Overlapped/Non-Overlapped Speech Transition Point Detection Using Bag-of-Audio-Words

Shikha Baghel[1], S. R. Mahadeva Prasanna[2], and Prithwijit Guha[1]

[1] Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Assam 781039, India
[2] Department of Electrical Engineering
Indian Institute of Technology Dharwad, Dharwad-580011, India
Email: {shikha.baghel, prasanna, pguha}@iitg.ac.in

***Abstract*—Overlapped speech refers to an audio signal which contains speech of two or more speakers speaking simultaneously. Overlapped speech is one of the main sources of error for speaker diarization systems. This work presents an initial study to identify the transition points of overlapped to non-overlapped speech and vice-versa. Characteristics of overlapped and non-overlapped speech are examined in terms of the vocal tract system, excitation source, and modulation spectrum. The Hilbert envelope (HE) of Linear Prediction (LP) residual signal represents the excitation source characteristics of speech signal. The Sum of Ten Largest Peaks (STLP) of the spectrum and Mel-Frequency Cepstral Coefficients (MFCCs) represent the vocal tract shape information. The modulation spectrum energy (ModSE) captures the information of slowly varying temporal envelope of speech. A Bag-of-Audio-Words (BoAW) based approach is used to detect the transition points. News debates are one of the main sources of naturally occurred overlapped speech. Therefore, the present work is evaluated on Indian news debate scenario. A high Identification Rate (IR) and low Spurious Rate (SR) is observed when all the features are used simultaneously as a $16d$ feature(13-MFCCs, HE of LP residual, STLP and ModSE) for the detection task.**

***Index Terms*—Overlapped speech, MFCCs, excitation source, Hilbert envelope, vocal tract system, modulation spectrum, Bag-of-Audio-Words**

I. Introduction

Overlapped speech is produced when two or more speakers speak simultaneously. It is considered as one of the main sources of error for diarization systems [1], [2]. Ryant et al. [3] discussed the importance of handling overlapped speech for speaker diarization. Conventional speech processing applications such as speech and speaker recognition, consider speech only from a single speaker. Hence, the overlapped speech regions need to be identified and processed separately. This requires the detection of transition points from non-overlapped to overlapped speech and vice-versa. In this work, non-overlapped speech refers to the audio signal containing the speech of only one speaker at a time. This study aims to detect such transition points in news debate audio. The frequent presence of overlapped speech in news debates makes it appropriate to consider for this study.

978-1-7281-8895-9/20/$31.00 ©

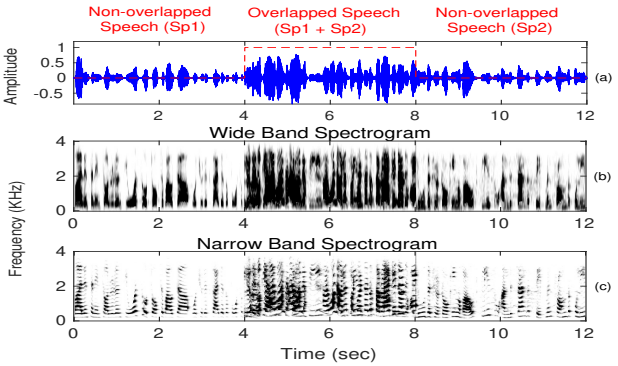

Fig. 1. Illustrating the Spectrogram. (a) Speech signal, High intensity of (b) wideband spectrogram, and (c) narrowband spectrogram for overlapped speech than non-overlapped speech.

Overlapped speech can be produced in a competitive or a non-competitive scenario [4]. Competitive overlapped speech is produced when two or more speakers are in a competition to grab the opportunity for speaking, and thus they continue to speak simultaneously for a significant duration [4]. Overlapped speech present in news debates is considered as the competitive one. In a non-competitive scenario, speakers co-operate with each other and allow others to speak. In such cases, overlapping occur for a small duration [4]. Overlaps present in conversational speech are an example of a non-competitive scenario.

The signal characteristics of non-overlapped speech vary significantly from overlapped speech. These deviations in the characteristics can be observed from the spectrograms shown in Fig. 1. Spectrum for overlapped speech is harmonically richer than non-overlapping speech due to the presence of more than one fundamental frequency (F_0). This can be observed from the narrowband spectrum shown in Fig. 1(c). A wideband spectrogram is shown in Fig. 1(b), which shows higher energy for overlapped speech than non-overlapped regions. Fig. 1(a) shows a 12 sec long speech signal contains non-overlapped speech for first and last 4 sec, and overlapped speech for middle 4 sec. These 4 sec long overlapped and non-overlapped speech segments are taken from a news debate audio.

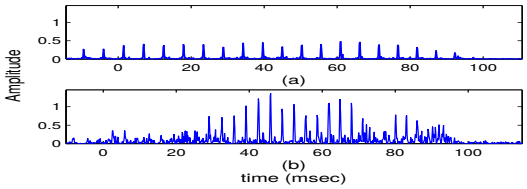

Fig. 2. Illustrating HE of LP residual. (a) Non-overlapped speech contains a lower residual, (b) a higher residual is exhibited for overlapped speech.

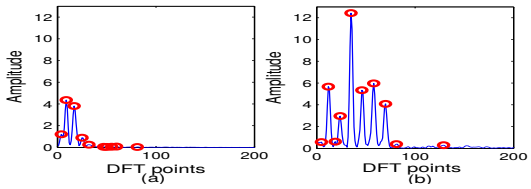

Fig. 3. Illustrating spectral peaks for one frame of (a) non-overlapped speech, which shows lower spectral peaks amplitude, and (b) overlapped speech with higher spectral peaks amplitude.

Different nature of spectrograms for overlapped and non-overlapped speech motivates to study the speech characteristics for transition point detection. An enhanced time-frequency based representation called pyknogram has been used for tracking harmonic patterns present in speech signal for detecting overlapped speech [2]. Yousefi et al. [5] proposed two features derived from online Convolutive Non-negative Matrix Factorization (CNMF). Boakye et al. [6] explored spectral flatness, harmonic energy ratio, modulation spectrogram features, and MFCC features for overlapping speech detection in distant microphone audio. The usefulness of Fundamental frequency (F_0) and related features have also been explored [7], [8]. Some works have also utilized prosodic and voice quality features such as loudness, voice-probability Jitter, shimmer, and logarithmic harmonics-to-noise ratio (logHNR) [8]. Linear prediction (LP) residual energy and LP coefficients also show discrimination between non-overlapped and overlapped speech [9].

This work presents an initial study done in the direction of overlapped/non-overlapped transition point detection in news debate scenarios. Excitation source, vocal tract and modulation spectrum characteristics of a speech signal are explored for this work (section II). The Hilbert Envelope (HE) of LP residual (sub-section II-A), Sum of Ten Largest Peaks (STLP) (sub-section II-B), Modulation Spectrum Energy (ModSE) (sub-section II-C) and MFCC (sub-section II-B) features are studied. The Bag-of-Audio-Words (BoAW) approach is used to transform these speech features into distribution based representation (section III). Transition points are detected based on the dissimilarity of these distributions (section IV). The proposed approach is evaluated on a news debate dataset and the results are discussed in section V. Section VI concludes the present work and discusses the future directions.

II. Features for Overlapped/Non-overlapped Speech Transition Point Detection

This section explains the speech features used for transition point detection. Speech signal (sampling rate, $F_s = 8$ kHz) is processed with a frame size of 20 ms, and a shift of 10 ms to extract features.

A. Excitation Source Feature

The LP residual signal captures the excitation source information of a speech signal [10]. The LP residual represents the error in predicting the current sample based on the past p samples. This error is expected to be higher in overlapped speech due to the presence of more than one speaker's speech. Thus, the LP residual signal exhibits different behavior for both the speech cases.

HE of LP Residual– A 12^{th} order LP analysis is performed to obtain the corresponding LP residual signal. The time-varying changes of the excitation characteristics are smeared in the LP residual due to its bipolar nature. These changes are further enhanced by computing the HE of the LP residual [10]. Fig. 2 illustrates the higher residual error for overlapped speech (Fig. 2(b)) than that in non-overlapped one (Fig. 2(a)). Similarly, Fig. 5(b) represents the HE of LP residual (blue color), which shows the higher values for overlapped speech compared to the non-overlap one.

B. Vocal Tract System Features

Vocal tract shape can be represented in terms of the formants and the envelope of the short-time power spectrum of the speech signal. MFCCs capture the information of power spectrum envelope by taking human perception into consideration. Short-time spectra of speech signal show spectral peaks corresponding to formant locations [11]. Thus, the first ten largest peaks of the spectrum can be considered for representing the formant information. Spectra of overlapped speech are expected to have sufficiently higher energies distributed up to high frequencies due to the superimposition of two speech signals. However, spectrum energies are mostly concentrated towards low frequencies for non-overlapped speech (Fig. 1(b)). Therefore, the features such as MFCCs and Sum of Ten Largest Peaks (STLP) are worth exploring in this study.

Sum of Ten Largest Spectral Peaks (STLP)– Ten largest peaks are picked from the magnitude spectrum of each frame and summed to obtain the STLP feature. Fig. 3 illustrates the ten largest peaks (highlighted in red circles) in spectrum for one frame of overlapped (Fig. 3(b)) and non-overlapped (Fig. 3(a)) speech. This figure shows the discriminative behavior of the STLP feature for both the classes. The STLP feature (blue color) is plotted in Fig. 5(c) for a 12 sec long speech signal. This figure shows higher values of STLP for overlapped speech than that in the non-overlapped regions.

Mel Frequency Cepstral Coefficients (MFCC)– The power spectrum of each frame is mapped onto the mel scale using a mel filter bank with 26 overlapping triangular filters. The first 13 coefficients of MFCCs are considered for the detection of transition points.

C. Modulation Spectrum Feature

Modulation refers to the slowly varying temporal envelope of speech. The envelope of speech can be varied according

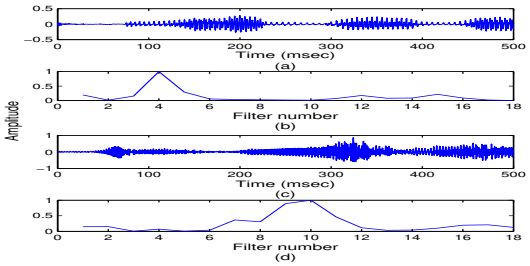

Fig. 4. Illustrating Modulation spectrum energy. (a) Non-overlapped speech, (b) Modulation spectrum energy components from the critical band filters for non-overlapped speech, (c) Overlapped speech, (d) Modulation spectrum energy components from the critical band filters for overlapped speech which is higher than non-overlapped speech.

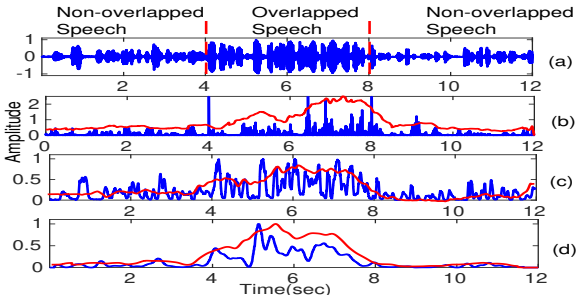

Fig. 5. Illustrating speech specific features. (a) Speech signal of 12 sec duration, (b) HE of LP residual, (c) Sum of ten largest spectral peaks (STLP), (d) Modulation spectrum energy (ModSE). The blue color plots ((b), (c) and (d)) represent raw features, while the corresponding smoothed features are plotted in red color.

to the number of sound units spoken per unit time, which is known as the syllabic rate of speech. In the case of overlapped speech, the syllabic rate is expected to be higher due to the superimposition of more than one speech signal.

Modulation Spectrum Energy (ModSE)– The syllabic rate can be represented in terms of the modulation spectrum energy. The extraction of modulation spectrum energy can be found in detail in [12]. A higher modulation spectrum energy is expected for overlapped speech (Fig. 4(d)) than non-overlapped one (Fig. 4(b)). Fig. 4 illustrates the modulation energy distribution of 4 Hz component for a frame. However, Fig. 5(d) illustrates the ModSE feature (blue color) for a 12 sec long speech signal.

III. Methodology: Bag of Audio Words

The Bag-of-Audio-Words (BoAW) approach is motivated by the Bag-of-Words (BoW) representation used in text analysis. The BoW approach is basically used to represent text documents. The words appearing in the natural language are considered as the units in BoW (text file), and thus these units are discrete one. While, in BoAW (audio file) approach, audio words are not discrete. However, audio words are obtained by a clustering method to demonstrate the original feature space perfectly. The BoAW approach provides a fixed size histogram as a feature.

The BoAW approach is described in Fig. 6. First, K-Means clustering is performed on the extracted features to obtain representative audio words (basic units) for the BoAW approach (Fig. 6(a)). The number of clusters (k) is varied from 2 to 10 and finalized the one at which the minimum detection error is achieved. For this work, $k = 5$ is used. The set of k centroids act as the code-book of the system which is given as

$$CB = \{\mathbf{c_1}, \mathbf{c_2}, \ldots \mathbf{c_k}\} \tag{1}$$

where, $c_j; j = 1, 2, \ldots K$ are the K centroids. These centroids act as the primary words (basic units) that are considered to be present in an input signal. These words (centroids) are termed as audio words to highlight the fact that they are associated with atomic and perceptual units of hearing, and not to the linguistic units. Further, the learned code-book is used to label an input data which is mathematically given as

$$l_i = arg \min_j |f_i - c_j|; \quad j = 1, 2, \ldots k \tag{2}$$

where, a label l_i is assigned to the feature vector f_i. This step is termed as the vector quantization. A second level of feature extraction is needed to execute the transition point detection. After vector quantization, a fixed size feature vector is generated by considering the frequency of each code-word in a given speech signal. This results in a histogram representing the word vector. Mathematically, the histogram is constructed as

$$\mathbf{w_j} = \sum_{i=1}^{N} \delta(l_i, j); \quad j = 1, 2, \ldots K \tag{3}$$

where, $\delta(.)$ denotes the Kronecker delta, l_i is the label of i^{th} feature vector and N is the total number of feature vectors. The frequency of occurrence of j^{th} code-word in a given duration is represented by $w_j; j = 1, 2, \ldots K$ (Fig. 6(b)). Histograms are considered as the second level representation of features. The dissimilarity between two consecutive histograms are calculated for transition point detection. A higher similarity between two histograms indicates their belongingness to thr same category.

IV. Overlapped/Non-Overlapped Speech Transition Point Detection

Section II illustrates the different behavior of features for overlapped and non-overlapped speech. The evidences from all the features need to be combined for the effective transition point detection. The HE of LP Residual, STLP, and ModSE are combined to create a 3 dimensional feature ($3d$ feature). MFCCs($13d$) and $3d$ feature are combined to create a $16d$ feature for the detection task. In Fig. 5(b), (c) and (d), some lower feature values (blue color) are observed in the overlapped speech region than that in the non-overlapped regions. These lower values observed in overlapped speech may be attributed to the presence of silence or non-speech regions. This may affect the overall detection performance. Therefore, smoothing is performed over a period of 1 sec on the raw features to remove such spurious regions. This is plotted by red color in Fig. 5(b), (c) and (d).

In the BoAW approach, the transition point detection is based on the dissimilarity between the distribution of two audio files. The detection is performed using $3d$ features, 13-MFCCs, and $16d$ feature, separately. These features are the

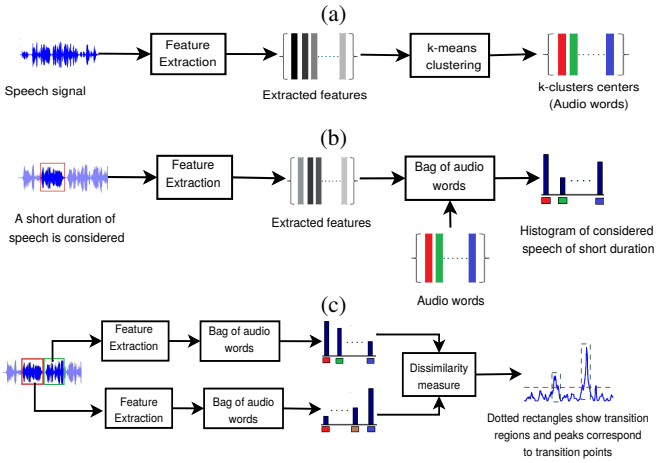

Fig. 6. Illustrating Bag-of-Audio-Words approach. (a) Code-book generation from the extracted features of the entire speech, (b) Histogram generation with the help of code-book and features extracted from short-duration speech, (c) Transition point detection based on the dissimilarity between two consecutive histograms.

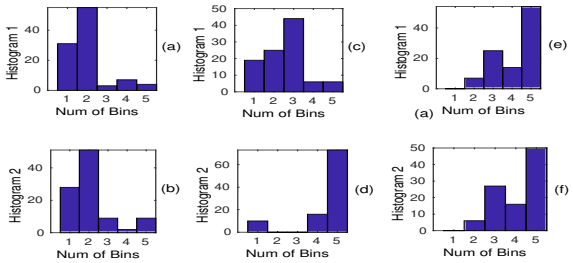

Fig. 7. Demonstrating histograms for, non-overlapped speech which shows similar shape for (a) histogram 1 and (b) histogram 2, transition from non-overlap to overlapped speech which shows different shape for both the histograms (c) and (d), overlapped speech which shows similar shape for both histograms (e) and (f).

first level of feature representation and considered as inputs for the BoAW approach. Fig. 6(c) illustrates the transition point detection approach used in this work. The BoAW approach transforms first level features into the labeled data using vector quantization. The labeled data is further processed in the following manner. A time instance t is considered as a counter at which the decision of transition point is to be made. This counter is moved across the entire speech file with an increment of 10 ms. A duration of 1 sec is considered on both sides of this t^{th} instance to compute the histograms for both of these 1 sec intervals (Fig. 6(c)). If these histograms belong to the same speech region i.e., either overlapped or non-overlapped speech, then the shape of these histograms is expected to be similar. This results in a low dissimilarity value between the two histograms. The dissimilarity value is expected to increase as the counter t moves from one speech region to others and is maximum when t^{th} time instance is the transition point. At such points, the shape of both of the histograms is different as they belong to different speech categories. Fig. 7 shows the similar shape for histogram 1 and histogram 2 for both the categories i.e. overlapped (Fig. 7(e) and (f)) and non-overlapped speech (Fig. 7(a) and (b)). Fig. 7(c) and (d) show two histograms for the transition from non-overlapped to overlapped speech. The shape of these histograms is different since histogram 1 (Fig. 7(c)) corresponds to non-overlapped speech, and histogram 2 (Fig. 7(d)) belongs to the overlapped speech. Therefore, a higher dissimilarity value is expected during the transitions than the homogeneous region (i.e., either overlapped or non-overlapped region), and is maximum at the transition point.

The dissimilarity between two distributions p and q is calculated by the Bhattacharyya distance which is given as

$$BD(p,q) = 1 - (\sum_{j=0}^{N} \sqrt{p(j)q(j)}) \quad (4)$$

where, $BD(p,q)$ is the Bhattacharyya dissimilarity, N is the total number of bins in histograms. The Otsu thresholding, an adaptive thresholding approach, is performed on the dissimilarity values calculated by using Eq. 4. The regions over which dissimilarity crosses the threshold are considered as the regions of interest i.e., the transition regions. The differentiation for the detected transition regions is performed to detect the peaks in the dissimilarity values and mark these peak locations as the transition points. A tolerance window of 50 ms is used to declare a detected point as a true transition point.

V. Results and Discussion

The proposed approach is evaluated on news debates broadcasted in an Indian news channel. Short audio files of 12 sec duration are generated from broadcast news debates. The first and last 4 sec of each audio file contains the non-overlapping speech of two different speakers (say, Sp_1 and Sp_2). The middle 4 sec (i.e. the speech from 4 to 8 sec) contains overlapped speech of two speakers (say, $\text{Sp}_1 + \text{Sp}_2$). This 4 sec of overlapped speech is taken from naturally occurred instances of overlapping speech of the same news debate. Such 12 sec long speech files (Fig. 1(a)) are synthetically generated by concatenating 4 sec of non-overlapped speech of Sp_1, 4 sec of overlapped speech of $\text{Sp}_1 + \text{Sp}_2$ and 4 sec of non-overlapped speech of Sp_2. For the evaluation of current work, 256 such files are used. The motive of the synthetic generation of speech files is to make sure the presence of overlapped speech for sufficient duration for analysis purposes. The present work studies the overlapped speech containing only two simultaneous speakers. Speech signals are resampled to 8 kHz.

The performance of the proposed method is measured in terms of the Identification Rate (IR) and Spurious Rate (SR). IR is the percentage of correctly identified transition points, and SR is the percentage of falsely identified transition points. The performance in terms of IR and SR for different features with respect to different threshold values is mentioned in the TABLE I. The present work is evaluated for three different thresholds η_1, η_2, and η_3 as $1.5, 1.3$, and 1.1 times of the Otsu threshold of respective features and observed the similar performance for these three thresholds. The IR for the $3d$ feature is lower than the $13d$ feature and $16d$ features. Since $3d$

TABLE I
RESULTS IN TERMS OF IDENTIFICATION RATE (IR) AND SPURIOUS RATE (SR)

Threshold ↓	Speech specific features					
	$3d$ feature		$13d$ feature		$16d$ feature	
	IR	SR	IR	SR	IR	SR
$\eta_1 = 1.5 \times \eta_{otsu}$	60.18	39.81	72.88	27.11	74.85	24.85
$\eta_2 = 1.3 \times \eta_{otsu}$	57.79	42.20	72.06	27.93	74.56	25.43
$\eta_3 = 1.1 \times \eta_{otsu}$	57.58	42.41	70.68	29.31	74.17	25.82

feature contains STLP as the vocal tract feature which is one dimensional and may not capture all the aspects of vocal tract shape which may be captured by the 13 dimensional MFCCs. The IR for $13d$ feature is almost comparable but lower than the IR of $16d$ feature. Since, $16d$ feature considered excitation source and modulation spectrum along with the vocal tract shape while $13d$ feature considered only vocal tract shape information. The SR is highest for the $3d$ feature and lowest for the $16d$ features. Therefore $16d$ feature is preferable for the overlapped/non-overlapped speech transition point detection than the other two features.

The present approach is also evaluated on a short segment (32 sec duration) of news debate. This short segment contains naturally occurred transitions in a news debate scenario. The speech signal of this short segment is shown in Fig. 8(a), where red solid vertical lines represent the actual transition points. Fig. 8(b), (c) and (d) show the detected transition points by solid vertical blue lines using $3d$ feature, $13d$ feature and $16d$ feature, respectively. Some spurious transition points are also detected by the proposed approach. Those are highlighted by the green dotted rectangles (Fig. 8(b), (c) and (d)). It can be observed that the number of spurious transition points is large in case of $3d$ feature in comparison with the other two features. All actual transition points are detected by using $13d$ feature and $16d$ feature, but the number of spurious transition points is more in case of $13d$ feature than $16d$ feature. This shows that the $16d$ feature is more suitable for the task than the other two features.

VI. CONCLUSION AND FUTURE DIRECTIONS

Speech specific features are explored for the overlapped/non-overlapped speech transition point detection for news debate scenario. The HE of LP residual, Sum of Ten Largest spectral Peaks (STLP), Modulation Spectrum Energy (ModSE), and MFCCs are studied for this work. This work utilizes the excitation source, modulation spectrum, and vocal tract characteristics for the transition point detection task.

This work is an initial attempt in the direction of characterizing overlapped speech and detecting the transition points. The present work studies overlapped speech of only two simultaneous speakers. As such, we intend to extend the work by doing analysis on a large data taken from news debates considering overlapped speech of more than two simultaneous speakers. The harmonic patterns present in the spectra of non-overlapped speech are expected to be disturbed in overlapped speech. Therefore, harmonic patterns present in speech signals need to be explored as an extension of this work.

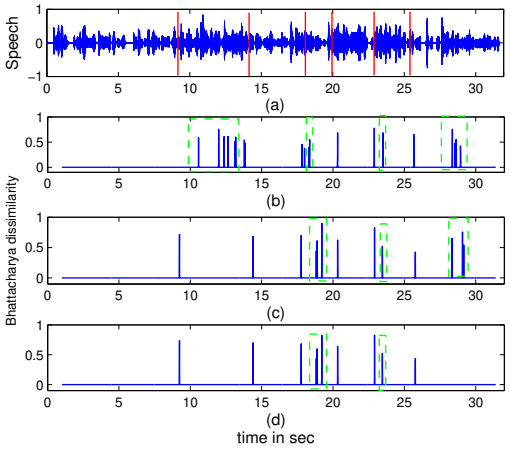

Fig. 8. Illustrating the transition point detection in a 32 sec long news debate segment containing naturally occured transitions: (a) Speech signal with solid red lines represent actual transition points, transition point detection using (b) $3d$ feature which shows more number of spurious detection highlighted by green dotted rectangles, (c) $13d$ feature which shows comparatively lesser number of spurious detection than $3d$ feature, and (d) $16d$ feature which shows least number of spurious detection.

REFERENCES

[1] M. Moattar and M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, no. 10, pp. 1065–1103, 2012.

[2] N. Shokouhi and J. H. L. Hansen, "Teagerkaiser energy operators for overlapped speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1035–1047, May 2017.

[3] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First dihard challenge evaluation plan," *2018, tech. Rep.*, 2018.

[4] S. A. Chowdhury, M. Danieli, and G. Riccardi, "Annotating and categorizing competition in overlap speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5316–5320.

[5] M. Yousefi, N. Shokouhi, and J. H. Hansen, "Assessing speaker engagement in 2-person debates: Overlap detection in united states presidential debates," in *Interspeech*, 2018, pp. 2117–2121.

[6] 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, pp. 32–35.

[7] Yang Shao and DeLiang Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, vol. 2, 2003, pp. II–205.

[8] E. Kurti, G. J. Brown, and B. Wells, "Resources for turn competition in overlapping talk," *Speech Communication*, vol. 55, no. 5, pp. 721 – 743, 2013.

[9] 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, pp. 1668–1672.

[10] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 309–319, Aug 1979.

[11] B. K. Khonglah and S. R. M. Prasanna, "Speech / music classification using speech-specific features," *Digital Signal Processing*, vol. 48, pp. 71–83, 2016.

[12] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 1997, pp. 1647–1650.