# Overlapping speaker segmentation using multiple hypothesis tracking of fundamental frequency

Aidan O. T. Hogg, *Student Member, IEEE,* Christine Evers, *Senior Member, IEEE,* Alastair H. Moore, and Patrick A. Naylor, *Fellow, IEEE*

***Abstract*—This paper demonstrates how the harmonic structure of voiced speech can be exploited to segment multiple overlapping speakers in a speaker diarization task. We explore how a change in the speaker can be inferred from a change in pitch. We show that voiced harmonics can be useful in detecting when more than one speaker is talking, such as during overlapping speaker activity. A novel system is proposed to track multiple harmonics simultaneously, allowing for the determination of onsets and end-points of a speaker's utterance in the presence of an additional active speaker. This system is bench-marked against a segmentation system from the literature that employs a bidirectional long short term memory network (BLSTM) approach and requires training. Experimental results highlight that the proposed approach outperforms the BLSTM baseline approach by 12.9% in terms of HIT rate for speaker segmentation. We also show that the estimated pitch tracks of our system can be used as features to the BLSTM to achieve further improvements of 1.21% in terms of coverage and 2.45% in terms of purity.**

***Index Terms*—speaker segmentation, pitch tracking, Kalman filter**

## I. Introduction

SPEAKER diarization answers the question of "who spoke when?" [1]. Diarization is performed without any prior knowledge of the number of speakers or the amount of speech the recording contains. Accurate diarization has become increasingly important in recent years for a multitude of tasks including voice control of smart devices and robot audition [2], [3]. Diarization is also required for applications such as speaker indexing [4], automatic speech recognition (ASR) [5] and enabling the use of single speaker-based algorithms in multi-speaker domains [6]. The diarization process is commonly separated into two independent tasks; the first task is segmentation and the second task is the clustering of those segments. This paper focuses on the segmentation task, i.e. the identification of the onsets and end-points of speakers.

The segmentation task is often complicated by the presence of overlapping speech where multiple speakers are active at the same time [7], [8]. Overlapping speech commonly occurs in conversational speech due to interruptions and backchannel vocalisations [9]. Environmental factors, such as reverberation [10] and noise [11], also render the task of accurate segmentation difficult to achieve. Various methods have been proposed to solve the problem of overlapping speaker segmentation including hidden markov model (HMM)-based methods that use Mel-frequency cepstral coefficients (MFCCs), linear predictive coding (LPC) and root mean square (RMS) energy features [12]. Methods that use long-term conversational features [13] have also been put forward along with multimodal techniques that use multiple microphone and camera systems [14]. More recently, deep learning approaches have become increasingly prevalent [15]–[18] which often require large amounts of labelled training data.

This paper proposes a novel approach to speaker segmentation that exploits the temporal variations and harmonic structure of the fundamental frequency ($F_0$) of voiced speech as outlined in Fig. 1. The advantage of this new approach is that it does not require any model training and can be employed even when only a single distant microphone (SDM) is used to record a conversation.

It is well known that voiced speech contains strong harmonic components, e.g. [19]. The harmonic characteristic has been exploited by many $F_0$ estimators in order to produce reliable pitch[1] estimates [20], [21]. These systems, however, assume that only one speaker is active at any given time. Multi-pitch tracking methods that estimate the pitch of multiple periodic signals have been proposed in [22]–[27]. However, this paper utilises multi-pitch tracking for the task of overlapping speaker segmentation.

In using multi-pitch tracking to address overlapping speaker segmentation, one clear difficulty arises; that of processing both voiced and unvoiced speech [28]. The intervals of unvoiced speech in a recording do not possess an $F_0$ or any harmonic characteristics [29]. Solutions to this problem have been proposed in the past [30] when pitch features have been used for diarization. This paper intends to deal with the problem of unvoiced regions of speech by use of the tracking approach. This allows spanning of short unvoiced intervals by continuing tracks for short periods even when no observations are detected. This does not, however, solve the problem caused by the presence of unvoiced speech either at the onset or

The support of the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS) grant no. EP/L016796/1 and EPSRC Fellowship grant no. EP/P001017/1 are gratefully acknowledged.

A. O. T. Hogg, A. H. Moore and P. A. Naylor are with the Dept. Electrical and Electronic Engineering, Imperial College London, Exhibition Road, SW7 2AZ, UK (e-mail: aidan@aidanhogg.co.uk, alastair.h.moore@imperial.ac.uk and p.naylor@imperial.ac.uk).

C. Evers is with the Sch. of Electronics and Computer Science, University of Southampton, Burgess Road, SO17 1TU, UK and was previously with the Dept. Electrical and Electronic Engineering, Imperial College London while carrying out this research (e-mail: c.evers@soton.ac.uk).

[1]Here the term 'pitch' is used for the instantaneous value of the fundamental frequency ($F_0$) of a voiced speech signal.

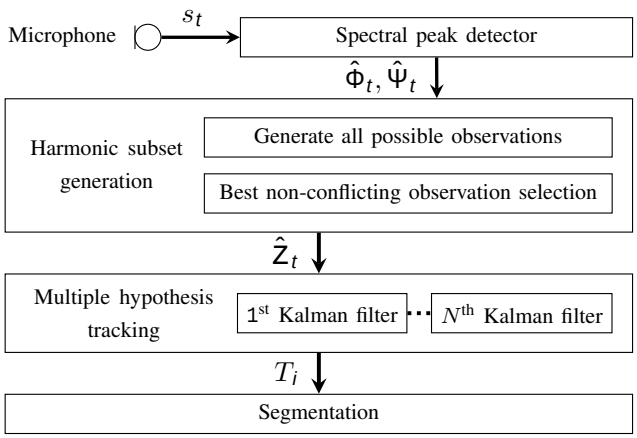



Fig. 1. Proposed system architecture with $s_t$: input signal, $\hat{}_t$: peak detections, $\hat{}_t$: detection reliabilities, $\hat{\mathbf{z}}_t$: selected observations, $T_i$: selected track hypotheses

end-point of a given speaker which will have the effect of delaying or shortening a pitch track. In this work, these errors will be safely ignored as they will be less than 50 ms [31] which is smaller than most collars used to account for human annotation imprecision [32]. The temporal aspect makes the approach different from [33] which utilises the fundamental frequency variation (FFV) spectrum for speaker identification [34], [35].

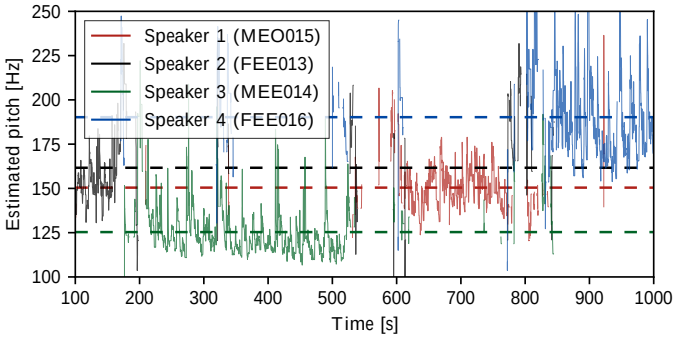
This paper is structured as follows: Section II presents an investigation leading to new insights regarding the $F_0$ of voiced speech. This investigation consists of two parts: (i) a detailed analysis of how variations in $F_0$ can be used to detect when speaker changes occur; (ii) a study of the role that the harmonic structure of voiced speech can play in identifying regions of overlapping speech. Section III presents the proposed method. Section IV explains the experimental setup and Section V presents the results.

## II. Speaker Segmentation Using Fundamental Frequency

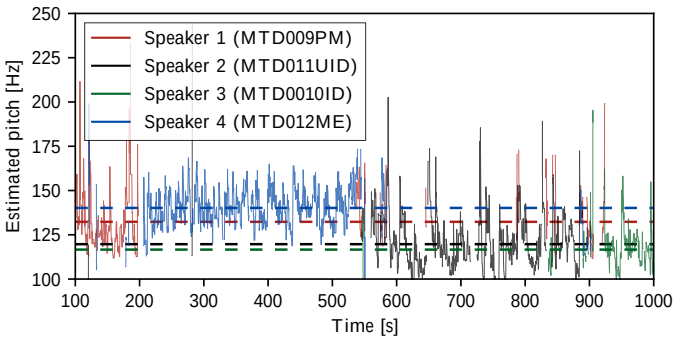
This section explores the relationship between speaker changes and pitch changes. We show that voiced pitch is a meaningful feature for speaker segmentation, including cases of overlapping speech.

### A. Average Variations in Fundamental Frequency

We know that speakers have different pitch variations which can be a useful feature in speaker identification [34]. To investigate we use a Kalman filter applied to pitch estimates calculated using [21], on recordings obtained from individual headset microphones (IHMs). The clean pitch tracks generated for each speaker are shown in Fig. 2 for two meetings taken from the AMI corpus [36]. The pitch tracks for a meeting where there is a variation in $F_0$ of over 50 Hz between the speakers is provided in Fig. 2(a). To exploit these $F_0$ variations, FFV features were proposed in [33]. However, for meetings where the $F_0$ mean and variance are similar between different speakers, FFV features would likely not provide enough information by themselves to detect reliably




(a) Kalman filter pitch tracks where 'ES2004b' is the input.




(b) Kalman filter pitch tracks where 'TS3003b' is the input.

Fig. 2. The individual pitch tracks generated from PEFAC by using the Kalman filter on the individual headset microphones separately. The dashed horizontal lines represent the mean of each speaker. The AMI speaker labels are also given in brackets where the first letter relates to the gender of the speaker, i.e. M: male and F: female.

when there is a change in the active speaker (see Fig. 2(b)). This limitation motivates the need for an investigation into the temporal variations in pitch to determine speaker changes. We hypothesise that pitch change events are a useful indicator of speaker changes.

### B. Temporal Variations in Fundamental Frequency

In this paper, we target the question of "how often is a change in pitch indicative of a change in speaker?". To explore this question on the AMI corpus [36], pitch estimates were calculated using the method of [21] applied to the IHM mixed-down stream of 16 AMI meetings. A Kalman filter [37] was used to track the pitch of the IHM mixed-down single channel stream as proposed in [38]. The Kalman track relies on the smooth variation of a speaker's pitch due to physiological constraints [39]. Table I shows the percentage of pitch changes detected by the Kalman filter that coincide with changes in the speaker. In this experiment a change in pitch was defined to have occurred if the prediction error of the Kalman filter is greater than a threshold experimentally set at 10 Hz. A standard collar of 250 ms [40] was set around each AMI label. A speaker change was deemed to have coincided with a change in pitch if one or more changes in pitch, detected by the Kalman filter, fell within the given collar. The results demonstrate that, of the 16 meetings investigated, a change in speaker will result in a pitch change with 69.41% probability. Therefore, pitch changes can be exploited constructively for speaker change detection. However, it is possible for a speaker

TABLE I
SPEAKER AND PITCH CHANGE ANALYSIS FOR THE AMI CORPUS.

| Meeting | SC $j$ PC |
|---|---|
| ES2004a | 94.49% |
| ES2004b | 89.25% |
| ES2004c | 95.21% |
| ES2004d | 91.85% |
| IS1009a | 96.12% |
| IS1009b | 98.94% |
| IS1009c | 97.67% |
| IS1009d | 98.55% |
| EN2002a | 92.35% |
| EN2002b | 87.01% |
| EN2002c | 79.37% |
| EN2002d | 86.00% |
| TS3003a | 76.54% |
| TS3003b | 76.59% |
| TS3003c | 75.82% |
| TS3003d | 81.34% |
| **Mean** | 88.57% |

| Meeting | PC $j$ SC |
|---|---|
| ES2004a | 78.76% |
| ES2004b | 68.60% |
| ES2004c | 70.22% |
| ES2004d | 73.38% |
| IS1009a | 68.91% |
| IS1009b | 64.27% |
| IS1009c | 59.38% |
| IS1009d | 66.60% |
| EN2002a | 88.59% |
| EN2002b | 83.40% |
| EN2002c | 87.70% |
| EN2002d | 81.02% |
| TS3003a | 52.08% |
| TS3003b | 48.46% |
| TS3003c | 56.47% |
| TS3003d | 62.68% |
| **Mean** | 69.41% |

PC | SC The probability that there is a 'pitch change' given that there is a 'speaker change'

SC | PC The probability that there is a 'speaker change' given that there is a 'pitch change'

change to occur without a change in pitch. As a consequence, better results would be achieved if pitch was not used as a sole feature for speaker segmentation but instead as a significant feature as part of a multimodal approach.

### C. Harmonic Structure of Fundamental Frequency

To extend this study, an analysis of overlapping speech segments is presented. Pitch estimation is not feasible during overlapping speech using single speaker pitch estimation algorithms [7]. Therefore, we propose to exploit the harmonic structure of speech for pitch estimation from overlapping speech signals. Fig. 3 shows two speakers from the TIMIT corpus [41] that have been overlapped in the interval between [0.91, 1.58] s; the spectrogram is the resulting mixture with the pitch tracks, obtained from the individual recordings, being overlaid. This illustrative example suggests that, if multiple $F_0$s can be tracked reliably, then overlapping speakers can be segmented. Fig. 3 also shows that the signals from different speakers differ not only in terms of the mean $F_0$ values of each track but also in the temporal shape of their trajectories. Overlapping speakers may, therefore, be segmented by identifying the onsets and end-points of each speaker's utterance.

## III. METHOD

The proposed method, shown in Fig. 1, includes the following components: (i) Measurements of the harmonic frequencies are first obtained using a spectral peak detector (see Section III-A). (ii) Subsets of the detected peaks are used to generate possible harmonically related observations (see Section III-B). (iii) The best non-conflicting observations (see Section III-C) are used to track the voice pitch of multiple, overlapping speakers using a Kalman filter (see Section III-D). However, the association of subsets of peak detections with the pitch of a specific speaker is unknown *a priori*. This problem is exacerbated by some subsets being related to false alarms (FAs)[2]. (iv) A multiple hypothesis tracking (MHT) method [42], [43], [44] is, therefore, proposed (see Section III-E) that probabilistically updates the pitch tracks using physically feasible subsets of peak detections (see Section III-F). (v) The complete segmentation is obtained from the resulting tracks where the start of a track corresponds to the onset of a speaker and the end of a track corresponds to the end-point of a speaker (see Section III-G).

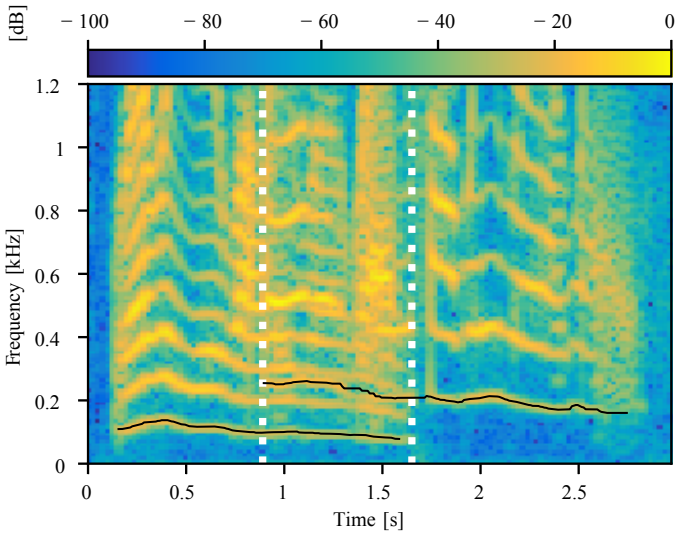



Fig. 3. $F_0$ tracks of a male and female speaker from the TIMIT corpus. The two speakers overlap between the two white dashed vertical lines.

### A. Spectral Peak Detector

The harmonics of voiced speech are estimated using a spectral peak detector that generates a set of $P_t$ peaks, $_t = f\phi_{t;1}, \quad , \phi_{t;P_t}g$ in the short-time Fourier transform (STFT) of the microphone signal at every time frame, $t$. Each peak, $\phi_{t;p}$, for $p \; 2 \; f1, \quad , P_tg$ is associated with a peak amplitude, $\psi_{t;p}$, which captures the detection reliability. A peak, $\phi_{t;p}$, is deemed to be reliable if $\psi_{t;p}$ is greater than a threshold, $\xi$, where $\xi$ can be determined experimentally. Only the reliable peak detections, $\hat{}_t = f\hat{\phi}_{t;1}, \quad , \hat{\phi}_{t;Q_t}g$, where $Q_t \quad P_t$, are retained for each time frame, $t$. The cardinality of $\hat{}_t$ is, therefore, less than or equal to the cardinality of $_t$ and varies over time.

### B. Generate All Possible Observations

Due to the harmonic nature of voiced speech, subsets of elements in $\hat{}_t$ may correspond to integer multiples of the $F_0$ of a speaker. For brevity, the remainder of this paper refers to peak detections corresponding to integer multiples of $F_0$ as 'harmonically related' detections. For multiple speakers, the subsets of elements in $\hat{}_t$ corresponding to harmonically related detections are unknown *a priori*. The association

[2]A FA in this context is when an observation/subset does not relate to the pitch of a speaker.

**Algorithm 1** Generation of observations from pitch measurements.

**for** $(\phi, \psi)$ **in** ( , ) **do** ▷ *inputs: and ; output:* $\mathbf{Z}_t$
  **if** $\psi >$ $\xi$ **then** ▷ *remove unreliable measurements*
    $\hat{}$.append($\phi$)
$Z$ = {} ▷ *all observations,* $\mathbf{Z}_t$*, for a given time frame* $t$
**for** $\hat{\phi}$ **in** $\hat{}$ **do**
  $n$ = 1; $F_0$ = $\hat{\phi}$ ▷ *initialisation*
  **while** $F_{\min} < F_0 < F_{\max}$ **do** ▷ *ignore* $F_0$ *if unrealistic*
    $F_0$ = $\hat{\phi}/n$
    $n$ = $n$ + 1
    $z$ = {} ▷ *possible subset,* $\mathbf{z}_{t;n}$*, for a given time frame* $t$
    **for** $\hat{\phi}$ **in** $\hat{}$ **do**
      **if** $|(\text{round}(\hat{\phi}/F_0) * F_0) - \hat{\phi}| < F_{\text{tol}}$ **then** ▷ $\pm F_{tol}$
        $z$.append($\hat{\phi}$)
    **if** $z$ **not in** $Z$ **and** length($z$) $>$ 1 **then**
      $Z$.append($z$) ▷ *remove* $\mathbf{z}_{t;n}$ *containing 1 harmonic*

between peaks and $F_0$ of a speaker needs to be resolved in order to track the $F_0$ of multiple speakers simultaneously. This is further complicated by the fact that the pitch of two speakers may correspond to integer multiples of each other. Therefore, the association between peak detections and the pitch of each speaker may be ambiguous in some time frames. The task is particularly problematic if the audio signal contains reverberation and noise.

To address the association problem, a probabilistic perspective is adopted. To determine the unknown subsets of harmonically related detections, all possible subsets of $\hat{}_t$, corresponding to integer multiples of $F_0$, within a tolerance of $F_{\text{tol}}$, are computed. Each resulting subset is denoted as an 'observation', $\mathbf{z}_{t;n}$. The generation of these subsets is of a combinatorial nature. To reduce the computational complexity of the problem, an $F_0$ estimate is only considered if it satisfies $F_{\min}$ $F_0$ $F_{\max}$, which defines the physical range of the human speech production system.

Consider the following illustrative example for a given frame: $P_t$ = 5, $_t$ = $\{100, 200, 350, 400, 450\}$ and $_t$ = $\{6.3$ $10^7, 4.5$ $10^7, 4.9$ $10^6, 2.3$ $10^6, 8.2$ $10^4\}$ where $_t$ = $\{\psi_{t;1},$ $, \psi_{t;P_t}\}$ and $_t$ = $\{\phi_{t;1},$ $, \phi_{t;P_t}\}$. If $\xi$ = 1 $10^6$ then $\hat{}_t$ = $\{100, 200, 350, 400\}$. Three alternative associations are feasible if $F_{\min}$ = 50 Hz and $F_{\max}$ = 300 Hz the observations could be interpreted such that $F_0$ is 50 Hz and $\mathbf{z}_{t;0}$ = $[100, 200, 350, 400]^T$. Alternatively $F_0$ could be $F_0$ is 100 Hz and $\mathbf{z}_{t;0}$ = $[100, 200, 400]^T$. Lastly $F_0$ could be 200 Hz and $\mathbf{z}_{t;1}$ = $[200, 400]^T$. Algorithm 1 outlines the generation of the observations from the peak detections.

### C. Best Non-conflicting Observation Selection

At each time frame, $N_t$ observations are generated, such that $\mathbf{Z}_t \triangleq \{\mathbf{z}_{t;0}, \mathbf{z}_{t;1},$ $, \mathbf{z}_{t;N_t}\}$. A particular problem arises when the observations result in multiple tracks for a single speaker at harmonics or subharmonics of $F_0$. The subsequent detection errors are normally classified as harmonic or octave errors, which are also common in $F_0$ estimation algorithms [45]. An $F_0$ observation is said to conflict if it is a harmonic or subharmonic of any other observation. To reduce such errors in our method, only one $F_0$ observation is tracked if multiple $F_0$ observations conflict. At each time frame, $t$, an iterative selection process is utilised to select the best non-conflicting observations. An empty set, $\hat{\mathbf{Z}}_t$, is first initialised. Then, at each iteration, the observation composed of the most measurements, that is the observation vector, $\hat{\mathbf{z}}_{t;0}$, of longest length is appended to $\hat{\mathbf{Z}}_t$ and all observations conflicting with $\hat{\mathbf{z}}_{t;0}$ are removed. If two or more observation vectors have the same length then the one associated with the highest $F_0$ is appended. This iterative process continues until all observations are either appended to $\hat{\mathbf{Z}}_t$, or removed. The $M_t$ $N_t$ selected observations, $\hat{\mathbf{Z}}_t$ = $\{\hat{\mathbf{z}}_{t;0}, \hat{\mathbf{z}}_{t;1},$ $, \hat{\mathbf{z}}_{t;M_t}\}$, at each time frame are then used to form tracks.

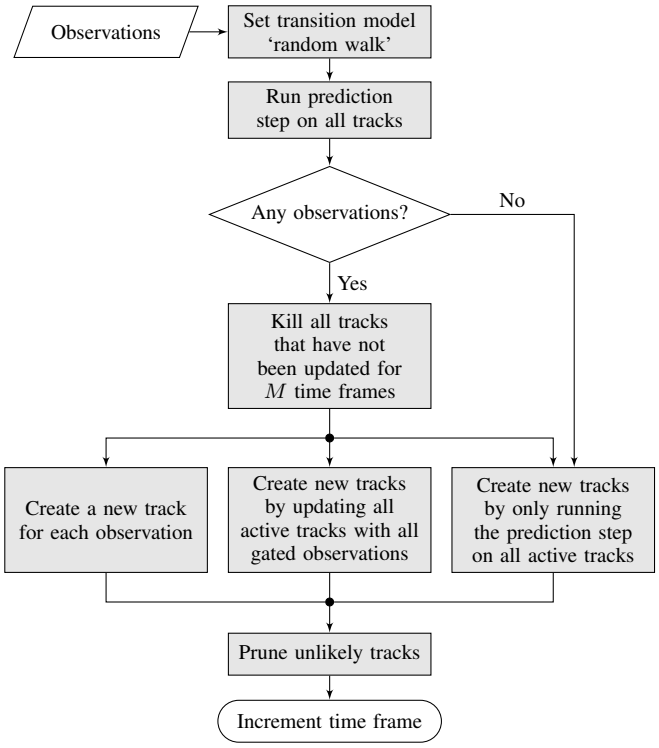



Fig. 4. Multiple hypothesis tracking (MHT) procedure at each time frame

### D. Kalman Filter for Pitch Tracking

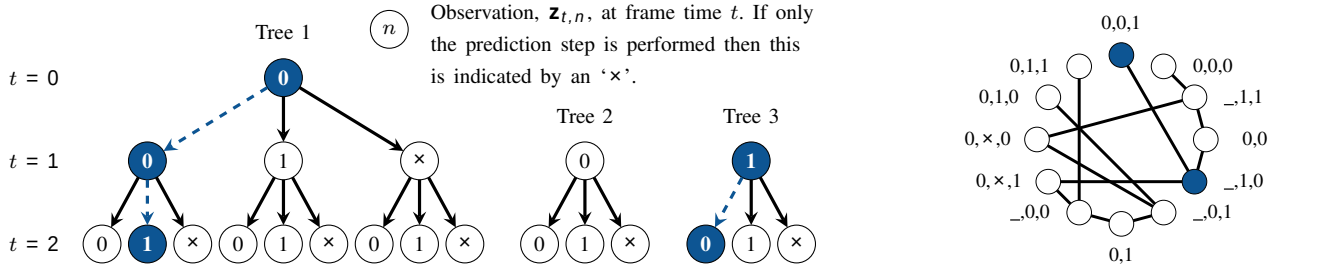
The pitch of active speakers is tracked by multiple Kalman filters [37] at each time frame, $t$. The input observations, $\hat{\mathbf{z}}_{t;n}$, each contain subsets relating to possible harmonically related detections. The Kalman filters track all possible pitch trajectories, $\mathbf{x}_t$ = $\{x_{1;t}, \ldots, x_{I;t}\}$, where $x_{i;t}$ corresponds to the pitch of speaker, $i$ 2 $\{1, \ldots, I\}$. The pitch, $x_{i;t}$, for time frame, $t$, is modelled as

$$x_{i;t} = x_{i;t\ \ 1} + w_{i;t}, \quad w_{i;t} \ 2\ \mathcal{N}(0, \sigma_w^2), \tag{1}$$

where the pitch at $t$ deviates from the pitch at $t$ 1 by a process noise term with a variance of $\sigma_w^2$. The observations, $\hat{\mathbf{z}}_{t;n}$, associated with speaker, $i$, are modelled conditionally on $x_{i;t}$ as

$$\begin{aligned} \hat{\mathbf{z}}_{t;n} &= \mathbf{h}_{t;n} x_{i;t} + \mathbf{v}_t, \quad \mathbf{v}_t \ 2\ \mathcal{N}(0, \mathbf{R}_t), \\ \mathbf{R}_t &= \text{diag}[\sigma_v^2, \quad , \sigma_v^2], \end{aligned} \tag{2}$$

where the covariance, $\mathbf{R}_t$ 2 $\mathbb{R}^{N_t \ \ N_t}$, and variance, $\sigma_v^2$, models the uncertainty in the observations, and $\mathbf{h}_{t;n}$ is a column vector




(a) Possible track hypotheses.

(b) Maximum weighted clique (MWC) in blue.

Fig. 5. Multiple hypothesis tracking (MHT) illustration (a) Track hypotheses generated. (b) An undirected graph, $G$, where each node is a track hypothesis and each edge connects two tracks which are not conflicting. The nodes are indexed using the observations that make up each track.

of the integer multiples of $F_0$ that maps the current state to the harmonic components contained in the current observation. For example, if $\hat{\mathbf{z}}_{t;0} = [100, 200, 400]^T$ (as in the example of Section III-B) where $F_0$ is 100 Hz then $\mathbf{h}_{t;0} = [1, 2, 4]^T$. The Kalman filter operates by estimating the state of the system and then acquiring feedback from the noisy measurements using a prediction step and an update step. The predicted pitch estimate, $\hat{x}_{i;tjt-1}$, and predicted estimation variance, $p_{i;tjt-1}$, are given by

$$\hat{x}_{i;tjt-1} = \hat{x}_{i;t-1jt-1}, \tag{3}$$

$$p_{i;tjt-1} = p_{i;t-1jt-1} + \sigma_w^2. \tag{4}$$

The updated pitch estimate, $\hat{x}_{i;tjt}$, and updated estimation variance, $p_{i;tjt}$, are given by

$$\hat{x}_{i;tjt} = \hat{x}_{i;tjt-1} + \mathbf{k}_{i;t}(\hat{\mathbf{z}}_{t;n} - \mathbf{h}_{t;n}\hat{x}_{i;tjt-1}), \tag{5}$$

$$p_{i;tjt} = (1 - \mathbf{k}_{i;t}\mathbf{h}_{t;n})^2 p_{i;tjt-1} + \mathbf{k}_{i;t}\mathbf{R}_t\mathbf{k}_{i;t}^T. \tag{6}$$

The optimal Kalman gain, $\mathbf{k}_{i;t}$, is a row vector given by

$$\mathbf{k}_{i;t} = p_{i;tjt-1}\mathbf{h}_{t;n}^T\mathbf{S}_{i;t}^{-1}, \tag{7}$$

where innovation variance, $\mathbf{S}_{i;t}$, is a matrix given by

$$\mathbf{S}_{i;t} = \mathbf{h}_{t;n}p_{i;tjt-1}\mathbf{h}_{t;n}^T + \mathbf{R}_t. \tag{8}$$

Therefore, the error between measurement and prediction follows as

$$\mathbf{e}_{i;tjt} = \hat{\mathbf{z}}_{t;n} - \mathbf{h}_{t;n}\hat{x}_{i;tjt}. \tag{9}$$

### *E. Multiple Hypothesis Tracking*

The flowchart in Fig. 4 demonstrates the MHT process at each time frame. At $t = 0$, all observations, $\hat{\mathbf{Z}}_0$, are used to generate $N_0$ new active Kalman filter tracks. For $t > 0$, each observation, $\hat{\mathbf{z}}_{t;n}$, could be interpreted as one of three alternatives: (i) a FA; (ii) the start of a new track or (iii) related to a currently active track. To resolve this uncertainty, all possibilities are expanded and MHT is utilised [46]. Fig. 5(b) shows how new tracks are generated from active tracks and observations. In order to reduce the computational complexity of the problem, gating is also applied to each observation when it is interpreted as being related to an active track. This gating is required in order to stop observations with a low probability of belonging to the active track being used to update the track and generating a new track hypothesis. This gating is dependant on the error between the measurement and the prediction, $\mathbf{e}_{i;tjt}$, whenever an update is performed. To reduce the computational complexity, gating is applied where a track is only updated by an observation if $\bar{e}_{i;tjt}$ is below a threshold, $\zeta$, otherwise the update is rejected. $\bar{e}_{i;tjt}$ is defined as the mean of the absolute values of the estimation error $\mathbf{e}_{i;tjt}$ for time frame $t$. This is because if the observation is too far from the predicted estimate it is considered unlikely to have originated from the active track.

### *F. Maximum Weighted Clique*

The number of generated tracks grows exponentially as new observations become available, as illustrated in Fig. 5(a), for practicality, therefore, pruning, e.g. [44], is required to reduce exponential growth in the number of track hypotheses. Not all tracks can be valid as they may conflict with other tracks, e.g. when more than one track uses one or more of the same observations in their history. The maximum weighted clique (MWC) method [47], [48] is, therefore, used to find the most likely set of tracks which contain no conflicts. An undirected graph, $G = (V, E)$, is shown in Fig. 5(b) where each hypothesis track, $T_i$, is represented by the node set $V = \{T_0, T_1, \ldots, T_L\}$, and the set of edges is $E \subseteq V \times V$, consisting of $M$ edges. A clique is a subgraph of $G$ with pairwise adjacent vertices, meaning that all pairwise vertices $T_i$ and $T_j$ are connected by an edge $(i, j)$. To find the MWC, each node is assigned a score, $w_i$, which is calculated by taking the average value of all previous estimation errors, $\mathbf{e}_{i;tjt}$, evaluated at each update. The MWC solution is the clique that maximises the following optimisation problem:

$$\begin{aligned} \max\ f(q) &= \sum_{i=0}^{L} w_i q_i, \\ \text{s.t. } q_i + q_j &\leq 1, \forall (i,j) \in E, \\ q_i &\in \{0,1\}, \text{ for } i \in \{0,1,\ldots,L\}, \end{aligned} \tag{10}$$

where $q_i = 1$ if the node $T_i$ belongs to the clique and $q_i = 0$ otherwise. In this case $E$ denotes the edge set of the complementary graph of $G$. The pruning technique then operates by calculating the MWC after $k$ time frames and discarding all other tracks.

### *G. Segmentation*

The result of this process is a complete segmentation, obtained from the analysis of the individual tracks. The onsets and end-points of a speaker correspond to times of initialisation and termination respectively of the corresponding tracks.

## IV. Experimental Setup

This section summarises two experiments that are carried out to evaluate the performance of the proposed system. Two baselines are introduced for performance and complexity comparison. Code available at [50].

### A. Exp-1: Full Segmentation using Proposed System

Exp-1 evaluates the performance of the proposed method as a complete segmentation system. The proposed method is compared against two baselines: baseline-1, previously presented by the authors in [49] and baseline-2, a state-of-the-art deep learning approach presented in [51].

*1) System Architecture of Baseline-1:* Baseline-1, shown in Fig. 6, and the proposed approach, shown in Fig. 1, differ in that baseline-1 keeps and tracks all generated subset information and, therefore, requires post-processing. The post-processing required by baseline-1 is due to the possibility that multiple tracks can be generated by a single speaker if conflicting observations are not removed as described in Section III-C. The post-processing consists of two components: 1) overlapping speech detection and 2) speaker change detection.

**Overlapping speech detection:** To determine the activity of multiple, simultaneously active speakers, uncorrelated tracks need to be identified. Any changes in the track cardinality are first detected. Fig. 7 shows an illustrative example where the track cardinality changes from 3 to 5 at time frame, $o_t$. The harmonic relation between tracks is compared to confirm overlapping speech between the track cardinality changes. A two step process is utilised to test whether two tracks are harmonically related and belong to the same speaker. Step-1 compares the mean $F_0$ of corresponding tracks, within a tolerance of $F_{\text{tol}}$, to an integer-multiple of any other tracks. Candidates that are not harmonically related indicate overlapping speech, such as $\{100, 124, 197, 259, 299\}$ (see Fig. 7 for an illustrative example). If the mean $F_0$ of two tracks is deemed to be harmonically related, such as $\{102, 199, 303\}$ in Fig. 7, then step-2 compares the similarity of the estimated trajectories. The trajectories are compared by evaluating the

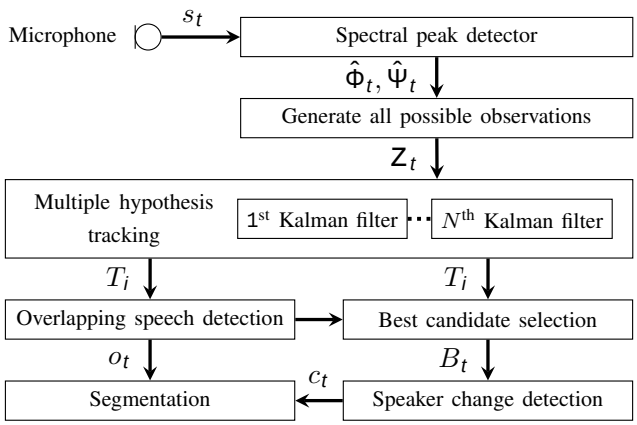



Fig. 6. Baseline-1 system architecture presented in [49] with $s_t$: input signal, $\hat{}_t$: peak detections, $\hat{}_t$: detection reliabilities, $\mathbf{Z}_t$: generated observations, $T_i$: selected track hypotheses, $o_t$: overlapping speech onsets, $B_t$: strongest candidate track and $c_t$: speaker change onsets.

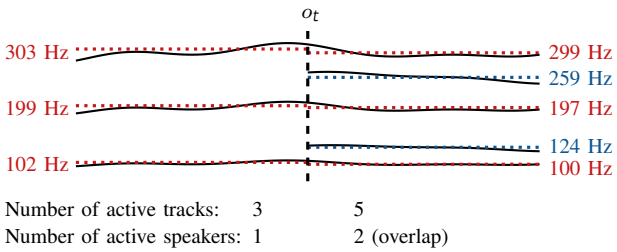



Fig. 7. An illustrative example of overlapping speech detection using the baseline system architecture

mean squared error (MSE) between all pairs in the set. Pairs corresponding to a MSE above a threshold, $\eta$, indicate a speaker overlap. The overlapping speech onset, $o_t$, can also be thought of as an onset of a new speaker prior to the previous speaker's end-point. This process solves the problem of one speaker generating multiple tracks since all the tracks generated by the same speaker will have the same trajectory and will also be harmonically related.

**Speaker change detection:** To segment overlapping speech, it is also necessary to detect the speaker changes, $c_t$, i.e. the onset of a new speaker after the end-point of the previous speaker. Speaker change detection is achieved by exploiting the temporal variations in the pitch. To accomplish this the method of [38], that operates using a single track, is utilised here. It was shown in Section III-C that multiple tracks can be generated for the same speaker. Therefore, in this investigation, the tracks are pruned to leave only the track that corresponds to the $F_0$ of the speaker. First, the $\hat{\psi}_{t;p}$ values relating to all measurements, $\hat{\phi}_{t;p}$, that correspond to a given track are summed,

$$B_i = \sum_{t=0}^{T} \sum_{q=0}^{Q} \hat{\psi}_{t;q}, \tag{11}$$

for the periods where multiple tracks are active. $B_i$ is a summary statistic for each track, $T_i$, where $i \in \{0, 1, \ldots, L\}$. The tracks, $T_i$, are then pruned by selecting the track corresponding to the maximum value of $B_i$

**Segmentation:** The complete speaker segmentation is finally determined as the union of the sets of speaker changes, $c_t$, with the overlapping speech onsets, $o_t$.

*2) Parameters of Baseline-1 and the Proposed System:* In the Exp-1 and Exp-2, a modified version of PEFAC [21], [52] was developed as the spectral peak detector for both the proposed system and baseline-1. The modification removes the restriction, in PEFAC, that the filter used to detect the harmonics is centred in the limited range of 0.9 - 1.1 times the fundamental. The parameters selected for the proposed system and baseline-1 in the experiments, as shown in this paper, are given in Table II where initialisation uses both physiological constraints of $F_{\text{min}}$, $F_{\text{max}}$, $F_{\text{tol}}$, $\sigma_w$, $\sigma_v$ and empirical tuning of $P$, $\xi$, $\zeta$, $k$, $\eta$. This empirical tuning was achieved through experimentation on the development set of the AMI corpus.

TABLE II
Parameter Setting for the Proposed and System Baseline-1

| $P$ | $\xi$ | $F_{\text{min}}$, $F_{\text{max}}$, $F_{\text{tol}}$ | $\sigma_w$, $\sigma_v$ | $\zeta$ | $k$ | $\eta$ |
|---|---|---|---|---|---|---|
| 20 | $1.0 \quad 10^6$ | 50, 300, 5 | 10, 400 | 25 | 1 | 7 |

*3) Computational Complexity of Baseline-1 and the Proposed System:* The baseline-1 approach is computationally demanding as all generated subset information needs to be kept and tracked. In contrast, the proposed approach has a lower computational complexity making it a more efficient implementation. To highlight the efficiency of the proposed approach, consider the computational complexity for a given time frame, $t$, with $A_t$ active tracks and $N_t$ observations. If all tracks are updated with all observations, then $A_t \quad N_t$ possible new tracks are created. It is also possible that all $N_t$ observations are wrong and, therefore, for each active track only the prediction step is performed, creating $A_t$ further tracks. Assuming in this analysis that no tracks are terminated and no observations are discarded due to gating, the total number of possible new tracks is $(A_t \quad N_t) + A_t$. The MWC is then computed using the Bron-Kerbosch algorithm [53], an enumeration algorithm for finding MWCs in an undirected graph, where each possible track represents a node. Calculating the MWC at each time step is the most expensive operation and, therefore, the Bron-Kerbosch algorithm dominates the complexity $O$-number. In the worst case, the time complexity for the Bron-Kerbosch algorithm is $O(3^{\frac{L}{3}})$ for an $L$-node graph. Accordingly, the complexity for each time frame is

$$f = O(3^{(A_t N_t + A_t)=3}). \tag{12}$$

Therefore, the number of tracks, $A_t$, and observations, $N_t$, at each time frame determine the computational complexity. The reduced computation of the proposed method over baseline-1 is achieved due to the early pruning in order to reduce the number of observations at each stage. Moreover, the proposed approach benefits from not requiring a post-processing step since each speaker corresponds to exactly one track.

*4) Baseline-2 in Exp-1:* Baseline-2 is a state-of-the-art deep learning approach. This task can be formulated as a sequence labelling task where the input is the sequence of feature vectors

$$\mathsf{X} = f\mathsf{x}_1, \mathsf{x}_2, \quad , \mathsf{x}_T g, \tag{13}$$

where $\mathsf{x}_t$ is a sequence of frame features extracted on a short overlapping sliding window and $T$ is the total number of features. The output is denoted by the corresponding sequence of labels

$$\mathsf{y} = f y_1, y_2, \quad , y_T g \ 2 \ f0, 1g^T. \tag{14}$$

The `pyannote.audio` [51] framework was used to train a neural network $f : \mathsf{X} \ ! \ \mathsf{y}$ that matches a feature sequence $\mathsf{X}$ to the corresponding label sequence $\mathsf{y}$. If there is a speaker change at frame $t$ then $y_t = 1$ otherwise $y_t = 0$.

**Feature extraction:** The waveform is used directly where $\mathsf{x}_t$ is SincNet learnable features [32].

**Network architecture:** The model stacks 2 bidirectional long short term memory networks (BLSTMs) and a multi-layer perceptron, each with 128 units in both forward and backward directions, and a final classification layer (2 units, softmax activation).

**Training:** The network was trained for 1000 epochs on the AMI database using the training set given in the 'full-corpus partition of meetings' [36]. The training configuration is the same as the original paper [32], [51]. To address the class imbalance problem and to account for human annotation imprecision, a positive neighbourhood of 100 ms (50 ms each side) is used around each speaker change event. The training is implemented using the Keras toolkit [54].

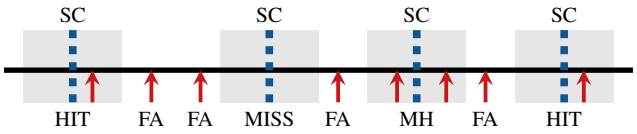



Fig. 8. Illustrative example of the evaluation framework used in Exp-1 where the blue dashed lines represent the oracle speaker change boundaries and the grey regions correspond to the given collar. A 'HIT' is where a speaker change has been detected once. A 'MISS' is when a speaker change has not been detected and multi-hit, 'MH', is where a speaker change has been detected multiple times within its collar. A FA is when a detection falls outside of any speaker change collars.

**Targets:** The speaker change labels are obtained from the ground-truth AMI annotation files.

*5) Evaluation Framework used for Exp-1:* To evaluate the performance of the proposed system, the following metrics have been defined. A 'HIT' is when a speaker change has been detected once. A 'MISS' is when a speaker change has not been detected and a multi-hit (MH) is when a speaker change has been detected multiple times within a time collar applied around every ground-truth speaker change in order to account for possible inaccuracies [40]. The HIT rate is given by

$$\frac{\text{HITs + MHs}}{\text{HITs + MHs + MISSs}}. \tag{15}$$

The false alarm (FA) rate is given by

$$\frac{\text{FAs}}{\text{HITs + MHs + FAs}}. \tag{16}$$

The MH rate is given by

$$\frac{\text{MHs}}{\text{HITs + MHs}}. \tag{17}$$

These detection types are defined graphically in Fig. 8 where the scores would be as follows: HIT rate: 75%, MISS rate: 25%, MH rate: 25%, FA rate: 57%. A standard collar of 250 ms is used for Exp-1. The MSE in time is also calculated for all the HITs and the closest MH detections to the ground-truth given by the AMI annotation files.

### *B. Exp-2: Full Segmentation using the Pitch Tracks as Features in Pitch-BLSTM*

*1) Proposed Pitch-BLSTM Model:* Exp-2 evaluates the performance of the proposed method when used as an input feature to the pitch-BLSTM, the same BLSTM used for baseline-2. This Section shows how pitch can be used as an input feature to a deep neural network (DNN) along with standard MFCC features.

**Feature extraction:** Two features are extracted for use in Exp-2. A 59 dimensional MFCC feature is extracted using `librosa` [55] as shown in Table III, and a 26 dimensional pitch feature is extracted using the proposed method. To extract this pitch feature using the proposed system at each frame, 26 bins are created at intervals of 10 Hz in the

TABLE III
MFCC FEATURE.

| 59 | MFCC Feature |
|---|---|
| 19 | Cepstral coefficients |
| 19 | Delta cepstral coefficients |
| 19 | Delta delta cepstral coefficients |
| 1 | Delta energy coefficients |
| 1 | Delta delta energy coefficients |

range from 50 to 300 Hz. Each bin is assigned a value '1' if a track is contained in that bin and '0' otherwise. This makes it possible to create a feature vector of fixed length that captures the number of $F_0$ tracks and their frequency information at each frame. A comparison is made by performing segmentation using the 59 dimensional MFCC feature alone and in conjunction with the 26 dimensional pitch feature.

**Network architecture:** The pitch-BLSTM uses the same network architecture as baseline-2 (see Section IV-A4).

**Training:** The network was trained for 200 epochs on the same AMI training set used in Exp-1. The same training configuration as Exp-1 was also used.

**Targets:** The same targets as Exp-1 are used which were obtained from the ground-truth AMI annotation files.

*2) Exp-2: Evaluation Framework:* To evaluate the performance of the pitch-BLSTM when trained on different features, `pyannote.metrics` [56] was utilised. Two metrics are calculated: 1) the segment-wise coverage, which is the ratio of the duration of the intersection with the most co-occurring hypothesis segment, and the duration of the reference segment; 2) the purity, which would be the same as the coverage if the reference and hypothesis segments were to switch roles and indicates how pure the hypothesis is for each segment. The results presented in Section V-B are a duration-weighted average over each segment.

## V. EXPERIMENTAL EVALUATION

In this Section the performance of Exp-1 and Exp-2 will be evaluated on the AMI corpus.

### A. Exp-1

Exp-1 is evaluated (see Section IV-A5) on AMI and the statistical results are given in Table IV. An illustrative example is also provided in Fig. 9 to compare the proposed method against baseline-1.

*1) Illustrative Example on AMI:* To illustrate the operation of the proposed and baseline-1 methods, a speech segment from meeting 'TS3003b' in the AMI corpus [36] was selected in Fig. 9. Fig. 9(a) shows $\hat{}_t$ generated from PEFAC. Fig. 9(b) shows the observations, $\mathsf{Z}_t$, generated from the pitch measurements, $\hat{}_t$, (see Section III). Fig. 9(c) highlights how multiple tracks, $T_i$, can be generated for the same speaker when baseline-1 is used. Fig. 9(c) also shows how all these tracks are still harmonically related to each other. Fig. 9(d) shows the effect of choosing the best non-conflicting observation used in the proposed method. Lastly, Fig. 9(e) demonstrates how post-processing is not needed as the single speaker no longer has multiple tracks associated with it. To compare the performance of the proposed method, Fig. 9(f), and baseline-1, Fig. 9(g), for this speech segment the results for both methods are included. Baseline-1 results in 3 errors

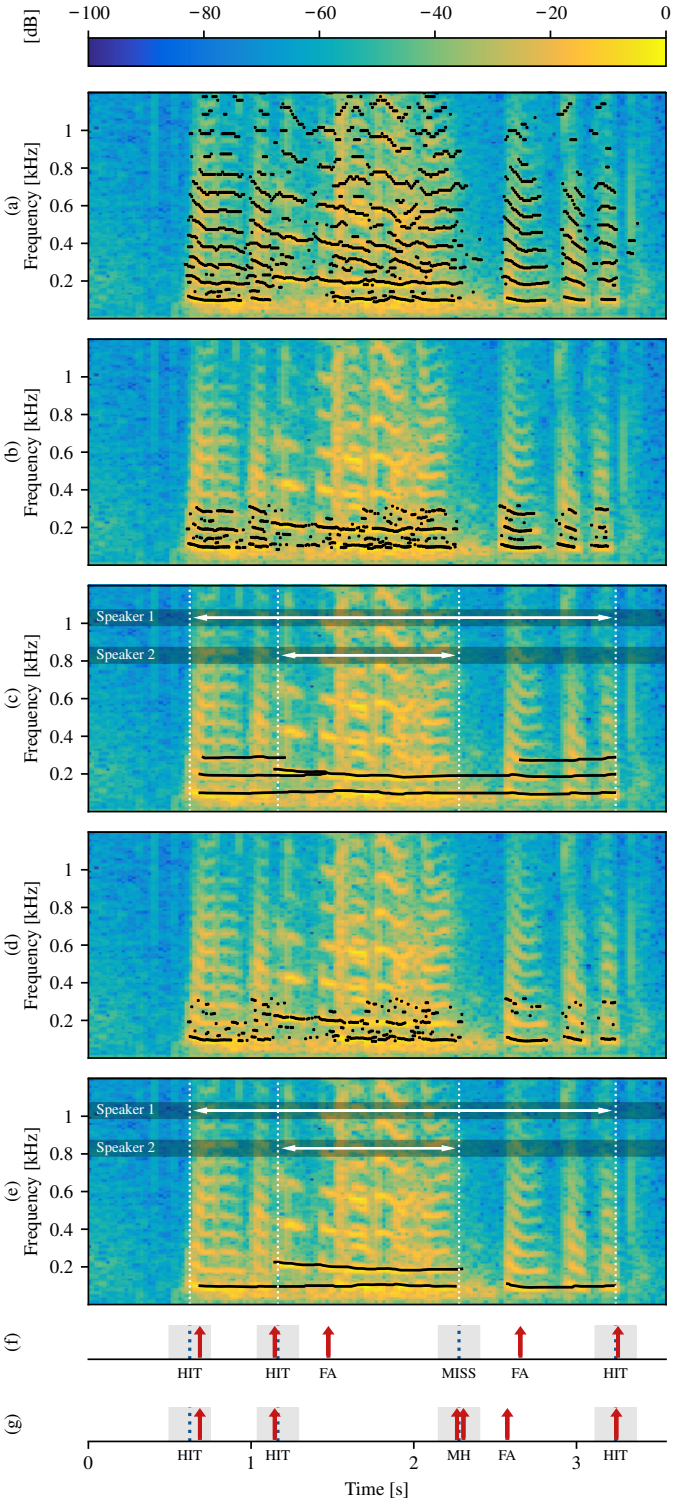



Fig. 9. Illustrative AMI example. Top-to-bottom: (a) Generated output from PEFAC, (b) generated observations where the black crosses show the $F_0$ value for each observation, (c) generated tracks from all possible observations (baseline-1 before post-processing), (d) pruned observations and (e) generated tracks from best non-conflicting observations, (f) baseline-1 performance (g) proposed method performance.

TABLE IV
PERFORMANCE COMPARISON OF BOTH THE IHM MIXED-DOWN STREAM AND THE SDM STREAM ON THE MULTI-SPEAKER MEETINGS IN THE AMI CORPUS ALONG WITH THE BLSTM APPROACH USING A COLLAR OF 250 MS. SUPPLEMENTARY RESULTS FOR BASELINE-1 AVAILABLE AT [50].

| Meeting | Proposed Pitch System | | | | | | | | | | BLSTM Baseline System (Baseline-2) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IHM Mixed-Down Stream | | | | | SDM Stream | | | | | IHM Mixed-Down Stream | | | | | SDM Stream | | | | |
| | HIT | MISS | MH | MSE | FA | HIT | MISS | MH | MSE | FA | HIT | MISS | MH | MSE | FA | HIT | MISS | MH | MSE | FA |
| **EN2002a** | 78.7% | 21.3% | 52.1% | 0.014 | 64.9% | 80.0% | 20.0% | 53.9% | 0.014 | 64.7% | 83.6% | 16.4% | 38.1% | 0.008 | 47.2% | 67.4% | 32.6% | 67.4% | 0.011 | 55.6% |
| **EN2002b** | 83.0% | 17.0% | 58.4% | 0.014 | 68.5% | 80.2% | 19.9% | 55.2% | 0.015 | 70.4% | 81.5% | 18.5% | 36.0% | 0.008 | 53.6% | 69.0% | 31.0% | 69.0% | 0.011 | 56.3% |
| **EN2002c** | 82.1% | 17.9% | 57.8% | 0.014 | 71.5% | 80.0% | 20.0% | 54.0% | 0.015 | 73.8% | 81.5% | 18.6% | 35.3% | 0.007 | 52.2% | 70.3% | 29.7% | 70.3% | 0.012 | 61.2% |
| **EN2002d** | 78.3% | 21.7% | 54.4% | 0.014 | 66.6% | 80.8% | 19.1% | 54.5% | 0.015 | 66.9% | 84.8% | 15.2% | 40.0% | 0.008 | 51.1% | 71.9% | 28.1% | 71.9% | 0.012 | 55.3% |
| **ES2004a** | 72.8% | 27.2% | 48.9% | 0.015 | 73.6% | 75.0% | 25.0% | 53.0% | 0.015 | 76.1% | 60.7% | 39.3% | 22.5% | 0.010 | 54.0% | 51.3% | 48.7% | 51.3% | 0.015 | 65.2% |
| **ES2004b** | 70.8% | 29.2% | 42.1% | 0.015 | 81.0% | 75.8% | 24.1% | 51.7% | 0.014 | 81.3% | 61.8% | 38.2% | 19.3% | 0.010 | 54.8% | 58.5% | 41.5% | 58.5% | 0.014 | 73.6% |
| **ES2004c** | 72.9% | 27.1% | 43.2% | 0.016 | 78.2% | 77.2% | 22.9% | 49.4% | 0.015 | 79.5% | 63.4% | 36.6% | 24.3% | 0.008 | 45.0% | 66.8% | 33.2% | 66.8% | 0.012 | 72.9% |
| **ES2004d** | 75.2% | 24.8% | 49.1% | 0.014 | 71.0% | 76.2% | 23.8% | 50.6% | 0.013 | 74.7% | 59.4% | 40.6% | 20.0% | 0.010 | 48.4% | 66.1% | 33.9% | 66.1% | 0.012 | 61.5% |
| **ES2014a** | 61.6% | 38.4% | 33.7% | 0.015 | 85.3% | 61.6% | 38.4% | 33.7% | 0.015 | 85.3% | 50.8% | 49.2% | 21.4% | 0.014 | 61.9% | 46.9% | 53.1% | 15.7% | 0.012 | 81.8% |
| **ES2014b** | 67.3% | 32.7% | 42.3% | 0.015 | 83.4% | 67.3% | 32.7% | 42.3% | 0.015 | 83.4% | 43.8% | 56.2% | 14.9% | 0.010 | 61.7% | 47.4% | 52.6% | 15.3% | 0.013 | 80.0% |
| **ES2014c** | 67.3% | 32.7% | 38.7% | 0.016 | 82.0% | 67.3% | 32.7% | 38.7% | 0.016 | 82.0% | 45.3% | 54.7% | 15.0% | 0.012 | 61.0% | 51.6% | 48.4% | 17.5% | 0.013 | 79.5% |
| **ES2014d** | 69.2% | 30.8% | 43.6% | 0.015 | 77.1% | 69.2% | 30.8% | 43.6% | 0.015 | 77.1% | 52.8% | 47.2% | 17.6% | 0.013 | 61.4% | 49.6% | 50.4% | 12.9% | 0.013 | 76.5% |
| **IS1009a** | 72.7% | 27.3% | 44.9% | 0.014 | 77.0% | 73.5% | 26.5% | 49.6% | 0.014 | 77.0% | 72.8% | 27.1% | 31.3% | 0.008 | 52.1% | 66.5% | 33.5% | 66.5% | 0.008 | 64.7% |
| **IS1009b** | 69.2% | 30.8% | 41.0% | 0.016 | 79.2% | 72.2% | 27.8% | 44.0% | 0.016 | 80.7% | 73.6% | 26.4% | 29.2% | 0.008 | 47.3% | 67.2% | 32.8% | 67.2% | 0.010 | 67.3% |
| **IS1009c** | 74.5% | 25.5% | 43.5% | 0.014 | 84.5% | 70.3% | 29.7% | 40.6% | 0.015 | 86.3% | 65.3% | 34.7% | 22.4% | 0.008 | 63.3% | 68.8% | 31.2% | 68.8% | 0.009 | 76.2% |
| **IS1009d** | 72.3% | 27.7% | 45.0% | 0.014 | 75.8% | 76.2% | 23.8% | 48.7% | 0.015 | 76.0% | 64.5% | 35.5% | 23.0% | 0.010 | 49.3% | 67.0% | 33.0% | 67.0% | 0.010 | 63.4% |
| **TS3003a** | 68.5% | 31.5% | 39.5% | 0.015 | 85.2% | 70.7% | 29.3% | 43.0% | 0.015 | 88.4% | 43.6% | 56.4% | 11.1% | 0.011 | 71.9% | 45.2% | 54.8% | 45.2% | 0.014 | 81.6% |
| **TS3003b** | 76.0% | 24.0% | 43.2% | 0.015 | 83.8% | 80.7% | 19.3% | 53.5% | 0.016 | 85.8% | 51.8% | 48.2% | 23.2% | 0.017 | 51.7% | 50.3% | 49.7% | 50.3% | 0.021 | 79.7% |
| **TS3003c** | 74.2% | 25.8% | 42.8% | 0.015 | 86.4% | 76.8% | 23.2% | 49.5% | 0.015 | 89.0% | 51.6% | 48.4% | 14.9% | 0.018 | 68.9% | 58.7% | 41.3% | 58.7% | 0.021 | 81.1% |
| **TS3003d** | 75.0% | 25.0% | 43.2% | 0.016 | 74.7% | 80.8% | 19.2% | 57.9% | 0.014 | 75.1% | 64.7% | 35.3% | 24.8% | 0.015 | 48.3% | 59.9% | 40.1% | 59.9% | 0.015 | 61.2% |
| **TS3007a** | 68.2% | 31.9% | 39.4% | 0.015 | 77.5% | 68.2% | 31.9% | 39.4% | 0.015 | 77.5% | 61.9% | 38.1% | 22.8% | 0.011 | 49.6% | 52.2% | 47.8% | 16.8% | 0.009 | 75.5% |
| **TS3007b** | 80.1% | 19.9% | 57.9% | 0.014 | 84.9% | 80.1% | 19.9% | 57.9% | 0.014 | 84.9% | 71.2% | 28.8% | 22.5% | 0.007 | 41.9% | 62.4% | 37.6% | 14.2% | 0.009 | 80.5% |
| **TS3007c** | 74.0% | 26.0% | 50.0% | 0.014 | 76.6% | 74.0% | 26.0% | 50.0% | 0.014 | 76.6% | 77.4% | 22.6% | 33.3% | 0.009 | 51.4% | 67.8% | 32.2% | 23.8% | 0.009 | 73.1% |
| **TS3007d** | 79.4% | 20.6% | 57.9% | 0.014 | 68.2% | 79.4% | 20.6% | 57.9% | 0.014 | 68.2% | 81.3% | 18.6% | 34.4% | 0.009 | 46.6% | 71.2% | 28.8% | 22.4% | 0.010 | 67.8% |
| **Mean** | 73.5% | 26.5% | 46.4% | 0.015 | 77.4% | 74.7% | 25.3% | 48.9% | 0.015 | 78.4% | 64.5% | 35.5% | 24.9% | 0.010 | 53.9% | 60.6% | 39.4% | 47.6% | 0.012 | 70.5% |

(1 MISS and 2 FA) caused by the $F_0$ track of speaker 2 being very similar to the first harmonic of speaker 1. This highlights a disadvantage of baseline-1 where all the observations are tracked. The proposed method contains two errors (1 MH and 1 FA) due to an unvoiced region of speech from Speaker 1 between [2.28, 2.58] s. The proposed method attempts to deal with this problem of unvoiced speech by continuing tracks even after the $F_0$ is no longer detectable. This can be seen earlier in the example between [1.21, 1.45] s where there are gaps in the detection of the pitch track from Speaker 1.

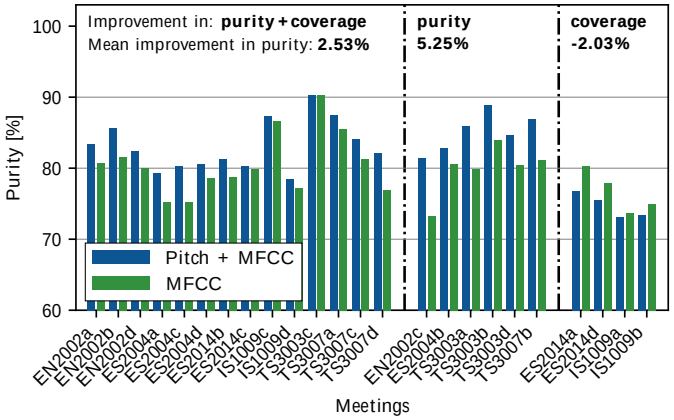
*2) Statistical Results on AMI:* To evaluate the performance, the proposed method is compared against a commonly used baseline-2 method [51] where the model, provided in the paper, was trained on the AMI corpus [36]. Two types of microphone inputs were evaluated on 24 meetings in the AMI corpus: (i) a mixed-down IHM stream containing the sum of close talking microphones without significant reverberation or noise; (ii) a SDM containing room reverberation and ambient noise. The results in Table IV show that the proposed method achieves a better HIT rate using only the pitch information compared to baseline-2 which uses SincNet learnable features extracted from the raw waveform. One explanation for the improvement in the HIT rate seen on the proposed system is an improvement in detecting overlapping speech events in the spoken backchannel, such as 'um' and 'uh-huh'.

The proposed method also achieves an improvement of 1.2% on the HIT rate on the SDM stream compared against the performance on the IHM stream. The improvement in HIT rate does, however, need to be traded off against a degradation of 1.0% in the FA rate. What should be noted is that the 1.2% improvement in HIT rate and a slight increase in FA rate does show that the proposed method is not massively affected by the presence of noise and reverberation. Baseline-2 is more affected by these degradations which can be seen from a drop in HIT rate of 3.9% and an increase in the FA rate by 16.6% between the IHM and SDM stream.

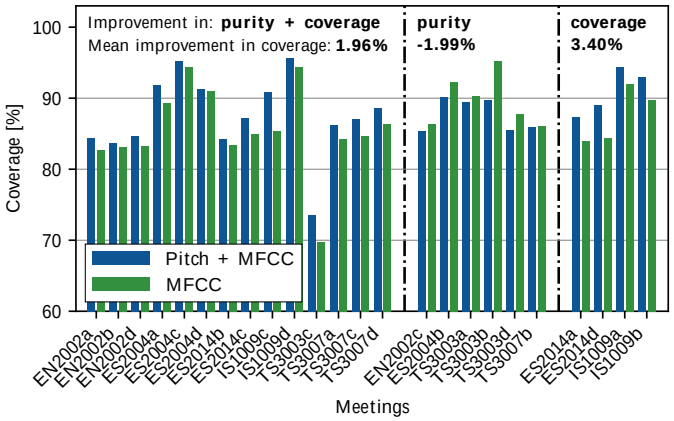
### *B. Exp-2*

In its simplest, direct form, the proposed system only relies on pitch information. This motivates consideration of the proposed method as part of a multimodal approach. Such a multimodal approach can be formulated by using the proposed system to calculate pitch features and using those features in conjunction with other audio features as an input to the pitch-BLSTM system. The network architecture and feature extraction of the pitch-BLSTM system along with a description of the data used for training is elaborated on in Section IV-B1.

*1) Statistical Results on AMI:* The performance is evaluated on the AMI SDM stream and the results for the coverage and purity can be seen in Fig. 10. The increase in performance by including pitch features generated from the proposed system is evident. 20 out of 24 meetings have an improvement in purity; the largest improvement being 5.51% for meeting 'IS1009c'. The coverage is also improved by incorporating pitch with improvements seen in 18 out of 24 meetings; the largest improvement being in meeting 'EN2002c' which increased by 8.15%. In 14 out of 24 meetings, both the purity and the coverage is improved by incorporating the proposed pitch features whereas there are no meetings where the performance is worse for both measures.




(a) Purity performance improvement where the mean improvement for each subgroup is also given. The mean improvement for all meetings is 2.45%.




(b) Coverage performance improvement where the mean improvement for each subgroup is also given. The mean improvement for all meetings is 1.21%.

Fig. 10. Performance comparison of the pitch-BLSTM system using pitch and MFCCs as input features on the SDM stream. The meetings are ordered alphabetically in three subgroups. The first group sees an improvement in both metrics; the second group only sees an improvement in purity and the last group only sees an improvement in coverage.

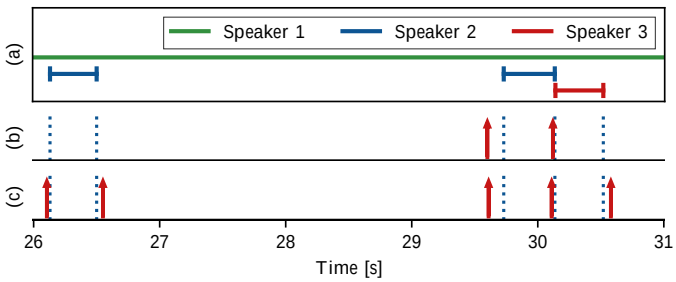



Fig. 11. AMI meeting 'IS1009c' between [18:26, 18:31] mins. (a) Reference given by the AMI labels where speaker 1 is 'FIO084', speaker 2 is 'FIO089' and speaker 3 is 'FIE088'. (b) Segmentation generated from pitch-BLSTM using only MFCCs as input features. (c) Segmentation generated from pitch-BLSTM using both MFCCs and pitch, extracted from the proposed method, as input features.

This improvement is best explained by better overlapping speech detection. To illustrate this for meeting 'IS1009c', an example is given in Fig. 11 that shows the improvements to overlapping speech detection when pitch features from the proposed system are utilised. In Fig. 11 Speaker 1 is active for the full 6 seconds shown while Speakers 2 and 3 voice the phatic expressions of 'mm-hm' to signify they are listening.

TABLE V
PERFORMANCE COMPARISON OF THE PITCH-BLSTM SYSTEM USING PITCH AND MFCCS AS INPUT FEATURES ON THE SDM STREAM WITH A COLLAR OF 250 MS.

| Meeting | Pitch + MFCC Features | | | | | MFCC Features | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | MISS | MH | MSE | FA | HIT | MISS | MH | MSE | FA |
| **EN2002a** | 80.9% | 19.1% | 60.1% | 0.015 | 24.4% | 80.7% | 19.3% | 64.3% | 0.016 | 27.6% |
| **EN2002b** | 81.0% | 18.9% | 54.3% | 0.015 | 28.7% | 80.1% | 19.9% | 61.9% | 0.015 | 31.8% |
| **EN2002c** | 74.2% | 25.8% | 50.6% | 0.014 | 29.5% | 69.0% | 30.9% | 51.0% | 0.017 | 34.5% |
| **EN2002d** | 81.8% | 18.2% | 60.3% | 0.016 | 23.4% | 81.7% | 18.3% | 64.2% | 0.018 | 27.4% |
| **ES2004a** | 59.1% | 40.9% | 36.6% | 0.021 | 29.9% | 57.5% | 42.5% | 43.8% | 0.020 | 32.9% |
| **ES2004b** | 59.2% | 40.8% | 36.9% | 0.017 | 24.8% | 53.9% | 46.1% | 32.0% | 0.016 | 29.4% |
| **ES2004c** | 53.0% | 47.0% | 29.5% | 0.020 | 24.4% | 52.4% | 47.6% | 30.9% | 0.021 | 27.9% |
| **ES2004d** | 63.3% | 36.7% | 38.4% | 0.016 | 27.0% | 61.1% | 38.9% | 39.9% | 0.019 | 30.9% |
| **ES2014a** | 47.3% | 52.7% | 25.4% | 0.023 | 44.1% | 60.5% | 39.5% | 50.0% | 0.020 | 41.5% |
| **ES2014b** | 49.7% | 50.3% | 31.1% | 0.020 | 40.7% | 48.0% | 52.0% | 34.9% | 0.023 | 47.6% |
| **ES2014c** | 57.2% | 42.8% | 33.7% | 0.023 | 33.9% | 60.9% | 39.1% | 42.2% | 0.020 | 37.0% |
| **ES2014d** | 44.5% | 55.5% | 23.7% | 0.019 | 36.2% | 53.3% | 46.7% | 36.8% | 0.018 | 41.6% |
| **IS1009a** | 45.5% | 54.5% | 25.1% | 0.022 | 33.6% | 45.0% | 55.0% | 22.8% | 0.016 | 39.6% |
| **IS1009b** | 46.2% | 53.8% | 22.0% | 0.020 | 25.5% | 50.1% | 49.9% | 30.1% | 0.021 | 28.9% |
| **IS1009c** | 50.1% | 49.9% | 32.0% | 0.016 | 34.5% | 40.0% | 60.0% | 22.9% | 0.016 | 60.4% |
| **IS1009d** | 46.5% | 53.5% | 27.3% | 0.016 | 23.9% | 38.6% | 61.4% | 18.4% | 0.025 | 35.3% |
| **TS3003a** | 42.2% | 57.8% | 24.9% | 0.026 | 58.8% | 28.5% | 71.5% | 16.7% | 0.020 | 64.1% |
| **TS3003b** | 48.1% | 51.9% | 32.1% | 0.019 | 45.6% | 27.2% | 72.8% | 15.6% | 0.026 | 52.8% |
| **TS3003c** | 65.1% | 34.9% | 41.2% | 0.018 | 60.3% | 60.1% | 39.9% | 39.1% | 0.020 | 65.4% |
| **TS3003d** | 64.0% | 36.0% | 39.7% | 0.018 | 38.1% | 55.3% | 44.7% | 35.5% | 0.018 | 34.5% |
| **TS3007a** | 67.5% | 32.5% | 44.6% | 0.016 | 26.2% | 66.6% | 33.4% | 51.9% | 0.015 | 28.9% |
| **TS3007b** | 67.9% | 32.1% | 43.1% | 0.014 | 23.9% | 53.9% | 46.1% | 34.2% | 0.018 | 32.1% |
| **TS3007c** | 73.9% | 26.1% | 48.0% | 0.018 | 28.2% | 72.0% | 28.0% | 54.7% | 0.017 | 32.6% |
| **TS3007d** | 78.1% | 21.9% | 53.8% | 0.018 | 23.0% | 73.1% | 26.9% | 52.0% | 0.019 | 30.4% |
| **Mean** | 60.3% | 39.7% | 38.1% | 0.018 | 32.9% | 57.1% | 42.9% | 39.4% | 0.019 | 38.1% |

'TS3003b' is an example of a meeting where including the pitch features generated by the proposed system does not improve the performance in terms of coverage (see Fig. 10). One possible explanation is that FAs can be caused by hesitation markers (e.g. 'um', 'er', or 'uh') which are words that are spoken in conversation by the active speaker to indicate that they have not finished speaking.

An example of these FA errors can be seen in meeting 'TS3003b' between [10:24, 10:35] mins for the utterance "you can put uh a lot uh of uh functions uh in one uh yeah" from speaker 'MTD011UID'. These hesitation markers cause FAs due to the speaker voicing them at a different pitch to the rest of his utterance. The proposed system incorrectly identifies the 'uh' hesitations as a different speaker. These FA detections result in errors being present in the generated pitch feature.

The inclusion of the proposed pitch features in a segmentation system can be thought of as a trade-off between improved overlapping speech detection vs less FAs caused by hesitation markers. Another way to view this trade-off is: not including pitch features will lead to mostly 'under segmentation' errors and the inclusion of pitch features will likely lead to 'over segmentation' errors. The results (see Fig. 10) show that in the AMI corpus, pitch inclusion leads to better overall performance in most cases.

It could also be argued that in a complete diarization system 'over segmentation' errors are easier to handle as two segments belonging to the same speaker can be merged back together. If an audio recording is 'under segmented', however, it becomes a much more challenging problem to solve.

The performance is also evaluated using the metrics for Exp-1 and the results are given in Table V. It can be seen

that the four 'TS3003' meetings give a HIT rate performance of 42.8% on average when only MFCC features are used. However, Table V also shows how the performance on the 'TS3003' meetings can be improved by 12.1% on average by the addition of pitch features. It is also important to note that the performance in Table V when both pitch and MFCC features are used is lower at 60.3% in terms of the HIT rate than the proposed pitch system on the SDM stream in Table IV which achieves a HIT rate of 74.7%. This is likely due to the BLSTM system not being optimised for this type of pitch input feature in terms of its loss function and architecture. The purpose of including the BLSTM approach is to have an 'off-the-shelf' comparison and, therefore, the configuration is precisely as presented in the authors' original paper [51].

## VI. Conclusion

This paper showed that the harmonic structure of voiced speech can be exploited for the task of speaker segmentation. An investigative study on a well-established corpus of conversational speech showed how changes in pitch and changes in speaker are related. A novel method has been proposed that relies on a MHT framework to track multiple speakers even when they are talking simultaneously. The proposed method outperformed a BLSTM, in terms of HIT rate, by 12.9% on the AMI corpus SDM stream. We also showed that the pitch estimates obtained by the proposed system can be used as input features for neural networks. We showed that the segmentation performance of the baseline BLSTM can be improved by 1.21% in terms of coverage and 2.45% in terms of purity, by incorporating the pitch estimates as input features, in addition to MFCCs.

## References


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

[2] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1484–1498, Sept. 2018.

[3] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1620–1643, Apr. 2020.

[4] M. Sinclair, "Speech segmentation and speaker diarisation for transcription and translation," PhD Thesis, The University of Edinburgh, June 2016.

[5] P. Cerva, J. Silovsky, J. Zdansky, J. Nouza, and L. Seps, "Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives," *Speech Commun.*, vol. 55, no. 10, pp. 1033–1046, Nov. 2013.

[6] C. Kwan, J. Yin, B. Ayhan, S. Chu, X. Liu, K. Puckett, Y. Zhao, K. C. Ho, M. Kruger, and I. Sityar, "Speech separation algorithms for multiple speaker environments," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, June 2008, pp. 1644–1648.

[7] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. Eur. Conf. on Speech Commun. and Technol.*, 2001, pp. 1359–1362.

[8] A. Hogg, C. Evers, and P. Naylor, "Multichannel overlapping speaker segmentation using multiple hypothesis tracking of acoustic and spatial features," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, June 2021.

[9] M. Kunešová, M. Hrúz, Z. Zajíc, and V. Radová, "Detection of overlapping speech for the purposes of speaker diarization," in *Proc. Speech and Comput.*, ser. Lecture Notes in Computer Science, A. A. Salah, A. Karpov, and R. Potapova, Eds., vol. 11658. Cham: Springer, July 2019, pp. 247–257.

[10] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer-Verlag, 2010.

[11] X. A. Miró, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[12] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2008, pp. 4353–4356.

[13] S. H. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2013, pp. 7746–7750.

[14] V. Rozgic, K. J. Han, P. G. Georgiou, and S. Narayanan, "Multimodal speaker segmentation in presence of overlapped speech segments," in *Proc. IEEE Int. Symp on Multimedia (ISM)*, Dec. 2008, pp. 679–684.

[15] J. Zhong, P. Zhang, and X. Li, "A combined feature approach for speaker segmentation using convolution neural network," in *Advances in Multimedia Inform. Process. – PCM*, ser. Lecture Notes in Computer Science, B. Zeng, Q. Huang, A. El Saddik, H. Li, S. Jiang, and X. Fan, Eds. Cham: Springer International Publishing, 2018, pp. 550–559.

[16] L. Sarı, S. Thomas, M. Hasegawa-Johnson, and M. Picheny, "Pre-training of speaker embeddings for low-latency speaker change detection in broadcast news," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019, pp. 6286–6290.

[17] R. Wang, M. Gu, L. Li, M. Xu, and T. F. Zheng, "Speaker segmentation using deep speaker vectors for fast speaker change scenarios," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2017, pp. 5420–5424.

[18] R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast tv using bidirectional long short-term memory networks," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017.

[19] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, May 1995, pp. 756–759.

[20] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "A Kalman-based fundamental frequency estimation algorithm," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2017, pp. 314–318.

[21] S. Gonzalez and D. M. Brookes, "PEFAC - A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, pp. 518–530, Feb. 2014.

[22] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[23] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation using harmonic MUSIC," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, 2006, pp. 521–524.

[24] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "The multi-pitch estimation problem: Some new solutions," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 3, 2007, pp. 1221–1224.

[25] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2008, pp. 101–104.

[26] R. Peharz, M. Wohlmayr, and F. Pernkopf, "Gain-robust multi-pitch tracking using sparse nonnegative matrix factorization," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2011, pp. 5416–5419.

[27] M. Wohlmayr, R. Peharz, and F. Pernkopf, "Efficient implementation of probabilistic multi-pitch tracking," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2011, pp. 5412–5415.

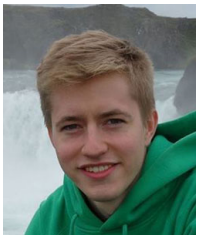
[28] D. Wang and G. Hu, "Unvoiced speech segregation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 5, May 2006.

[29] S. Ahmadi and A. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 333–338, May 1999.

[30] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 985–993, July 2009.

[31] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech

recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 3, pp. 201–212, June 1976.
[32] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Language Technol. Workshop (SLT)*, Dec. 2018, pp. 1021–1028.
[33] K. Laskowski, M. Heldner, and J. Edlund, "The fundamental frequency variation spectrum," in *Proc. FONETIK*, 2008.
[34] K. Laskowski and Q. Jin, "Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2009, pp. 4541–4544.
[35] Q. Jin, R. Li, Q. Yang, K. Laskowski, and T. Schultz, "Speaker identification with distant microphone speech," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2010, pp. 4518–4521.
[36] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proc. Intl. Conf. Machine Learning for Multimodal Interaction (ICMI)*, Berlin, Heidelberg, 2006, pp. 28–39.
[37] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. of the ASME J. of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, Mar. 1960.
[38] A. O. T. Hogg, P. A. Naylor, and C. Evers, "Speaker change detection using fundamental frequency with application to multi-talker segmentation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019.
[39] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *J. Acoust. Soc. Am.*, vol. 111, no. 3, pp. 1399–1413, Mar. 2002.
[40] S. W. McKnight, A. O. T. Hogg, and P. A. Naylor, "Analysis of phonetic dependence of segmentation errors in speaker diarization," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2020.
[41] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium (LDC), Philadelphia, USA, Corpus, 1993.
[42] D. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 6, pp. 843–854, Dec. 1979.
[43] K. Yoon, Y. Song, and M. Jeon, "Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views," *IET Image Process.*, vol. 12, no. 7, pp. 1175–1184, June 2018.
[44] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House, 1998.
[45] W. Hess, *Pitch Determination of Speech Signals*. Springer-Verlag, 1983.
[46] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *IEEE Intl. Conf. on Computer Vision (ICCV)*, Washington, DC, USA, 2015, pp. 4696–4704.
[47] D. J. Papageorgiou and M. R. Salpukas, "The maximum weight independent set problem for data association in multiple hypothesis tracking," in *Optimization and Cooperative Control Strategies*. New York, NY, USA: Springer, 2009, pp. 235–255.
[48] P. R. J. Östergård, "A new algorithm for the maximum-weight clique problem," *Nordic J. of Comput.*, vol. 8, no. 4, pp. 424–436, Dec. 2001.
[49] A. O. T. Hogg, C. Evers, and P. A. Naylor, "Multiple hypothesis tracking for overlapping speaker segmentation," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2019.
[50] A. Hogg, 2021. [Online]. Available: https://github.com/ahogg/Overlapping_speaker_segmentation_using_multiple_hypothesis_tracking_of_fundamental_frequency
[51] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2020.
[52] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[53] C. Bron and J. Kerbosch, "Algorithm 457: Finding all cliques of an undirected graph," *Commun. ACM*, vol. 16, no. 9, pp. 575–577, Sept. 1973.
[54] F. Chollet, "Keras," 2015. [Online]. Available: https://keras.io
[55] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, *et al.*, "Librosa/librosa: 0.7.2," Jan. 2020.
[56] H. Bredin, "Pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017.



**Aidan O. T. Hogg** (M'18) is currently a Ph.D. student at Imperial College London developing professional interests in Speech and Audio Signal Processing. He received the M.Eng. degree in Electronic and Information Engineering from Imperial College London in 2017. He has also worked in various engineering roles, including software and hardware development at Broadcom, Dialog Semiconductor and Nuance Communications. His current research focuses on speaker diarization, multichannel speaker change detection and statistical signal processing for audio applications. He has also worked on non-intrusive speech quality estimation and machine learning.

**Christine Evers** (M'14, SM'16) Christine Evers (M'14-SM'16) is a lecturer in the School of Electronics and Computer Science at the University of Southampton. She was the recipient of an EPSRC Fellowship, hosted at Imperial College London between 2017-2019. She worked as a research associate at Imperial College London between 2014-2017; as a senior systems engineer at Selex Electronic Systems between 2010-2014; and as a research fellow at the University of Edinburgh between 2009-2010. She received her PhD from the University of Edinburgh in 2010; her MSc degree in Signal Processing and Communications from the University of Edinburgh in 2006; and her BSc degree in Electrical Engineering and Computer Science from Jacobs University, Germany, in 2005. Her research focuses on Bayesian learning for machine listening. She is currently member of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing and serves as an associate editor of the EURASIP Journal on Audio, Speech, and Music Processing.

**Alastair H. Moore** is a post-doctoral researcher at Imperial College London and spatial audio consultant with Square Set Sound. He received the M.Eng. degree in Electronic Engineering with Music Technology Systems in 2005 and the Ph.D. degree in 2010, both from the University of York, York, U.K. He spent 3 years as a Hardware Design Engineer for Imagination Technologies PLC designing digital radios and networked audio consumer electronics products. In 2012, he joined Imperial College, where he has contributed to a series of projects in the field of speech and audio processing applied to voice over IP, robot audition, and hearing aids. Particular topics of interest include microphone array signal processing, modeling and characterization of room acoustics, dereverberation, and spatial audio perception. His current research is focused on signal processing for moving, head-worn microphone arrays.

**Patrick A. Naylor** (M'89, SM'07, F'20) is Professor of Speech and Acoustic Signal Processing at Imperial College London. He received the BEng degree in Electronic and Electrical Engineering from the University of Sheffield, UK, and the PhD degree from Imperial College London, UK. His research interests are in speech, audio and acoustic signal processing. His current research addresses microphone array signal processing, speaker diarization, and multichannel speech enhancement for application to binaural hearing aids and robot audition. He has also worked on speech dereverberation including blind multichannel system identification and equalization, acoustic echo control, non-intrusive speech quality estimation, and speech production modelling with a focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several collaborative links with industry. He is currently a member of the Board of Governors of the IEEE Signal Processing Society and President of the European Association for Signal Processing (EURASIP). He was formerly Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing. He has served as an associate editor of IEEE Signal Processing Letters and is currently a Senior Area Editor of IEEE Transactions on Audio Speech and Language Processing.