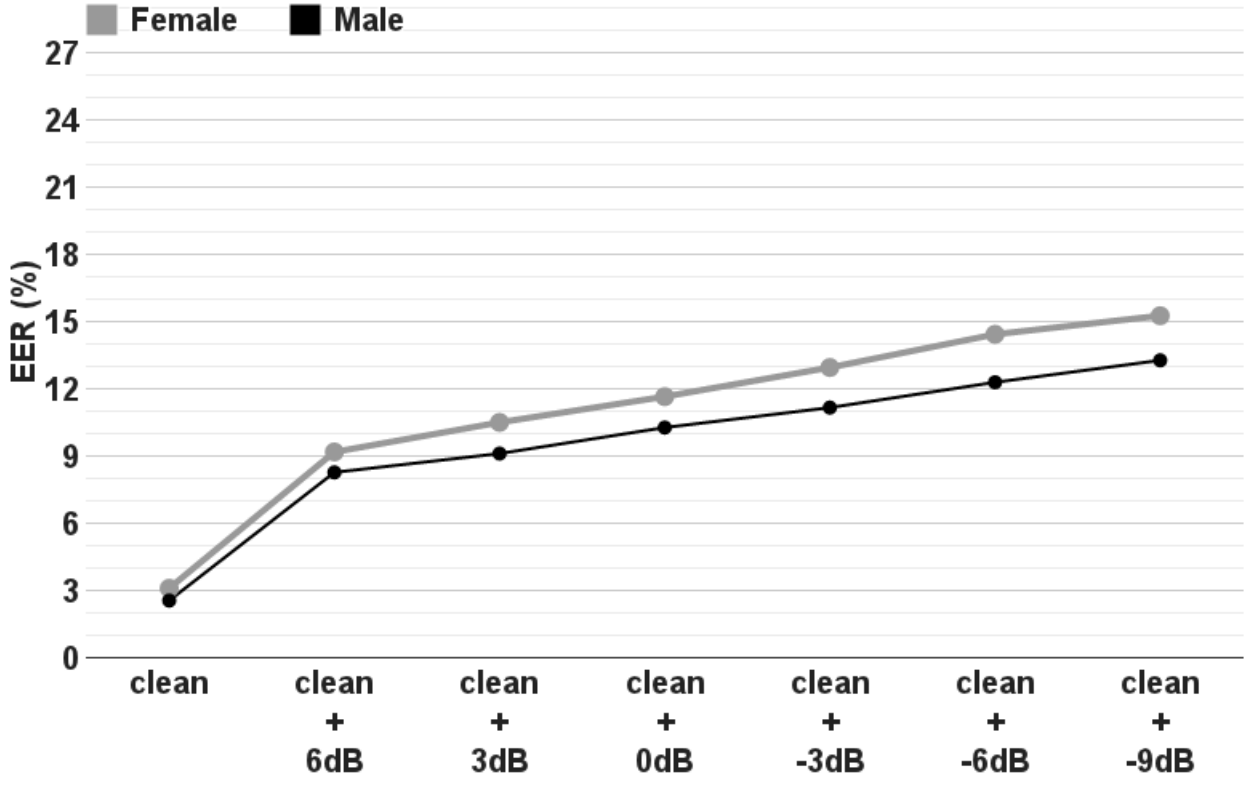
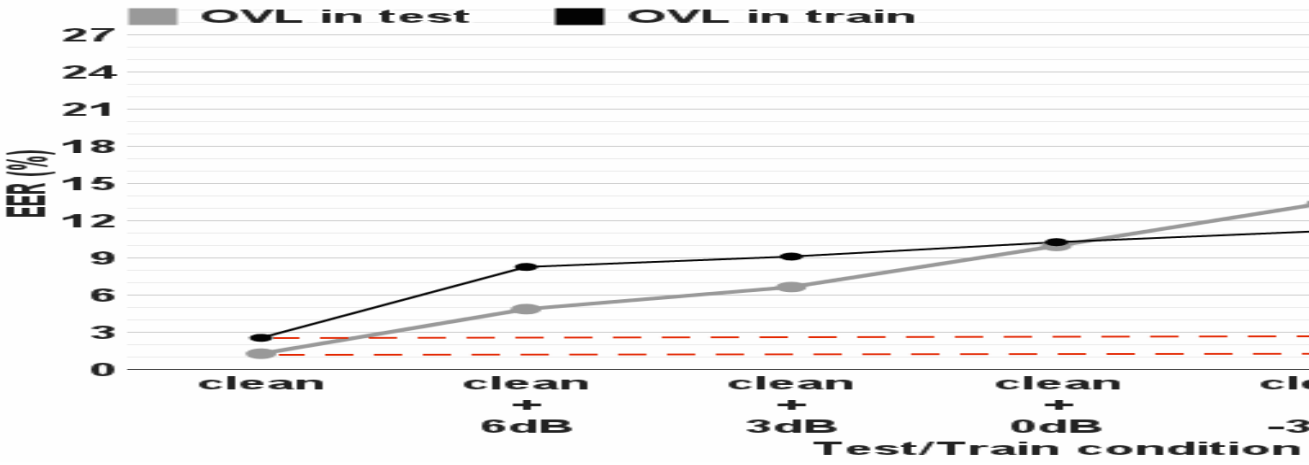


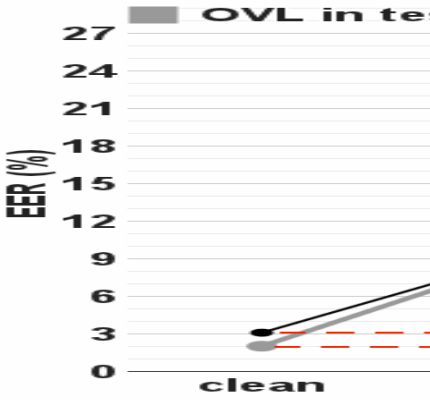
Figure 1.10.
[t]0.5





[b]0.3

Figure 1.13. male



[b]0.3

Figure 1.14. female

Figure 1.15. Comparing the impact of increasing overlap (OVL) in train vs. test data by decreasing SIR values. Experiments for male (a) and female (b) speakers. Lower SIR drops the performance more rapidly when applied to test data.

CHAPTER 2

CONCLUSION

SAMPLE SOLO APPENDIX

VITA

Navid Shokouhi