Co-channel Speech Detection via Spectral Analysis of Frequency Modulated Sub-bands

Navid Shokouhi, Seyed Omid Sadjadi, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3021, USA
{navid.shokouhi,sadjadi, john.hansen}@utdallas.edu

Abstract

Overlapped-speech is known to degrade performance in automatic speech systems. In this study, a sub-band speech analysis technique is proposed to detect overlapped-speech segments in single-channel multi-speaker scenarios (i.e., co-channel speech). Sub-band signals are obtained by decomposing the input speech using a gammatone filterbank. Filterbank outputs are then used to modulate the frequency argument of a sinusoidal carrier. We show that the spectra of these frequency-modulated signals, namely Gammatone Sub-band Frequency Modulation (GSFM) features, are more disperse in overlapped-speech segments compared to single-speaker regions. We quantify the dispersion rate to obtain a measure for the amount of overlapped speech in a given speech segment. Overlap detection experiments are conducted using the speech separation challenge corpus and GSFM features are compared to commonly used overlap detection features. Detection errors are reduced by a relative 50% across different signal-to-interference values ranging from 0 to 9dB.

Index Terms: co-channel speech, overlapped-speech detection, sub-band analysis, gammatone filterbank

1. Introduction

Co-channel speech is referred to a monophonic audio recording in which at least two speakers are present. Separating the speakers in co-channel speech has traditionally been a challenging task for automatic speech applications [1]. In the past decade, due to vast developments in automatic detection systems such as speaker identification (SID) and diarization, a growing trend in recognizing regions of overlapped speech has been observed. In speaker identification, for example, the presence of co-channel speech in conversational audio reduces the reliability of trained speaker-models. SID scoring performance is also extenuated for co-channel speech regions in test segments (e.g., the concept of usable speech in test recordings as presented in [2, 3]). In addition, state-of-the-art speaker diarization systems have currently reached a stage where one of the main sources of error is overlapped-speech [4, 5, 6]. In this study, the focus is solely on developing an overlapped-speech detection system. Such a system may be used in any of the aforementioned applications as a data purification step or a signal processing front-end.

A majority of previous studies have adopted the concept of spectral harmonicity as a key component to detect overlapped speech. The motivation comes from the fact that the presence of two fundamental frequencies in overlapped speech disarranges the harmonic structure observed in single speaker speech. In [7], spectral autocorrelation peak-valley ratios are introduced as a highly discriminating feature for co-channel speech detection based on the same assumption [8]. Spectral flatness measure, which is defined as the geometric to arithmetic mean of spectral bins in a speech frame, has also been used as a feature to capture harmonicity in overlapped-speech detection [9]. Fundamental frequency estimates are highly correlated with harmonicity and have often been considered in overlapped-speech detection systems [10, 11]. The adjacent pitch period comparison (APPC) method, proposed in [11], employs temporal variation of the estimated “pitch” period as a measure to detect “usable” speech (i.e., non-overlapping speech segments). This technique is based on the assumption that the temporal variation of adjacent pitch periods is significantly higher in overlapped-speech regions. Another pitch-based method is presented in [12, 13], where a multi-pitch tracking algorithm is developed to estimate the fundamental frequency in the presence of multiple speakers. Regions at which more than one fundamental frequency is estimated are marked as co-channel speech1. The multi-pitch tracking technique proposed in [13] decomposes the speech signal into multiple channels (i.e., sub-bands) and only the reliable sub-bands are considered for pitch estimation. Another aspect that has been utilized in distinguishing overlapped from single-speaker speech is the difference in the statistical characteristics of these two classes of speech [14, 15, 16]. This is done by using kurtosis [15] and signal entropy [14] as features. In our study, we incorporate the use of sub-band signals to design a feature extraction framework suitable for overlapped-speech detection. The purpose is to simultaneously analyze different regions of the spectrum in small time segments, since the impact of an overlapping speaker’s interference is neither uniform across different sub-bands nor over time intervals. The resulting time-frequency (T-F) units are employed in a frequency modulation paradigm to magnify the non-coherent harmonic structure observed in overlapped speech segments. We show that the spectra of the frequency-modulated T-F units are more disperse in overlapped-speech segments compared to single-speaker regions. These spectral characteristics of the frequency modulated signals are used as input features to extract scores for a binary classifier. The remainder of the paper is presented in the following order. The theory and motivation behind the feature extraction procedure are described in the next section. Section 3 describes the classification framework as well as the baseline systems where we use commonly adopted features for overlapped-speech detection [17, 18]. Section 4 provides a description of the experimental setup and further analyses on the presented system.

---

1 In this study the terms co-channel speech and overlapped-speech are used interchangeably.
Finally, conclusions are drawn in Section 5.

2. Proposed features: Gammatone Sub-band Frequency Modulation Spectral Roll-off

We begin this section with a brief analysis of frequency modulated sinusoids and their spectral characteristics [19]. Modulating the frequency of a sinusoid results in a signal with more frequency components than the original sinusoid. For example, the spectrum of a single-tone frequency modulated carrier contains frequency components that depend both on the amplitude and frequency of the modulating signal, which both contribute to the modulation index, $\beta$ [19]. Fig. 1(a) shows the spectrum of a single-tone FM signal. This signal is typically simplified and defined as,

$$x_c(t) = A_c \cos(2\pi f_c t + (\beta \sin(2\pi f_m t)))$$

(1)

where $f_m$ is the frequency of the modulating sinusoid and $A_c$ and $f_c$ are the carrier amplitude and frequency, respectively. In Fig. 1(a), the amplitude of the $n^{th}$ frequency component of $x_c(t)$, considering $f_c$ as the origin, is the $n^{th}$ order Bessel coefficients at $\beta$ (i.e., $J_n(\beta)$).

Since overlapped speech consists of two speech signals, a closer analogy to the problem of overlapped speech is observed in the case where the modulating signal has more than one sinusoid [20], as seen in (2). Fig. 1(b) and (c) compare the spectra of double-tone FM signals in two scenarios. In Fig. 1(b), the two tones are harmonically related (i.e., one is an integer multiple of the other), while in Fig. 1(c), the frequencies of the modulating tones do not share a common integer factor. Consequently, the spectrum in Fig. 1(c) is more dispersed than that in Fig. 1(b). The same conclusion can be made from (3) where the number of frequency components of the Fourier transform of $x_c(t)$ is greater when $f_1$ and $f_2$ are not harmonically related. An FM signal for a double-tone modulating signal with frequencies $f_1$ and $f_2$ can be represented as,

$$x_c(t) = A_c \cos(2\pi f_c t + [\beta_1 \sin(2\pi f_1 t) + \beta_2 \sin(2\pi f_2 t)])$$

(2)

Defining $J_n(.)$ as the $n^{th}$ order Bessel function, the Fourier series expansion of (2) can be compactly defined as follows [19],

$$x_c(t) = A_c \sum_{n} J_n(\beta_1) J_m(\beta_2) \cos(2\pi (f_c + nf_1 + mf_2) t)$$

(3)

This recent observation is of particular interest, since in overlapped speech segments one of the major confusions is in distinguishing related harmonics (i.e. those that belong to the same speaker) versus non-related harmonics. Here, (3) suggests that the number of frequency components of $x_c(t)$ in a given range is greater when caused by non-harmonically related frequencies. On the other hand, if $f_1$ and $f_2$ are related, (3) can be simplified as,

$$x_c(t) = A_c \sum_{n} (K_{m,n}) \cos(2\pi (f_c + n'f_1) t)$$

(4)

where for each $m$ and $n$ pair,

$$K_{m,n} = J_n(\beta_1) J_m(\beta_2)$$

(5)

However, a substantial difference between the spectral characteristics in multi-tone and single-tone FM signals versus speech is that in speech the spectrum consists of multiple harmonic components across its bandwidth. In order to interact with only a few sinusoidal components, a natural solution is to decompose the signals into multiple sub-bands (i.e., channels) by means of a filterbank. The gammatone filterbank has been widely used in computational auditory scene analysis (CASA) literature to simulate the auditory periphery processing [21]. In [22], a segmentation-grouping approach was employed to detect and separate overlapped-speech T-F units in an unsupervised fashion. In our study, sub-band decomposition is realized through a gammatone filterbank and sub-band outputs are used to modulate the instantaneous frequency of a sinusoidal carrier. Detecting overlapped speech with the use of multiple sub-bands results in multiple decisions for each speech segment. If a specific channel does not have the sufficient information to distinguish overlapped speech from single-speaker speech, the lack of information can be compensated by other channels. We expect that this framework will result in a more consistent detection compared to a scenario where the system only relies on a single decision per frame. The feature extraction procedure is as follows:

1. Apply a gammatone filterbank to the speech signal
2. Demodulate each channel to base-band (since the output

![Figure 1: Comparing the dispersity of different FM signals. From top to bottom. a) Single-tone FM Spectral Magnitude $f_1 = 10$. b) Harmonically related double-tone FM Spectral Magnitude $f_1=10$, $f_2=20$. c) Not harmonically related double-tone FM Spectral Magnitude $f_1=10$, $f_2=25$](image_url)
3. Classification System

The classification system employed to detect overlapped speech uses a combination of the information in all sub-bands, since each band can provide a local decision as to whether a frame contains speech from more than one speaker. We use the overall count of highly dispersive T-F units (i.e., T-F units with higher GSFM roll-off) per time unit as a decision score for the amount of overlapped speech in a given speech segment. By applying a threshold to the GSFM roll-off values (i.e., \( R(t, f) \)) for

\[
R(t, f) = \sum_{i=2}^{N} \frac{P_i(t, f)}{P_i(t, f)},
\]

where \( t \) and \( f \) denote the time and frequency of a T-F unit, respectively. \( P_i(t, f) \) corresponds to the \( i^{th} \) peak in the GSFM spectrum with respect to the center peak (i.e., \( P_1(t, f) \)) that lies on the carrier frequency and is equal to the 0\(^{th}\) Bessel coefficient (see Fig. 5). We call \( R(t, f) \) the GSFM roll-off factor computed at T-F unit \((t, f)\). The collection of GSFM roll-off factors are used as a time-frequency representation for any given speech segment. In the following section, we describe how such a representation can be used to determine the likelihood that a signal consists of overlapped speech.

![Figure 2: GSFM feature extraction block-diagram. Steps from left to right are: 1) Decomposing the signal into sub-bands using a gammatone filterbank. 2) Demodulate all the channels to base-band. 3) Framing. 4) Apply the FM function.](image)

![Figure 3: Comparison of GSFM spectra for overlapped speech and single-speaker speech. a) The GSFM of an overlapped speech segment at the 13\(^{th}\) gammatone sub-band with center frequency 428 Hz. b) The GSFM of single-speaker speech at the same sub-band.](image)

3.1. Baseline features

For baseline overlapped-speech detection systems, a combination of several features that have previously proven to be effec-
tive for this task are used [9, 17, 18]. The features include spectral flatness measure (SFM) [9], [17], and signal kurtosis [16], and spectral autocorrelation peak-valley ratios (SAPVR) [7].

4. Experiments

We conducted overlapped-speech detection experiments using data from the speech separation challenge [23]. This database consists of clean and overlapped utterances. Overlapped data are categorized based on average signal-to-interference ratios (SIR) (0, 3, 6, 9 dB). Sentences comprise of a selection of identical phrases uttered by one of 34 speakers (see [23]). We use 8kHz as our selected sampling rate. In overlapped segments, the target speech is mixed with another speaker’s utterance in one of the 4 SIR categories. Note that utterances are artificially summed, making the data different from natural overlapping speech that occurs in conversations; an inevitable compromise required to gain access to sufficient overlapped data. We intend to investigate three parameters in the experiments:

- **Accuracy**: which is implicitly estimated by calculating performance errors.
- **Precision**: how short can the overlapped segments be before significantly dropping performance?
- **Robustness**: can the systems perform consistently well for different SIR values?

Experiments are standard detection tasks where each detection system outputs a score for a sample utterance. At the end, all scores are evaluated under different thresholds, and the equal-error-rate (EER)² is reported as a measure for system performance. Since our main goal is to evaluate detection system performance, regardless of a specific data configuration, we use a synthetic example set where the number of target and non-target examples are equal (i.e. prior = 50%), which is not the case in real-life conversations [24].

Scores are estimated for all baseline systems by using the average feature value over all frames. In the case of SAPVR, we observed that applying an energy-based voice activity detection (VAD) helps improve performance. GSFM scores are obtained by applying 8 to the roll-off factors extracted for each T-F unit.

4.1. Performance across different SIRs

A main challenge in overlapped-speech detection is having to deal with a large range of interference ratios, even within a single recording. Ideally, an overlapped-speech detection system should perform consistently well in any SIR. However, this is neither feasible nor necessary. As mentioned earlier, the database used in this study consists of 4 different average SIR conditions. Fig. 4 compares the performance of our proposed GSFM spectral roll-off feature with the four baseline systems described in Section 3.1.

4.2. The effect of duration

All algorithms described in this study use the collective knowledge obtained from a given speech segment to estimate a score that represents the likelihood of overlapped speech. The precision of the overlap detection system tells us how short a given segment could be before we observe a substantial drop in performance. We estimate the EER across different signal time-lengths in 0dB average SIR (Fig. 5). It is observed that the EER varies most in kurtosis for different signal durations. This is expected, since kurtosis is calculated from 3rd and 4th order moments which are less accurately estimated with insufficient data. From Fig. 5, we also conclude that the breaking point for all systems is at approximately 0.6 seconds, which implies that overlap detection performance drops dramatically for segments less than 0.6 seconds long. The performance drop is more severe for GSFM features, since it uses the distribution of roll-off values in calculating the final score (8).

5. Conclusions

A sub-band analysis technique was developed to detect overlapped speech segments throughout audio streams. The algorithm uses the sub-band outputs from a gammatone filterbank as the modulating signal of a sinusoidal carrier to magnify the presence of multiple speakers. The proposed GSFM features were motivated by an investigation on the spectral structures in single-tone and multi-tone FM signals and extending their properties to harmonics in sub-band speech. Our experiments indicated that the features extracted using this method were highly discriminative for overlapped-speech detection. On the average, GSFM features reduce system errors by 52% across different SIR values considered in this study.
6. References


