International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

# Overlapped Speech Detection in Multi-Party Meetings

Thein Htay Zaw[a]*, Mie Mie Thaw[b]

[a,b]*University of Computer Studies, Mandalay (UCSM), Mandalay, Myanmar*
[a]*Email: theinhtayzaw@ucsm.edu.mm*
[b]*Email: miemiethaw@ucsm.edu.mm*

**Abstract**

Detection of simultaneous speech in meeting recordings is a difficult problem due both to the complexity of the meeting itself and the environment surrounding it. The system proposes the use of gammatone-like spectrogram-based linear predictor coefficients on distant microphone channel data for overlap detection functions. The framework utilized the Augmented Multiparty Interaction (AMI) conference corpus to assess model performance. The proposed system offers enhancements over base line feature set models for classification.

***Keywords:*** overlapping speech; gammatone like spectrogram; linear predictor coffeicients.

## 1. Introduction

Overlapping speech is a natural part of the behavior of human communication. For this reason, audio recordings of meetings usually contain regions where speech overlaps. It is an effect that adds challenge to existing state-of-the-art speech analysis systems such as diarizing speakers and recognizing speech. Meetings are known to be the most difficult application area due to high speech spontaneity, room reverberation and variable microphone signal quality. Many studies have been performed on identification of overlap and its effect on diarisation. According to some of the studies [1, 2], the presence of speaker overlap can directly correlate a portion of the performance deterioration on real meeting results. Nevertheless, accurate identification of overlapping speech segments is a difficult problem due both to the nature of the speech and the environmental impacts such as reverberation noise, background and echo[3]. A number of research on overlap analysis and speech overlap in speaker diarization have been performed. [4] used the HMM-based segmenter to detect speech, non-speech and overlap speech from audio meetings, where the models are trained using cepstral features along with instantaneous and LPC residual energies and subsequent entropy diarization from ground truth alignments.

---

* Corresponding author.

In[5] an attempt is proposed to model overlap speech as a non-linear transformation of features in the cepstral domain, and [6] is proposed sparse coding approach to convolution. In addition to using short-term segments (20 30 ms), [7] showed improved overlap detection by increasing silence and change of speaker statistics with short-time characteristics computed over 3-4 s. The more conventional approaches are focused on a careful range of handcrafted features that can be input into an HMM decoder [8, 7] or a neural network [9, 10]. A more recent alternative is to allow a neural network to retrieve the information in the relevant "raw" input form, as in an acoustic signal spectrogram[10]. The methods mentioned in [11] earned up to 76 percent single-speaker speech, 60.6 overlap speech, and 68.4 average speech accuracy, respectively. The performance of classification of nonspeech and noise, though, is not definedThe previous algorithms are based on Gauussian mixture modeling[12], a segmenter based on HMM[13], pyknogram analysis[14], and LSTM[15]. The best performance was recorded in [9] with an F-score equal to 0.8 for intervals of 500 ms, but the findings applied only to artificially mixed recordings and male speakers. Thus, although the performance of the existing solutions is quite decent, there are still considerable opportunities for improvement. Obviously the main aim is to enhance the detection for short frames of up to 10 ms. Using new feature set, the system aimed to achieve high precision. The paper's key focus is on enhancing the identification of overlaps in conversational speech. The method uses the Augmented Multiparty Interaction (AMI) meeting corpus for assessment[17]. The system is structured as follows. Sec.2 states the data set used in the research, the techniques of feature engineering and the motivation for selecting each characteristic. Sec.3 outlines the approach used to detect overlaps. The findings of the various studies are presented in Section 4, followed by a conclusion in Section 5.

## 2. Proposed Features

The system developed new feature extraction method based on GTF and LPC for overlap speech detection is developed. The feature extracted using proposed method is a vector instead of the matrix. The flow chart of the GTF-LPC feature vector calculation algorithm is shown in Figure 1.

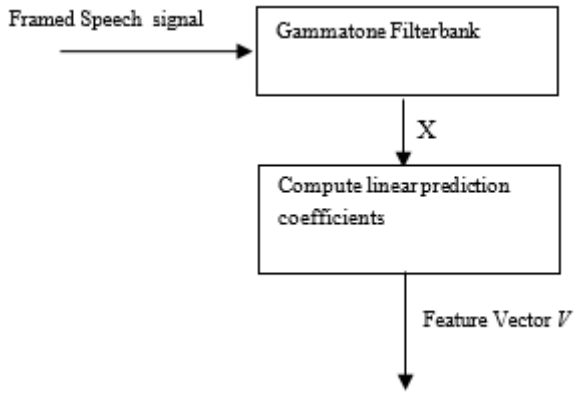


**Figure 1:** The flow chart of the GTF-LPC feature vector calculation

From the above figure, the GTF-LPC feature vector is extracted as following:

Initially, the framed speech signal is passed to a filterbank of Gammatone, assuming the number of filters is N.

Thus the output of this filterbank is a vector x with N columns, and each column, i=1, ... ,N, represents the output in time domain of each Gammatone filterbank bandpass filter. LPC is then rendered to vector x. The LPC formula is defined in section 2.2. Feature vector demonstrates the statistical structure of its framed speech signal in various frequency bands that are designed to mimic human hearing frequency resolution. So although LPC provides relatively reliable representation of the data observed. So assume the feature vector V generated from the speech signal will concentrate on the difference between the speakers in the statistical structures. Hence it is being used to model the individual speaker's distribution.

### *2.1. Gammatone like spectrogram*

Gammatone auditory filter banks are non-uniform bandpass overlap filters, designed to mimic human hearing frequency resolution [20]. Each filter's impulse response is in the shape of the Gammatone function. The GammaTone Filter impulse response is characterized as

$$h(t) = \begin{cases} ct^{n-1}exp(-2\pi bt)\,cos(2\pi f_0 t + \emptyset)\,, t > 0 \\ 0, \qquad\qquad\qquad\qquad t < 0 \end{cases} \qquad (1)$$

Where n is the filter order (affects the filter skirts slope), b is the filter bandwidth (Impacts the length of the impulse response), c is simply a scalar specifying filter gain, is the center frequency of the filter, Ø is the phase. The bandwidth b of order 4 can be computed at the center frequency as the following formula

$$b = 1.019*2*pi*ERB \qquad (2)$$

And, bandwidth b of order 5 can be computed as following formula at centre frequency

$$b = 1.164*2*pi*ERB \qquad (3)$$

The equivalent rectangular bandwidth (ERB) of the filter is given with the equation

$$ERB = 24.7(4.37\ fc/1000 + 1). \qquad (4)$$

Even so, it can still be computationally expensive to process a signal with a bank of M Gamma- tone filters. Ellis proposed an alternative approach using an approximation based on a fast Fourier trans- form (FFT) [16]. In this method, a conventional fixed-bandwidth spectrogram has been computed, then these frequency bins are aggregated via a weighting function into Gam-matone responses with coarser resolutions. This mimics very closely matches Slaney's accurate method, despite neglecting each frequency bin's phase information when summing them up [16].

### *2.2. Linear Predictor Coefficients*

The Another common way of getting spectral information is through LPC. The signal u is estimated by a linear combination of its previous values

$$\boldsymbol{u_n} = \sum_{i=1}^{p} a_i \, u_{n-i} \qquad (5)$$

Where $ai$ is the coefficients of linear predictor, $u(n-1)$ is the previous detected values, with and $p$ is the LPC order. The coefficients are calculated by using the Levinson-Durbin recursion to minimize residual error energy.

### *2.3. Dataset*

The AMI corpus[17] is a meeting corpus that contains almost 100 hours of multi-talker conversational meeting data. A session has at least three speakers, and a total of 171 speakers (114 males and 57 females) are part of the entire corpus. The each meeting is captured using a set of different devices, namely a microphone array consisting of eight remote microphones, headset and lapel microphones. The system uses the first microphone channel from the array of microphones (the Array-1 specifies). In addition to the naturally occurring overlapping regions of speaker, the microphone array channel is distant recording, thus involving reverberation of which background noise. Human annotators use the headset recordings to annotate the AMI corpus). 7 to 9 per cent of the total spoken speech frames (at 10 ms) consist of overlapped speech when analyzing the annotations. The rest of the recording is either unlabeled single speaker speech / silence.

## 3. Experimental Setup

The proposed overlap detection system is illustrated in Figure 2.

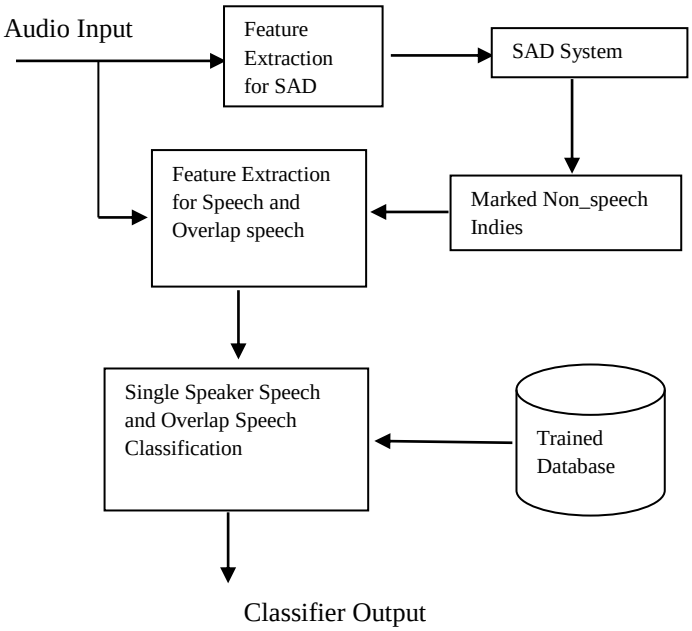


**Figure 2:** Flowchart of the system

The organization formed a training dataset using the audio from AMI corpus. It consists of five main parts. Firstly, features are extracted for SAD system. Part 2 classified noise_silence interval and speech interval. In part 3, extracts proposed features for overlap detection system and marked the interval of noise_silence frames

in part 4. Finally, extracted audio features are fed to the ANN classifier to generate single speaker speech and overlap speech.

### *3.1. Speech Activity Detection*

A speech activity detection (SAD) system aims to classify speech and non-speech segments within a given audio stream. There are two significant ways in which the errors made by a SAD system affect diarisation. The missed speech segments and false-alarm speech detections contribute in the form of missed and false alarm errors directly to the diarization error rate (DER). Hence, a weak SAD method would adversely affect the metric DER diarization measurement. The speech portions that a SAD system misses reduce the voice data available to speaker clusters. Both occurrences result in clustering error on the increase. The MFCCs features, first and second derivatives and the spectral flatness feature are used for the SAD system. The result for experiments is shown in table 1. The details of the SAD algorithm can be found in our previous work [20].

**Table 1:** SAD error rates of model based classification

| Meeting ID | Miss Speech | False Alarm | SAD Error |
|---|---|---|---|
| ES2004a | 5.97 | 21.55 | 27.53 |
| ES2006d | 2.12 | 13.62 | 15.75 |
| ES2007d | 6.08 | 13.91 | 19.99 |
| ES2009b | 4.37 | 9.42 | 13.8 |
| ES2010b | 6.99 | 14.89 | 21.88 |
| ES2011c | 5.56 | 14.61 | 20.17 |
| ES2012d | 7.32 | 12.09 | 19.42 |
| ES2013a | 7.41 | 32.74 | 40.16 |

### *3.2. Detecting Overlap Segments*

#### *3.2.1. Features Combination*

In feature combination set, 17-MFCC, 17- dimension of the first derivatives and the 17-dimension of the second derivatives and spectral flatness feature are used. Moreover, the proposed features gammatone like spectrogram based LPC coefficients are add to feature set for more accurate results. Waveform X (at sample rate 16000) is passed via the auditory filterbank model 64 channel gammatone with the smallest 50HZ frequency and the largest 8000HZ frequency. Each band's outputs then have their energy integrated over windows of 0.025secs, 0.010sec frame step for successive columns. The feature vector X is returned to these magnitudes. Feature vector X is fed to LPC 10 order to remove coffeics from linear predictors. All features are calculated over a 25 ms speech frame with a 10 ms frame step. ANN classification system processes the selected feature vectors.

#### *3.2.2. Classification Method*

The system choosen for an ANN for its discriminating characteristics, its ability to represent non-linear

frameworks and the convenience of the subsequent probabilities it produces for experiments on classification. They are brain-inspired systems, as the "neural" part of their name suggests, that are intended to replicate the way we humans learn. Artificial neural network consists of input and output layers and (in most cases) a hidden layerof units that convert the input into anything that can be used by the output layer [18]. Multiple artificial neuron groups connected together form a neural network. ANNs are flexible and adaptable in nature, meaning that they alter their configuration depending on the info (internally or externally) that passes through the network. Artificial neural networks consist of simple elements, known as nodes, that function in parallel. Training shall be performed until a reasonable output for a particular input is created. The network is adjusted and stops based on the difference between target and output when the difference between targets and output is zero or minimum; i.e. output matches input[19]. The research work used Neural Network Toolbox 11.1 to detect overlap speech and single speaker speech. It is ANN perceptron of multiple layers. It includes layers of inputs, hidden nodes, and outputs. Back propagation algorithm is used to train classifier ANN.

## 4. Result And Discussion

The performance of the overlap speech detection system was evaluated on the AMI dataset. Overlap speech detection results are as shown in tables. These tables showed maximum, average and minimum accuracy results of 8 audios in test set. Figure 3 and 4 showed the gammatone like spectrogram image. Spectrogram was computed in the frequency range from 50 to 8000 Hz using 25 ms Hanning window with a 15ms overlap between windows. This frequency range was chosen because harmonic traces could be seen most clearly in it. To enhance harmonic visibility, we plotted the gammatone like spectrogram with color level from -90 dB to -30 dB.

**Table 2:** The performance of ANN system (Mfcc,Delta,Delta_Delta,SFM)

| Meeting ID | Single | Overlap | Average |
|---|---|---|---|
| ES2004a | 66.58 | 51.95 | 64.18 |
| ES2006d | 70.27 | 66.36 | 69.29 |
| ES2007d | 64.84 | 58.89 | 64.03 |
| ES2009b | 61.78 | 57.59 | 61.43 |
| ES2010b | 64.46 | 60.35 | 64.11 |
| ES2011c | 56.69 | 74.17 | 59.31 |
| ES2012d | 72.81 | 49.92 | 69.14 |
| ES2013a | 69 | 61.51 | 68.38 |

In the first part of experiments, the system used Mfcc, Delta, Delta_Delta and SFM to detect single speaker speech and overlap speech. The result for experiments is shown in table 2. Overlap speech accuracy is higher than ordinary four features plus GTF_LPC order 5 feature set classification. Accuracy of single speaker speech and average are lower than gammarone order 5 feature set classification.

In the second part of experiments, the system used ordinary four features plus GTF_LPC gammatone order 4 to

detect single speaker speech and overlap speech. The result for experiments is shown in table 3. Accuracy drops in single speaker speech and average than ordinary four features plus GTF_LPC order 5 feature set classification.But overlap speech accuracy is high in some audios. Figure 3 show gammatone like spectrogram for order 4. The harmonics are not clearly distinguishable on the spectrogram.

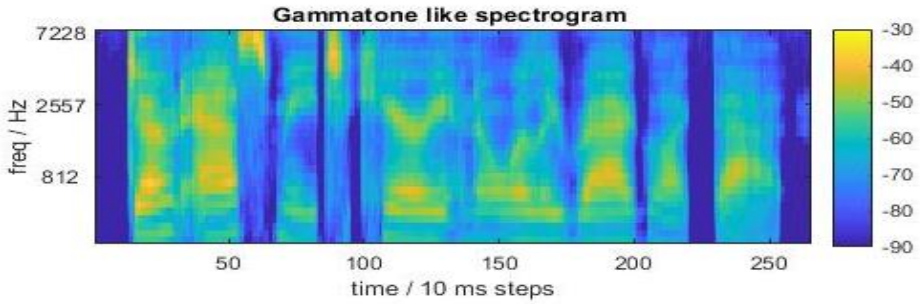


**Figure 3:** Gammatone like spectrogram for order 4

**Table 3:** The Performance of ANN system (4 features plus GTF_LPC with gammatone order 4 )

| Meeting ID | Single | Overlap | Average |
|---|---|---|---|
| ES2004a | 60.99 | 58.44 | 60.57 |
| ES2006d | 63.15 | 71.84 | 65.33 |
| ES2007d | 62.93 | 61.36 | 62.72 |
| ES2009b | 56.62 | 62.59 | 57.1 |
| ES2010b | 59.73 | 64.39 | 60.13 |
| ES2011c | 57.92 | 73.15 | 60.21 |
| ES2012d | 75.01 | 46.99 | 70.51 |
| ES2013a | 69.68 | 57.25 | 68.64 |

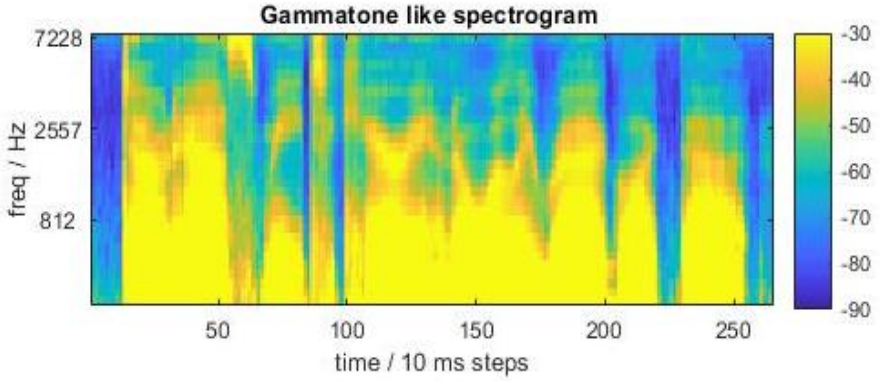


**Figure 4:** Gammatone like spectrogram for order 5

Finally, the system used ordinary four features plus GTF_LPC order 5 feature set to detect single speaker speech and overlap speech. The result for experiments is shown in table 4. The table demonstrates higher

accurate result than other two methods in single speaker speech and average.Overlap speech accuracy is drop. But system aims to use subsequence jobs such as speaker change detection and speaker clustering process. Miss detection of single speaker speech effects on speaker change detection and speaker clustering process . Figure 4 show gammatone like spectrogram for order 5. The harmonics are clearly distinguishable than order 4 on the spectrogram.

**Table 4:** The performance of ANN system (4 features plus GTF_LPC with gammatone order 5 )

| Meeting ID | Single | Overlap | Average |
|---|---|---|---|
| ES2004a | 67.38 | 50.91 | 64.68 |
| ES2006d | 71.11 | 65.85 | 69.8 |
| ES2007d | 65.63 | 57.34 | 64.49 |
| ES2009b | 62.65 | 56.96 | 62.18 |
| ES2010b | 65.56 | 59.4 | 65.03 |
| ES2011c | 57.62 | 73.41 | 60.01 |
| ES2012d | 73.56 | 48.76 | 69.58 |
| ES2013a | 69.40 | 60.28 | 68.64 |

## 5. Conclusion

In summary, research work has developed an single speaker speech and overlap speech detection system based on three feature set with neural network models. Using the model built with the optimal subset of features, the proposed method achieved higher accurate result than other two baseline methods in single speaker speech and average. All training data are used scenario audios, so the accuracy may drop in testing using non scenario audios. It is observed that gammatone like spectrogram and LPC features play a significant role in differentiating both classes. Moreover, combining the proposed method with detection of background noise, laughter and hesitation would be beneficial to prevent the algorithm from misinterpreting those sounds.

## References

[1]. S. Otterson and M. Ostendorf, “Efficient use of overlap information in speaker diarization,” in 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU) (pp. 683-686). IEEE.

[2]. M. Huijbregts and C. Chuck, "The blame game: Performance analysis of speaker diarization system components," In Eighth Annual Conference of the International Speech Communication Association. 2007.

[3]. S.N. Wrigley, G.J. Brown, V. Wan and S. Renals, "Speech and crosstalk detection in multichannel audio," in IEEE Transactions on speech and audio processing 13, no. 1 (2004): 84-91.

[4]. K. Boakye, Kofi, B. Trueba-Hornero, O. Vinyals, and G, Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4353-4356. IEEE, 2008..

[5]. P. Dighe, M. Ferras, and H. Bourlard, "Detecting and labeling speakers on overlapping speech using vector taylor series," in Fifteenth Annual Conference of the International Speech Communication

Association. 2014.

[6]. R. Vipperla, D. Wang, S. Bozonnet, and N. Evans, "Speech overlap detection using convolutive non-negative sparse coding," in (2011).

[7]. S. H. Yella, and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," in IEEE/ACM Transactions on Audio, Speech, and Language Processing 22, no. 12 (2014): 1688-1700.

[8]. K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in Ninth Annual Conference of the International Speech Communication Association. 2008.

[9]. V. Andrei, H. Cucu, and C. Burileanu, "Detecting Overlapped Speech on Short Timeframes Using Deep Learning," in INTERSPEECH, pp. 1198-1202. 2017.

[10]. M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolíková, O. Novotný et al. "BUT System for DIHARD Speech Diarization Challenge 2018," in Interspeech, pp. 2798-2802. 2018.

[11]. N. Sajjan, S. Ganesh, N. Sharma, S. Ganapathy, and N. Ryant, "Leveraging LSTM models for overlap detection in multi-party meetings," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5249-5253. IEEE, 2018.

[12]. N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, and J. H. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2834-2838. IEEE, 2013.

[13]. K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4353-4356. IEEE, 2008..

[14]. N. Shokouhi, A. Ziaei, A. Sangwan, and J. H. HL, "Robust overlapped speech detection and its application in word-count estimation for Prof-Life-Log data," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4724-4728. IEEE, 2015.

[15]. J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France. 2013.

[16]. D. P. W. Ellis, "Gammatone-like spectrograms. web resource." URL: http://www. ee. columbia. edu/~ dpwe/resources/matlab/gammatonegram (2009).

[17]. J. Carletta, A. Simone, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec et al, "The AMI meeting corpus: A pre-announcement," In International workshop on machine learning for multimodal interaction, pp. 28-39. Springer, Berlin, Heidelberg, 2005.

[18]. L. Dormehl,. "What is an artificial neural network? Here's everything you need to know," in Digital Trends (2019).

[19]. T. G. Dietterich, "Ensemble methods in machine learning." In International workshop on multiple classifier systems, pp. 1-15. Springer, Berlin, Heidelberg, 2000.

[20]. T.H. Zaw and M. M. Thaw, "Speech Activity Detection in Multi Party Meetings," in International Journal of Scientific Research and Engineering Development-– Volume 3 Issue 3, May – June 2020