DETERMINING THE NUMBER OF SPEAKERS IN A MEETING USING MICROPHONE ARRAY FEATURES

Erich Zwyssig[1,2]*, Steve Renals*[1] *and Mike Lincoln*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, Scotland UK
[2]EADS IW, Appleton Tower, 6th Floor, Edinburgh, EH8 9LE, Scotland UK

ABSTRACT

The accuracy of speaker diarisation in meetings relies heavily on determining the correct number of speakers. In this paper we present a novel algorithm based on time difference of arrival (TDOA) features that aims to find the correct number of active speakers in a meeting and thus aid the speaker segmentation and clustering process. With our proposed method the microphone array TDOA values and known geometry of the array are used to calculate a speaker matrix from which we determine the correct number of active speakers with the aid of the Bayesian information criterion (BIC). In addition, we analyse several well-known voice activity detection (VAD) algorithms and verified their fitness for meeting recordings. Experiments were performed using the NIST RT06, RT07 and RT09 data sets, and resulted in reduced error rates compared with BIC-based approaches.

Index Terms— Speaker diarisation in meetings, microphone array, time difference of arrival (TDOA), speech segmentation and clustering, BIC, voice activity detection (VAD)

1. INTRODUCTION

Speaker diarisation aims to find the number of active speakers in a recording and identify when each speaker was talking. Speaker diarisation is typically carried out in three steps: (i) Detecting when speech is present in the recording; (ii) Splitting the speech segments where the speaker changes mid-segment; (iii) Identifying and clustering speech segments from the same speaker. Current speaker diarisation systems, when evaluated on meeting recordings such as those used for the NIST RT evaluations[1], achieve speech activity detection error rates of less than 10% and diarisation error rates of less than 15%. These results are, however, highly dependent on the system correctly identifying the number of active speakers in the meeting—if too few or too many speakers are detected, the error rates increase significantly.

Pardo et al. [1], Sun et al. [2] and Vijayasenan et al. [3] have successfully demonstrated how microphone array beamforming features can be used to improve the speaker diarisation performance in meetings. These systems do not use the speaker location information which may be estimated from the array to explicitly identify the number of speakers. Instead they implicitly estimate the active speaker count during the clustering process. This may in part be due to the fact that knowledge of the array geometry, which is required for speaker localisation, may not be used in the NIST evaluations. This constraint is somewhat artificial and unrealistic, since microphone arrays of known geometry—such as that used in the AMI/DA recordings [4] and a number of commercial products—are increasingly used for meeting recording. Lathoud [5] has previously used array geometry information for segmenting multiple concurrent speakers in meetings.

In this paper we use time difference of arrival (TDOA) features from an array of known geometry to determine active speaker locations. We then use this information with the Bayesian information criterion algorithm to explicitly determine the number of active speakers in a meeting, and this information is used for diarisation. As a precursor to this, we analyse several well-known voice activity detection (VAD) algorithms on the RT corpus from 2006, 2007 and 2009, since VAD is a pre-processing stage to our diarisation system.

2. VOICE ACTIVITY DETECTION

2.1. VAD algorithms

Accurate VAD is a crucial first stage for many speech processing algorithms. A number of approaches to VAD have been developed, and we have compared the accuracy of five systems on the RT meeting data: ITU-T P.56 [6], Sohn [7], Aurora [8], SHOUT [9] and AZR [10]. When tuning the thresholds of our implementation of the AZR algorithm, we found that the maximum peak of the normalised autocorrelation (MaxPeaks) component was ineffective and performance improved using only the zero-crossing rate of the autocorrelation (CrossCorr). The results reported for the AZR algorithm therefore only include the CrossCorr algorithm. In addition to the five algorithms mentioned, as a baseline, we present results for classifying all segments as speech.

[1]http://www.itl.nist.gov/iad/mig/tests/rt/

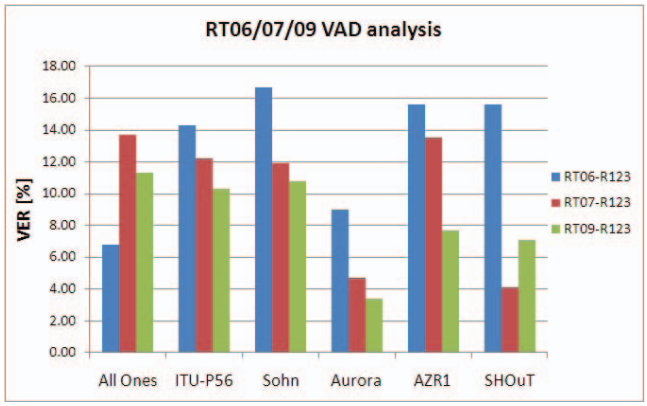

Fig. 1. Voice activity detection error rate for all algorithms

2.2. VAD results

Voice activity detection systems are evaluated in terms of the VAD error rate (VER). The VER penalises both missed speech and false alarms, and is fully described in [11]. For these experiments, Wiener-filter-based noise reduction [8] was first applied to the individual microphone signals. The BeamformIt[2] microphone array processing toolkit was then used to perform delay-sum beamforming on the signals, after which VAD was carried out. Scoring was performed using the standard NIST VAD scoring tools and Figure 1 shows the voice activity detection error rate for each of the algorithms when tested on the complete NIST RT06, RT07 and RT09 data sets.

On the RT06 test set, perhaps surprisingly, classifying all segments as speech outperforms all the other algorithms with 6.8% VER. This implies that for this particular set of meetings there are very few non-speech intervals leading to few false alarm errors for the 'all-speech' algorithm. For RT07 and RT09, which contain more non-speech segments, the all-speech, ITU, Sohn and AZR algorithms have similar results, and are consistently outperformed by Aurora and SHOUT, with Aurora having the lowest overall error when averaged over all three test sets.

3. SEGMENT SPLITTING AND THE BAYESIAN INFORMATION CRITERION

The speech segments identified by the voice activity detection algorithms may contain speech from more than one speaker. In order to avoid the entire segment being incorrectly assigned to a single speaker during diarisation, we must determine whether a segment contains one or more speakers, and the Bayesian information criterion (BIC) [12] has been found to be reliable and has been used in a number of state-of-the-art diarisation systems (e.g. [13]).

[2] http://www.xavieranguera.com/beamformit/

The Bayesian information criterion for an audio cluster $\mathcal{C}_k$ is defined as

$$BIC(\mathcal{C}_k) = \sum_{i=1}^{k} \{ -\frac{1}{2} n_i \log |\Sigma_i| \} - \lambda P, \quad (1)$$

where n_i is the number of samples in the cluster and Σ_i is the sample covariance matrix. The penalty P is defined as

$$P = \frac{1}{2}(d + \frac{1}{2} d (d+1)) \log N, \quad (2)$$

where $N = \sum_i n_i$ is the total sample size and d the number of parameters per cluster. Note that λ, the penalty weight, is usually set to 1.

The Bayesian information criterion can now be used to calculate whether a speech segment contains one or more different speakers and to determine whether two speech segments are from the same speaker. Using the BIC for both is best explained for the latter. The increase in the BIC value for merging two segments s_1 and s_2 is defined as:

$$BIC = n \log \Sigma - n_1 \log \Sigma_1 - n_2 \log \Sigma_2 - \lambda P. \quad (3)$$

If the BIC value is greater than zero then the information content of the merged segments is higher than the individual segments and the two segments are likely to belong to the same speaker and should be merged. Similarly, a speaker change is indicated by a positive peak of the BIC value when calculating a series of BIC values for a sliding split point of a speech segment.

4. SPEAKER DIARISATION

Speaker diarisation is the process of determining 'who spoke when'. As mentioned previously, identifying the correct number of speakers is important for good diarisation performance, and here we present a novel method for speaker diarisation which explicitly calculates the number of speakers by estimating their location using a microphone array.

4.1. Speaker diarisation using TDOA analysis

TDOA estimation seeks to identify the time difference between signals from a given sound source arriving at two different microphones. An established method for performing TDOA estimation from the microphone signals is the generalised cross correlation with phase transform (GCC-PHAT [14, 15]), and the estimates produced may be further improved by Viterbi smoothing [16]. If the relative location of the microphones is known then, given the TDOA values for a pair of microphones, simple geometry may be used to calculate the angle of arrival of the signal in relation to the microphones. In fact, due to rotational symmetry, for two microphones, a single delay estimate results in 2 angles of arrival—the correct one, and another reflected on the axis of the two microphones

A subset of the NIST RT meetings (those recorded at the University of Edinburgh, IDIAP and TNO) were recorded using an 8 element circular microphone array of 20 cm diameter. These are the only meetings in the NIST RT data set for which the relative locations of the microphones is known, and therefore our diarisation results are presented for this subset.

The directivity pattern of a MVDR superdirective beamformer [17] for an 8-element microphone array with a diameter of 20 cm and a sample rate of 16 kHz (the conditions used in the NIST recordings), shows a main lobe width of $10\,^{\circ}$. In order to identify the angle of the speakers in relation to the array, we therefore created a sector activity (SA) map of $N = 36$ possible sectors, one every $10\,^{\circ}$. Every 256ms, the TDOA values for each microphone pair are estimated and the angle of arrival values calculated. A count in the sector corresponding to that angle is then incremented. We accumulate counts in 5s windows with 1s overlap, and the highest scoring sector for each window is recorded, so that for every second of recording we have the sector with the most activity (the active sector) calculated over 5s.

We then take the segmentation output of a VAD, and assign each speech segment to the active sector corresponding to the window. If the speech segment overlaps two windows, and the windows have different active sectors, then the segment is split and the resulting two segments assigned the corresponding active sectors from the two windows.

Having assigned each speech segment to a sector, the longest segment for each sector is selected as that sector's reference segment. We then make a second pass over the data—for each speech segment, we calculate its BIC score (cf. Eq.3) with each of the reference segments, and increment a count in a size N^2 sector matrix (SM) at location (i, j), where i corresponds to the sector the segment was originally assigned to, and j to the sector of the reference segment with the highest BIC score. Ideally, this matrix would only have entries on its diagonal because the originally assigned sector would be the same as the sector with the highest BIC score; the indices of the entries would then correspond to sectors with speakers. In reality this is not the case, however we have observed that entries do tend to cluster around locations on the diagonal. In order to identify the sectors with speakers we look for peaks in the entries on the diagonal of the sector matrix. The indices of the peaks correspond to the sector number in which we estimate a speaker is located—these are the speaker sectors. Finally, we make a third pass over the data and assign each speech segment to the speaker sector closest to the sector it was originally assigned. The complete process is shown in Figure 2.

4.2. Diarisation results

Diarisation system are evaluated in terms of the diarisation error rate, or DER. In addition to missed speech and false alarms, DER (see [11]) also takes into account the speaker

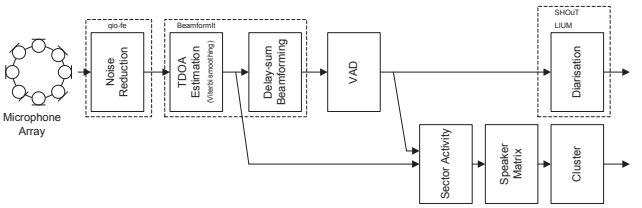

Fig. 2. Processing diagram for VAD and speaker diarisation

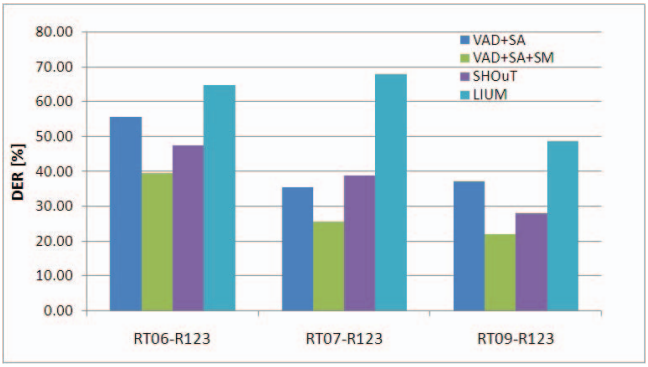

Fig. 3. Diarisation error rate DER for all algorithms

to whom each segment is assigned, and penalises segments assigned to the wrong speaker. In order to account for errors in the reference labels and slight variations in automatic processing, a tolerance of ± 250ms is permitted at the edge of each speech segment.

Four diarisation experiments on the EDI, IDI and TNO meetings from the RT06, RT07 and RT09 data sets were conducted, using the output from the Aurora [8] VAD process. First the DER of the direct output of the sector activity map (VAD+SA) was calculated, i.e. diarisation with fixed number of 36 speakers. Next the new algorithm (VAD+SA+SM) was evaluated. Finally, in order to provide baseline results, two open source diarisation systems—the SHOUT speech recognition toolkit[3] and the LIUM speaker diarisation system[4] were used. The results are given in Figure 3 and Table 1. The results show that the basic VAD+SA method achieves an improvement of 26% / 15% absolute compared to the LIUM tool. The VAD+SA+SM outperforms both SHOUT and LIUM, giving an improvement of 51% relative / 29% absolute compared to LIUM and 22% relative / 8% absolute compared to SHOUT. In addition, the number of speakers estimated by the VAD+SA+SM system is significantly closer to the actual number of speakers in the meeting than either of the other systems.

[3] http://shout-toolkit.sourceforge.net/

[4] http://lium3.univ-lemans.fr/diarization/doku.php

Table 1. VER, DER, and estimated number of speakers for each meeting. FA denotes false alarm, MS denotes missed speech.

			VAD+SA (basic)					VAD+SA+SM		SHOUT		LIUM	
	Meeting	spkrs	DER	VER	FA	MS	spkrs	DER	spkrs	DER	spkrs	DER	spkrs
RT06-R123	EDI_20050216-1051	4	48.06	9	2.9	6.1	32	31.02	4	45.28	10	65.68	7
	EDI_20050218-0900	4	53.18	8.9	2.5	6.4	32	30.37	5	49.84	9	67.2	7
	TNO_20041103-1130	4	65.26	8	1.8	6.2	32	57.07	7	46.81	12	61.65	2
	avg (RT06)		55.50	8.64	2.40	6.24		39.41		47.35		64.87	
RT07-R123	EDI_20061113-1500	4	44.63	4.7	4.2	0.5	32	31.82	4	56.74	14	72.49	1
	EDI_20061114-1500	4	27.45	6.3	5.7	0.6	32	20.34	4	23.43	10	64.06	6
	avg (RT07)		35.33	5.57	5.01	0.55		25.61		38.71		67.93	
RT09-R123	EDI_20071128-1000	4	34.63	3.9	3.2	0.7	32	16.65	4	23.43	8	56.2	3
	EDI_20071128-1500	4	45.95	5.3	4.7	0.6	32	35.11	5	30.95	13	82.32	2
	IDI_20090128-1600	4	27.81	1.6	0.7	0.9	32	11.96	4	23.82	9	19.74	19
	IDI_20090129-1000	4	41.73	4.9	4	0.9	32	25.54	4	34.38	14	41.67	12
	avg (RT09)		37.18	3.85	3.07	0.78		21.89		28.02		48.68	
	avg (all)		42.27	5.59	3.23	2.36		27.78		35.75		57.06	

5. DISCUSSION AND FUTURE WORK

We have proposed a TDOA-based algorithm to determine the number of active speakers in a meeting, and applied this to the diarisation task. The proposed algorithm outperforms BIC-based diarisation tools, due to its improved estimation of the number of speakers in the meeting. The algorithm is computationally less expensive than BIC based methods and can be easily adapted to require only a single pass over the data, making it applicable to online processing.

The algorithm performs well on NIST RT data, in which speakers are typically in a fixed location and do not move around. Such a restriction is not realistic, and moving speakers would significantly increase the error rate for TDOA-based systems, such as ours. Future work will include experiments involving moving speakers and combining our proposed method with BIC segmentation and clustering.

6. REFERENCES

[1] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. Computers*, pp. 1212–1224, 2007.

[2] H. Sun, T. L. Nwe, B. Ma, and H. Li, "Speaker diarization for meeting room audio," in *Proc. Interspeech*, 2009, pp. 900–903.

[3] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382 – 1393, 2009.

[4] S. Renals, T. Hain, and H. Bourlard, "Recognition and interpretation of meetings: The AMI and AMIDA projects," in *Proc. IEEE ASRU*, 2007.

[5] G. Lathoud, I. McCowan, and D. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *Eurospeech*, 2003.

[6] ITU-T, "ITU P.56, Objective Measurement of Active Speech Level," `http://www.itu.int/rec/T-REC-P.56/e`, 2011.

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

[8] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proc. ICSLP*, 2002.

[9] M. Huijbregts and F. de Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. 53, no. 2, pp. 143–153, 2011.

[10] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *Proc. Interspeech*, 2010.

[11] X. Anguera, *Robust Speaker Diarization for Meetings*, Ph.D. thesis, Universitat Politècnica de Catalunya, 2006.

[12] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[13] G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Trans. Audio, Speech, and Language Processing*, 2012, In press.

[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[15] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE ICASSP*, 1997.

[16] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[17] J. Bitzer and K. Uwe Simmer, "Superdirective microphone arrays," in *Microphone arrays: signal processing techniques and applications*, M. Brandstein and E Ward, Eds. Springer Verlag, 2001.