Assessing Speaker Engagement in 2-person Debates: Overlap Detection in United States Presidential Debates

Midia Yousefi\(^1\), Navid Shokouhi\(^2\), John H.L. Hansen\(^3\)

\(^1\)Center for Robust Speech Systems(CRSS)
The University of Texas at Dallas, Richardson, Texas, USA
\(^2\)University of Melbourne

Abstract

Co-channel speech recordings typically contain significant amounts of overlap in which the intelligibility and quality of the desired speech is degraded by interference from a competing talker. Convolutive Non-negative Matrix Factorization (CNMF) has been shown to be a successful approach in detecting overlap by extracting specific acoustic basis dimensions for each speaker from an audio stream. While the results of CNMF have been successful, it requires isolated single speech recordings for each speaker to derive their corresponding bases functions/dimensions. In our previous work, The Teager-Kaiser Energy Operator (TEO)-based Pyknogram has been introduced which does not require prior information concerning the speakers. In this study, Pyknogram and CNMF based solutions for overlap detection within audio streams have been examined using the GRID dataset. TEO-based Pyknogram is shown to achieve a relative 8-10% lower Equal Error Rate (EER) compared to CNMF features. Another drawback of CNMF is that its performance drops considerably when dealing with spontaneous speech that has not been considered for extracting bases in the training step. In addition to the experiments on GRID corpus, a secondary evaluation is also performed based on naturalistic audio streams with overlap. Specifically, we collected a real-world audio database of US Presidential debates stemming from the last 12 years that are challenging due to overlap, changing Signal to Interference Ratio (SIR), and environmental noise, etc. Our experiments indicate that TEO-based Pyknogram is well suited for detecting overlap in challenging real world scenarios such as the US presidential debates.

Index Terms: overlap detection, co-channel speech processing, speaker diarization

1. Introduction

In many occasions such as conference meetings, debates, and pilot-air traffic controller conversations, a mixture of speech utterances are recorded over a single communication channel. These recordings are referred to as co-channel speech. One of the main challenges of co-channel recordings is detection/separation of overlapped speech segments.

Over recent years, development of state-of-the-art automatic speech applications has reached a point where overlapping speech can be considered as a significant source of error \([1][2][3][4]\). Segments of overlap in training data have serious consequences for processing models in all applications such as speech/speaker recognition, speaker diarization and dialog modeling \([5]\). In speaker diarization where the main goal is to determine “who speaks when”, current systems are generally able to detect only one single speaker per segment. Hence, intervals where more than one active speaker is present can contribute to diarization error. Moreover, overlapping speech can lead to impurities in the speaker model which directly degrades clustering performance. Removing detected overlap segments from the decision process in speaker diarization can lead to improved performance. Also, in Automatic Speech Recognition (ASR), transcribing what has been said in co-channel recordings is difficult due to overlapping intervals. In \([4]\) Shriberg et al., they reported 9% absolute higher Word Error Rate (WER) for overlapping segments versus non overlapping ones. Speaker identification is another field that can benefit from overlap detection. It is found that about 40% of the 0dB Target to Interference Ratio (TIR) recordings have sufficient information concerning the target speaker to do trusted speaker identification \([6]\). Implementing an overlap detection/separation system can be advantageous for the aforementioned applications as a preprocessing step which leads to improved trained speaker models, and obtaining higher scores in decision processes.

Traditional attempts to process co-channel speech is to separate target speech while suppressing competing talkers speech \([6][7][8][9]\). However, recent approaches attempt to identify “Usable” segments in co channel recordings. The term “Usability” is context dependent and should be determined based on the application goal. In these techniques, the focus is on features that are well-suited for detecting overlap. One of the basic features considered by many researchers is the existence of two fundamental frequencies in overlapping segments. Spectral Autocorrelation Peak Valley Ratio (SAPVR) is another tool that detects the structure in the spectral autocorrelation domain \([10]\). The autocorrelation function is periodic due to spectrum harmonicity in single talker speech. Consequently, the SAPVR would have a smaller value in overlapped segments. However, the main drawback of these features is that they only detect the voiced-voiced situations. Another feature related to the speech spectrum, is spectral flatness which is primarily used for separating voiced and unvoiced portions of speech \([11]\). This, ultimately reveals information about the number of speakers \([12]\). A different feature that looks into the statistics of segments is Kurtosis. This feature measures Gaussianity of a random variable. Speech signals are modeled with Gamma or Laplacian distribution. According to the central limit theorem, combinations of speech signals tend to have Gaussian distribution \([13]\). Hence, Kurtosis can serve as an effective feature for overlap detection. In \([14]\), Convolutive Non-negative Matrix Factorization is proposed which has proven to have effective results. CNMF with a sparsity constraint is capable of projecting segments of overlapping speech into lower dimensional speaker-specific bases.

In our previous work, an enhanced time-frequency repre-
sentation called Pyknogram has been introduced which was shown to be successful in tracking harmonic patterns in speech signals [15] [16]. The Pyknogram can suppress non harmonic structure by detecting resonant time-frequency bins in a speech spectrum. To the best of our knowledge, there is no comparison between results of CNMF and TEO-based Pyknogram in the literature. In this study, we aim to compare the results of the aforementioned techniques.

The contributions of this paper are three-fold: First, we derive two features suitable for overlap detection based on online CNMF. Second, the Pyknogram-based feature is extracted using Teager-Kaiser Energy operators and compared with CNMF features. Third, the best approach is used to analyze a real world audio database of US Presidential debates.

The remainder of this paper is as follows: The details of Online CNMF is described in the Sec.2. Section 3 presents Pyknogram features for overlap detection. In Sec.4, Experiments and results are provided. Finally, conclusion and discussion are presented in the last section.

2. Online Convolutive Non-negative Matrix Factorization-based features

Non-negative, multi-variate data can be represented by an additive combination of lower rank bases. Non-negative Matrix Factorization (NMF) is one of the most popular approaches to factorize data into its latent structure [17]. However, NMF fails to capture any dependency between successive batches of the input signal. To tackle this problem, CNMF has been introduced [18] which is capable of revealing time varying patterns of the signal. The factorization can be expressed as:

\[ X = \sum_{t=0}^{T-1} W_t H_t, \]

where \( X \), \( W \) and \( H \) are the input data, basis and coefficient matrices. \( T \) is the convolution range that defines the time dependence duration of the bases. It has also been reported [19] that adding a sparsity constraint to NMF and its extensions improves the quality of trained bases. Sparsity is controlled by a weighted minimization of the \( L_1 \) norm of the coefficients. \( W \) and \( H \) are calculated by minimizing the following cost function:

\[ (\hat{W}, \hat{H}) = \arg \min_{W,H} ||D - WH||^2 + \lambda \sum_{ij} H_{ij}, \]

where \( \lambda \) is the regularization parameter for controlling sparsity. This minimization requires all data to be available in the memory during training which is an issue when dataset is large. Another drawback of conventional algorithms for CNMF is that if data is continuously generated by a sensor or microphone, they fail to extract the bases in real time. An online version of CNMF algorithms addresses these problems. In the first step, the bases are initialized randomly and then for each utterance, the bases are updated by the following equations:

\[ \hat{H} = \hat{H} \cdot \frac{W(t)X_{t:t+T} - W(t)H_t}{W(t)^T(W(t)H_t) + \lambda}, \] (3)

\[ U^1(t) = X_{t:t+T} - W(t)H_t, \] (4)

\[ V^1(t, t') = W(t)H_{t:t+T}, \] (5)

\[ U = U + U^1, \]

\[ V = V + V^1, \] (6)

\[ W(t) = W(t) \cdot \frac{U}{V}, \] (7)

where the multiplication (\( \cdot \)) and the division are element wise operators. \( U \) and \( V \) are the accumulated model parameters that are updated throughout the algorithm. Using these equations, speaker-specific bases are extracted and then concatenated to form a single matrix as shown in Fig.1. Spectral magnitudes of overlapped speech are then decomposed according to (3) with \( W \) kept fixed. The coefficient matrix \( H \) reveals considerable information about each speaker [20]. The energy of each speaker is calculated as:

\[ E_{\text{speaker}_{i th}} = \sum_{i \in I_{\text{speaker}_{i th}}} H_i, \] (8)

where \( I_{\text{speaker}_{i th}} \) is the speaker-specific rows in the coefficient matrix. In segments of overlap, the energy of multiple speakers are high, hence their ratio would be close to one. The first feature for overlap detection is as follows:

\[ E_{\text{ratio}} = \frac{E_{\text{speaker}_{1 th}}}{E_{\text{speaker}_{2 th}}}, \] (9)

where \( E_{\text{speaker}_{1 th}} \) and \( E_{\text{speaker}_{2 th}} \) denote speakers with the two highest energies.

The second feature deals with variance of the energy difference in each frame. In order to compute this measure, the energy of each speaker is normalized to the highest energy among all speakers. Energy difference is calculated between all speaker pairs. Finally, the inverse of the variance of energy differences is the second CNMF feature:

\[ m_{ij} = \frac{1}{\text{var}(\text{energy difference})}, \] (10)

These two features tend to detect overlap segments based on the energy information of active speakers.

3. Teager-Kaiser Energy based Pyknogram

A linear combination of \( N \) AM-FM signals can represent speech signal resonances [21] as:

\[ S[n] = A[n]\cos(W[n]), \] (11)
where \( A[n] \) and \( W[n] \) are time varying amplitude and frequency. Extracting \( A[n] \) and \( W[n] \) from the speech signal is done in two steps:

1. Apply a bandpass filterbank on speech signal.
2. Apply discrete energy separation algorithms on each filterbank output.

In the first step, the speech signal is transformed into the spectro-temporal domain via a Gammatone filterbank which is based on the human auditory system. Each of these filters are tuned to a distinct frequency range. In the next step, amplitude and frequency for each bandpass signal is computed using an energy estimation operator. The energy of the filtered signals can be estimated using the Teager-Kaiser energy operator as follows:

\[
E_{teo}(S(n)) \propto A^2(n)W^2(n) = S^2(n) - S(n-1)S(n+1)
\]

(12)

TEO estimates the energy needed to generate the present sample using only three samples of the signal. This attribute provides us with the ability to capture the energy fluctuations in the signal. Additionally, the short window (of three samples) required by the algorithm implies great temporal resolution. Another advantage of TEO over conventional Fourier analysis is its capability in estimating energy in a nonlinear manner which makes it suitable for processing speech signals. The energy required by a system to generate speech at higher frequencies is expected to be higher than energy required to generate the signal at lower frequencies. As we know, voiced parts of speech have more energy in lower frequencies, while energy of unvoiced phonemes is concentrated in higher frequencies. Consequently, TEO is an excellent tool for detecting overlap in speech segments that involves voiced/unvoiced or unvoiced/unvoiced phonemes from both speakers due to capability of capturing sharp changes of energy in a nonlinear manner. Using (12), the time varying amplitude and frequency for each filterbank are calculated as follows:

\[
W(n) = \frac{1}{2\pi} \arccos(1 - \frac{E_{teo}(S(n)) - S(n-1))}{E_{teo}(S(n))},
\]

(13)

\[
A(n) = \sqrt{\frac{E_{teo}(S(n))}{\sin^2(2\pi W(n))}},
\]

(14)

Next, the frequency candidates of spectrogram peaks can be extracted as a weighted average of \( W(n) \):

\[
F(n, i) = \frac{\sum W(n)A^2(n)}{\sum A^2(n)},
\]

(15)

Here, the summation is over all samples in a 25 ms frame, \( i \) is the filterbank index and \( n \) is the time sample. If the extracted frequencies are aligned with a filterbank center frequency, they are selected as frequencies of the spectrogram peaks. The amplitude of these peaks is the average of \( A(n) \) over all samples of a frame. This time-frequency representation, denoted by \( S_{pykno}(n, i) \), is called the Pyknogram which has been very successful in detecting overlapped speech. As indicated in [15], sudden jumps in the harmonics of speech is an indicator of interfering speech. The difference of Pyknogram between successive frames can be used as a feature for overlap detection, since overlapping speech segments tend to have higher \( D_{pykno} \):

\[
D_{pykno} = \sqrt{\sum_i ((S_{pykno}(n, i) - S_{pykno}(n-1, i))^2)},
\]

(16)

4. Experiments, Results and Discussions

4.1. CNMF and Pyknogram Experimental Setup

In order to compare CNMF features with Pyknogram, the GRID database [23] has been used for experiments. GRID is a multi-speaker corpus consisting of 18 male and 16 female speakers. Overlapping speech has been generated by adding pairs of sentences with Signal-to-Interference-Ratio (SIR) ranging from 0 to 9dB. An example of the generated overlapped speech in time and frequency domain is shown in Fig.2. Since the overlapped speech is generated in a controlled manner, the average SIR throughout the file can be considered constant which is a great advantage over conversational speech corpora. A real world audio dataset of presidential debates has also been used for evaluation of the best overlap detection approach on the GRID corpus base experiments.

The first step of extracting CNMF features, is to train speaker-specific bases. The number of the bases is set to 10 for each talker. Convolutional range of 5 is chosen for the time span of the bases. The sparsity weight, \( \lambda \) is set to 0.01. In order to obtain the bases for each speaker, 1000 files are processed with 100 iterations of CNMF algorithm. Overall, for all 34 speakers, 34×1000×100 CNMF iterations are needed. The bases are initialized randomly for the first utterance. Next, they are gradually updated by processing all utterances. Two additional parameters, \( U \) and \( V \) are calculated as each file enters the process. \( W, U \) and \( V \) are fed back to the algorithm to be updated by the next speech file. Finally, the bases and model are
saved for each speaker. The main advantage of this approach for learning bases is that, if the bases are already trained for a particular amount of data, newly arrived data can enter the process and the bases are incrementally updated to reflect the changes. While, conventional techniques need the whole dataset (including the previous and newly arrived data) to reside in the memory and be processed again. In the next step, speaker-specific bases are concatenated in a matrix. Temporal coefficients are calculated by mapping the overlapped speech into the bases. Since, 10 bases are learned for each speaker, every 10 rows of the matrix \(H\), correspond to one talker. Eventually, the two CNMF features are derived according to (9) and (10).

For extracting Pyknograms of speech files, a 4\(^{th}\) order log-arithmically spaced Gammatone filterbank with 128 channels is used. Using (14) and (15), the amplitude and frequency candidates of peaks are derived. Among the frequency candidates, those which are aligned with their corresponding filterbank center frequencies are selected. The Pyknogram feature is calculated according to (16). We use Equal Error Rate (EER) for evaluation of the features performance in detecting overlap.

### 4.2. Results and Discussions

The results of CNMF and Pyknogram-based solutions are shown in Fig.3. The EER for speaker energy ratio is very high. Since, the bases are normalized, speaker energies are simply the sum of speaker-specific bases coefficients. High correlation between bases of different talkers leads to the spread of the main two speakers energy across all speaker-specific bases. Consequently, this feature is not a suitable detector for overlapped speech. Variance of energy differences among all existing speakers in the dataset has a better performance, since it measures the variations in speaker energies which is lower in non overlapping segments. The Pyknogram feature has the best performance in terms of EER. This feature tracks the harmonicity in the frequency domain, hence can detect overlapped segments with frames of sudden jump in harmonic structure. Generally, the Pyknogram shows a 8-10% better performance versus CNMF features.

So far, artificially generated overlapped speech has been used to evaluate the performance of the features. In addition, we have also collected a real-world audio database of US presidential debates stemming from the last 12 years. This database contains three rounds of debates in four presidential elections. This dataset is a challenging one due to varying SIR and environmental noise. Presidential debates contain various kinds of overlap due to the controversial topics discussed by the candidates. Short overlaps of agreement or disagreement are very common in this database. Also, In the Trump-Clinton debate in 2016, there are many occasions which both candidates are talking simultaneously for a longer duration. Overlap decision in short length segments are less reliable. According to [15], different features such as kurtosis, Spectral Flow measure (SFM), Spectral Autocorrelation Peak-Vally Ratio and Pyknogram need segments with at least 2 seconds of duration for reliable overlap detection. Hence, these approaches may not be able to detect very short words of acknowledgment or disagreement in the conversations. According to the previous experiments on the GRID database, the Pyknogram has better performance, so we have used this approach to detect overlapping segments of speech in US presidential debates. The results of the overlap percentage detected in each debate is presented in Fig.4. As expected, Trump-Clinton debates (from 2016) have highest percentage of overlap. This can be interpreted in two ways. First, the number of segments which the competing talker interferes with the main speaker is more. Second, the duration of the interfering speech is more. Fig.5 presents the aggregated overlap in all three rounds of debates. Debates of 2004, 2008 and 2012 have close amounts of overlap which is very common in all 2-person conversations while debate of 2016 has a significant overlap percentage.

### 5. Conclusion

This study aimed at identifying segments of overlapping speech in co-channel recordings. Convolutional Non negative Matrix Factorization has been a widely used technique to address this problem. CNMF has two main drawbacks: First it needs single speaker audio files in order to train speaker-specific bases, second the whole dataset should be available in the memory to form the input matrix. For the latter problem we have presented an online approach which gradually updates the derived bases by processing each file. TEO-based Pyknogram is another method that has been introduced lately. To the best of our knowledge, our study is the first to compare CNMF and Pyknogram for overlap detection. The Pyknogram shows 8-10% improvement in EER compared to CNMF solution. Further, a significant advantage of Pyknogram is that it processes the overlapped audio without requiring the corresponding single speaker utterances. In addition, we have also used U.S. presidential debate dataset over the last 12 years for overlap detection. This database is challenging due to noise, varying SIR and difference in length of the overlapped segments. Our experiments reveal that TEO-base Pyknogram is a suitable feature for overlap detection in the real world co-channel audio recordings.
6. References


