SPEAKER CHANGE DETECTION USING FUNDAMENTAL FREQUENCY WITH APPLICATION TO MULTI-TALKER SEGMENTATION

Aidan O. T. Hogg, Christine Evers† and Patrick A. Naylor*

Department of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

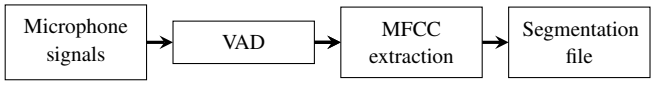
This paper shows that time varying pitch properties can be used advantageously within the segmentation step of a multi-talker diarization system. First a study is conducted to verify that changes in pitch are strong indicators of changes in the speaker. It is then highlighted that an individual's pitch is smoothly varying and, therefore, can be predicted by means of a Kalman filter. Subsequently it is shown that if the pitch is not predictable then this is most likely due to a change in the speaker. Finally, a novel system is proposed that uses this approach of pitch prediction for speaker change detection. This system is then evaluated against a commonly used MFCC segmentation system. The proposed system is shown to increase the speaker change detection rate from 43.3% to 70.5% on meetings in the AMI corpus. Therefore, there are two equally weighted contributions in this paper: 1. We address the question of whether a change in pitch is a reliable estimator of a speaker change in multi-talk meeting audio. 2. We develop a method to extract such speaker changes and test them on a widely available meeting corpus.

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

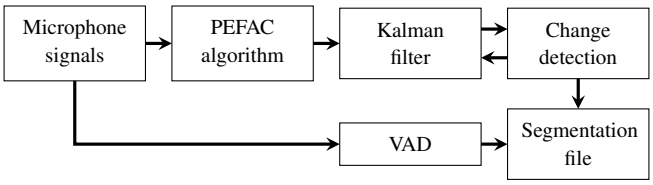
1. INTRODUCTION

It is often desirable to keep records of speech, for example, during conference calls and at meetings. To store these discussions in a more useful manner, automatic speech recognition (ASR) can be used to generate transcripts. However, whereas ASR addresses the question of what is said, it cannot answer the question of who spoke at any given time. Accurate knowledge of the identity of the speaker is typically required for speaker indexing [1]; improved ASR [2] and to bring single speaker-based algorithms into multi-speaker domains. The task of identifying a speaker within an audio recording or stream is often referred to as diarization, which has the end goal of answering the question: "who spoke when?" [3]. The process of audio diarization consists of two tasks: the first task is segmentation which establishes when a new talker starts speaking and the current speaker stops; the second task is clustering, where every segmented part of the audio containing speech is assigned to an individual speaker. This whole process is also often complicated by the presence of reverberation [4] and noise [5].

In the past, systems have been proposed that perform the diarization task, some of the most commonly used being: LIUM [6], DiarTk [7], ALIZÉ [8] and SIDEKIT [9]. These systems all contain different segmentation subsystems which can be grouped together into a set of categories. The first type uses Mel-frequency cepstral coefficients (MFCCs) [10] to perform Bayesian information criterion (BIC) segmentation [6], [9]. The second type uses a uniform segmentation [7]. The last type performs a one-step segmentation and clustering algorithm in the form of an evolutive hidden Markov model (E-HMM) [8]. Various other segmentation algorithms have also been proposed in the literature [11], [12], [13].



(a) Typical ('SIDEKIT') segmentation system.



(b) Pitch-based segmentation system.

Fig. 1. Diarization system architecture comparison.

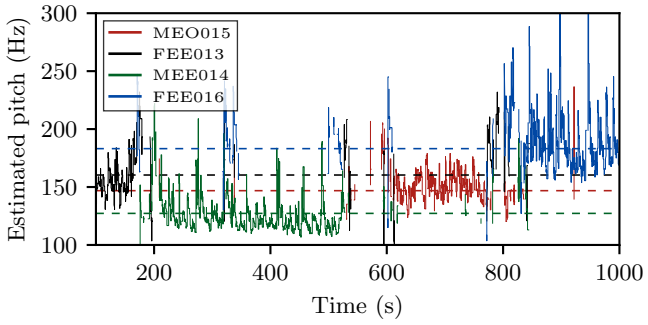
The SIDEKIT system in particular consists of three steps, shown in Fig. 1a. The first step merges together all of the voice active regions of the VAD output that are in close proximity to each other. The second step performs Gaussian divergence segmentation where the MFCCs are used to segment the voice active regions that contain multiple speakers. Lastly, linear Bayesian information criterion (BIC) based segmentation is performed which also uses the MFCCs to fuse consecutive voice active regions of the same speaker. It is important to note that none of these popular systems, including SIDEKIT, use pitch information to perform segmentation. Pitch is ordinarily only utilised as a feature [14] to improve the performance of the clustering component in a diarization system.

This paper will focus on the task of segmentation and show why temporal variation in pitch can be used advantageously for speaker change detection. We will also present a novel method using pitch to improve the segmentation process. It must be noted that throughout this paper we have used the term pitch to be synonymous with fundamental frequency even though pitch strictly refers to a perceptual phenomenon.

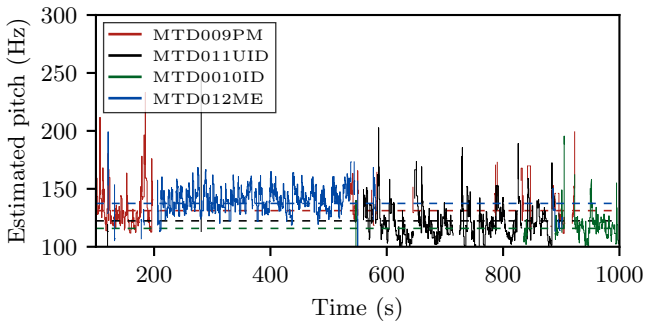
In the past methods have been proposed that use pitch to improve the segmentation process. In [15] pitch is used along side LSP [16] and MFCC features to calculate a divergence distance threshold to detect speaker change boundaries. There have also been methods

*The support of the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged.

†The research leading to these results has received funding from the UK EPSRC Fellowship grant no. EP/P001017/1.



(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.

Meeting	SC | 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%

Meeting	PC | 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%

$^{PC \mid SC}$ The probability that there is a 'pitch change' given that there is a 'speaker change'

$^{SC \mid PC}$ The probability that there is a 'speaker change' given that there is a 'pitch change'

Table 1. Speaker and pitch change analysis for the AMI corpus.

that have been developed for real-time diarization which use pitch as the sole feature [17], [18]. None of these previous methods, however, attempt to model the pitch which has two major advantages. First, the model can be exploited to remove errors in the pitch estimates. Second, the errors in the pitch prediction given by the model can be utilised to detect speaker changes instead of using the delta pitch change between two frames [17], [18].

Figure 1b shows the novel system that is proposed in this paper which takes advantage of pitch modelling when performing segmentation. This method uses a Kalman filter to predict the future pitch of the speaker. Kalman filters have been used in the past to perform pitch estimation, for example [19], [20], [21]. In contrast, the proposed system only uses the Kalman filter for future pitch prediction and not pitch estimation. Hence, it could be used in conjunction with any pitch estimator; for the purposes of this paper we used the pitch estimator PEFAC [22]. The main idea behind this method is that the pitch prediction made by the Kalman filter can be used to decide if there has been a change in speaker. It does this by assuming that the pitch of a speaker should be predictable whereas, if the pitch cannot be predicted, then a speaker change may be the cause.

2. SPEAKER PITCH TRACKS

First to address the question of whether changes in pitch are reliable estimators of speaker changes in multi-talk meeting audio, the following investigation has been carried out in Sections 2 and 3.

Figure 2 is generated by first running PEFAC on the headset microphone recordings taken from AMI [23]. Then the Kalman filtering method described in Section 4 was applied to the result to generate smooth pitch estimates. The measurements obtained are the best ground-truth available of the individual speaker pitch tracks. It is clear to see from Fig. 2a that the four individuals in AMI meeting 'ES2004b' speak at a very different pitch. However, AMI meeting 'TS3003b', in Fig. 2b, highlights that some individuals speak at a very similar pitch. This is most likely due to the fact that in this particular meeting all the speakers are male. The dotted lines show the mean pitch of each speaker in AMI where the first letter of the speaker label relates to the gender of the speaker i.e. M: male and F: female. It can also be observed in both figures that the average variation in pitch is very similar for most speakers. This result demonstrates that the mean of the pitch considered in isolation does not contain enough information to identify the speaker.

3. VARIATIONS OF PITCH OVER TIME

It has been seen in Section 2 that some speakers do indeed have a very similar mean pitch for their voice and, therefore, this section will show that even under these conditions it is still possible to identify when there is a change in the speaker using information about the way in which pitch varies over time.

It has been previously shown that the pitch of an individual speaker only varies in a smooth manner due to physiological constraints [24]. Accordingly it is possible to predict the future pitch of the speaker based on their current pitch. Thus if the pitch cannot be predicted then this could be an indication that there has been a change in speaker. In this paper, this prediction is attained by means of a Kalman filter which is described in detail in Section 4.

Table 1 is generated using the headset microphone recordings taken from AMI and shows the probability that there is a speaker change given that there is a pitch change and vice-versa. These results demonstrate that if there is a change in speaker then there is a very high probability that there will be a change in pitch. Thus this result verifies that the detection of pitch changes can be exploited constructively for speaker change detection. Table 1, however,

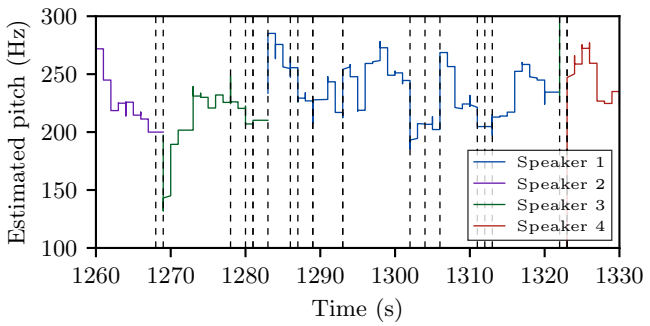


Fig. 3. Estimated speaker pitch tracks (solid lines) from 'IS1009b' along with the actual speaker changes (dotted lines).

shows that there can be a speaker change without a change in pitch.

Figure 3 provides a visualization example for the results shown in Table 1. This plot is generated using the method from Section 4 with a single distance microphone (SDM) recording from AMI. It is here to highlight that in this particular meeting, 'IS1009b', when a pitch change occurs, it always coincides with a change in the speaker. The plot also shows that many speaker changes go undetected.

4. METHOD

We now present a method that utilises the time varying properties of the pitch to detect speaker changes within a multi-talker scenario. A block diagram of the proposed system is provided in Fig. 1b.

4.1. Pitch Estimation

The first step of the proposed method is the pitch estimator; for this paper the PEFAC [22], [25] algorithm was chosen.

4.2. Kalman Filter

The next step is to use a Kalman filter [26] to estimate the pitch trajectories from PEFAC. Pitch, $x(n)$, for time frame, n, is modelled here as a random walk with zero-mean, normally distributed increments such that

$$x(n+1) = x(n) + w, \quad w \in \mathcal{N}(0,\, \sigma_w^2)\,, \tag{1}$$

where the pitch at $n+1$ deviates from the pitch at n with variance of σ_w^2. The PEFAC estimates $z(n)$ are modelled as

$$z(n) = x(n) + v, \quad v \in \mathcal{N}(0,\, \sigma_v^2)\,, \tag{2}$$

where the measurement noise, v, in this case models the errors in the pitch estimates from PEFAC.

The Kalman filter estimates the state of the system and then acquires feedback from noisy measurements using a prediction step and an update step. The predicted pitch estimate, $\hat{x}_{n|n-1}$, and predicted estimate variance, $P_{n|n-1}$, are given by

$$\hat{x}_{n|n-1} = \hat{x}_{n-1|n-1}, \tag{3}$$

$$P_{n|n-1} = P_{n-1|n-1} + \sigma_w^2. \tag{4}$$

The updated pitch estimate, $\hat{x}_{n|n}$, and updated estimate variance, $P_{n|n}$, are given by

$$\hat{x}_{n|n} = \hat{x}_{n|n-1} + K_n(z_n - \hat{x}_{n|n-1}), \tag{5}$$

$$P_{n|n} = (1 - K_n)^2 P_{n|n-1} + K_n^2 \sigma_v^2. \tag{6}$$

Where the innovation variance, S_n, and optimal Kalman gain, K_n, are given by

$$S_n = P_{n|n-1} + \sigma_v^2, \tag{7}$$

$$K_n = \frac{P_{n|n-1}}{S_n}. \tag{8}$$

Thus the error between measurement and prediction follows as

$$\tilde{y}_{n|n} = z_n - \hat{x}_{n|n}. \tag{9}$$

In our method we utilise two useful outputs from PEFAC: the pitch estimate of each frame and the corresponding probability that the frame is voiced.

The prediction step is carried out on every frame, however, the update step is only performed if the frame is voiced. This is considered to be the case if the probability that the frame is voiced is above a threshold ξ.

Thus if an unvoiced frame is observed then the pitch remains constant in (3) with the predicted estimate variance being increased in (4). This outcome is desirable as it makes the prediction less reliable as time goes on without a voiced frame being observed.

Given that Kalman gain, K_n, trades-off the measured pitch against the predicted pitch for the frame, K_n increases as the time between the update steps increases. This means that as the time elapsed since the last update frame increases the result will be more influenced by the measurement, otherwise it will be more influenced by the prediction from the model. This is seen in (5) if $K_n = 1$ then $\hat{x}_{n|n} = z_n$ (only the measurement) else if $K_n = 0$ then $\hat{x}_{n|n} = \hat{x}_{n|n-1}$ (only the prediction).

4.3. Speaker Change Detection

Our approach for speaker change detection utilises the error between the measurement and the prediction (9). If the error is above a threshold ϕ then that implies that the error is large and the pitch could not be predicted. An experiential threshold ϕ is acceptable in this case as the pitch of different speakers can be easily anticipated. Therefore, this change detection approach works by attributing a large prediction error to a change in the speaker.

A Kalman filter is initialised and tracks the first speaker. Subsequently, when (9) exceeds a threshold of ϕ a new Kalman filter is initialised to track the second speaker. On detection of the next speaker change, the measurement of the pitch is compared with all previously generated Kalman filter pitch tracks to find the track closest to the current measurement of the pitch. If the difference between the current measurement and the last pitch value of the closest Kalman pitch track is below a threshold of ρ, the previous Kalman filter is continued. If on the other hand, the closest Kalman filter to the measurement does not satisfy this threshold then a new Kalman filter would be generated.

The reasoning behind this Kalman filter birthing approach is that if the speakers do indeed have a different mean pitch, e.g. AMI meeting 'ES2004b' shown in Fig. 2a, then the different Kalman filter pitch tracks should correspond to the different speakers in the audio recording of the meeting.

4.4. Voice Activity Detection

To generate the final segmentation, the detected speaker changes from the pitch are merged with the results from voice activity detection (VAD) [27].

Meeting	Performance Evaluation							
	Proposed Pitch Segmentation				MFCC Segmentation (SIDEKIT)			
	Hit	Miss	Multi-Hit	MSE	Hit	Miss	Multi-Hit	MSE
ES2004a	72.80%	15.20%	12.00%	0.0334	50.40%	36.00%	13.60%	0.0287
ES2004b	74.89%	15.15%	9.96%	0.0431	46.75%	39.39%	13.85%	0.0337
ES2004c	64.65%	27.27%	8.08%	0.0409	42.42%	50.00%	7.58%	0.0404
ES2004d	69.10%	23.61%	7.30%	0.0379	46.78%	41.20%	12.02%	0.0270
IS1009a	65.12%	27.91%	6.98%	0.0442	34.88%	62.79%	2.33%	0.0487
IS1009b	72.53%	22.53%	4.95%	0.0543	36.81%	52.20%	10.99%	0.0219
IS1009c	75.31%	16.67%	8.02%	0.0452	38.27%	53.70%	8.02%	0.0280
IS1009d	61.86%	24.58%	13.56%	0.0558	36.44%	50.85%	12.71%	0.0335
EN2002a	63.76%	27.18%	9.06%	0.0520	36.59%	54.70%	8.71%	0.0349
EN2002b	66.03%	23.75%	10.21%	0.0606	38.48%	51.78%	9.74%	0.0393
EN2002c	68.46%	22.88%	8.67%	0.0530	41.42%	49.57%	9.01%	0.0330
EN2002d	59.12%	35.14%	5.74%	0.0480	33.11%	59.46%	7.43%	0.0333
TS3003a	76.19%	14.29%	9.52%	0.0247	38.10%	28.57%	33.33%	0.0237
TS3003b	87.10%	2.76%	10.14%	0.0362	58.53%	10.60%	30.88%	0.0249
TS3003c	76.52%	4.92%	18.56%	0.0406	57.95%	11.36%	30.68%	0.0290
TS3003d	73.95%	10.08%	15.97%	0.0404	56.30%	19.75%	23.95%	0.0317
Mean	70.46%	19.62%	9.92%	0.0444	43.33%	42.00%	14.68%	0.0320

Table 2. Performance of both the proposed system and SIDEKIT on multi-talker meetings in the AMI corpus.

The VAD output detects active speech regions, therefore, as part of a preprocessing step if these regions have small pauses between them then they are merged together. Subsequently, both the onsets of speech detected by the VAD and the speaker changes detected by the pitch are concatenated. If a VAD onset and a detected speaker change are within ζ of each other, then only the detected speaker change is included in the segmentation file.

5. COMPARATIVE EVALUATION

In order to evaluate the performance of this newly proposed system shown in Fig. 1b, it is compared against a typical segmentation system SIDEKIT [9] as illustrated in Fig. 1a.

Both the accuracy and the reliability of speaker change detection are compared for both the proposed system and SIDEKIT with the results shown in Table 2. The hit rate is defined as the number of speaker changes that are detected by a single detection. In contrast, the miss rate is given by the number of speaker changes that go undetected and the multi-hit rate is specified as the number of speaker changes that are detected multiple times. When evaluating segmentation performance, it is common practice to apply a time collar around every ground-truth speaker change in order to account for possible inaccuracies. The results in Table 2, therefore, incorporate a collar of 50 ms applied to each ground-truth speaker change.

Through experimentation it was found that the implementation of the proposed method should use a process noise, v, of 0.01 and a pitch variation, w, of 20. It was also decided that a frame should be considered voiced if ξ was greater than 95% and the thresholds ϕ and ρ should be set to 10 Hz and 50 Hz respectively. In a similar manner it was determined that a VAD onset should only be incorporated into the segmentation file if ζ was greater than 5 ms.

It can be seen in Table 2 that the percentage of speaker changes that are detected increases from 43.3% for SIDEKIT to 70.5% for our system. Thus, the proposed pitch system is far more likely to detect a speaker change within the given 50 ms collar. It is important to note for both systems that increasing the collar decreases the miss rate, increases the multi-hit rate and does not change the hit rate.

The mean squared error (MSE) in time was also calculated in Table 2 for all the hits and the closest multi-hit detections to the oracle speaker changes, against the ground-truth given by the label files from AMI. The results show that when a speaker change is detected by both systems the use of MFCCs in SIDEKIT gives slightly more accurate temporal segmentation (MSE = 30 ms) compared to the use of pitch (MSE = 40 ms).

To realise the significance of this improvement, the whole diarization process should be considered. In a typical diarization system after the segmentation process, clustering is performed and then Viterbi alignment is exploited as previously reported in [28]. Consequently, mediocre performance in the segmentation system is tolerated. However, if the clustering algorithm is given a better segmentation, where almost all segments just contain one speaker, then it will achieve a far better clustering result; improving the performance of the given diarization system which is highly desirable. This is verified in [29] where an evaluation is undertaken which shows that improving the segmentation performance leads to better diarization accuracy and a lower diarization error rate.

6. CONCLUSION

A study of meetings in the AMI corpus has shown that a pitch change is a strong indicator of a speaker change. This finding motivates the use of pitch change as a feature - possibly combined with other features - in speaker segmentation as used, for example, in the first step of speaker diarization. It was also verified that pitch from an individual speaker is smoothly varying and can be predicted by a Kalman filter. Therefore, in this paper, a Kalman filtering approach was proposed to identify speaker change boundaries based on a model of the temporal variation of pitch.

The proposed Kalman filter prediction error-based approach performed well when compared against a previous MFCC-based method. An evaluation on the AMI corpus showed a speaker change detection increase from 43.3% to 70.5%.

7. REFERENCES

[1] M. Sinclair, "Speech segmentation and speaker diarisation for transcription and translation," Ph.D. thesis, The University of Edinburgh, Jun. 2016.

[2] 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 Communication*, vol. 55, no. 10, pp. 1033–1046, Nov. 2013.

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

[4] P. A. Naylor and N. D. Gaubitch, (Eds.), *Speech Dereverberation*, PUB-SV, 2010.

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

[6] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2013.

[7] D. Vijayasenan and F. Valente, "DiarTk : An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 2170–2173.

[8] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "ALIZE/SpkDet: A state-of-the-art open source software for speaker recognition," *Proc. Odyssey IEEE Workshop*, 2008.

[9] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in Python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5095–5099.

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[11] R. Li, T. Schultz, and Q. Jin, "Improving speaker segmentation via speaker identification and text segmentation.," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2009, pp. 3073–3076.

[12] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, no. 1, pp. 111–126, Sep. 2000.

[13] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna, "Unsupervised speaker change detection using probabilistic pattern matching," *IEEE Signal Processing Letters*, vol. 13, no. 8, pp. 509–512, Aug. 2006.

[14] K. Laskowski, M. Heldner, and J. Edlund, "The fundamental frequency variation spectrum," in *Proceedings of FONETIK*. 2008, vol. 2008, pp. 29–32, Citeseer.

[15] L. Lu and H.-J. Zhang, "Unsupervised speaker segmentation and tracking in real-time audio content analysis," *Multimedia Systems*, vol. 10, no. 4, pp. 332–343, Apr. 2005.

[16] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *Acoustical Society of America Journal*, vol. 57, pp. S35, Jan. 1975.

[17] M. Yang, Y. Yang, and Z. Wu, "A pitch-based rapid speech segmentation for speaker indexing," in *Seventh IEEE International Symposium on Multimedia (ISM'05)*, Dec. 2005.

[18] B. Abdolali, "A Novel Method for Speech Segmentation Based on Speaker's Characteristics," *Signal & Image Processing : An International Journal*, vol. 3, no. 2, pp. 65–78, Apr. 2012.

[19] . Salor, M. Demirekler, and U. Orguner, "Kalman filter approach for pitch determination of speech signals," *St. Petersburg*, p. 4, Jun. 2006.

[20] O. Das, J. O. S. Iii, and C. Chafe, "Real-time pitch tracking in audio signals with the extended complex kalman filter," *Proc. Conf. on Digital Audio Effects*, vol. 20, pp. 118–124, Sep. 2017.

[21] 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 Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, Oct. 2017, pp. 314–318, IEEE.

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

[23] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. Intl. Conf. Machine Learning for Multimodal Interaction (ICMI)*, Berlin, Heidelberg, 2006, pp. 28–39.

[24] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 3, pp. 1399–1413, Mar. 2002.

[25] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997.

[26] 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.

[27] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[28] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech & Language*, vol. 20, no. 2, pp. 303–330, Apr. 2006.

[29] S. Cheng, H. Wang, and H. Fu, "BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 141–157, Jan. 2010.