Exploring the Large-Scale TDOA Feature Space for Speaker Diarization

Yi Yang and Jia Liu

Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing, P.R. China
{yangyy,liuj}@mail.tsinghua.edu.cn

Abstract. Using Time-Delay-Of-Arrival (TDOA) features has been proven greatly beneficial to the conventional acoustic feature-based speaker diarization systems by linking the speakers with their localization information. However, most state of-the-art speaker diarization systems depend on (relatively) limited distant microphones, which might not be sufficient in completely exploring the spatial information of speakers. In this study, the feature space spanned by TDOAs from (up to) 64 distant microphones is explored for the purpose of improving the performance of speaker classification, as an important branch of speaker diarization. Additionally, observing the intrinsic correlations of the high-dimensional feature space spanned by large-scale TDOAs, we compare several dimensionality reduction algorithms to explore an effective low-dimensional representation of TDOAs. Experimental results of speaker classification show consistent improvements when expanding the TDOA feature space by increasing the number of distant microphones. Furthermore, dimensionality reduction with the manifold information has been proven to be necessary for large-scale TDOAs.

Keywords: Speaker diarization, time-delay-of-arrival, dimensionality reduction, regularization.

1 Introduction

Using Multiple Distant Microphones (MDM) [1] has been proven effective in improving the performance of Speaker Diarization(SD) [2] which aims to find out "who spoke when" in audio recordings. The underlying reason is that the spatial features, namely Time Delay of Arrival (TDOA), convey the discriminative information of speaker locations, which offers different aspects of speaker identity information from the conventional acoustic features. Therefore, plenty of works are concentrated on the applicability of TDOA features [1, 3, 4, 5], all of which have consistently proven TDOA features to be complementary to the acoustic features.

Nevertheless, most previous speaker diarization systems maximally make use of sixteen microphones. Meanwhile, there are rarely papers discussing the relationship between the number of microphones and the system performance. Triggered by these two observations, we design an experiment which records the audio signals with

C. Stephanidis (Ed.): HCII 2014 Posters, Part I, CCIS 434, pp. 551–556, 2014.

large-scale (64) distant microphones (LSDM) for the purpose of speaker diarization, by which how the classification performance varies from the number of microphones could be investigated.

In this paper, a series of eigenvalue decomposition (EVD) based dimensionality reduction methods [7] with various regularization terms are evaluated to characterize the correlational information of TDOA features for the purpose of enhancing the discriminative power. This paper is organized as follows. In Section 2, we describe how to record the speech data with 64 distant microphones, following by the section introducing the derivation of TDOAs. Section 4 analyzes the properties of TDOA feature space and then present several specified regularized dimensionality reduction methods for the data. Experimental results on our corpus are shown in Section 5 and this paper is concluded in Section 6.

2 Audio Recording by Large-Scale Distant Microphones (LSDM)

2.1 The 64-microphone Audio Recording System

Fig. 1 is a multiple microphones device which we design as 64-channel elements and select part of asymmetric elements to compose one LSDM system. The 64-channel digital signals are collected with PXI-4496 multi-channel collection board produced by National Instruments.

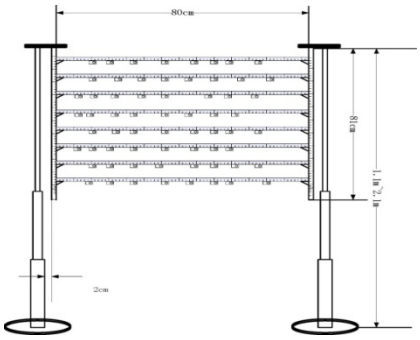

Fig. 1. The 64-channel multiple microphones device

2.2 Derivation of Large-Scale TDOA Feature Space

TDOA features are computed for all couples of distant microphones (in this paper there are in sum $\frac{64*63}{2} = 2016$ TDOA features) by the conventional Generalized Cross Correlation with PHAse Transform (GCC-PHAT) algorithm [8]. The following equations show the procedure of computation. Firstly, for i-th and j-th microphones, the Fourier representations of the snapshots ($x_i[n]$ and $x_j[n]$) should be derived as $x_i(f)$ and $x_j(f)$, with which the GCC-PHAT could be defined by:

$$G_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|} \quad (1)$$

in which the operator $[X(f)]^*$ denotes the complex conjugate of $X(f)$. The TDOA feature of such pair of microphones is estimated by $\hat{d}_{PHAT}(i,j) = \mathrm{argmax}_d(\hat{R}_{PHAT}(d))$.

The feature vector consists of $\frac{N*(N+1)}{2}$ TDOA features.

3 Regularized Dimensionality Reduction for Correlated TDOAS

Suppose there are n training samples, each of which is composed of $D = 2016$ TDOA features as $\vec{x}_i, i = 1,2,\dots,n$. Each sample is assigned by a class label $y_i \in \{1,2,\dots,C\}$ for C different speakers. As introduced in the first section, the feature space of $\vec{x}_i$ is highly correlated and abound in the data redundancy. Therefore, the dimensionality reduction will be addressed. The graph-based dimensionality reduction [9] is perhaps the most commonly-used approach which has been proven effective in various pattern classification tasks, including the MDM systems with TDOA features [8]. The basic idea is to find out the projection $W \in R^{D\times d}, d < D$ with which the low-dimensional representation $\vec{z} = W^T\vec{x}$ keeps the most discriminative information in the original high-dimensional feature space. The general optimization problem is to maximize the following objective function:

$$\overrightarrow{W} = \mathrm{argmax}_w \frac{\mathrm{tr}(W^T S^b W)}{\mathrm{tr}(W^T((1-\alpha)S^w + \alpha I)W)} \quad (2)$$

where $S^b$and S^w stand for the between-class and within-class scatter matrices [10]. They describe the separability among different classes and the class compactness, respectively.

3.1 Design of Scatter Matrices and Regularization Terms

The fundamental idea to capture such data structures is to design the scatter matrices by $S^w = \frac{1}{2}\sum_{ij} w_{ij}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^t$ and $S^w = \frac{1}{2}\sum_{ij} w_{ij}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^t$, where the operator $\vec{x}^t$ refers to the transpose of the vector $\vec{x}$. Two weights, namely w_{ij} and b_{ij}, are expected to describe the connectivity between two data points $\vec{x}_i$ and $\vec{x}_j$, which formulate the within/between-class affinity graphs. Such subtleties could be well captured by connecting a limited number of neighbors. This can be realized by setting $w_{ij}$and b_{ij}as 1 if $\vec{x}_i/\vec{x}_j$ is one of nearest neighbors of $\vec{x}_j/\vec{x}_i$, which results in two scatter matrices S_M^w and S_M^b [11].

In the view of Bayesian interference, the theoretical assumption of this regularization term is that the each row vector of the projection matrix T conform to a multivariate Gaussian distribution with the identity covariance matrix whose variance is controlled by the smoothness parameter α. Observing these facts, we investigate

the manifold regularizer to model the subtleties and protrusions over the entire training sets regardless of the speaker identity. This means that the inherent data structures of TDOA features are expected to be explored. The calculation of this regularizer is similar with that of within-class scatter matrix (Eq. 3.1). Symbolized by R_M, the regularizer is derived by $R_M = \frac{1}{2}\sum_{ij} r_{ij}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^t$. If $\vec{x}_i$ and $\vec{x}_j$ are neighbors, r_{ij} is set as 1, otherwise 0.

3.2 Compared Dimensionality Reduction Algorithms

Since we have reviewed several ways of estimating between-class scatter matrix, within-class scatter matrix, and the regularizer with different interpretations on the TDOA feature space, we compare several ways of derivation T of by defining S^b, $S^w$and R: LDA with L2-norm regularization, LDA with manifold regularization, LDA without regularization, manifold dimensionality reduction with L2-norm regularization, manifold dimensionality reduction without regularization.

4 Experimental Setup

Our data contain 50 different speakers. The evaluation method is the average pair-wise identification error rate of all possible pairs of speakers. There are in total 5000 samples segmented from the recorded speech signals, in which each speaker (class) has the same number of points. We choose 3000, 1000, and 1000 samples consisting of the training, developing, and testing sets: the projection T is trained by the training samples and the relevant parameters (e.g. the size of neighbors and the smoothing parameters) are tuned by the samples from the developing set. The performance is mainly measured by the average identification error rate for all pairs of speakers.

5 Experimental Results and Discussions

5.1 Experimental Design

To answer two research questions raised in the first section of this paper, we design two experiments. One is to explore the relationship between the identification rate and the number of used microphones. To realize this, we randomly select N; $8 < N \leq 64$ microphones and evaluate the performance of using the TDOA features solely and TDOA+MFCC features. We repeat these two steps 10 times for each possible N and compute the average rate. The other experiment is to explore whether the dimensionality reduction on the large-scale TDOA features is indispensable. Moreover, we compare the approaches mentioned in Section 3.3 to investigate the best way to exploit the structural information of TDOA feature space among these methods.

5.2 Relationship between the Number of Microphones and Speaker Identification Rate (Error Rate)

To find out the relationship between N and the identification rate, we compare the average error rate using different numbers of microphones from 9 to 64 Fig. 2 shows the result. Obviously, the performance in both cases (TDOA only and TDOA+MFCC) is approximately enhanced when N goes larger, which definitely indicates that adopting more distant microphones benefits the speaker identification. Moreover, the improvement in both cases strongly suggests that additional discriminative power from spatial information with more microphones is also helpful to the system with acoustic features.

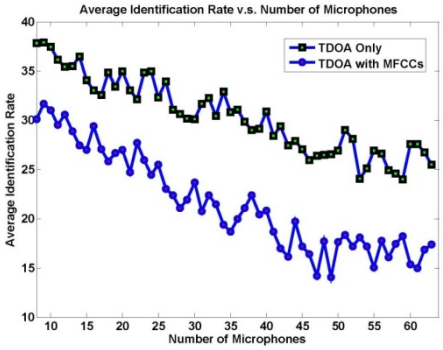

Fig. 2. Performance using different numbers of microphones based on TDOA and features merged by TDOA and MFCCs

Table 1. Performance comparison among five approaches

Methods	Error Rate
No Dim-Reduction	18.4%
LDA without Regu.	17.0%
LDA with L2-Norm Regu.	16.6%
LDA with Manifold Regu.	16.3%
Manifold Dim-Reduction without Regu.	15.2%
Manifold Dim-Reduction with L2 Norm Regu.	15.3%

5.3 Dimensionality Reduction of TDOA Feature Space

As mentioned in Section 3.3, several dimensionality reduction algorithms are evaluated with 64-microphone TDOA features. Since in the previous experiment we found that the merged feature vector outperforms the TDOA feature vector, in this part, we compare the different reduction algorithms by combining the reduced features and MFCCs, which yields the results shown in Table 1.

6 Conclusions and Future Works

In this paper, we designed a novel experiment which collected the data from (up to) 64 distant microphones, which makes it possible to answer several meaningful research questions about MDM-based speaker classification, as an important component of speaker diarization. In the near future, we will design and experiment the evaluation of the large-scale distant microphone system on the speaker diarization system.

Acknowledgements. Thanks to NSFC (61105017) agency for funding.

References

1. Pardo, J., Anguera, X., Wooters, C.: Speaker diarization for multiple-distant microphone meetings using several sources of information. IEEE Transaction on Computers 56, 1212–1224 (2007)
2. Tranter, S., Reynolds, D.: An overview of automatic speaker diarization systems. IEEE Transaction on Audio, Speech, and Language Processing 14, 1557–1565 (2006)
3. Vijayasenan, D., Valente, F., Bourlard, H.: An information theoretic combination of mfcc and tdoa features for speaker diarization. IEEE Transaction on Audio, Speech, and Language Processing 19, 431–438 (2011)
4. Vijayasenan, D., Valente, F., Motlicek, P.: Multistream speaker diarisation through information bottleneck system outputs combination. In: Proceeding of International Conference of Acoustics, Speech and Signal Processing, pp. 4420–4423 (2011)
5. Vijayasenan, D., Valente, F., Bourlard, H.: Multistream speaker diarization of meetings recordings beyond mfcc and tdoa features. Speech Communication 54, 55–67 (2012)
6. Evans, N., Fredouille, C., Bonastre, J.: Speaker diarization using unsupervised discriminant analysis of interchannel delay features. In: Proceeding of International Conference of Acoustics, Speech and Signal Processing, pp. 4601–4604 (2009)
7. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extension: A general framework for dimensionality reduction. IEEE Transaction on Pattern Analysis and Machine Intelligence 29, 40–51 (2007)
8. Anguera, X., Wooters, C., Hernando, J.: Speaker diarization for multi-party meetings using acoustic fusion. In: 2005 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 426–431. IEEE (2005)
9. Niyogi, X.: Locality preserving projections. In: Neural Information Processing Systems, vol. 16, p. 153 (2004)
10. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188 (1936)
11. Chen, H.T., Chang, H.W., Liu, T.L.: Local discriminant embedding and its variants. In: Computer Vision and Pattern Recognition, vol. 2, pp. 846–853 (2005)