SPEAKER CHANGE DETECTION AND SPEAKER DIARIZATION USING SPATIAL INFORMATION

Mathieu Hu[1]*, Dushyant Sharma*[2]*, Simon Doclo*[3]*, Mike Brookes*[1]*, Patrick A. Naylor*[1]

[1] Department of Electrical and Electronic Engineering, Imperial College London, UK
[2] Voicemail-To-Text Research, Nuance Communications Inc. Marlow, UK
[3] Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg, Germany
mathieu.hu12@imperial.ac.uk

1. ABSTRACT

In this paper, we present a novel speaker change detection and speaker diarization algorithm using spatial information in the form of features derived from estimated Room Impulse Response (RIR)s. A blind system identification approach is used to obtain an estimate of the RIRs, from which the C5 feature is derived and used in the labeling algorithm. Experimental results using 2 speakers for different locations within a fixed room show that our approach achieves a higher hit rate in the speaker change detection task and a lower variance in the diarization error rate when compared with a baseline algorithm.

Index Terms— Blind system identification, speaker diarization, speaker change detection

2. INTRODUCTION

Beamforming is a common technique in hearing-aids and assitive listening technologies to improve speech intelligibility [1]. It exploits the spatial diversity of the signals at different microphones and combines the multi-channel input into a single-channel output so that the signal coming from the steering direction is enhanced. However, the accuracy of the estimated Direction-of-Arrival (DOA), which decreases as the level of noise and reverberation increases [2], has a significant impact on the performance [3].

In a multi-speaker scenario, such as a meeting, knowing when the identity of the active speaker changes is a valuable piece of information for assistive listening devices as it can be used to re-steer a beamformer. Determining 'who spoke when?' is the goal of speaker diarization. That consists of detecting speaker changes and labeling with a unique label speech segments spoken by the same person.

Spatial-information-based diarization has been investigated in [4] and [5]. The diarization system in [5] is based on Time-Difference-of-Arrival (TDOA) features: an Unsupervised Discriminant Analysis (UDA) is applied to estimated TDOA between every pair of microphones to separate the speakers in the new feature space as it is known that the TDOA estimates obtained from the Generalized Cross-Correlation (GCC)-Phase Transform (PHAT) algorithm are sometimes spurious. This, however, requires at least 3 microphones.

In this paper, we propose a novel application of Blind System Identification (BSI) which performs speaker change detection and diarization by exploiting the room acoustic information encapsulated in the estimated RIRs. The proposed diarization system relies on spatial features extracted from estimated RIRs. The robustness to BSI errors of the proposed method is also evaluated.

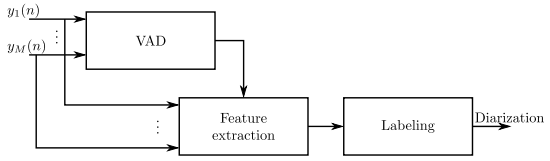

Fig. 1: Block diagram of the diarization system

The remainder of this paper is organized as follows. In section 3, the diarization system is described. In section 4, the experimental setup is detailed and the results shown in section 5.

3. THE SPEAKER DIARIZATION SYSTEM

3.1. Signal model

In a typical meeting scenario, only one speaker is active at any given moment in time. Even though several speakers are present in the audio stream, the system in practice has a Single-Input-Multiple-Output (SIMO) structure. Hence, for P speakers and M microphones, at any time n, the signal $y_m(n)$ recorded at the m^{th} microphone is given by eq. (1):

$$y_m(n) = h_{m,p}(n) * s_p(n) + \nu_m(n) \tag{1}$$

where p represents the identity of the active speaker, $h_{m,p}(n)$ is the RIR relating the p^{th} speaker to the m^{th} microphone and $\nu_m(n)$ the additive noise present at the m^{th} microphone.

3.2. The overall diarization system

We used the diarization system described in [5]. Its block diagram is shown in Fig. 1. It consists of a Voice Activity Detector (VAD) detecting the non-speech parts, a feature extraction algorithm and a labeling step.

The VAD, based on the P.56 standard [6, 7], takes the summed microphone signals as the input and detects active speech segments. The output consists of estimated time instants indicating the beginning and the end of segments of active speech. A post-processing step is added so that the estimated active speech segments separated by less than 100 ms of estimated pause are merged together.

A window of duration t_e sliding with an offset t_s is then applied within each of these active speech segments to obtain frames. Features are extracted from each of these frames. The type of features as well as the method to obtain them from the input signals will be described in section 3.3 and section 3.4.

The features are then labeled using a k-means initialized Hidden Markov Model (HMM), the details of which will be given in section 3.5.

3.3. Spatial feature extraction of the baseline

The method described in [5] is taken as the baseline as diarization is also achieved based on spatial features only. More precisely, the TDOAs between every pair of microphones are estimated using the GCC-PHAT algorithm. Therefore, for each frame, $\binom{M}{2} = \frac{M(M-1)}{2}$ estimated TDOAs are obtained. In the case where M is greater or equal to 3, dimension reduction techniques aiming at reducing the impact of estimation noise are possible. In [5], a UDA [8] is used for that purpose.

In the implementation of the feature extraction scheme, the estimates of the TDOAs were obtained by computing the cross-correlation function in the frequency domain. To improve the noise robustness of the algorithm, $y_m(n)$ is processed so that the cross-correlation function is computed only on the the 60% largest samples in absolute values. If we denote by $\tilde{y}_m(n)$ the processed $y_m(n)$ and $\underline{\tilde{y}}_m(f)$ its Fourier transform, the cross-correlation function between the i^{th} and j^{th} microphones is given by:

$$\underline{g}_{i,j}(f) = \frac{\underline{\tilde{y}}_i(f)[\underline{\tilde{y}}_j(f)]^*}{|\underline{\tilde{y}}_i(f)[\underline{\tilde{y}}_j(f)]^*|} \tag{2}$$

where $*$ and $|.|$ respectively represent the complex conjugate and the module operators and f is the frequency bin index.

The TDOA is then obtained by finding the position of the peak of the inverse Fourier transform of $\underline{g}_{i,j}(f)$.

3.4. Spatial Feature extraction of the suggested method

The microphone signals $y_m(n)$ can be viewed as a combination of two independent quantities: the dry speech signal $s_p(n)$ and the RIRs $\{h_{m,p}(n)\}, m \in \{1, 2, \ldots, M\}$. While the dry speech contains the characteristics of the speaker, the set of RIRs holds information about the relative position of that speaker to the microphone array. Therefore spatially characterizing localized speakers is possible by blindly estimating the RIRs.

Because of the SIMO structure, BSI is theoretically possible provided that the conditions for the system to be identifiable are fulfilled [9]. Examples of algorithms tackling the problem can be found in [10], [11] or [12].

Nevertheless, since the estimated RIRs are not accurate nor consistent enough to directly use them for diarization, we suggest to extract a feature, referred as C_x, which is analogous to the well-known C_{50}. It represents the ratio between the energy in the first x ms of a RIR and that of the remaining taps, i.e.

$$C_x(\hat{\boldsymbol{h}}_m) = \frac{\sum_{j=0}^{n_x-1} \hat{h}_m^2(j)}{\sum_{j=n_x}^{L_m-1} \hat{h}_m^2(j)} \tag{3}$$

where n_x is the sample corresponding to x ms, $\hat{\boldsymbol{h}}_m$ the estimated RIR at the m^{th} microphone and L_m its length in samples.

Diarizations based on C_x for $x \in \{5, 10, 15, 50\}$ showed that C_5 yields the best speaker discrimination. This may be due to its similarity to the Direct-to-Reverberant Ratio (DRR) [14] which is well correlated with the distance between a speaker and a microphone [15].

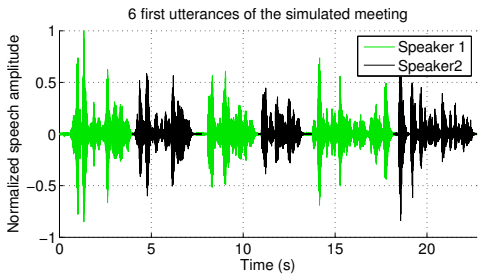

Fig. 2: Example speech from simulated meeting

3.5. Feature labeling

The extracted features were then labeled using a k-means initialized P-state HMM [16]. The features belonging to each state were modeled by a single Gaussian distribution with a diagonal covariance matrix. The initial guesses of the transition and prior probabilities followed a uniform distribution.

An iterative scheme was then used to estimate the most likely path:

1. Given an assignment of each feature to a state, compute the observation likelihood.
2. Given the prior and transition probabilities as well as the observation likelihood, compute the most likely path using the Viterbi algorithm.
3. Given the new feature-to-state assignment, update the parameters of the Gaussian distributions of each state.
4. Given the estimated path and the new statistics of each state, update the prior and transition probabilities using the Baum-Welch algorithm.

4. SIMULATIONS

4.1. Speech input generation

The simulated meeting data were obtained using 20 utterances spoken by 2 speakers from the test set of the TIMIT database [17]. Two sets of RIRs were generated using the image method [18], one set corresponding to each speaker. Each utterance was then convolved with the corresponding set of RIRs. The reverberant utterances were then combined to produce an interleaved signal, where the speakers speak in turn. The whole speech data had a duration of approximately 60 s. The simulated data were free of instants where both speakers are talking at the same time.

White Gaussian noise was added to the dry reverberant meeting signal to achieve a Signal-to-Noise Ratio (SNR) of 30 dB. An excerpt of the simulated speech signal at the first microphone is shown in Fig. 2.

4.2. Experimental setup

The considered room is of dimension 5 m $\times$ 6 m $\times$ 3 m. Throughout the experiments, the reverberation time is set to $T_{60} = 0.5$ s, leading to RIRs of length $L = 4000$ for a sampling frequency $f_s = 8000$ kHz. The microphones of the microphone array with $M = 2$ were placed at coordinates $(2 \pm 0.2, 3, 1.5)$ expressed in a Cartesian system.

For each of the VAD based estimated active speech segments, a sliding analysis window of $t_e = 1$ s is applied with a sliding offset

of $t_s = 100$ ms. This leads to frames of duration $t_e = 1$ s overlapping by 900 ms. A given frame contains either no speaker, only one of the speakers or both. When no speaker is present in the frame, i.e. the VAD failed in detecting the pause, the estimation of the RIRs should not correspond to any the ground truth RIRs. Therefore, the estimated RIRs are given by one of the two sets of ground truth RIRs, randomly chosen and corrupted by additive noise following the model described in [13] so that the Normalized Projection Misalignment (NPM) has a small value (10^{-6} dB). As shown in eq. (4), such a low value means that the estimation is almost orthogonal to the RIRs and therefore holds no information.

$$NPM(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \frac{||\boldsymbol{h} - \frac{\boldsymbol{h}^T\hat{\boldsymbol{h}}}{\hat{\boldsymbol{h}}^T\hat{\boldsymbol{h}}}\hat{\boldsymbol{h}}||_2^2}{\boldsymbol{h}^T\boldsymbol{h}} \tag{4}$$

where $\boldsymbol{h}$ is the stacked true RIRs, $\hat{\boldsymbol{h}}$ an estimate of $\boldsymbol{h}$.

In the case where only one speaker is present, the estimated RIRs were given by the ground truth RIRs corresponding to that speaker corrupted by additive noise so that a desired NPM ϵ_s is achieved. In the case where both speakers are present, the estimated RIR at each microphone is given as an average impulse response weighted by the proportion of the active time of each speaker in the considered frame. Noise was also added in the latter case to achieve an NPM of ϵ_s. A different realization of the additive noise is computed for each frame so that RIR estimates obtained from a BSI algorithm periodically reinitialized are simulated. Estimates of C_5 are then obtained from these sets of RIRs, one per microphone.

Accuracy of BSI. In the first experiment, the speakers were respectively localized at coordinates $(3.18, 4.88, 1.57)$ and $(2.33, 3.98, 1.53)$. For that particular configuration, the true TDOAs were -0.398 ms and 0.199 ms for the first and second speaker respectively. The values of C_5 were respectively $(-5.82, -4.37)$ and $(-1.05, -3.50)$, in dB, for the first and second speakers. In that setup, the robustness of the proposed method to BSI errors was investigated by evaluating the Diarization Error Rate (DER) for ϵ_s taking 20 linearly spaced values between -10 dB and -1 dB.

Monte-Carlo simulation. In the second experiment, the locations of the speakers were randomly drawn under the constraint that they had to be at least 50 cm away from the walls, the microphone array and each other. The accuracy of the estimated RIRs for frames effectively containing speech was set to achieve $\epsilon_s = -10$ dB. The performance of the system was evaluated over 100 different speaker locations.

Since the implemented method to estimate the TDOA features operates in the frequency domain, a Hamming window was applied to reduce windowing artifacts. As that method outputs integers and that the UDA cannot be applied due to the small number of microphones ($M = 2$), it is not always possible to directly fit a Gaussian distribution model over the estimated TDOA in the HMM. To overcome that issue, a small amount of white Gaussian noise, the variance σ^2 of which was equal to 0.01, was added.

4.3. Evaluation

The performance of the diarization system was evaluated in terms of DER as defined in [19]. The score represents the fraction of duration attributed to a wrong speaker or non-speech. To take the inaccuracy of the hand labels into account, a tolerance threshold of 250 ms was used.

Hit, miss and false alarm rates were used to evaluate the performance of the system for speaker change detection. These were defined as follows:

- The Hit Rate (HR) corresponds to the percentage of estimated speaker changes lying within 250 ms around a true speaker change
- The Miss Rate (MR) corresponds to the percentage of true changes not estimated within 250 ms
- The False Alarm Rate (FAR) is the percentage of estimated speaker changes that do not correspond to a true speaker change

A key point in the success of the diarization system is the separability of the features. When these features follow a Gaussian distribution, which is assumed in our HMM, that separability can be measured by the Bhattacharyya distance [20]:

$$\begin{aligned} B(D_i, D_j) =& \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}(i,j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ &+ \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}(i,j)|}{\sqrt{|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|}}\right) \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance matrix of the cluster D_k. $|.|$ is the determinant operator and $\boldsymbol{\Sigma}(i,j)$ is given by $\boldsymbol{\Sigma}(i,j) = \frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}$. The Bhattacharyya score between two clusters increases these clusters are more separable.

5. EXPERIMENTAL RESULTS

5.1. Fixed location, varying NPM

Figure 3 shows the evolution of the DER of the proposed method for each value of NPM from -10 dB to -1 dB . Although the values of the NPM decreases from -10 dB to -1 dB, the proposed diarization system seems to be strongly affected for values of NPM below -2 dB. However, the Bhattacharyya score decreases as the NPM increases as shown in Fig. 4.

Figure 5 is an example of the C_5 feature points for an NPM of -5 dB. As the NPM increases, the clusters seem to merge, which results in a higher DER and a lower Bhattacharyya score.

The TDOA based diarization system achieved a DER of 37% with a Bhattacharyya distance of 0.8.

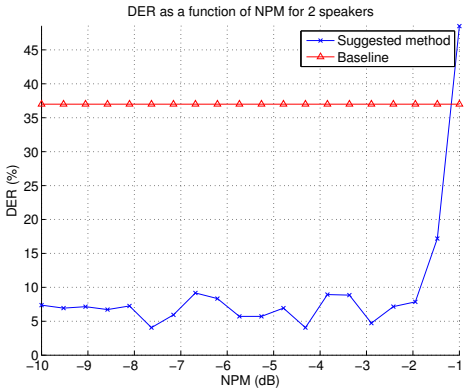

Fig. 3: DER as a function of NPM for 2 speakers at given locations

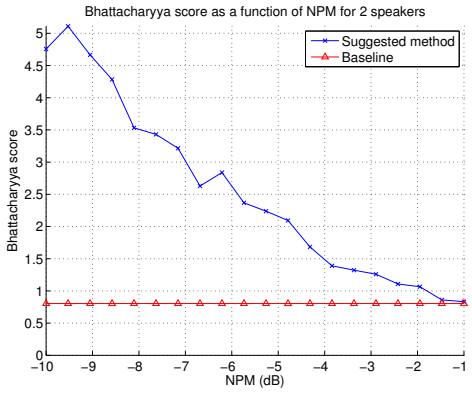

Fig. 4: Bhattacharyya score as a function of NPM for 2 speaker at given locations

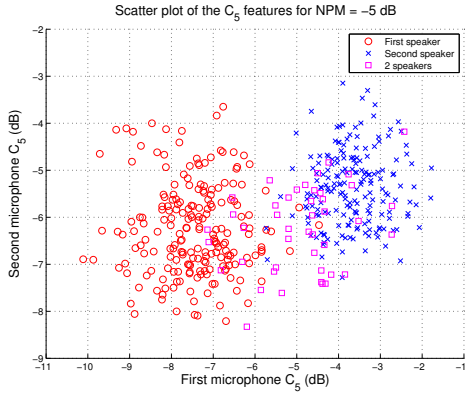

Fig. 5: Scatter plot of the C_5 features for NPM $= -5$ dB and $M = 2$

5.2. Fixed NPM, changing locations

Figure 6 shows the DER obtained from 100 different speaker locations for an NPM of -10 dB. The proposed method leads to less variability of the DER than that of the approach using TDOA features only and has a mean DER of 8.95% against 17.5% for the baseline.

Table 1 shows the mean and standard deviation of the diarization system evaluated using the hit, miss and false alarm rate metrics. It can be seen that on average the proposed method yields a higher HR and a lower MR and FAR than that of the baseline method while consistently yielding a smaller standard deviation.

	Method	HR	MR	FAR
mean	Suggested	72.8%	32.83%	27.82%
	Baseline	69.06%	39.15%	41.39%
std.	Suggested	6.54%	6.55%	6.18%
	Baseline	8.25%	7.64%	16.13%

Table 1: Performance of the diarization system in terms of speaker change detection

Fig. 6: Box diagram of the DER obtained from 100 different speaker locations. The estimated RIRs had an NPM of $-10\,dB$

6. DISCUSSION AND CONCLUSION

In this paper, a novel use of spatial features from estimated RIRs for speaker change detection and diarization was proposed and compared with a baseline approach using TDOA features. Our approach was shown to outperform the baseline on average and shown to have a lower error variance. Furthermore, the proposed method was evaluated with different levels of errors in the BSI. The proposed method was shown to be robust to BSI errors up to an NPM of -2 dB.

7. ACKNOWLEDGMENT

The authors would like to thank Ms. Felicia Lim for her input in the topic of BSI errors.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° ITN-GA-2012-316969.

8. REFERENCES

[1] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K.J. Ray Liu, Eds., chapter 9. Wiley, 2008.

[2] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

[4] D. Ellis and J.C. Liu, "Speaker turn segmentation based on between-channel differences," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, 2004.

[5] N.W.D. Evans, C. Fredouille, and J.-F. Bonastre, "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 4061 – 4064.

[6] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2013.

[7] ITU-T, "Objective measurement of active speech level," Mar. 1993.

[8] Jian Yang, D. Zhang, Zhong Jin, and Jing-Yu Yang, "Unsupervised discriminant projection analysis for feature extr," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006.

[9] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[10] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Processing*, vol. 82, no. 8, pp. 1127–1138, Aug. 2002.

[11] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[12] M.A. Haque and M.K. Hasan, "Noise robust multichannel frequency-domain LMS algorithms for blind channel identification," *IEEE Signal Process. Lett.*, vol. 15, pp. 305–308, 2008.

[13] F. Lim and P. Naylor, "Statistical modelling of multichannel blind system identification errors," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Antibes, France, 2014.

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

[15] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2110–2117, 2002.

[16] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[17] 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," Corpus LDC93S1, Linguistic Data Consortium, Philadelphia, 1993.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[19] NIST, "Spring 2006 (rt-06s) rich transcription meeting recognition evaluation plan," http://www.itl.nist.gov/iad/mig//tests/rt/2006-spring/docs/rt06s-meeting-eval-plan-V2.pdf, February 2006.

[20] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. 15, no. 1, pp. 52–60, Feb. 1967.