SPEAKER DIARIZATION USING UNSUPERVISED DISCRIMINANT ANALYSIS OF INTER-CHANNEL DELAY FEATURES

Nicholas W. D. Evans

EURECOM, Sophia Antipolis, France
and Swansea University, Swansea, UK
email: nick.evans@eurecom.fr

Corinne Fredouille and Jean-François Bonastre

Laboratoire Informatique d'Avignon (LIA)
University of Avignon, France
emails: {firstname.lastname}@univ-avignon.fr

ABSTRACT

When multiple microphones are available estimates of inter-channel delay, which characterise a speaker's location, can be used as features for speaker diarization. Background noise and reverberation can, however, lead to noisy features and poor performance. To ameliorate these problems, this paper presents a new approach to the discriminant analysis of delay features for speaker diarization. This novel and nonetheless unsupervised approach aims to increase speaker separability in delay-space. We assess the approach on subsets of four standard NIST RT datasets and demonstrate a relative improvement in diarization error rate of 25% on a separate evaluation set using delay features alone.

Index Terms— Speaker diarization, multiple distant microphones

1. INTRODUCTION

Speaker diarization involves the detection of speaker turns within an audio document (segmentation) and the grouping together of all same-speaker segments (clustering). Over recent years the community has placed emphasis on the conference meetings scenario, where there is typically more than one, table-top microphone. The question then is how best to use this additional information? There are two general approaches in the literature: the first involves the beamforming of the multiple signals to create a single enhanced signal prior to conventional acoustic feature extraction; the second involves the use of the inter-channel delay estimates themselves as features for diarization. This paper is concerned with the second approach.

Previous work has assessed the use of inter-channel delay features which may be used either on their own [1, 2], or combined with acoustic features [3, 4, 5]. Delay features are now very popular and of the six sites that submitted results for the NIST Rich Transcription (RT) 2007 evaluation [6] all report some experiments which use estimates of inter-channel delay. Background noise and reverberation, however, commonly lead to inaccurate delay estimates and this could account for the relatively poor performance that is achieved with delay features alone or the only modest improvement when they are combined with acoustic features. This has led to a number of recent proposals to improve the reliability of inter-channel delay features. Anguera *et al.* [4] report a complete front end comprised of N-best inter-channel delay estimation, noise thresholding and a dual pass Viterbi decoding scheme. More recently Otterson [7] investigated delay estimates from all-microphone pairs used in combination with energy ratios and principal component analysis (PCA). None of these approaches, however, is discriminant and thus they do not necessarily increase speaker separability. Discriminant approaches generally require supervision, i.e. the very speaker labels that diarization aims to discover, thus they are not suitable for our work.

Inspired by the recent work of Yang *et al.* [8] this paper presents a new approach to the discriminant analysis of inter-channel delay features which allows, for the first time, the learning of a discriminant feature transformation that is suitable for speaker diarization. This novel and nonetheless unsupervised approach aims to increase speaker separability in delay-space and hence the effectiveness of inter-channel delay features. We report speaker diarization experiments to assess the merit of the approach using delay features independently of acoustic features, i.e. not combined with acoustic features.

The remainder of this paper is organised as follows. Section 2 describes our baseline features. The unsupervised discriminant analysis (UDA) approach is described in Section 3. Our experimental work is described in Section 4 before our conclusions are presented in Section 5.

2. FEATURE EXTRACTION

In this section we describe our baseline feature extraction which involves speech activity detection and delay estimation.

2.1. Speech activity detection

Speech activity detection (SAD) is an essential element of any diarization system and one that could be performed using delay features. Whilst this is the subject of our ongoing research this was not the objective of the research reported here and so we opted to use the SAD system that was used for LIA's submission to the NIST RT'07 evaluation. The SAD system uses summed channels as its input and the features are 12 un-normalised linear frequency cepstral coefficients (LFCC) plus energy augmented by their first and second derivatives. The classifier is based on iterative Viterbi decoding and model adaptation applied to a two-state HMM, where one state is for speech and the other is for non-speech. Each state is initialised with a 32-component GMM trained on separate data using an EM/ML algorithm and state transition probabilities are fixed to 0.5. Finally, some state duration rules are applied in order to refine the speech/non-speech segmentation. The SAD stage is the only one to use acoustic features; inter-channel delay features are used everywhere else.

2.2. Delay estimation

We assume an unknown number of stationary speakers who are seated around a conference table on which are placed a number, m, of stationary microphones. There are thus ${}^{m}C_2 = m!/(2*(m-2)!)$

unique combinations of microphone pairs, so for 4 microphones there are 6 delay features and for 7 microphones there are 21 delay features. The inter-channel delay is estimated during speech periods from sliding windows of 0.5 seconds in length and with a window rate of 10 Hz using a standard generalised cross correlation with phase transform (GCC-PHAT) algorithm [9]. The set of feature vectors is represented by $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_P$. Each $\mathbf{x}_p$ is composed of ${}^mC_2$ features representing the delay between different microphone pairs, i.e.:

$$\mathbf{x}_p^T = (x_{p_{1,2}}, x_{p_{1,3}}, ..., x_{p_{1,m}}), (x_{p_{2,3}} ... x_{p_{2,m}}), ..., (x_{p_{m-1,m}}),$$

where ^{T} indicates the transpose and $x_{p_{k,l}}$ represents the estimated delay between the speech signals from microphones k and l in sample window p. The delay features are optionally processed by UDA, as is described in the next section, and are then fed directly into the diarization system as is described later in Section 4.

3. UNSUPERVISED DISCRIMINANT ANALYSIS

We seek a projection $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ in which (i) the inter-cluster (speaker) separability is maximised and (ii) the intra-cluster (speaker) separability is minimised. A suitable projection may be obtained by maximising the Fisher criterion:

$$J(\mathbf{w}) = \frac{J_B(\mathbf{w})}{J_W(\mathbf{w})} = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}, \quad (1)$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between-class and within-class scatter respectively. $J_B(\mathbf{w})$ and $J_W(\mathbf{w})$ are the corresponding scatters in the projected space. This is linear discriminant analysis (LDA), and we rely upon the sample labels being available in order to estimate cluster variance; LDA is a *supervised* approach and as such is not suitable for speaker diarization. Thus we require an *unsupervised* approach and in this paper we investigate a variant of LDA known as unsupervised discriminant analysis (UDA) which was proposed in 2006 by Yang *et al.* [8]. Some of this work originates from He *et al.* [10].

Rather than using the sample labels to estimate the between and within-class scatter, as is done in LDA, UDA estimates local scatter, $\mathbf{S}_L$, in place of $\mathbf{S}_W$ and non-local scatter, $\mathbf{S}_N$, in place of $\mathbf{S}_B$ in Equation 1. This is achieved by way of an adjacency matrix, $\mathbf{H}$, which is calculated according to:

$$H_{ij} = \begin{cases} 1 & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|^2 < \delta, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

which indicates whether or not any two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ have a Euclidean distance less than a pre-defined, empirically optimised threshold δ. The adjacency matrix, $\mathbf{H}$, defines pairs, or groups of samples which are to be regarded as within-class and negates the need for class labels thus allowing the calculation of an LDA-like projection without supervision. In [8] it is shown how the adjacency matrix is used to estimate the local scatter as follows:

$$\begin{aligned} \mathbf{S}_L &= \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P} H_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \\ &= \frac{1}{2}\left(\sum_{i=1}^{P}\sum_{j=1}^{P} H_{ij}\mathbf{x}_i\mathbf{x}_i^T + \sum_{i=1}^{P}\sum_{j=1}^{P} H_{ij}\mathbf{x}_j\mathbf{x}_j^T \right. \\ &\quad \left. -2\sum_{i=1}^{P}\sum_{j=1}^{P} H_{ij}\mathbf{x}_i\mathbf{x}_j^T\right) \\ &= (\mathbf{X}\mathbf{D}\mathbf{X}^T - \mathbf{X}\mathbf{H}\mathbf{X}^T) = \mathbf{X}\mathbf{L}\mathbf{X}^T, \end{aligned} \quad (3)$$

where the diagonal elements of $\mathbf{D}$ are equal to the column sums of $\mathbf{H}$. Each element D_{ii} indicates the number of samples within a Euclidean distance of δ to each sample i and acts to normalise their contribution to the resulting local scatter matrix $\mathbf{S}_L$. $\mathbf{L} = \mathbf{D} - \mathbf{H}$ is referred to as the Laplacian matrix [8, 10].

The non-local scatter matrix is calculated in exactly the same way as in Equation 3 except that each element H_{ij} in $\mathbf{H}$ is replaced by $1 - H_{ij}$ to give a new matrix $\mathbf{H}_N$ which for any sample i indicates all other samples j which are to be considered as between-class. Thus the non-local scatter is estimated according to $\mathbf{S}_N = \mathbf{X}\mathbf{L}_N\mathbf{X}^T$ where $\mathbf{L}_N = \mathbf{D}_N - \mathbf{H}_N$ and where each element of $\mathbf{D}_N$ is equal to the column sum of $\mathbf{H}_N$.

In the transformed space the local and non-local scatter are $\mathbf{w}^T\mathbf{S}_L\mathbf{w}$ and $\mathbf{w}^T\mathbf{S}_N\mathbf{w}$ respectively and thus, in now identical fashion to LDA, we determine a projection from the following generalised eigenvalue problem:

$$\mathbf{S}_N\mathbf{w} = \lambda\mathbf{S}_L\mathbf{w}. \quad (4)$$

In the usual manner the projection $\mathbf{w}$ into a $d-$dimensional space ($d < {}^mC_2$) is formed from the generalised eigenvectors which correspond to the d largest positive eigenvalues.

In addition to the locality defining threshold, δ, we have found it necessary to introduce one additional parameter. We have found that GCC-PHAT can produce a small number of anomalous delay estimates in the delay vectors $\mathbf{x}$. These are simply noisy delay estimates which can lead to inaccuracies in the adjacency matrix $\mathbf{H}$. This in turn produces poor estimates of the local and non-local scatter matrices and ineffective UDA projections.

We introduce a locality defining mask, $\mathbf{\Theta}$, which is used to reduce the effect of noisy delay estimates and thus to improve the resulting UDA projection. With our modification the calculation of Equation 2 takes into account only a percentage, θ, of the dimensions in $\mathbf{x}_i$ and $\mathbf{x}_j$ which have the smallest Euclidean distance; the remaining values are essentially treated as missing data. For any $\mathbf{x}_i$ and $\mathbf{x}_j$, the top half of Equation 2 is thus modified to $\|\Theta\cdot(\mathbf{x}_i - \mathbf{x}_j)\|^2$ where the $\cdot$ symbol indicates element wise multiplication and where the elements of $\mathbf{\Theta}$ are equal to one for the $\theta \times {}^mC_2$ dimensions with the smallest Euclidean distance and zero otherwise. The locality defining mask could be introduced into the diarization system itself, namely for the likelihood calculation in the Viterbi decoding stage. Whilst it may well lead to instability problems with iterative Viterbi decoding and adaptation this is nonetheless something that we intend to investigate.

In Section 4.4 we present diarization results as a function of the two UDA parameters, δ and θ, and have achieved a significant improvement in performance with this modification to the original UDA algorithm that was proposed in [8].

4. EXPERIMENTAL WORK

Here we describe the datasets used for our experimental work, the baseline diarization system, and our results. To assess the merit of the approach whilst avoiding interactions between delay and acoustic features, we report speaker diarization experiments using delay features independently of acoustic features, i.e. not combined.

4.1. Databases

Our experiments were performed using subsets of four standard NIST RT speaker diarization datasets. The '04 and '05 datasets were used for development and the '06 and '07 datasets were used for evaluation. We removed from each dataset all shows with fewer

than 3 microphones; 2 microphone channels give a single-order feature vector for which there is no capacity for feature transformation. As illustrated in Tables 1 and 2 there then remain 12 shows in the development set and 10 shows in the evaluation set.

4.2. Diarization system

The diarization system follows a fairly standard agglomerative clustering approach. We start with a 16-state (speaker) HMM which is initialised with a k-means algorithm. Each state has only a single Gaussian and a diagonal covariance matrix. Iterative Viterbi decoding and adaptation are performed until the model converges. State merging is controlled with a ΔBIC algorithm and is controlled with a conventional penalty parameter, λ, which is the same for every show within any one experiment. If state merging is performed then the process is repeated starting from the Viterbi decoding and adaptation stage and this continues until no further states warrant merging. The aim is for one state of the HMM to eventually represent a single speaker. Except for minor modifications to accommodate differing feature dimensions the back-end diarization system is identical for all experiments. Except where explicitly stated otherwise, all system parameters are empirically optimised on the development set and are applied without modification to the evaluation set.

4.3. Baseline results

Here we present a number of experiments where speaker diarization performance is reported in terms of diarization error rates (DERs) as specified by NIST, e.g. [11]. The DER incorporates both SAD and speaker error rates. Average development set SAD scores of 3.0% and 1.3% for the miss probability and false alarm rate respectively give an indication of performance. Differences between DER scores and average SAD scores give an indication of speaker error rates which might be more suited to assess the work presented here. Nonetheless, the reporting of DER scores rather than speaker error rates is more appropriate to facilitate the comparison of our results to those of others.

Here we report two sets of experiments. The first set aims to assess speaker diarization performance using the raw delay features described in Section 2. The results of this experiment form a baseline against which may be compared performance using UDA-derived feature transformations. The second set of experiments use conventional PCA-derived feature transformations and provide another baseline to help evaluate the benefit of the discriminative attributes of UDA.

For the first set of experiments the size of the feature vector depends on the number of available microphones. The second columns of Tables 1 and 2 show the number of microphones and corresponding raw feature dimensions for each show. The performance of the baseline system is illustrated in the third columns of Tables 1 and 2 for the development and evaluations sets respectively. For the development set the DER varies greatly with a minimum of 9% and a maximum of 68% DER. The average DER is 34%. With the same configuration minimum, maximum and average DERs of 10%, 50% and 32% respectively are achieved on the evaluation set (Table 2).

For the second set of experiments the raw full-order features are processed with conventional PCA. For these experiments (and those below using UDA) we decided to evaluate performance using the same number of feature dimensions for each audio show and investigate 1 and 2 dimension feature transformations. Results are illustrated in the fourth and fifth columns of each Table for 1 and 2 dimension transformations respectively. Average diarization performances of 38% and 40% are worse than for the baseline which shows that accounting soley for the global variance does not produce good perfomance and that a more discriminatory approach is required.

4.4. UDA results

In order to optimise UDA performance we ran a first set of experiments to choose a stable ΔBIC penalty factor λ and then, with a fixed λ, a second set of experiments in order to optimise the locality defining parameters δ and θ. An illustration of performance obtained from 2-dimension UDA feature transformations is given in Figure 1. The top plot (a) shows DER performance against the locality defining threshold, δ, for an empirically optimised locality defining threshold of θ=0.6. The solid profile shows performance for the development set and shows a minimum DER of 21% for δ=5. The dashed profile shows performance for the evaluation set. Performance is noticeably worse across the range but the optimum value of δ=2.5 compares well to that for δ=5 where the DER is 24%. Both profiles are reasonably flat near to δ=5.

Figure 1 (b) shows DER performance against the local defining mask, θ, for optimised δ=5. The solid profile illustrates performance for the development set and again shows a DER of 21% at θ=0.6. The dashed profile illustrates performance for the evaluation set and again shows a DER of 24% at θ=0.6. These values correspond to those in Figure 1 (a) for δ=5. The evaluation set profile shows a second trough around θ=0.85. This may be caused by a poorly tuned locality defining threshold which does not translate well from the development set to the evaluation set. The profile is nonetheless below the corresponding baseline DER of 32% between θ values of 0.55 and 0.95 and shows merit in the UDA approach.

Both 1 and 2 dimension UDA transformations were optimised in the same fashion and performance is summarised in Tables 1 and 2. Whilst UDA does not bring improvements for each show, development set DERs of 20% and 21% for 1 and 2 UDA dimensions respectively compare favourably with the baseline DER of 34% and those of PCA. However, the 1 dimension system does not translate well to the evaluation set and brings an improvement of only 1% absolute in DER (31% cf. 32%). The 2 dimension transformation fares much better and reduces the DER to 24%. This corresponds to a relative improvement of 25%.

5. CONCLUSIONS

When signals from multiple microphones are available it is possible to use estimates of inter-channel delay as features for diarization. However, whilst acoustic and delay features have been successfully combined improvements in performance coming solely from delay features is often small. This is thought to be caused by background noise and reverberation which can lead to inaccurate estimates of delay and could account for the performance gap between diarization performance with acoustic-only and delay-only features.

This paper describes what is believed to be the first assessment of a recently proposed unsupervised approach to the discriminant analysis of delay features which aims to increase speaker separability in a reduced-dimension delay-space. The approach shows merit on four standard NIST RT datasets. Using two-dimension UDA-derived feature transformations reasonably consistent improvements are observed on both development and evaluation datasets on which a baseline DER of 32% is shown to be reduced to 24% through the proposed approach and delay features alone. This amounts to a relative improvement of 25%.

Show	# mics/ feats.	Base-line	PCA d1	PCA d2	UDA d1	UDA d2
CMU_20050228-1615	3/3	36	26	8	9	8
CMU_20050301-1415	3/3	45	10	10	10	10
ICSI_20000807-1000	6/15	9	62	62	18	17
ICSI_20010531-1030	6/15	40	45	45	31	34
ICSI_20011030-1030	6/15	30	58	68	21	23
ICSI_20011113-1100	6/15	38	83	83	53	50
LDC_20011121-1700	10/45	39	14	39	13	13
LDC_20011207-1800	4/6	68	16	34	10	23
NIST_20030623-1409	7/21	31	19	20	20	21
NIST_20030925-1517	7/21	41	44	60	31	28
NIST_20050412-1303	7/21	15	32	33	19	11
NIST_20050427-0939	7/21	19	51	37	12	13
Average	**6/15**	**34**	**38**	**40**	**20**	**21**

Table 1. Development dataset summary: the number of microphones and raw feature dimensions for each show and their respective baseline, PCA and UDA (both 1 and 2 dimensions) DER performances. Average DERs are time weighted. The shows are subsets of the NIST RT'04 and RT'05 evaluation data.

Show	# mics/ feats.	Base-line	PCA d1	PCA d2	UDA d1	UDA d2
CMU_20061115-1030	3/3	34	19	21	21	21
CMU_20061115-1530	3/3	37	46	11	49	16
NIST_20051024-0930	7/21	25	34	41	16	18
NIST_20051102-1323	7/21	18	28	30	12	14
NIST_20051104-1515	7/21	10	21	19	27	7
NIST_20060216-1347	7/21	23	30	47	22	24
VT_20050408-1500	4/6	36	36	42	40	36
VT_20050425-1000	7/21	41	52	27	56	15
VT_20050623-1400	4/6	50	35	48	35	52
VT_20051027-1400	4/6	50	44	43	30	53
Average	**5/10**	**32**	**34**	**33**	**31**	**24**

Table 2. As for Table 1 except for the evaluation dataset where data are subsets of the NIST RT'05 and RT'06 evaluation data.

The feature transformation approach proposed here is intended to derive more robust features in a reduced-dimension delay-space. Using inverse transforms it might be possible to adapt the approach in order to attenuate noise in the original space. This would have obvious application in beamforming which could also contribute to improved speaker diarization performance.

6. REFERENCES

[1] D. P. W. Ellis and J. C. Liu, "Speaker turn detection based on between-channel differences," in *NIST meeting recognition workshop at Proc. ICASSP*, May 2004, pp. 112–117.

[2] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," *Proceedings of Machine Learning and Multimodal Interaction (MLMI '06), Washington. Lecture Notes in Computer Science*, vol. 4299, pp. 257–264, May 2006.

[3] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: Icsi rt06s evaluation system," in *Proc. ICSLP*, Sept. 2006, pp. 1674–1677.

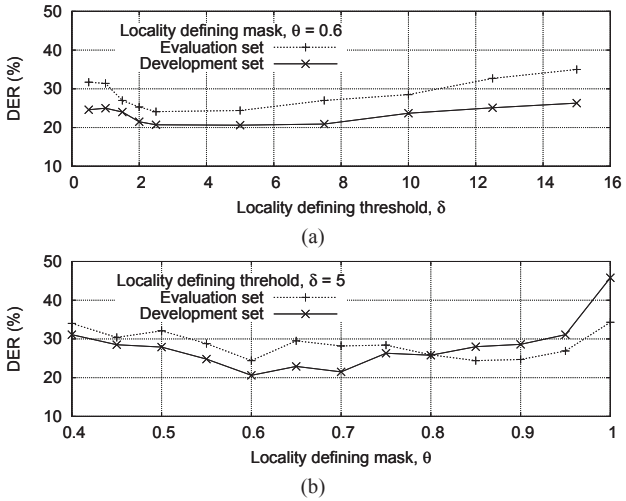

Fig. 1. Plots of DER against (a) the locality defining threshold, δ, (optimised θ=0.6) and (b) the locality defining mask, θ, (optimised δ=5) obtained from two-dimension UDA-derived feature transformations. In each case respectively θ and δ are optimised on the development set. Profiles are shown for both for the development and evaluation datasets.

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

[5] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *R. Stiefelhagen et al. (Eds): CLEAR 2007 and RT 2007, LNCS*, vol. 4625, pp. 509–519, 2008.

[6] R. Stiefelhagen, R. Bowers, and J. Fiscus (Eds.), "Multimodal technologies for preception of humans, international evaluation workshops: CLEAR 2007 and RT 2007," *Lecture Notes on Computer Science*, vol. 4625, 2008.

[7] S. Otterson, "Improved location features for meeting speaker diarization," in *Proc. Interspeech*, August 2007, pp. 1849–1852.

[8] J. Yang, D. Zhang, Z. Jin, and J.-Y. Yang, "Unsupervised discriminant projection analysis for feature extr," *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 904–907, 0-0 2006.

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

[10] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328–340, March 2005.

[11] NIST, "Spring 2007 (RT'07) Rich Transcription meeting recognition evaluation plan," `http://www.nist.gov/speech/tests/rt/2007/docs/rt07-meeting-eval-plan-v2.pdf`, February 2007.