Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations

Sree Harsha Yella, *Student Member, IEEE*, and Hervé Bourlard, *Fellow, IEEE*

***Abstract*—Overlapping speech has been identified as one of the main sources of errors in diarization of meeting room conversations. Therefore, overlap detection has become an important step prior to speaker diarization. Studies on conversational analysis have shown that overlapping speech is more likely to occur at specific parts of a conversation. They have also shown that overlap occurrence is correlated with various conversational features such as speech, silence patterns and speaker turn changes. We use features capturing this higher level information from structure of a conversation such as silence and speaker change statistics to improve acoustic feature based classifier of overlapping and single-speaker speech classes. The silence and speaker change statistics are computed over a long-term window (around 3-4 seconds) and are used to predict the probability of overlap in the window. These estimates are then incorporated into a acoustic feature based classifier as prior probabilities of the classes. Experiments conducted on three corpora (AMI, NIST-RT and ICSI) have shown that the proposed method improves the performance of acoustic feature-based overlap detector on all the corpora. They also reveal that the model based on long-term conversational features used to estimate probability of overlap which is learned from AMI corpus generalizes to meetings from other corpora (NIST-RT and ICSI). Moreover, experiments on ICSI corpus reveal that the proposed method also improves laughter overlap detection. Consequently, applying overlap handling techniques to speaker diarization using the detected overlap results in reduction of diarization error rate (DER) on all the three corpora.**

***Index Terms*—Meeting room recordings, simultaneous speakers, speaker diarization, spontaneous conversations, spontaneous overlapping speech.**

I. Introduction

OVERLAPPING speech occurs when there is more than one speaker speaking at any given instant of time in an audio recording. This is a very common phenomenon in spontaneous conversations like meeting room discussion, telephone conversations, television chat shows and other similar media [1], [2]. The factors causing overlapping speech in a multi-party conversation are diverse [3], [4], [5], [6]. It can occur when listeners use back-channels to show their involvement and also to convey their agreement with the foreground speaker. It also most commonly occurs when one or more participants try to interrupt the foreground speaker and take the conversation floor. In informal conversations among multiple participants, there are situations where multiple parallel conversations (schism) [7] take place involving several participants in each sub-conversation. Also, it has been observed that overlaps are a common phenomenon during conversation floor exchanges among speakers. Apart from the above mentioned broad patterns of occurrences of overlaps, there are always some idiosyncrasies specific to a particular conversation or participant that can cause overlapping speech at any place during the conversation. Previous studies have shown that the error rates of automatic speech processing systems increase when processing speech from multiple simultaneous speakers [8], [3]. Several diagnostical studies on speaker diarization systems have also shown that overlapping speech is one of the main sources of error in state of the art speaker diarization systems [9], [10], [11].

Several previous works have proposed methods to detect overlapping speech in meeting room conversations. Earlier works have concentrated on detecting overlapping speech in audio captured using head/lapel microphones worn by the participants in the conversation [12], [13], [14]. These works have focussed on issues arising from cross-talk, breath noise and channel variations across different close talking microphones used to capture the audio. Pfau *et al.* [13] have proposed a hidden Markov model (HMM) based approach to infer the sequence of hidden states speech and non-speech in each participant's channel. They have also proposed a method to detect overlapping speech segments by putting a threshold on the cross correlation value between speech signals of multiple channels.

Wrigley *et al.* [12] proposed a more generalized approach to multi-channel speech activity detection, where a HMM with four states, single-speaker speech, cross-talk, overlapping speech and non-speech was used. They explored various acoustic features useful for this task and found kurtosis, 'fundamentalness' and cross-correlation related features to be most effective. Laskowski *et al.* [14] have proposed a method to improve multi-channel speech activity detection in overlapping speech regions by modelling the turn taking behavior of participants in the conversation. These works were mainly concerned with improving speech activity detection on the

Manuscript received January 13, 2014; revised June 01, 2014; accepted July 16, 2014. Date of publication August 08, 2014; date of current version August 23, 2014. This work was supported by the Swiss National Science Foundation (SNSF) by an SNF-RODI grant. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Brian Kingsbury.

The authors are with Idiap Research Institute, Martigny CH-1920, Switzerland (e-mail: shyella@idiap.ch; bourlard@idiap.ch).

Digital Object Identifier 10.1109/TASLP.2014.2346315

meeting room data captured by close-talking microphones with an aim to facilitate reliable automatic speech recognition (ASR). Speaker diarization in its standard setup is evaluated on distant microphone speech where there is no one-to-one correspondence between the speakers and the channels used in the recording. So, recent works on overlap detection and speaker diarization have focussed on detecting the overlapping speech in recordings captured using distant microphones.

Otterson *et al.* [15] trained overlapping speech models using synthesized overlapping speech obtained by adding multiple single-speaker speech utterances. But, experiments revealed that though the trained models were effective in detecting artificially synthesized overlaps, they did not generalize well to naturally occurring overlaps in meeting conversations. The same authors [16] proposed a two step method to handle overlapping speech in speaker diarization assuming oracle overlap detection. Boakye *et al.* [17], [18], [19] explored various acoustic features for overlapping speech detection in distant microphone audio. They found features such as Mel-frequency cepstral coefficients (MFCC), energy, spectral flatness as being the most useful features for overlap detection. Experiments on meeting recordings from AMI meeting corpus [20] have shown that overlap detection is possible with reasonable accuracy which in turn reduced the diarization error rate (DER).

Huijbregts *et al.* [21] proposed a method for overlapping speech detection in a two pass speaker diarization system by training overlapping speech models from speech surrounding speaker changes hypothesized by an automatic speaker diarization system. The trained overlap model was used in the second pass along with speaker models in Viterbi decoding to identify overlapping speech segments. Zelenak *et al.* [22], [23], [24] have proposed the use of time delay of arrival (TDOA) based features extracted from the cross-correlation of speech signals captured by multiple distant microphone channels to improve short-term spectral feature based overlap detection. The cross-correlation based features were used along with short-term spectrum based features as two parallel feature streams in a multistream overlapping speech detection system. To reduce the dimensionality of the cross-correlation based feature vector, and also to make it independent of the number of distant microphone channels used for the recording, they have proposed a way to transform these features to a fixed dimension feature space by applying principal component analysis (PCA) or artificial neural networks (ANNs). Experiments have shown that the cross correlation based features improve the overlapping speech detection. Zelenak *et al.* [25] have also shown improvements in overlapping speech detection by the use of prosodic features. More recently, convolutional non-negative sparse coding based approaches have been successfully applied to the problem of overlap detection [26], [27].

All the methods described above rely directly on features computed from the acoustic signal captured by distant microphones. However, in the recordings of meeting room conversations by distant microphones, the speech signal is often corrupted by background noise resulting in a recording with low signal to noise ratio (SNR). In such scenarios, it is important to explore higher level information present in the structure of a conversation which carries useful cues in modelling the occurrence of overlapping speech. Such information if captured effectively, could potentially be more robust to noisy conditions in the recording. It can also be easily transferred across different corpora of multi-party conversations irrespective of the recording conditions and meeting room setup, which have a direct influence on acoustic features. Studies on conversational analysis have shown that overlaps are more likely to occur at specific locations in a conversation such as speaker turn changes [8], [4] and have also shown that single-speaker speech, silence and overlapping speech patterns are related to each other [28], [29], [30].

Motivated by these studies, the present work proposes the use of features that can be easily extracted automatically from conversations such as silence [31] and speaker change statistics [32] to capture higher level information in a conversation that is relevant for overlap detection. These features are extracted from a long context of about 4 seconds surrounding a given time instant and are used to estimate the probability of occurrence of overlap at that instant. We also explore methods to combine the complementary information present in the two features, silence and speaker change statistics, to improve upon the probability estimates obtained from either of the individual features. These probability estimates are incorporated into the acoustic feature based classifier as prior probabilities of the overlapping and single-speaker speech classes. We report experiments on the AMI [20], NIST-RT [33] and ICSI [34] meeting corpora to validate the generalizability of the proposed method in overlap detection. In these experiments, we demonstrate that the model that is trained to estimate the probability of occurrence of overlap using the meetings from AMI corpus generalizes to other meeting corpora such as NIST-RT and ICSI. We also report speaker diarization experiments on AMI and NIST-RT data sets to evaluate the subsequent improvements in speaker diarization achieved due to improvements in overlap detection.

The paper is organized as follows. Section II provides a brief overview of the various datasets used in the present work. Section III describes the baseline systems for speaker diarization, overlap detection and briefly explains overlap handling mechanism typically used in speaker diarization systems. Section IV explains the proposed conversational features based on silence and speaker change statistics to estimate the probabilities of single-speaker and overlapping speech classes along with the details of how these features are combined to improve the probability estimates from individual features and how these probabilities are incorporated into acoustic feature based system. Section V presents the experimental results on overlap detection studies and their impact on speaker diarization and Section VI summarizes the conclusions of the present work.

II. Data-sets Used

For the experiments in the present work, we use meeting room recordings from three different corpora namely, AMI, NIST-RT and ICSI. The audio captured using multiple distant microphones is enhanced by performing beamforming using BeamformIt toolkit [35]. The acoustic features for speaker

diarization and overlap detection are extracted from this enhanced signal.

A. AMI Meeting Corpus

The AMI corpus [20] consists of about 100 hours of meeting recordings recorded at multiple sites (Idiap, TNO, Edinbrough). The corpus contains both natural and scenario based meetings. The corpus is annotated at multiple levels with several low-level and higher-level information such as speakers segments, word transcripts, summaries, dialogue acts etc. For the experiments reported in the present work we have randomly selected 35 meetings for training set and 25 meetings for the test set (referred as AMI-test set) and 10 meetings for the development set. Meetings from all the recording sites are present in training, testing and development sets.

B. NIST RT Meeting Corpus

The NIST RT meeting corpus [33] consists of meetings belonging to NIST RT speaker diarization evaluation campaigns of 2005, 2006, 2007 and 2009. Each set consists of meeting recordings from multiple sites and varying number of speakers. The corpus also contains ground-truth speaker segmentation obtained by force-aligning the manual transcripts of individual head-microphone channels. The meetings from RT 05, 06, 07 are used as training set while the RT 09 set of meetings is used for testing.

C. ICSI Meeting Corpus

The ICSI meeting corpus [34] contains 75 meetings which contain on the whole around 72 hours of spontaneous multiparty conversation recordings recorded at ICSI, Berkeley. All meetings were recorded by both close talking and distant microphones and contain word level orthographic transcription and speaker information. For the experiments in the present work, we have used meetings from the groups Bmr, Bro, Bed. We have randomly picked 35 meetings as part of training set and 15 meetings as part of the test set (referred as ICSI-test set). Both training and test sets have meetings from all the three groups.

III. Baseline Speaker Diarization and Overlap Detection Systems

A. Speaker Diarization System and Overlap Handling

The baseline speaker diarization system is a non-parametric system based on the information bottleneck (IB) framework [36], [37], [38], [39]. The method has been shown to give similar performance to that of parametric systems based on the HMM/GMM framework [40] with the advantage of significantly less running time. The IB method of clustering is a distributional clustering algorithm that clusters items with similar distributions over a set of variables known as relevance variables. It was initially applied to the task of document clustering, using the set of words in the documents as a relevance variable set [41]. In this scenario, documents containing similar distributions over words were clustered together. In the case of speaker diarization, components of a background GMM estimated over speech regions of a given multi-party speech recording are used as a set of relevance variables. This is motivated from state-of-the-art methods in speaker identification where GMMs are used as universal background models (UBMs). Let $X = \{x_1, x_2, \ldots, x_M\}$ denote the set of input variables that need to be clustered and let $Y = \{y_1, y_2, \ldots, y_N\}$ denote the set of relevance variables that carry meaningful information about the desired clustering output $C = \{c_1, c_2, \ldots, c_P\}$. The IB method aims to find the optimal clustering by maximizing the function below:

$$\mathcal{F} = I(C,Y) - \frac{1}{\beta} I(C,X) \tag{1}$$

where β is a Lagrange multiplier, $I(C,Y)$ denotes mutual information between the set of relevance variables Y and the clustering output C, and similarly $I(C,X)$ denotes mutual information between the input variables set X and the clustering output C. By maximizing $\mathcal{F}$ in (1), the clustering algorithm aims at preserving as much information as possible about the relevance variables in the final clustering i.e., maximizing $I(C,Y)$ while being as compact as possible by minimizing mutual information $I(C,X)$ between the input variable set X and the clustering output C. The IB function $\mathcal{F}$ can be maximized in several ways, the current system [37] uses a greedy agglomerative solution to the optimization. The clustering starts with uniform (over) segmentation of speech regions, which are treated as set of input variables X. The set of relevance variables Y is denoted by components of background GMM estimated over these speech regions. The agglomerative clustering is initialized with each member of set X as an individual cluster and then at each clustering step of IB method, the two clusters that have most similar distributions over the relevance variables are combined. The similarity is obtained in the form of loss in the IB function $\mathcal{F}$, resulting due to the merge of two clusters c_i, c_j as,

$$\nabla\mathcal{F}(c_i,c_j) = JS(P(Y|c_i), P(Y|c_j)) - \frac{1}{\beta} JS(P(X|c_i), P(X|c_j)) \tag{2}$$

where $JS()$ stands for Jensen-Shannon divergence between two distributions and is given by,

$$\begin{aligned} JS(P(Y|c_i), P(Y|c_j)) = {} & \pi_i KL(P(Y|c_i), P(Y|c_{ij})) \\ & + \pi_j KL(P(Y|c_j), P(Y|c_{ij})) \end{aligned} \tag{3}$$

where KL() stands for Kullback-Leibler divergence between two distributions, $\pi_i = \frac{P(c_i)}{P(c_i)+P(c_j)}$, $\pi_j = \frac{P(c_j)}{P(c_i+P(c_j))}$ and c_{ij} is the cluster formed after the merge of the clusters c_i and c_j. The relevance variable distribution of the cluster formed due to the merge is obtained by averaging the relevance variable distributions of the individual clusters in the merge. At each step, two clusters that result in the lowest value of $\nabla\mathcal{F}$ are merged into one cluster. The stopping criterion is based on a threshold over the normalized mutual information $\frac{I(C,Y)}{I(X,Y)}$. Once the final clusters have been obtained, a re-alignment step is performed by estimating a GMM from the data assigned to each cluster and using these cluster models to perform Viterbi decoding with a minimum duration constraint. This step is intended to correct the errors in the segmentation introduced due to initialization of clustering by uniform segmentation of speech regions. The diarization output is evaluated using a metric called diarization error rate (DER), which is a standard metric used in

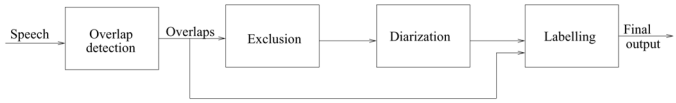

Fig. 1. Block diagram of overlap handling for speaker diarization: Overlaps obtained from overlap detection are first excluded from diarization process and later, appropriate speaker labels are assigned to the overlap segments based on methods described in Section III-A1 in the labeling stage, to generate the final diarization output.

NIST-RT evaluation campaigns [33]. DER is the sum of speech/non-speech error and speaker error. Speech/non-speech error is the sum of miss and false alarm errors by the automatic speech/non-speech detection system. Speaker error is the clustering error happening whenever speech segments of a speaker are attributed to a different one. A forgiveness collar of ± 0.25 seconds is applied around the reference segment boundaries while scoring the automatic systems' output.

1) Overlap Handling for Speaker Diarization: Overlapping speech causes errors in speaker diarization system in two ways. First, it introduces impure segments containing speech from multiple speakers into the clustering process. Secondly, the overlap segments are scored n times where n is the number of speakers in the overlap, which increases the missed speech error even if one of the speakers is not assigned correctly. To avoid these errors, Otterson *et al.* [16] have proposed a method to handle overlaps for speaker diarization that consists of two steps which are overlap exclusion and overlap labeling. In overlap exclusion, detected overlap segments are excluded from the clustering process so that they do not corrupt the speaker models. In overlap labeling, the segments detected as overlap are labeled by two speakers according to a heuristic such as assigning the overlap segment to two nearest speakers in time. This heuristic is based on observations from studies on NIST RT meeting data which revealed that assigning speakers based on proximity to the overlap gives a significant reduction in DER [16]. The two speakers can also be assigned based on cluster likelihoods, in which the segment is assigned to two clusters for which the data in the overlap segment has highest likelihood [17], [22]. The overlap handling method in a typical speaker diarization system is summarized in the Fig. 1.

B. Overlap Detection Using Acoustic Features

The baseline overlap detection system is based on acoustic features derived from short-term spectrum that have been shown to be effective in classifying single-speaker speech and overlapping speech in the literature [18]. In the present work, we use 12 Mel frequency cepstral coefficients (MFCCs) with log energy, spectral flatness, linear prediction (LP) residual energy computed from LP analysis of 12th order, along with their deltas resulting in a feature vector of dimension 30. All features were extracted from an acoustic frame of size 30 ms with a frame rate of 10 ms. Prior to overlap detection, automatic speech/non-speech detection is performed using the SHOUT algorithm [42] and non-speech regions are excluded from further processing. After this single-speaker, overlapping speech detection is performed on the detected speech regions in the recording. In baseline overlap detection system, the classes single-speaker and overlapping speech are represented by states of a hidden Markov model (HMM). The emission probability distributions of these states are modelled using Gaussian mixture models (GMMs). To control the tradeoff between the false overlap detections and total number of overlaps being detected, an overlap insertion penalty (OIP) is introduced which penalizes the overlap detections. This has a positive effect on the precision of the classifier while effecting the recall in a negative manner. Let $V \in \{ov, sp\}$ denote the set of HMM states representing overlapping (ov) and single-speaker (sp) speech classes, and let F denote the sequence of acoustic feature vectors then, the most likely sequence of the states in a given audio recording can be obtained by Viterbi decoding as,

$$V^* = \arg\max_V P(V|F) = \arg\max_V P(F|V)P(V) \quad (4)$$

The prior probability $P(V)$ of the state sequence is usually approximated by first order Markov model assuming that the current state is dependent only on its previous state and is represented by the state transition probability. In the baseline overlap detection system, these probabilities are fixed to a constant value based on the statistics observed in the training data. But studies on conversational analysis show that the probability of occurrence of an overlap is not constant across the recording, and there are places in a conversation where overlaps are more likely to occur [8], [4]. In the current work, we propose a method to capture this information and use it in the acoustic feature based overlap detection system.

IV. Conversational Features for Overlap Detection

In this section, we present in detail the long-term features extracted from the structure of a conversation that are supposed to carry relevant information about occurrence of overlap. We explore features that can be easily extracted automatically such as silence and speaker change statistics.

A. Silence Statistics for Overlap Detection

Several studies on multi-party conversational analysis have shown that single-speaker speech, silence and overlap patterns in a conversation are related to each other and carry useful information about the conversations and the participants in the conversations [28], [29], [30]. Motivated by these studies, we explore the relation between silence, single-speaker speech, overlapping speech durations in a segment to predict the occurrence of overlap in that segment. In particular we hypothesize that segments containing more silence are less likely to contain overlapping speech. To verify this hypothesis, we perform experiments using the AMI training set meetings.

Let D be a variable indicating the duration of silence in a segment and $n^S(D = d)$ be the number of segments that contain d seconds of silence where, the length of the segment is given by the variable S. Let $V \in \{ov, sp\}$ be a binary variable denoting the classes we are interested in detecting which are overlapping speech (ov) and single-speaker speech (sp). Let $n^S(V = ov, D = d)$ denote the number of segments of length S seconds that contain d seconds of silence and an occurrence of

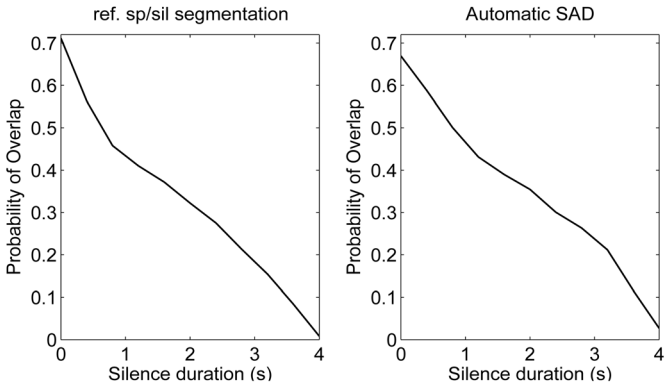

Fig. 2. Probability of overlap based on silence duration obtained using ground truth speech/sil segmentation and automatic speech activity detector output.

overlap. Given these counts, it is possible to estimate the probability of overlap in a segment conditioned on the amount of silence in that segment as:

$$P^S(V = ov|D = d) = \frac{n^S(V = ov, D = d)}{n^S(D = d)} \tag{5}$$

Fig. 2 shows $P^S(V = ov|D = d)$ for different values of d (duration of silence) for a segment length of four seconds (i.e., $S = 4$). In the left plot, speech/silence segmentation is obtained from the ground-truth segmentation and in the right plot, it is obtained from the automatic speech activity detector (SAD) output [42]. It can be noticed from Fig. 2 that the probability of overlap in a segment is inversely proportional to the amount of silence in the segment which supports our hypothesis that segments containing more silence are less likely to contain overlapping speech. In particular, from the left subplot of Fig. 2, it can be observed that when the amount of silence in a segment is zero, the probability of occurrence of overlap in that segment is around 0.7. In other words, this illustrates that it is possible to estimate the probability of occurrence of an overlap in a segment by the amount of silence present in that segment. This information is potentially useful as speech/silence detection is a simpler task compared to single-speaker/overlapping speech detection. The right subplot in Fig. 2, which uses SAD output to compute the amount of silence in a segment, also shows similar trends which means that ground-truth silence duration can be replaced by the estimates of silence from SAD output. The probability of a single-speaker speech within a segment can be estimated as:

$$P^S(V = sp|SIL = x) = 1 - P^S(V = ov|D = d). \tag{6}$$

To compute these statistics for the whole recording, the segment is progressively shifted by one frame at each step. The probabilities $P_i^S(ov|d_i), P_i^S(sp|d_i)$ are estimated $\forall i \in \{1 \ldots N\}$ where N is the total number of frames in the recording and d_i denotes the duration of silence in the segment centered around the frame i. This process is depicted in Fig. 3.

To verify how the estimated probabilities $P^S(ov_i|d_i)$ and $P^S(sp_i|d_i)$ generalize to sets of meetings that are different from those used during training, cross entropy between these estimated probabilities and true distribution of the classes in a development set of meetings is computed. The probabilities for

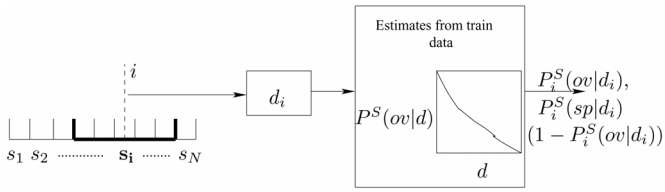

Fig. 3. Estimation of probabilities of single-speaker speech and overlapping speech states for a frame i based on duration of silence d_i present in the segment s_i centered around the frame i.

the true distributions are obtained for each frame $i \in \{1 \ldots N\}$ as follows, $P^T(ov_i) = 1, P^T(sp_i) = 0$ if the frame i is overlapped and $P^T(sp_i) = 1, P^T(ov_i) = 0$ if the frame i belongs to single-speaker speech. The knowledge of whether a frame i belongs to the overlapped ($\{OV\}$) or single-speaker ($\{SP\}$) class is obtained from the ground-truth segmentation of these meetings. The cross entropy between the true distribution and the estimated distribution is computed as follows:

$$C(S) = -\frac{1}{L}\left(\sum_{i\in\{OV\}} \log(P_i^S(ov|d_i)) + \sum_{j\in\{SP\}} \log(P_j^S(sp|d_j))\right) \tag{7}$$

where L is the total number of frames used in the computation. To eliminate the bias in the estimate of $C(S)$ resulting from uneven number of samples present in single-speaker speech and overlap classes, the cross entropy measure is computed by considering equal number of samples from each class.

It can be observed from Fig. 4 (a) that the cross entropy decreases as we increase the segment length (S) used to estimate the probabilities of overlap until some point and then starts to increase. This decrease in the cross entropy suggests that the estimated probabilities are much closer to the true probabilities when they are estimated from a longer context than a frame. The lowest value of cross entropy is found around a segment length of 4 secs. This indicates that a segment length of 4 secs is optimal to compute the silence statistics. Similar plots of cross entropy are also shown for RT 09 and ICSI meetings respectively in subplots (b) and (c) of Fig. 4. Note that the probability estimates for NIST-RT 09 and ICSI meetings are obtained using the model learned from AMI training set. These plots also show similar trends to that observed on AMI development set, which gives an indication that the estimated statistics have similar effects on other data sets. It also shows that model learned to estimate the probabilities of overlapping and single-speaker speech from AMI training set can be generalizable to other meeting corpora such as NIST-RT and ICSI.

B. Speaker Change Statistics for Overlap Detection

Studies on conversational analysis have shown that overlaps occur more often at some specific parts of conversations [8]. Especially, it was shown that a significant proportion of the overlaps occurs during speaker turn changes [8]. Motivated by these studies, the current work analyzes the relationship between the occurrence of overlap in a segment and the number of speaker changes in the segment. Specifically, the study hypothesizes that overlap probability in a segment is directly proportional to the number of speaker changes in the segment. In other words, segments containing more speaker changes are highly

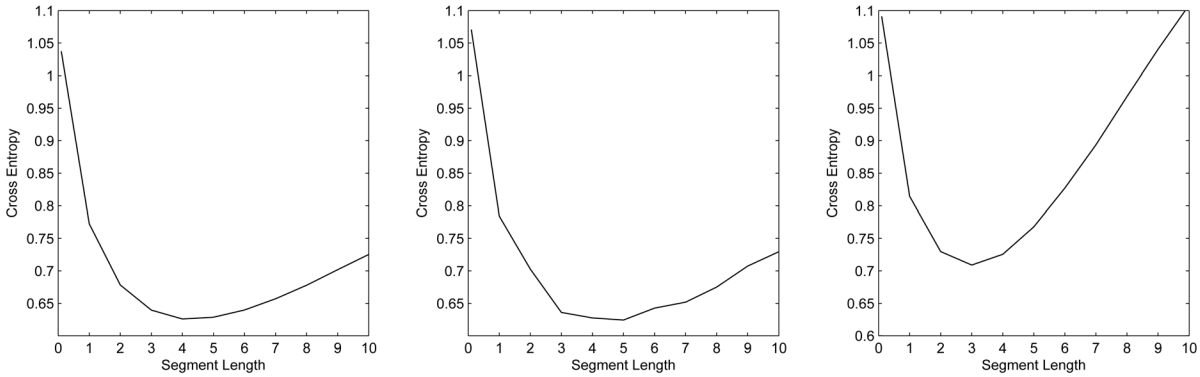

Fig. 4. Cross entropy between estimates based on silence duration (obtained from model learnt from AMI-train set) and true distribution on various datasets (a) AMI development set (b) NIST-RT 09 (c) ICSI.

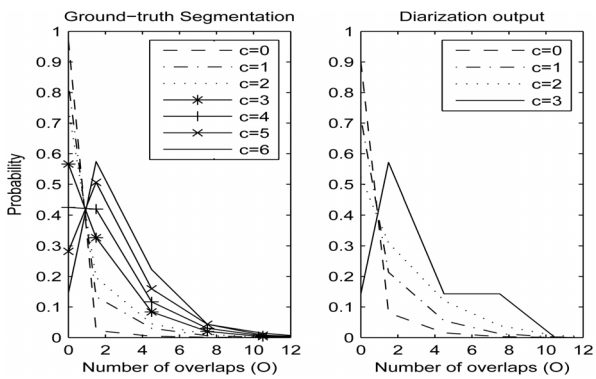

Fig. 5. Probability distributions of number of occurrences of overlaps (O) for different number of speaker changes (c) obtained using ground truth segmentation and diarization output.

probable to have more number of overlaps than those having fewer speaker changes. To verify this hypothesis, we perform experiments using the AMI training set.

In the first experiment, the distribution of the number of overlaps is analyzed for different number of speaker changes in a segment. Let C and O respectively denote the variables indicating number of speaker changes and overlaps in a segment. In the present work, an occurrence of overlap is defined as a contiguous segment of overlapping speech surrounded by single-speaker speech or silence regions. The number of overlaps is obtained by counting such occurrences in the segmentations (obtained by force-aligning close talking microphone audio with manual transcripts) provided by the corpus authors. Let $n^S(C = c)$ denote the number of segments of length S seconds which contain c number of speaker changes and, let $n^S(O = o, C = c)$ denote the number of segments containing o number of overlaps and c speaker changes. Then, the probability $P^S(O = o|C = c)$ of having o number of overlaps in a segment of length S seconds conditioned on the fact that it contains c number of speaker changes can be estimated as:

$$P^S(O = o|C = c) = \frac{n^S(O = o, C = c)}{n^S(C = c)} \tag{8}$$

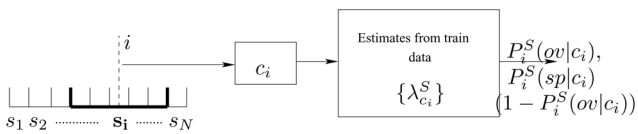

Fig. 6. Estimation of probabilities of single-speaker and overlapping speech classes for a frame i based on number of speaker changes c_i present in the segment s_i centered around the frame i.

Fig. 5 shows the distribution of $P^S(O|C = c)$ i.e., the distribution of the number of overlaps (O) in segments of length six seconds ($S = 6$) for different number of speaker changes (c). The speaker changes are obtained from the ground truth speaker segmentation and automatic diarization output for left and right subplots respectively. It can be observed from Fig. 5 that, as the number of speaker changes increases, the probability of occurrence of more overlaps also increases. Also, it can be observed that the distribution of $P^S(O|C = c)$ for different c seem to follow a Poisson distribution with a rate that is directly proportional to the number of speaker changes (c). Number of speaker changes in the diarization output is lower when compared to ground truth speaker segmentation due to constraints and errors introduced by the automatic system. Nevertheless, a similar phenomenon can also be observed for distributions estimated from diarization output. Fig. 5 supports our hypothesis that segments containing more speaker changes contain more overlaps. This information can be useful when incorporated into the baseline overlap detector which is based on acoustic features, since it does not contain evidence from the conversational patterns in the meetings.

Motivated from the empirical distributions in Fig. 5, we model the probability of number of occurrences of overlaps in a given segment by a Poisson distribution whose rate λ^S_c depends on the number of speaker changes c in the segment S i.e.,

$$P^S(O = o|c) = \frac{(\lambda^S_c)^o e^{-\lambda^S_c}}{o!} \tag{9}$$

where the rate parameter λ^S_c is a maximum likelihood estimate from the training set of meetings, which is simply the mean of the number of occurrences of overlaps in segments of length S seconds which contain c speaker changes. After estimating the

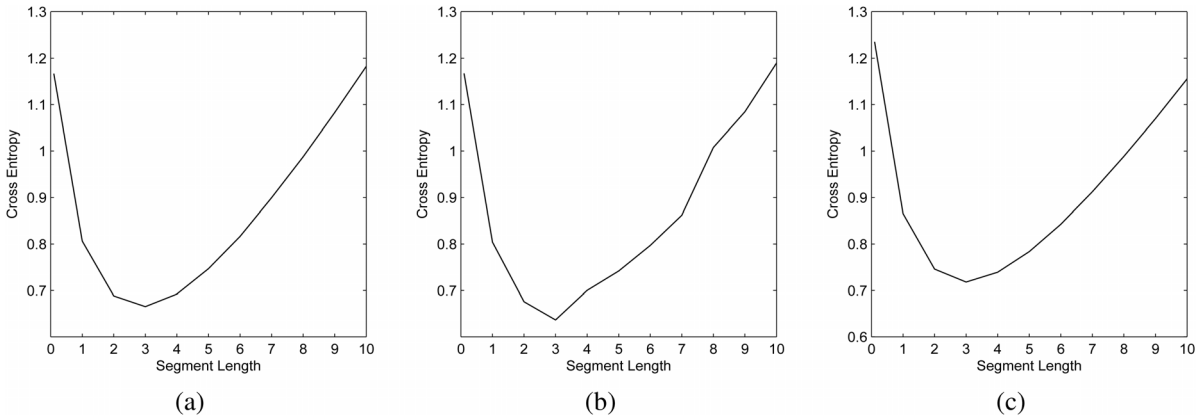

Fig. 7. Cross entropy measure between estimates based on speaker changes (obtained from model learnt from AMI-train set) and true distribution on various datasets (a) AMI development set (b) RT 09 (c) ICSI for various segment lengths.

set of rate parameters $\{\lambda_c^S\}$ for different values of c, the probability of occurrence of overlap in a segment conditioned on the number of speaker changes in the segment can be obtained as,

$$P^S(V = ov|c) = 1 - P^S(O = 0|c) \tag{10}$$
$$= 1 - e^{-\lambda_c^S} \tag{11}$$

and, the probability of single speaker speech can be obtained as,

$$P^S(V = sp|c) = 1 - P^S(V = ov|c) \tag{12}$$
$$= e^{-\lambda_c^S}. \tag{13}$$

The probabilities $P_i^S(V = ov|c_i)$ and $P_i^S(V = sp|c_i)$ are estimated for all the frames i in a given recording as depicted in Fig. 6. Cross entropy between the estimated probabilities and the true distribution is computed by replacing the probability estimates in (7) by the estimates based on speaker changes $P^S(ov|c)$ and $P^S(sp|c)$. Fig. 7 shows the cross entropy values for various segment lengths S computed on the AMI, NIST-RT 09 and ICSI data sets. It can be observed from Fig. 7(a) that a segment length of three seconds yields optimal estimates of the probabilities on the AMI development set. It can also be observed from subplots (b) and (c) in Fig. 7 that the estimated statistics generalize well to unseen meetings from different corpora such as NIST-RT 09 and ICSI.

C. Combination of Silence and Speaker Change Statistics

The probability estimates of single-speaker and overlapping speech classes obtained using silence and speaker change statistics are based on different conversational phenomenon and we hypothesize that combining the information captured by these two features might result in a better estimate of the class probabilities. Motivated by this hypothesis, we explore various combination strategies proposed in the literature to obtain an estimate that exploits the information captured by both features. In particular, we experimented with two types of combination methods, early combination and late combination.

In the early combination strategy, a combined feature vector is formed for each frame by appending the individual features, corresponding to the frame. Let d_i and c_i denote the duration of silence and number of speaker changes in a segment centered around frame i. Let b_i denote the combined feature vector formed by appending the features d_i and c_i corresponding to the frame i i.e., $b_i = [d_i, c_i]$. A logistic regression classifier is trained using these feature vectors on a training set of meetings with a binary target variable $0, 1$ denoting single-speaker and overlapping speech classes respectively. The target labels for each frame are obtained from the ground-truth segmentation. The output of this classifier is in the range [0,1] and can be treated as a probability estimate of the overlapping speech class i.e., $P(V = ov|b_i)$. Given this estimate, the probability of single speaker speech $P(V = sp|b_i)$ is obtained as $1 - P(V = ov|b_i)$.

In the late combination method, we explore different ways to combine the probability estimates $P(V = ov|d_i)$ and $P(V = ov|c_i)$ obtained from the individual features to get the final overlap probability estimate $P(V = ov|d_i, c_i)$. We experimented with the standard combination methods such as the sum and product rules with uniform weights for individual estimators as well as inverse-entropy based weighting scheme [43]. The combination according to sum rule can be written as,

$$P(V|x_i, y_i) = w_d P(V|d_i) + w_c P(V|c_i) \tag{14}$$

The combination according to product rule can be written as:

$$P(V|x_i, y_i) = \frac{1}{P(V)} P(V|d_i)^{w_d} P(V|c_i)^{w_c}. \tag{15}$$

where $P(V)$ denotes the prior probability of a class where $V \in \{ov, sp\}$. While using uniform weights, both w_d and w_c are set to 0.5 each. In the inverse entropy based weighting scheme, the weights are set as explained below. Let H_d^i and H_c^i respectively denote entropy of probability estimators based on silence and speaker change statistics for a frame i. The weights of individual probability estimators for a frame i are computed as, $w_d = \frac{1/H_d^i}{1/H_d^i + 1/H_c^i}$ and $w_c = \frac{1/H_c^i}{1/H_d^i + 1/H_c^i}$.

To evaluate the usefulness of different combination strategies, we use the cross-entropy measure computed on AMI development set. The cross entropy is computed based on (7) using the probability estimates of the classes obtained by the above mentioned combination strategies. Cross entropy based studies on development set of meetings revealed that the inverse-entropy based weighting gave similar results to the

TABLE I
CROSS ENTROPY MEASURE ON AMI DEVELOPMENT SET OF MEETINGS FOR VARIOUS COMBINATION STRATEGIES

method	cross entropy
prod-uniform	0.60
sum-uniform	0.63
log. reg.	0.62

uniform weighting of the individual estimators. Table I shows cross entropy measures obtained for different combination methods mentioned above. It can be observed from Table I that estimates obtained by the product rule of combination have the lowest cross-entropy. Therefore, based on these findings, in the current work, we use product rule based combination method to combine information captured by different conversational features.

D. Combination of Conversational and Acoustic Features

The probability estimates of single-speaker and overlapping speech classes obtained from silence and/or speaker change statistics are integrated into the acoustic feature based classifier as prior probabilities of these classes. As mentioned earlier, the prior probabilities of the classes in the acoustic feature based system usually are fixed to a constant value based on the proportion of samples in each class observed during the training phase. However, studies have shown that the probability of overlap occurrence is neither constant across different conversations nor within a conversation. Therefore, to address this issue in the acoustic feature based classifier, we introduce prior probabilities that are estimated based on the silence and the speaker change statistics in context of a frame. These probabilities encode information present in the long-term context of a frame and change depending on the context. Let F denote the sequence of acoustic features and let C denote the sequence of conversational features which can be either the individual features or their combination. Given these, the most probable state sequence V^* where the states in the sequence belong to the set $\{ov, sp\}$ can be estimated by Viterbi decoding as:

$$\begin{aligned} V^* &= \arg\max_V P(V|F,C) \\ &= \arg\max_V P(F|V,C)P(V|C) \\ &\doteq \arg\max_V P(F|V)P(V|C) \end{aligned} \tag{16}$$

assuming that given the state/class the observed acoustic features F are independent of the conversational features C. The term $P(F|V)$ is modelled using GMM distributions of the corresponding states and is only dependent on the current frame. The term $P(V|C)$ estimates the probability of the states based on the conversational features such as silence duration and/or number of speaker changes in a segment surrounding the current frame and captures the information present in the long-term context of the frame which is not present in the acoustic features. Therefore, we hypothesize that this combination will improve the performance of the classifier.

V. EXPERIMENTS AND RESULTS

In this section, we present the experimental results of overlapping speech detection using standard acoustic features as explained in section III-B and the proposed method of using probability estimates of the classes obtained from conversational features as prior probabilities of the states as explained in section IV. We also present experiments evaluating the effect of the acoustic feature based overlap detection system and the proposed method for overlap detection on speaker diarization. Fig. 8 summarizes the proposed method for overlap detection. The obtained overlaps are used for overlap handling in speaker diarization as explained in section III-A1.

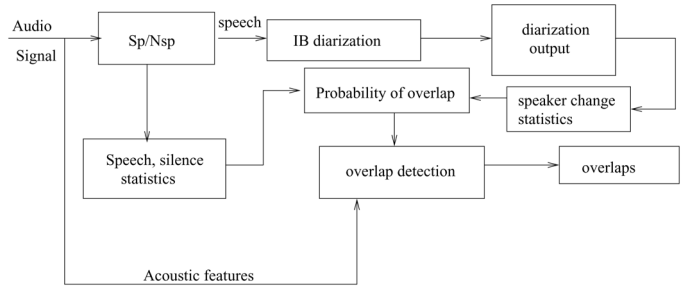

Fig. 8. Block diagram of the proposed method of overlap detection: Speech, silence statistics and speaker change statistics are obtained as by product of initial pass of IB speaker diarization. These statistics are used to estimate probability of occurrence of overlap as explained in sections IV-A, IV-B and IV-C. These probabilities are incorporated into acoustic feature based overlap detector (section III-B) as prior probabilities as detailed in section IV-D.

A. Experiments on Overlap Detection

We evaluate the performance of the acoustic feature based overlap detector and the proposed method of incorporating conversational information into acoustic feature based detector on three different meeting corpora namely, AMI, NIST-RT and ICSI. A HMM/GMM based overlap detector is trained for each corpus using training set of meetings for the respective corpus as described in section III-B. The parameters needed for estimating the probabilities of single-speaker and overlapping speech classes from the conversational features as described in sections IV-A, IV-B and IV-C are obtained from the training set of meetings from the AMI corpus. These parameters are then used to estimate the probabilities of the classes on the test set of meetings in all the corpora and are incorporated into the acoustic feature based HMM/GMM system as explained in section IV-D.

In the first experiment, we verify the hypothesis stated in the section IV-D that, incorporating the probabilities of overlapping and single-speaker speech classes estimated from the proposed conversational features such as silence and speaker change statistics as prior probabilities of the classes in the acoustic feature based classifier improves its performance. The systems are compared using the metrics such as recall, precision, f-measure and error. Recall of a system on overlap detection task is computed as the ratio between the duration of overlap that is correctly detected by the system to the total duration of actual overlap. Precision is computed as the ratio between duration of the overlap that is correctly detected by the system and total duration of overlapping speech detected by the system. F-measure is harmonic mean between recall and precision which is a good indicator of the classifier performance when the number of samples in each class is skewed as in the current study. The error rate is

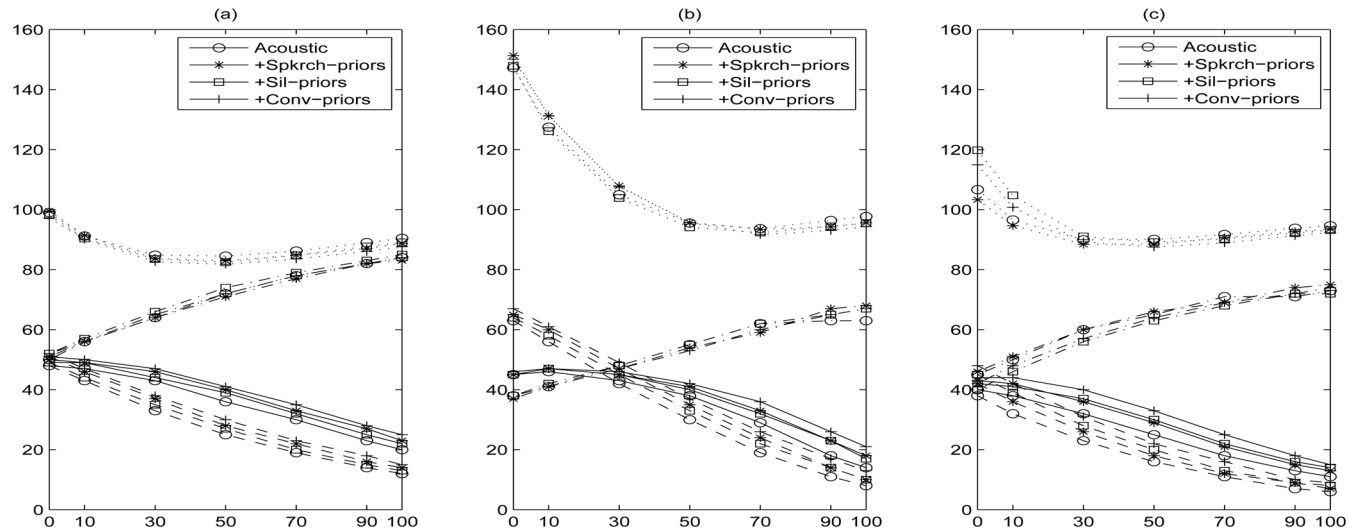

Fig. 9. Overlap detection evaluation on (a) AMI-test set (b) NIST-RT 09 set and (c) ICSI-test set. In each subplot (a), (b), (c), dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection in percentage on y-axis for various overlap insertion penalties (OIP) in x-axis.

computed as the ratio between the total duration of missed and false overlap detections to the total duration of overlap.

Fig. 9 shows the performance of the acoustic feature based overlap detector and the proposed method on the test sets of three meeting corpora in terms of recall (dashed line), precision (-.- line), f-measure (solid line) and error (dotted line) of the respective classifiers on the task of overlap detection. In Fig. 9, the acoustic feature based classifier (section III-B) is denoted by label Acoustic, the system using probability estimates from silence statistics (section IV-A) is denoted as +Sil-priors, the system using probability estimates from speaker change statistics (section IV-B) is denoted as +Spkrch-priors and, the system using the probability estimates from the combination of the individual features by product rule (section IV-C) is denoted as +Conv-priors. The figure plots various evaluation metrics for different values of overlap insertion penalty (OIP) that is introduced to have a trade-off between the true and the false overlap detections. In general, it can be observed that higher values of OIP tend to increase the precision of the classifier while sacrificing the recall.

The error rates achieved by the acoustic feature based overlap detector on AMI-test set are similar to the error rates obtained by prior works [17], [23], [27] in literature on the data set. It can be observed from the Fig. 9(a) that incorporating the probabilities of the classes estimated from the individual conversational features (Sil-priors, Spkrch-priors) improves the performance of the acoustic feature based system as they consistently achieve higher f-measure and lower error over the acoustic feature based classifier for all the values of OIP. Also, the combination (Conv-priors) of the conversational features (Sil-priors, Spkrch-priors) leads to further improvements in overlap detection. In the experiments reported here, the combination is performed based on the product rule as described in section IV-C since, the combination based on product rule obtained the lowest cross-entropy on AMI development set (see Table I). The improvements in terms of f-measure and decrease in the error rate achieved by the proposed method is mainly due to increase in the recall of the classifier when compared to the acoustic feature based system. This indicates that the proposed method is able to identify instances of overlaps which are not detected by the acoustic feature based system. Similar trends can be observed in all the corpora though the absolute values of the evaluation metrics are different. In general, it can be observed that incorporating conversational features into the acoustic feature based classifier improves the performance of the classifier as shown by the consistent higher f-measure and the lower error rates achieved by the proposed method on all the three corpora when compared to that of the acoustic feature based system. This result is particularly encouraging as it demonstrates that, model trained to estimate probability of overlap based on conversational features using one corpus (AMI), generalizes well to meetings from the other corpora (NIST-RT and ICSI).

1) Laughter Overlap Detection: Laughter is a very common phenomenon in human interactions [44]. Studies on spontaneous conversations have shown that overlaps and laughter occurrences are correlated with each other [45], [46]. Studies done on ICSI corpus have shown that 9% of speaking time contains laughter [45]. Based on these studies, we evaluate the performance of the proposed method for overlap detection in laughter segments of ICSI corpus. The start and end times of laughter segments and the corresponding speakers are obtained from the annotations done for analysis of laughter in [45]. Fig. 10 presents results of overlap detection on laughter segments based on acoustic features alone(Acoustic) and combination of acoustic and conversational features (+Conv-priors). To obtain the conversational features silence and speaker change statistics which are used to estimate the prior probabilities of the classes, we superimpose the ground-truth laughter segments of a meeting recording over speaker diarization output of the respective meeting. It can be observed from the Fig. 10 that the combination of conversational features improves the performance of the acoustic feature based overlap detector as shown by the f-measures for various OIPs.

2) Robustness to Speaker Diarization Errors: The conversational features proposed in the current work make use of the baseline IB diarization output to compute speaker change sta-

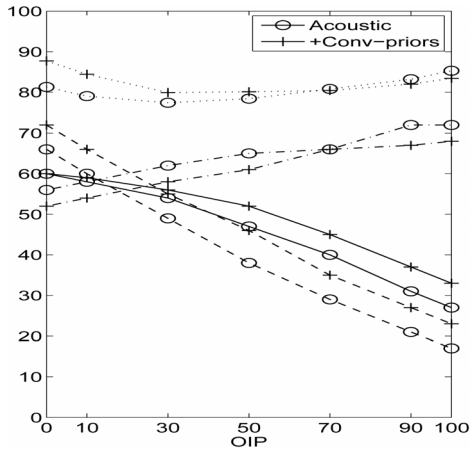

Fig. 10. Overlap detection evaluation on laughter segments from ICSI-test set: dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection in percentage on y-axis for various overlap insertion penalties (OIP) on x-axis.

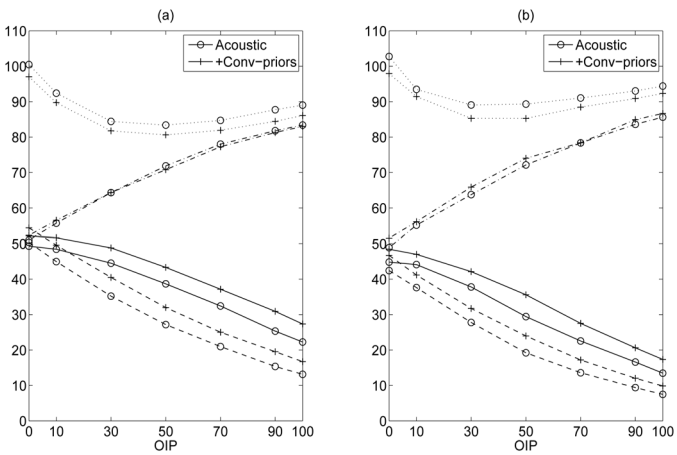

Fig. 11. Effect of diarization errors: (a) High error set (b) Low error set. Dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection on in percentage y-axis for various overlap insertion penalties (OIP) on x-axis.

tistics. These statistics computed over a long-term context of a frame are used to estimate the probability of overlap at that frame. To evaluate the effect of errors made by the clustering algorithm of speaker diarization on overlap detection, we compare the overlap detection performance on meetings with high and low diarization error. For this purpose, we divided the AMI-test set into two subsets, based on the speaker error of the clustering algorithm. All the meetings less than speaker error of 15% were put in low error set and rest of the meetings were put in high error set. Fig. 11 plots the evaluation metrics for overlap detection for acoustic feature based system (Acoustic) and the combination of acoustic and conversational feature based system (+Conv-priors) on high and low error sets. From Fig. 11, it can be observed that the performance of the acoustic feature based detector is improved on both the sets by the combination of conversational features. The low error set (Fig. 11(b)) has slightly high precision when compared to the high error set (Fig. 11(a)).

B. Speaker Diarization with Overlap Handling

In this section, we evaluate the effect of overlap detection on speaker diarization on three different corpora AMI-test set, NIST RT-09 and ICSI-test sets. For experiments on ICSI-test set, we included laughter segments in speech regions given as input to the diarization system to make the scenario as natural as possible. We used ground-truth speech/non-speech segmentation for experiments on ICSI corpus to avoid missing laughter segments because of automatic speech activity detection which might classifying laughter segments as non-speech. The detected overlaps are used in speaker diarization by performing overlap exclusion and labeling techniques explained in section III-A1. Using these methods, we compare the two overlap detection systems one based on just the acoustic features and the other incorporating the information from conversational features as prior probabilities of the classes. Since the combination of the conversational features showed the best performance in overlap detection (Fig. 9), we use the overlap detection hypothesis generated by the system using the combination of the conversational features based on the product rule (Conv-priors). To decide the optimal value of OIP to use in overlap detection system, we perform tuning experiments on AMI development corpus and pick an OIP that gives lowest DER. As proposed in earlier works [22], [18], we perform overlap exclusion and overlap labeling steps using different overlap detection hypothesis dependent on the value of OIP used. This is based on the rationale that, high precision overlap detection hypothesis is desirable for overlap labeling step to avoid increasing false alarm errors and high recall hypothesis is desirable for overlap exclusion, as it helps in avoiding as much overlapping speech as possible from corrupting the speaker (cluster) models.

1) Overlap Exclusion: To tune the overlap insertion penalty (OIP) for overlap exclusion, we ran experiments on a development set of meetings by performing overlap exclusion using the overlap hypothesis generated by various values of OIP. These experiments have revealed a similar trend to that observed in previous studies [23], where the DER reduction was not a smooth function of OIP. Therefore, for overlap exclusion we use the detection hypothesis obtained with OIP set to zero as done in previous studies [23], [27]. Table II reports DER and its components speech/non-speech error (SpNsp) and speaker error (Spkr) obtained on test set of meetings from AMI and NIST-RT 09 corpus using the baseline speaker diarization system in three scenarios; without any overlap exclusion (Baseline), overlap exclusion using the hypothesis generated by acoustic feature based system (Acoustic) and overlap exclusion using the hypothesis generated by the proposed method (Conv-priors). First of all, it can be observed from the Table II that performing overlap exclusion reduces the speaker error (Spkr) of the diarization as expected, since it avoids corruption of resulting speaker models. Also, it can be observed from Table II that on AMI-test set, the proposed method reduces the DER by around 17% relative to the baseline speaker diarization system that does not do any overlap exclusion. The acoustic feature based overlap detector reduces the DER by around 13% relative to the baseline. The table also reports the f-measures of both the overlap detection systems used to perform overlap exclusion. This reveals that higher reduction in DER achieved by the proposed method is due to the ability of the proposed method to detect more overlap at the given OIP (0).

TABLE II
OVERLAP EXCLUSION ON AMI-TEST SET, NIST-RT 09 AND ICSI-TEST DATA SETS: SPNSP REPRESENTS ERROR IN SPEECH/NON-SPEECH DETECTION, SPKR AND F-MEASURE OF THE RESPECTIVE OVERLAP DETECTION SYSTEMS AT THE OPERATING POINT (OIP $= 0$) USED FOR EXCLUSION

Corpus	System	SpNsp	Spkr	DER	f-measure
AMI	Baseline	13.5	16.9	30.4	-
	Acoustic	13.5	12.8	26.3	0.48
	+Conv-priors	13.5	11.6	25.1	0.51
RT 09	Baseline	12.7	21.2	33.9	-
	Acoustic	12.7	20.0	32.7	0.44
	+Conv-priors	12.7	19.0	31.7	0.46
ICSI	Baseline	17.7	15.6	33.3	-
	Acoustic	17.7	14.8	32.5	0.44
	+Conv-priors	17.7	14.2	31.9	0.49

Table II also presents results of similar experiments on NIST-RT 09 and ICSI-test data set. On RT-09 data set the acoustic feature based overlap detector reduces the DER by 3.5% relative and adding conversational features further reduces the DER by 6.4% relative. On ICSI-test set, the acoustic feature based overlap detector reduces the DER by 2.4% relative and the DER reduces by 4.2% relative when conversational features are added. The drop in the relative reductions in DER when compared to AMI data set (Table II) is mainly due to the performance drop in the acoustic feature based classifier as indicated by the f-measures of the classifiers on the corpora in Table II. Nevertheless, the proposed method (Conv-priors) achieves a lower DER than the system using overlaps from acoustic feature based system (Acoustic) on RT 09 and ICSI-test data sets also.

2) Overlap Labeling: To determine the optimal value of OIP for labeling task, we ran tuning experiments on AMI development set using overlap hypothesis obtained for different values for OIP. Fig. 12 plots the relative change in DER due to overlap labeling as a function of OIP used to generate the overlap hypothesis. In the present work, we use nearest neighbor based labeling as it gave similar results to the cluster likelihood based labeling. It can be observed from Fig. 12 that the OIP value of 90 gives the highest relative decrement in DER on the AMI development set. Based on this observation, we use the overlap hypothesis obtained by setting the value of OIP to 90 while performing overlap labeling. Table III presents the overlap labeling results on the test set of meetings in AMI and RT corpora. It can be observed from Table III that on AMI corpus, the proposed method (Conv-priors) decreases the DER by 2.9% relative to baseline diarization system that does not use any overlap information (Baseline). It also achieves lower DER than the system using overlaps detected by acoustic feature based system (Acoustic). The decrease in DER achieved by the proposed method is due to the decrease in speech/non-speech error (SpNsp). The speech/non-speech error (SpNsp) is reduced from 13.5% in the baseline diarization system to 11.9% in proposed method which is around 12% relative reduction. This reduction is due to the detection of overlapping speech and labeling it as speech. But the improvement in the final DER is not in the same range due to the errors introduced during labeling which increase the speaker error when the identified overlap segments are not assigned to correct speakers. This highlights the need for a novel speaker labeling method for the detected overlaps.

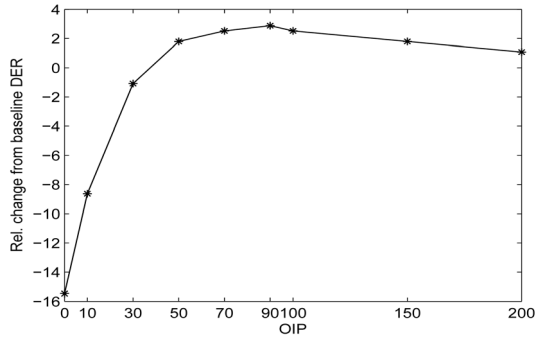

Fig. 12. Overlap labeling: Tuning for optimal overlap insertion penalty (OIP) on AMI development set.

TABLE III
OVERLAP LABELING ON AMI-TEST AND NIST-RT 09 AND ICSI-TEST DATA SETS: MISSED (MISS) AND FALSE-ALARM (FA) ERRORS IN SPEECH/NON-SPEECH DETECTION AND TOTAL SPEECH/NON-SPEECH ERROR (SPNSP), SPEAKER ERROR (SPKR), DIARIZATION ERROR RATE (DER) AND F-MEASURES OF THE RESPECTIVE OVERLAP DETECTION SYSTEMS AT THE OPERATING POINT (OIP $= 90$) CHOSEN TO DO LABELING

Corpus	System	Miss	FA	SpNsp	Spkr	DER	f-measure
AMI	Baseline	12.5	1.0	13.5	16.9	30.4	-
	Acoustic	10.9	1.4	12.3	17.4	29.7	0.23
	+Conv-priors	10.2	1.7	11.9	17.6	29.5	0.28
RT 09	Baseline	11.6	1.1	12.7	21.2	33.9	-
	Acoustic	11.0	1.5	12.5	21.4	33.9	0.18
	+Conv-priors	10.6	1.6	12.2	21.5	33.7	0.26
ICSI	Baseline	17.7	0	17.7	15.6	33.3	-
	Acoustic	16.6	0.1	16.7	15.9	32.6	0.18
	+Conv-priors	15.8	0.4	16.2	16.1	32.3	0.26

Similar trends can be observed in Table III on RT 09 data set also. The lower DERs achieved by the proposed method on the two data sets can be attributed to better detection of overlaps as indicated by the higher values of f-measure and lower error rate obtained by it on the data sets at the given OIP of 90 when compared to acoustic feature based system (see Fig. 9). On ICSI corpus also, overlaps obtained by the combination of acoustic and conversational features achieve the highest reduction of DER by labeling. The relative reduction on ICSI-test set is better than RT-09 even though the overlap detection performance is similar. This is due to the reason that ICSI-test set contains overlaps from laughter segments which tend to have more speakers than normal speech overlaps. Therefore, the labeling method is less prone to errors when labeling overlaps in laughter segments.

3) Overlap Exclusion Followed by Labeling: In this section we summarize the results of diarization experiments when overlap exclusion and labeling are performed alone and together where exclusion is followed by labeling as described in Fig. 1. Table IV summarizes the DERs obtained on all the three data sets AMI-test and NIST-RT 09 and ICSI-test while performing overlap handling techniques Exclusion, Labeling, and Both (exclusion followed by labeling) using overlaps detected by acoustic feature based system (Acoustic) and the proposed method of combining acoustic and conversational features(+Conv-priors). It can be observed from Table IV that the proposed method achieves highest reduction in DER on all the corpora. It increases the relative error reduction due to overlap detection to around 20% on meetings from AMI-test from around 15% achieved by acoustic feature based overlap

TABLE IV
DERs (WITH RELATIVE IMPROVEMENTS OVER BASELINE DIARIZATION WITHIN PARENTHESIS) OBTAINED WHILE PERFORMING OVERLAP HANDLING BY USING OVERLAPS DETECTED BY PROPOSED METHOD (+CONV-PRIORS) AND ACOUSTIC FEATURE BASED OVERLAP DETECTOR (ACOUSTIC)

Corpus	System	Exclusion	Labelling	Both
AMI	Acoustic	26.3 (+13.4%)	29.7 (+2.3%)	25.8 (+15.1%)
	+Conv-priors	25.0 (+17.7%)	29.5 (+2.9%)	24.2 (+20.3%)
RT 09	Acoustic	32.7 (+3.5%)	33.9 (-)	32.7 (+3.5%)
	+Conv-priors	31.7 (+6.4%)	33.7 (+0.6%)	31.5 (+6.7%)
ICSI	Acoustic	32.5 (+2.4%)	32.6 (+2.1%)	31.8(+4.5%)
	+Conv-priors	31.9 (+4.2%)	32.3 (+3%)	30.9 (+7.2%)

detection system. On RT 09 meetings, it increases the relative DER reduction to around 7% from 3.5% achieved by system based on acoustic features and on ICSI-test set it increases the relative error reduction to around 7% from 4% achieved by the acoustic feature based system.

VI. CONCLUSIONS

Motivated from the studies done on conversational analysis, in the present work, a method to improve the acoustic feature based overlap detector using long-term conversational features was proposed. Experiments done in the present work revealed that features extracted automatically from conversations such as silence and speaker change statistics carry relevant information about overlap occurrence in a conversation. The features were computed over a window of around 3-4 secs to capture the information present in the long-term context of a frame. These features were then used to estimate the probability of overlap and single-speaker speech classes in the window. Cross entropy measure based studies on development data revealed that the probability estimates of the classes are closer to the true distribution as the length of the context used to compute the features is increased and reaches an optimum around 4 secs. The probability estimates of the overlapping and single-speaker classes obtained from the long-term context of each frame in the conversation were incorporated into the acoustic feature based classifier as prior probabilities of the classes.

Experimental results on overlap detection using three different meeting corpora (AMI, NIST-RT, ICSI) revealed that the proposed method improves the performance of acoustic feature based classifier. These experiments also revealed that the model learnt to estimate the class probabilities using data from AMI corpus is generalizable to other meeting corpora such as NIST-RT and ICSI. Experiments were also done to evaluate the effect of overlap detection on speaker diarization using the standard methods of exclusion and labeling. These experiments revealed that the proposed method decreases the DER by 20% relative to the baseline speaker diarization system on AMI-test data set. Using overlap detection from only acoustic feature based system reduced the DER by 15% relative. Speaker diarization experiments on NIST-RT 09 and ICSI-test data sets have also revealed that the proposed method achieves higher reduction in DER when compared to the system using only acoustic features for overlap detection. These experiments also highlight the need for an effective overlap labeling mechanism to assign speakers to the detected overlap segments as the reductions obtained in speech/non-speech error are compensated to an extent by the increase in the speaker error due to the errors done during labeling.

ACKNOWLEDGMENT

The authors thank Dr. Fabio Valente for the helpful discussions and suggestions all along the work, Dr. Phil Garner for proof reading the manuscript and the reviewers whose comments have improved the paper significantly.

REFERENCES

[1] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. Eurospeech*, 2005, pp. 1781–1784.
[2] M. Adda-Decker, B. Claude, A. Gilles, P. Patrick, B. d. M. Philippe, and H. Benoit, "Annotation and analysis of overlapping speech in political interviews," in *Proc. 6th Int. Lang. Resources Eval. (LREC'08)*, Marrakech, Morocco, May 2008, European Language Resources Association (ELRA).
[3] O. Cetin and E. Shriberg, "Overlap in meetings: Asr effects and analysis by dialog factors, speakers, and collection site," in *Proc. 3rd Joint Workshop Multimodal and Rel. Mach. Learn. Algorithms*, Washington, DC, USA, 2006.
[4] O. Cetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *Proc. ICSLP*, Pittsburgh, PA, USA, 2006, pp. 293–296.
[5] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Lang. Soc.*, vol. 29, no. 1, pp. 1–63, 2000.
[6] E. Kurtic, G. J. Brown, and B. Wells, "Resources for turn competition in overlapping talk," *Speech Commun.*, vol. 55, no. 5, pp. 721–743, 2013.
[7] H. Sacks, E. A. Schegloff, and G. Jefferson, "Simplest semantics for the organization of the turn-taking in conversation," *Language*, vol. 50, pp. 696–735, 1974.
[8] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc.Eurospeech*, Aalborg, Denmark, 2001, pp. 1359–1362.
[9] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1857–1860.
[10] M. Huijbregts, D. van Leeuwen, and C. Wooters, "Speaker diarization error analysis using oracle components," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 393–403, Feb. 2012.
[11] M. T. Knox, N. Mirghafori, and G. Friedland, "Where did I go wrong?: Identifying troublesome segments for speaker diarization systems," in *Proc. Interspeech*, Portland, OR, USA, 2012.
[12] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 84–91, Jan. 2005.
[13] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the icsi meeting recorder," in *2001 ASRU '01. IEEE Workshop Autom. Speech Recogn. Understand.*, 2001, pp. 107–110.
[14] K. Laskowski and T. Schultz, "Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 993–996.
[15] S. Otterson, "Use of speaker location features in meeting diarization," Ph.D. dissertation, Univ. of Washington, Seattle, WA, USA, 2008.
[16] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. ASRU*, Kyoto, Japan, 2007.
[17] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 32–35.
[18] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 941–943.
[19] K. Boakye, "Audio segmentation for meetings speech processing," Ph.D. dissertation, Univ. of California, Berkeley, CA, USA, 2008.
[20] [Online]. Available: http://corpus.amiproject.org/
[21] M. Huijbregts, D. A. v. Leuwen, and F. M. G. d. Jong, "Speech overlap detection in a two-pass speaker diarization system," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 1063–1066.
[22] M. Zelenák, C. Segura, and J. Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2302–2305.
[23] M. Zelenák, C. Segura, J. Luque, and J. Hernando, "Simultaneous speech detection with spatial features for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 436–446, Feb. 2012.

[24] M. Zelenák, "Detection and handling of overlapping speech for speaker diarization," Ph.D. dissertation, Univ. Politecnica de Catalunya, Barcelona, Spain, 2011.
[25] M. Zelenák and J. Hernando, "The detection of overlapping speech with prosodic features for speaker diarization," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1041–1043.
[26] R. Vipperla, J. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 4181–4184.
[27] J. Geiger, R. Vipperla, S. Bozonnet, N. Evans, B. Schuller, and G. Rigoll, "Convolutive non-negative sparse coding and new features for speech overlap handling in speaker diarization," in *Proc. Interspeech*, Portland, OR, USA, 2012.
[28] K. Laskowski, M. Osterdorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party convrersation," in *Proc. 9th ISCA/ACL SIGdial*, Columbus, OH, USA, 2008, pp. 148–155.
[29] K. Laskowski, M. Osterdorf, and T. Schultz, "Modeling vocal interaction for text-independent classification of conversation type," in *Proc. 8th ISCA/ACL SIGdial Workshop Discourse and Dialogue*, Antwerpen, Belgium, 2007, pp. 194–201.
[30] K. Laskowski, "Modeling norms of turn-taking in multi-party conversation," Proc. 48th Ann. Meeting of the Association for Computational Linguistics ser. ACL '10. Uppsala Sweden, Association for Computational Linguistics, 2010, pp. 999–1008.
[31] S. H. Yella and F. Valente, "Speaker diarization of overlapping speech based on silence distribution in meeting recordings," in *Proc. Interspeech*, Portland, USA, 2012.
[32] S. H. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, 2013.
[33] [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/
[34] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," in *Proc. ICASSP*, Hong Kong, 2003, pp. 364–367.
[35] [Online]. Available: http://www.xavieranguera.com/beamformit/
[36] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1382–1393, Sep. 2009.
[37] D. Vijayasenan and F. Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Proc. Interspeech*, Portland, OR, USA, 2012.
[38] S. H. Yella and F. Valente, "Information bottleneck features for HMM/GMM speaker diarization of meetings recordings," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 953–956.
[39] S. H. Yella and H. Bourlard, "Information bottleneck based speaker diarization of meetings using non-speech as side information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, 2014, pp. 96–100.
[40] C. Wooters and M. Huijbregts, *Multimodal technologies for perception of humans*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin/Heidelberg, Germany: Springer-Verlag, 2008, pp. 509–519, ch. The ICSI RT07s Speaker Diarization System.
[41] N. Slonim, N. Friedman, and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT press, 1999, pp. 617–623.
[42] M. Huijbregts and F. de Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Commun.*, vol. 53, no. 2, pp. 143–153, 2011.
[43] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream asr," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP '03*, 2003, pp. 741–744.
[44] P. Glenn, *Laughter in Interaction*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[45] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings," in *Proc. Interspeech*, 2007, pp. 1258–1261.
[46] L. Kennedy and D. P. W. Ellis, "Laughter detection in meetings," in *Proc. NIST Meeting Recognition Workshop*, Montreal, QC, Canada, 2004, pp. 118–121.

Sree Harsha Yella (S'13) is a Ph.D. student at Idiap Research Institute, Martigny, Switzerland, and École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. His Ph.D. thesis is focused on speaker diarization of spontaneous meeting room conversations. He obtained his bachelors (Hons) and masters degrees in computer science and engineering from International Institute of Information Technology, Hyderabad (IIIT-H), in 2008 and 2010, respectively. His masters thesis investigated speech summarization techniques using structural and acoustic cues in audio recordings. His research interests mainly include speech and spoken language processing, and applied machine learning.

Hervé Bourlard (F'00) received both the Electrical and Computer Science Engineering degree and the Ph.D. degree in applied sciences from Faculté Polytechnique de Mons, Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Full Professor at the École polytechnique fédérale de Lausanne EPFL, and (Founding) Director of a Swiss NSF National Centre of Competence in Research on Interactive Multimodal Information Management. Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now a member of the ICSI Board of Trustees.

His research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modelling, multimodal interaction, augmented multi-party interaction, and distant group collaborative environments.

He is the author/coauthor/editor of 6 books and over 300 reviewed papers (including one IEEE paper award) and book chapters. He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of IEEE ICASSP 2002, General Chairman of Interspeech 2003) and on the Editorial Board of several journals (e.g., past co-Editor-in-Chief of Speech Communication). He is the recipient of several scientific and entrepreneurship awards.