Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 42 (2017) 81–99

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Unsupervised classification of speaker roles in multi-participant conversational speech

Yanxiong Li *, Qin Wang, Xue Zhang, Wei Li, Xinchao Li, Jichen Yang, Xiaohui Feng, Qian Huang, Qianhua He

School of Electronic and Information Engineering, South China University of Technology, 381 Wushan Road, Guangzhou, China

Received 11 May 2015; received in revised form 21 January 2016; accepted 13 September 2016
Available online 16 September 2016

Abstract

This paper proposes an unsupervised method for analyzing speaker roles in multi-participant conversational speech. First, features for characterizing the differences of various roles are extracted from the outputs of speaker diarization. Then, an algorithm of role clustering based on the criterion of maximizing the inter-cluster distance without using any convergence threshold is proposed to obtain the number of roles and to merge the utterances belonging to the same role into one cluster. The contributions of different combinations of individual feature subsets are compared for the proposed method on the outputs from speaker diarization, and the combined feature subsets obtain higher *F* scores than the individual ones for clustering speaker roles. The impacts of both speaker diarization errors and feature dimensions on the performance of the proposed method are also discussed. Experiments are done on the outputs of both manual annotations and automatic speaker diarization to compare the proposed method with both the state-of-the-art clustering method and the supervised method. Evaluations show that the proposed method is superior to the previous clustering method and close to the conventional supervised method in terms of *F* scores under two different experimental conditions.

Keywords: Speaker role; Speaker diarization; Role clustering; Multi-participant conversational speech

1. Introduction

Nowadays, the volume of recorded multi-participant conversational speech (e.g. TV news recordings, Press conference recordings, Lecture recordings, Summit recordings) has been rapidly increasing (Li et al., 2009; Ostendorf et al., 2008). How to efficiently organize, browse and retrieve these multi-participant speech documents has been received more and more attentions in the field of speech signal processing (Hain et al., 2012; Sinclair and King, 2013; Sun et al., 2010; Yella and Bourlard, 2014). Speaker role classification or analysis is a tool for alleviating the management of these huge mass of speech documents, and is a basis for story segmentation, for abstracting speech documents, for extracting high-level information in an indexing purpose and for a more effective access to data content (Bigot et al., 2010). Speaker role classification refers to identifying all utterances of the same role of speakers in a speech document and assigning a unique label of speaker role to them. Generally speaking, there are various speaker roles in different multi-participant conversational speech and one role consists of one speaker or many speakers. For

* Corresponding author at: School of Electronic and Information Engineering, South China University of Technology, Room 223, Shaw Science Building, 381 Wushan Road, Guangzhou, China. Fax: +86 20 87111435.
E-mail address: yanxiongli@163.com (Y. Li).

http://dx.doi.org/10.1016/j.csl.2016.09.002
0885-2308/

example, there are at least three roles in Lecture speech, i.e. *Orator*, *Host/Hostess*, *Questioner*, and the role of *Orator* or *Host/Hostess* generally consists of one speaker while the role of *Questioner* comprises many speakers. In this paper, the main task of speaker role classification is to find the number of roles in the speech documents and to merge the utterances belonging to the same role into one cluster.

Some previous studies have tackled the problem of speaker role recognition or identification from lexical features extracted from the outputs of Automatic Speech Recognition (ASR) (Barzilay et al., 2000; Damnati and Charlet, 2011; Liu, 2006; Wang et al., 2011), from acoustic features (Bigot et al., 2012, 2013; Salamin and Vinciarelli, 2012; Vinciarelli, 2007), or from a combination of lexical and acoustic features (Dufour et al., 2014; Hutchinson et al., 2010; Laurent et al., 2014; Sapru and Valente, 2012). The previous studies are detailed as follows.

Barzilay et al. (2000) used the transcriptions outputted from ASR and manual speaker boundaries as input, and assigned each speech segment with a role label among *Anchor*, *Journalist* or *Program Guest*. They applied two algorithms to the classification task, i.e. BoosTexter and Maximum Entropy Model (MEM). Their approach was based on lexical features and speaker introduction phrases. They evaluated on 35 audio recordings of broadcast shows (about 17 h), and obtained a classification accuracy of 80% of segments. Damnati and Charlet (2011) combined speaker clustering and analysis of ASR output for assigning speaker turns a role among: *Anchor*, *Reporter* and *Other*, and obtained 86% classification accuracy for automatically segmented speaker turns on a 6.5-hour test corpus of 14 TVBN shows mixing news and conversational speech. ICSIBOOST was used to distinguish between the role of *Reporter* and the role of *Other*, which was a large-margin classifier based on a boosting method of weak classifiers. Liu (2006) proposed a Hidden Markov Model (HMM) based approach and a MEM for speaker role labeling using Mandarin broadcast news speech. The algorithms achieved classification accuracy of about 80% using the human transcriptions and manually labeled speaker turns. It was found that the MEM performs slightly better than the HMM, and that the combination of them outperforms any model alone. Wang et al. (2011) presented a supervised approach for detecting speaker roles (*Host/Chair*, *Reporter/Commentator*, *Audience participant*, *Other*) in broadcast conversational shows in three languages: English, Arabic, and Mandarin. Various lexical, structural, and social network analysis based features were explored, and feature importance was analyzed across the three languages. They also compared the performance when using features extracted from automatically generated annotations against that when using human annotations. AdaBoost algorithm was used for classifying speaker roles. Their algorithm achieved speaker role recognition accuracy of more than 86% for all three languages.

Salamin and Vinciarelli (2012) proposed an approach for the automatic recognition of speaker roles in conversational broadcast data, in particular, news and talk shows. The approach made use of behavioral evidence extracted from speaker turns and applies Conditional Random Fields (CRF) to infer the roles played by different individuals. The experiments were performed over a large amount of broadcast material (around 50 h), and the results showed an accuracy higher than 85%. Bigot et al. (2012) detected speaker roles, i.e. *Anchor*, *Journalist* and *Other*. Their work relies on the assumption of the existence of clues about speaker roles in temporal, prosodic and basic signal features extracted from audio files and from speaker segmentations. Each speaker was therefore represented by a 36-feature vector. They investigated the influence of two dimensionality reduction techniques (Principal Component Analysis and Linear Discriminant Analysis) and different classification methods, i.e. Gaussian Mixture Models (GMM), K-Nearest Neighbors (KNN) and Support Vectors Machines (SVM). Experiments were done on the 13-h corpus of the ESTER2 evaluation campaign. The best result reached about 82% of well recognized roles. Later, Bigot et al. (2013) proposed a classification strategy, by observing how building show-dependent models improves speaker role recognition. They modeled speaker roles using a supervised classification approach and recognized 5 roles (*Anchorman*, *Punctual and Recurrent Journalists*, *Punctual and Recurrent Others*). Their system was able to classify correctly 86.9% of speaker roles while being applied on manual speaker segmentations and 74.5% on automatic speaker diarization outputs. Vinciarelli (2007) presented two approaches for speaker role recognition in multiparty audio recordings. The experiments were conducted over a corpus of 96 radio bulletins with total length of about 19 hours. He identified six roles: *Anchorman*, *Secondary anchorman*, *Guest*, *Interview participant*, *Abstract* and *Meteo*. The approaches first segmented automatically the recordings into single speaker segments, and then performed role recognition using different techniques. The first approach was based on social network analysis; and the second relied on the intervention duration distribution across different speakers. The results showed that around 85% of the recording time was labeled correctly in terms of role.

Laurent et al. (2014) tackled the problem of speaker role detection from broadcast news shows by combining both lexical and acoustic features, and by using small decision trees, denoted bonsai trees. Experiments showed that using

bonsai trees as weak learners for the boosting algorithm greatly improved both system error rate and learning time. Hutchinson et al. (2010) made use of structural features and lexical features for classifying speakers into three roles: *Host*, *Expert guest* and *Soundbite*. They used a simple heuristic: the cluster whose members had the largest average number of turns was the *Host* cluster, that with the smallest average number of turns was the *Soundbite* cluster, and the remaining cluster contained the *Expert guests*. Evaluated on English and Mandarin talk shows, their method obtained performance similar to that reported in the previous works for broadcast news. Sapru and Valente (2012) investigated various structural, lexical and prosodic features as well as dialog act tags for speaker role identification. They defined two types of speaker roles, i.e. formal roles (including *Project Manager*, *Marketing Expert*, *User Interface Designer*, and *Industrial Designer*) and social roles (including *Protagonist*, *Supporter*, *Neutral*, *Gatekeeper*, *Attacker*). They used multi-class boosting algorithm (i.e. Boostexter) to produce a single accurate classifier by combining many weak learning algorithms. Results revealed an accuracy of 74% in recognizing the formal roles and an accuracy of 66% in correctly identifying the social roles. Dufour et al. (2014) first inputted acoustic and linguistic features to a classifier (i.e. ICSIBOOST) for detecting the spontaneity of each speech segment, and then proposed to directly apply their two-step spontaneity detection mehtod for speaker role recognition based on the link between speech spontaneity and speaker roles. They found that some speaker roles have a predominant class of spontaneity (for example, the prepared class for the *Commentator* role). Experiments made on the EPAC corpus showed that features and approaches initially designed to detect speech spontaneity in audio documents could directly be applied to classify speaker roles. Their approach allowed the system to assign the correct role to 74.4% of the speakers with the semi-automated system and reached a correct labeling of 76.8% of the duration with the fully-automated system.

As can be seen from the introductions to the previous studies, the methods used for classifying speaker roles were implemented according to the following procedures: the features (e.g. lexical features, acoustic features, or both of them) were first extracted from the text transcriptions or audio segments, and then heuristic rules or classifiers (e.g. HMM, MEM, CRF, GMM, KNN, SVM, decision trees, Boostexter, ICSIBOOST) were used to classify different speakers as pre-given types of roles. What's more, the authors in the previous works assumed that the type of multi-participant audio documents, the number and the name of speaker roles in these audio documents were known a priori. These assumptions are true for one specific type of audio documents. For example, in TV news, the number of speaker roles is fixed, e.g. three (i.e. *Anchor*, *Reporter* and *Other*). However, different types of multi-participants conversational speech have different types and numbers of speaker roles, and the type of audio documents, the number of speaker roles and the number of speakers per role, may be all or partly unknown a priori in practice, especially in the huge mass of speech documents. Hence, the previous methods for classification of speaker roles are effective only when these aforementioned assumptions are true, and thus have significant limitations when they are applied to process multiple types of multi-participant conversational speech documents, e.g. the speech documents of TV news, Press conferences, Lectures and Summits. This motivates us to propose a universal method which is effective for classifying speaker roles from all types of multi-participant conversational speech documents.

To overcome the drawbacks of the above methods, we propose an unsupervised method for speaker role classification. Some effective features are first defined to characterize the differences of various speaker roles, which are extracted from the outputs of speaker diarization. Then, an algorithm of speaker role clustering based on the criterion of maximizing the inter-cluster distance without using any threshold is proposed for obtaining both the number of roles and the utterances belonging to different speaker roles. The minimum of the distances between any two clusters is iteratively calculated and the corresponding two clusters with the minimum distance (denoted as cluster A and cluster B) are found. If the inter-cluster distance of the new $N_c - 1$ clusters (i.e. with merging cluster A and cluster B) is larger than the inter-cluster distance of the old N_c clusters (i.e. without merging cluster A and cluster B), then cluster A and cluster B are merged and the number of cluster is accordingly decreased by one (i.e. $N_c = N_c - 1$); otherwise, the stated mergence of clusters is terminated and thus the number of roles and the utterances belonging to different roles are finally obtained. We prepare four different types of multi-participant conversational speech documents to demonstrate the effectiveness and universality of the proposed method. The advantages of the proposed method are as follows. First, it can be used for speaker role classification in all kinds of multi-participant speech documents without knowing the type of speech documents, the name and the number of speaker roles, the number of speakers per role and the type of languages a priori. Second, it can discover speaker roles and merge all utterances of the same role into one cluster without using any complex classifiers or statistical models. Third, it uses features automatically extracted from audio data instead of text outputs of ASR and thus will be not affected by high recognition errors from ASR.

The rest of the paper is organized as follows. Section 2 describes the proposed method in detail, including the definitions and normalizations of features, the definitions of distances between any two different clusters, and the procedure of speaker role clustering. Section 3 presents experimental results and discussions, and finally conclusions are drawn in Section 4.

2. Proposed method

The main modules of the proposed method are shown in Fig. 1. The features used for characterizing different speaker roles are first extracted from audio data outputted by speaker diarization, and then speaker role clustering are implemented on these features for discovering the number of speaker roles and merging utterances belonging to the same role into one cluster.

2.1. Speaker diarization

The main task of speaker diarization is to detect speech segments and then assign them belonging to the same speaker a unique speaker label. Speaker diarization generally consists of three parts, i.e. speech detection (distinguishing speech from other audio segments, such as silence, applause, laughter), speaker segmentation (or speaker change detection) and speaker clustering (Anguera Miro et al., 2012; Moattar and Homayounpour, 2012). Speech detection is implemented by the work in Li et al. (2010), and then the detected speech segments are split into shorter segments (e.g. 3 s) to do speaker clustering by using the work in Li et al. (2014). In this study, speaker diarization is just one preprocessing step instead of main contribution of this paper and thus we will not discuss it in detail here.

2.2. Features extraction

Speakers of different roles in conversation have different characteristics in the following aspects: utterance durations, start and end time to make speech, duration of one utterance, frequency of utterance, number of speech segments, fluency of utterance, duration of inter-utterance, speaking rate, and so on. For example, the *Host*/*Hostess* is generally the first and last speaker in Lecture speech or Summit speech, and speaks fluently and relatively fast with higher utterance frequencies due to preparation in advance and their expertise; whereas the *Guests* spontaneously answer questions and thus they are not the first speaker and speak rather slowly with lots of silences or long pauses. Hence, extracting the features which can represent the characteristic differences among different speaker roles is definitely helpful for speaker role classification. After speaker diarization, the number of speakers, the start and end time of each speech segment of every speaker are all obtained, which is the basis for extracting features.

Before introducing features used for classifying speaker roles, it is necessary to define three terms, i.e. *speech segment*, *silence segment* and *turn*. A speech segment is defined by one continuous utterance of one speaker without silences not shorter than 1 second. A silence segment is defined by one continuous audio segment without containing any words and not shorter than 1 second with low energy, e.g. the pauses not shorter than 1 second, silent background noises (with low energy and not shorter than 1 second). If a segment with low energy is shorter than 1 second (e.g. segment generated by shorter breathing, pause, etc.), then it is regarded as one part of its adjacent speech segments instead of a silence segment. A turn is defined by one complete utterance that contains at least one word spoken by one speaker in a speech document, without interrupting by other speakers. That is, the speech segments contained in one turn all belong to the same speaker. The change points of different speakers are also the change points of different turns. A turn generally consists of one or many speech segments and silence segments. If a turn contains only one speech segment, then this speech segment is equivalent to a turn. The relations of turn, speech segments, and silence segments are

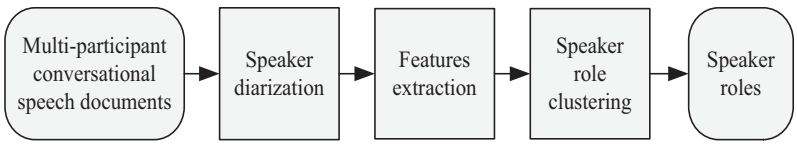

Fig. 1. The proposed method for unsupervised classification of speaker role.

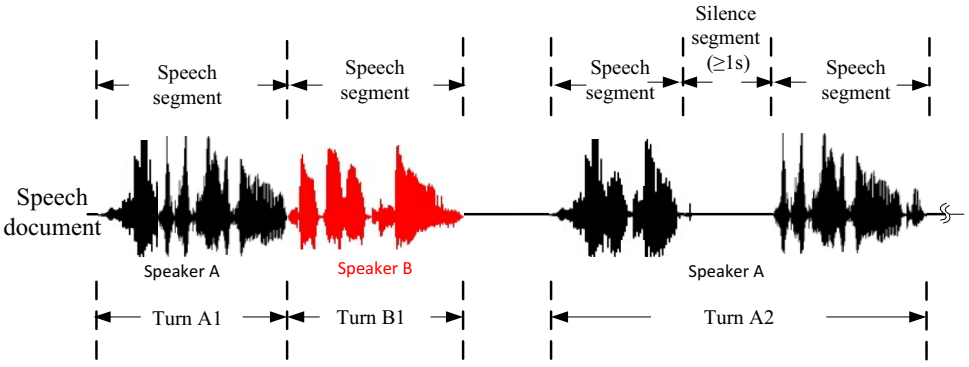

Fig. 2. The relations of turn, speech segments, and silence segments.

given in Fig. 2. The first turn (Turn A1) of speaker A consists of one speech segment, whereas the second turn (Turn A2) of speaker A includes two speech segments and one silence segment. The turn of speaker B (Turn B1) comprises one speech segment only.

Given one speaker has N turns in a speech document, and the start time and end time of the n-th turn of one speaker are denoted as $T_{n,s}$ and $T_{n,e}$, respectively, $1 \le n \le N$. It is also assumed that one speaker has M speech segments in a speech document, and the start time and end time of the m-th speech segment of one speaker are denoted as $S_{m,s}$ and $S_{m,e}$, respectively, $1 \le m \le M$. The features for characterizing different speaker roles are divided into three subsets, i.e. the feature subsets extracted from turns, the feature subsets extracted from speech segments, and the feature subsets extracted from both turns and speech segments.

2.2.1. *Feature subsets extracted from turns*

The features subsets extracted from turns, i.e. F_{Turn}, include 10 components: Number of Turns (NT), Largest Interval between two Adjacent Turns (LIAT), Total Intervals between two Adjacent Turns (TIAT), Mean of Intervals between two Adjacent Turns (MIAT), Duration of the Longest Turn (DLT), Total Duration of Turns (TDT), Mean of Durations of Turns (MDT), Start Time of the First Turn (STFT), End Time of the Last Turn (ETLT), Range of Utterance (RU).

NT: the total number of turns of one speaker in a speech document, and equal to N.
LIAT: the largest interval between two adjacent turns of one speaker in a speech document, defined below:

$$LIAT = \max_{1 \le n \le N-1} \left(T_{n+1,s} - T_{n,e}\right) \tag{1}$$

If one speaker utters only one time (i.e. $N = 1$), then LIAT is equal to zero.
TIAT: the total intervals between any two adjacent turns of one speaker in a speech document, defined below:

$$TIAT = \sum_{n=1}^{N-1} \left(T_{n+1,s} - T_{n,e}\right) \tag{2}$$

If one speaker utters only one time (i.e. $N = 1$), then TIAT is equal to zero.

MIAT: the mean of intervals between any two adjacent turns of one speaker in a speech document, defined below:

$$MIAT = \frac{1}{N-1} \sum_{n=1}^{N-1} \left(T_{n+1,s} - T_{n,e}\right) \tag{3}$$

If one speaker utters only one time (i.e. $N = 1$), then MIAT is equal to zero.

DLT: the duration of the longest turn of one speaker in a speech document, defined below:

$$DLT = \max_{1 \le n \le N} (T_{n,e} - T_{n,s}) \tag{4}$$

TDT: the total duration of all turns of one speaker in a speech document, defined below:

$$TDT = \sum_{n=1}^{N} (T_{n,e} - T_{n,s}) \tag{5}$$

MDT: the mean of durations of all turns of one speaker in a speech document, defined below:

$$MDT = \frac{1}{N} \sum_{n=1}^{N} (T_{n,e} - T_{n,s}) \tag{6}$$

STFT: the start time of the first turn of one speaker in a speech document, and equal to $T_{1,s}$.
ETLT: the end time of the last turn of one speaker in a speech document, and equal to $T_{N,e}$.
RU: the interval between the end time of the last turn and the start time of the first turn of one speaker in a speech document, defined below:

$$RU = T_{N,e} - T_{1,s} \tag{7}$$

2.2.2. Feature subsets extracted from speech segments

The feature subsets extracted from speech segments, i.e. F_{Seg}, consist of 4 elements: Number of Speech Segments (NSS), Duration of the Longest Speech Segment (DLSS), Total Duration of Speech Segments (TDSS), Mean of Durations of Speech Segments (MDSS). NSS, DLSS and MDSS, are similarly defined in the work of Bigot et al. (2012).

NSS: the total number of speech segments of one speaker in a speech document, and equal to M.
DLSS: the duration of the longest speech segment of one speaker in a speech document, defined below:

$$DLSS = \max_{1 \le m \le M} (S_{m,e} - S_{m,s}) \tag{8}$$

TDSS: the total duration of all speech segments of one speaker in a speech document, defined below:

$$TDSS = \sum_{m=1}^{M} (S_{m,e} - S_{m,s}) \tag{9}$$

MDSS: the mean of durations of all speech segments of one speaker in a speech document, defined below:

$$MDSS = \frac{1}{M} \sum_{m=1}^{M} (S_{m,e} - S_{m,s}) \tag{10}$$

2.2.3. Feature subsets extracted from turns and segments

The feature subsets extracted from both turns and segments, i.e. $F_{T/S}$, are composed of 6 parts: Ratio of total duration of Turns to number of Speech Segments (RTSS), Ratio of Total duration of turns to Total duration of speech segments (RTT), Ratio of total duration of Speech Segments to number of Turns (RSST), Number of Silence Segments in Turns (NSST), Total Duration of Silence Segments in Turns (TDSST), Mean Duration of Silence Segments in Turns (MDSST).

RTSS: the ratio of TDT (Total Duration of Turns) to NSS (Number of Speech Segments) of one speaker in a speech document, defined below:

$$RTSS = \frac{1}{M}\sum_{n=1}^{N}(T_{n,e} - T_{n,s}) \tag{11}$$

RTT: the ratio of TDT (Total Duration of Turns) to TDSS (Total Duration of Speech Segments) of one speaker in a speech document, defined below:

$$RTT = \frac{\sum_{n=1}^{N}(T_{n,e} - T_{n,s})}{\sum_{m=1}^{M}(S_{m,e} - S_{m,s})} \tag{12}$$

RSST: the ratio of TDSS (Total Duration of Speech Segments) to NT (Number of Turns) of one speaker in a speech document, defined below:

$$RSST = \frac{1}{N}\sum_{m=1}^{M}(S_{n,e} - S_{n,s}) \tag{13}$$

NSST: the total number of silence segments contained in turns of one speaker in a speech document. NSST is equal to NSS (Number of Speech Segments) minus NT (Number of Turns), as shown below:

$$NSST = M - N \tag{14}$$

TDSST: the total duration of silence segments contained in turns of one speaker in a speech document. TDSST is equal to TDT (Total Duration of Turns) minus TDSS (Total Duration of Speech Segments), as given below:

$$TDSST = \sum_{n=1}^{N}(T_{n,e} - T_{n,s}) - \sum_{m=1}^{M}(S_{m,e} - S_{m,s}). \tag{15}$$

MDSST: the mean duration of silence segments contained in turns of one speaker in a speech document. MDSST is equal to TDSST (Total Duration of Silence Segments in Turns) divided by NSST (Number of Silence Segments in Turns), as shown below:

$$MDSST = \frac{1}{M-N}\left(\sum_{n=1}^{N}(T_{n,e} - T_{n,s}) - \sum_{m=1}^{M}(S_{m,e} - S_{m,s})\right) \tag{16}$$

If the number of speech segments is the same as the number of turns for one speaker in a speech document, then MDSST is equal to zero.

2.2.4. Features normalization

As can be seen from the definitions of these aforementioned features, they have different units (e.g. seconds, number of occurrence) with large variations. To avoid the impact of their variations on measuring similarities between two clusters (i.e. speaker roles), they are normalized as shown below:

$$F_{ij} = \frac{F'_{ij} - \min(\boldsymbol{F}'_j)}{\max(\boldsymbol{F}'_j) - \min(\boldsymbol{F}'_j)}, \tag{17}$$

where F'_{ij} denotes the j-th feature of the i-th speaker in a speech document, and F_{ij} represents the normalized feature of F'_{ij}. $\boldsymbol{F}'_j$ is a feature vector consisting of the j-th feature of all speakers in a speech document, while $\max(\boldsymbol{F}'_j)$ and $\min(\boldsymbol{F}'_j)$ are the maximum and the minimum of feature vector $\boldsymbol{F}'_j$, respectively. Each normalized feature ranges from 0 to 1. Finally, a feature matrix, $\boldsymbol{F}$, consists of these normalized features above, as given below:

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}_1 \\ \cdots \\ \boldsymbol{F}_i \\ \cdots \\ \boldsymbol{F}_I \end{bmatrix} = \begin{bmatrix} F_{11}, F_{12}, \cdots F_{1j}, \cdots, F_{1J} \\ \cdots\cdots \\ F_{i1}, F_{i2}, \cdots F_{ij}, \cdots, F_{1J} \\ \cdots\cdots \\ F_{I1}, F_{I2}, \cdots F_{Ij}, \cdots, F_{IJ} \end{bmatrix}_{I\times J}, \tag{18}$$

where $\boldsymbol{F}_i$, I and J are the normalized feature vector of the i-th speaker, the number of speakers in a speech document and the dimension of features, respectively.

2.3. Speaker role clustering

In this Subsection, a distance metric is first defined for representing the similarities between any two speaker roles. Then, an algorithm of role clustering based on the criterion of maximizing inter-cluster distances is presented.

2.3.1. Feature distance

Features are extracted from the turns of the i-th speaker according to the method of features extraction described in Subsection 2.2, and the feature vector of the i-th speaker is denoted as $\boldsymbol{F}_i$ $(1 \le i \le I)$. A cluster is composed of feature vectors (i.e. $\boldsymbol{F}_i$) of some speakers which are expected to belong to the same role. Cluster $\boldsymbol{C}_k$ includes I_k feature vectors, $\boldsymbol{C}_k = \left[\boldsymbol{F}_1^k, \cdots, \boldsymbol{F}_i^k \cdots, \boldsymbol{F}_{I_k}^k\right]^T$; while cluster $\boldsymbol{C}_l$ is composed of I_l feature vectors, $\boldsymbol{C}_l = \left[\boldsymbol{F}_1^l, \cdots, \boldsymbol{F}_{i'}^l \cdots, \boldsymbol{F}_{I_l}^l\right]^T$, $(k, l = 1, 2, \ldots, N_C)$. $\boldsymbol{F}_i^k$ is the i-th feature vector in cluster $\boldsymbol{C}_k$ and $\boldsymbol{F}_{i'}^l$ is the i'-th feature vector in cluster $\boldsymbol{C}_l$, where $\boldsymbol{F}_i^k \ne \boldsymbol{F}_{i'}^l$ $(1 \le i \le I_k,\ 1 \le i' \le I_l)$. N_C represents the number of clusters.

The distance between feature vector $\boldsymbol{F}_i^k$ and feature vector $\boldsymbol{F}_{i'}^l$, $d\left(\boldsymbol{F}_i^k, \boldsymbol{F}_{i'}^l\right)$, is defined below:

$$d\left(\boldsymbol{F}_i^k, \boldsymbol{F}_{i'}^l\right) = \sqrt{\sum_{j=1}^{J}\left(F_{ij}^k - F_{i'j}^l\right)^2}, \tag{19}$$

where F_{ij}^k denotes the j-th feature of the i-th speaker in cluster $\boldsymbol{C}_k$ and $F_{i'j}^l$ denotes the j-th feature of the i'-th speaker in cluster $\boldsymbol{C}_l$. Distance matrix $\boldsymbol{D}(\boldsymbol{C}_k, \boldsymbol{C}_l)$ between $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$ is given as follows:

$$\boldsymbol{D}(\boldsymbol{C}_k, \boldsymbol{C}_l) = \begin{bmatrix} d\left(\boldsymbol{F}_1^k, \boldsymbol{F}_1^l\right), \cdots, d\left(\boldsymbol{F}_1^k, \boldsymbol{F}_{I_l}^l\right) \\ \cdots\cdots\cdots \\ d\left(\boldsymbol{F}_{I_k}^k, \boldsymbol{F}_1^l\right), \cdots, d\left(\boldsymbol{F}_{I_k}^k, \boldsymbol{F}_{I_l}^l\right) \end{bmatrix}_{I_k \times I_l} \tag{20}$$

Therefore, the distance $d(\boldsymbol{C}_k, \boldsymbol{C}_l)$ between two clusters is defined as the mean of all components of the distance matrix $\boldsymbol{D}(\boldsymbol{C}_k, \boldsymbol{C}_l)$, as given below

$$d(\boldsymbol{C}_k, \boldsymbol{C}_l) = \frac{1}{I_k \times I_l} \cdot \sum_{i=1}^{I_k}\left(\sum_{i'=1}^{I_l} d\left(\boldsymbol{F}_i^k, \boldsymbol{F}_{i'}^l\right)\right) \tag{21}$$

The smaller the distance $d(\boldsymbol{C}_k, \boldsymbol{C}_l)$ is, the more similar the two clusters are.

2.3.2. Role clustering algorithm

The objective of role clustering is to obtain the number of roles and to merge the turns belonging to the same role into one cluster. The algorithm of role clustering includes four steps, as described below.

Step 1: Each feature vector $\boldsymbol{F}_i^k$ is initially allocated to only one cluster $\boldsymbol{C}_k$, i.e. $N_C = I$. N_C and I are the initial number of clusters and the total number of speakers in a speech document, respectively. Each cluster $\boldsymbol{C}_k$ is composed of one feature vector only, i.e. $\boldsymbol{C}_k = [\boldsymbol{F}_i^k]$, $k = i$, $1 \le k \le N_C$, and $1 \le i \le I$.

Step 2: Compute distances between any two clusters from N_C clusters and obtain $0.5N_C \times (N_C - 1)$ different distances. The minimum distance, $d_{\min}(\boldsymbol{C}_k, \boldsymbol{C}_l)$, is then extracted from the $0.5N_C \times (N_C - 1)$ distances. That is, the corresponding two clusters with the minimum distance are $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$. Next, the old N_c clusters without merging $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$, are denoted as: $\{\boldsymbol{C}_k, \boldsymbol{C}_l, \boldsymbol{C}_{\overline{kl}}\}$, while the new N_c–1 clusters with merging $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$, are denoted as: $\{\boldsymbol{C}_{kl}, \boldsymbol{C}_{\overline{kl}}\}$. Here, $\boldsymbol{C}_{kl}$ consists of I_{kl} ($I_{kl} = I_k + I_l$) feature vectors from two clusters ($\boldsymbol{C}_k$ and $\boldsymbol{C}_l$), while $\boldsymbol{C}_{\overline{kl}}$ is a big background cluster (including N_c-2 clusters), comprising $I_{\overline{kl}}$ ($I_{\overline{kl}} = I - I_k - I_l$) feature vectors except the feature vectors of both the k-th cluster and the l-th cluster. $\boldsymbol{C}_{kl} = [\boldsymbol{F}_1, \cdots, \boldsymbol{F}_p, \cdots, \boldsymbol{F}_{I_{kl}}]^T$, $\boldsymbol{F}_p \in \{\boldsymbol{C}_k, \boldsymbol{C}_l\}, 1 \le p \le I_{kl}$, while $\boldsymbol{C}_{\overline{kl}} = [\boldsymbol{F}_1, \cdots, \boldsymbol{F}_q, \cdots, \boldsymbol{F}_{I_{\overline{kl}}}]^T$, $\boldsymbol{F}_q \notin \{\boldsymbol{C}_k, \boldsymbol{C}_l\}$, $1 \le q \le I_{\overline{kl}}$. Then, the inter-cluster distance for the old N_c clusters, $d(\boldsymbol{C}_k, \boldsymbol{C}_l, \boldsymbol{C}_{\overline{kl}})$, is defined in Eq. (22), while the inter-cluster distance for the new N_c-1 clusters, $d(\boldsymbol{C}_{kl}, \boldsymbol{C}_{\overline{kl}})$, is defined in Eq. (23).

$$d(\boldsymbol{C}_k, \boldsymbol{C}_l, \boldsymbol{C}_{\overline{kl}}) = \frac{d(\boldsymbol{C}_k, \boldsymbol{C}_l) + d(\boldsymbol{C}_k, \boldsymbol{C}_{\overline{kl}}) + d(\boldsymbol{C}_l, \boldsymbol{C}_{\overline{kl}})}{3}, \tag{22}$$

$$d(\boldsymbol{C}_{kl}, \boldsymbol{C}_{\overline{kl}}) = d(\{\boldsymbol{C}_k, \boldsymbol{C}_l\}, \boldsymbol{C}_{\overline{kl}}) = \frac{d(\boldsymbol{C}_k, \boldsymbol{C}_{\overline{kl}}) + d(\boldsymbol{C}_l, \boldsymbol{C}_{\overline{kl}})}{2} \tag{23}$$

Step 3: If the inter-cluster distance with merging $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$ is larger than that without merging $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$, i.e.:

$$d(\boldsymbol{C}_{kl}, \boldsymbol{C}_{\overline{kl}}) > d(\boldsymbol{C}_k, \boldsymbol{C}_l, \boldsymbol{C}_{\overline{kl}}), \tag{24}$$

that is,

$$d(\boldsymbol{C}_k, \boldsymbol{C}_l) < \frac{d(\boldsymbol{C}_k, \boldsymbol{C}_{\overline{kl}}) + d(\boldsymbol{C}_l, \boldsymbol{C}_{\overline{kl}})}{2}, \tag{25}$$

then $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$ are merged as a new cluster $\boldsymbol{C}_{kl}$ and the number of clusters is updated by decreasing one (i.e. $N_C = N_C - 1$), and then go to *Step 2*; otherwise, go to *Step 4*.

Step 4: Stop the procedure of clustering and N_C is regarded as the final number of clusters. The turns belong to the same role, if their corresponding feature vectors are merged in the same cluster.

The aforementioned role clustering algorithm is concisely outlined as shown in Fig. 3. Based on the criterion of maximizing the inter-cluster distances, the proposed role clustering algorithm iteratively merges two clusters with the minimum distance without using any threshold. The merging of clusters is continued until the inter-cluster distance does not increase.

3. Experiments

This Section gives experimental setups, including the introduction to experimental data and the definitions of performance metrics. Then, experimental results and discussions are presented.

3.1. Experimental setup

To the best of our knowledge, there is no official data set for speaker role classification up to now, and the experimental data used in the previous studies are different to each other. Hence, we construct a multi-participant conversational speech corpus which consists of 4 different types of conversational speech documents, including Press conferences, Lectures, Summits and TV news. The speech documents of Press conferences are composed of the press conference recordings of China premiers (e.g. Rongji Zhu, Jiabao Wen, Keqiang Li), the Press conference recordings of the

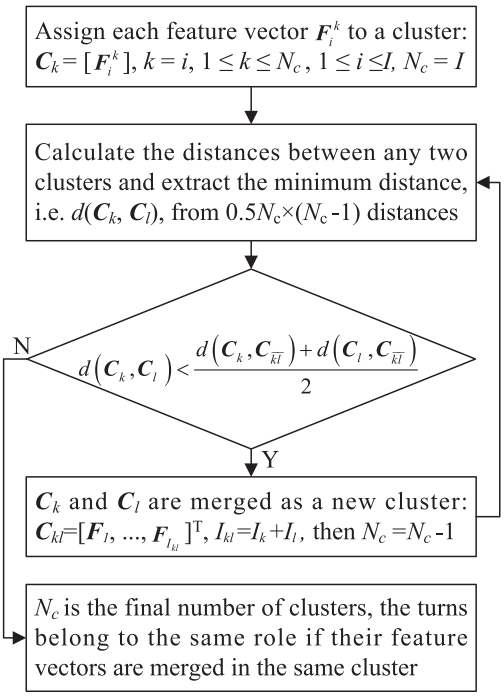

Fig. 3. The brief flowchart of the proposed algorithm for role clustering.

Table 1
The details of experimental data.

Data types	#F	#Spkrs/F	#Roles	(Roles, #Spkrs/R, ST/R)	Length (h)
Press conf.	40	8~20	4	(*Guest*, 1, ~ 31); (*Presider*, 1~2, ~ 5.5); (*Translator*, 1~2, ~ 29); (*Questioner*, 5~16, ~ 4.5)	70
Lectures	38	5~10	3	(*Orator*, 1, ~ 42); (*Presider*, 1, ~ 15); (*Questioner*, 3~8, ~ 8)	65
Summits	32	6~12	4	(*Guest*, 1, ~ 25); (*Presider*, 1~2, ~ 5.5); (*Translator*, 1~2, ~ 27); (*Questioner*, 5~16, ~ 2.5)	60
TV news	36	10~19	3	(*Anchor*, 1~2, ~ 10.8); (*Reporter*, 1~3, ~ 4.7); (*Interviewee*, 8~14, ~ 2.5)	18

#F: number of audio files; #Spkrs/F: number of speakers per file; #Roles: number of roles; (Roles, #Spkrs/R, ST/R): (speaker roles, number of speakers per role, speaking time per role in hour); length (h): total length of audio files in hour.

Ministry of Foreign Affairs of China and other official organizations of China. The speech documents of Lectures consist of speech recordings of national leaders, e.g. President Obama, President Jintao Hu, President Jinping Xi. The speech documents of Summits comprise speech recordings of economic summits held in China, e.g. APEC, Boao forum for Asia, Annual conference of China economy. The speech documents of TV news are the recordings of news of China Central Television. All speech documents are manually annotated by 10 postgraduate students (i.e. annotators) and then carefully inspected by other 10 postgraduate students (i.e. verifiers). Each audio file of experimental data is first annotated by one of 10 annotators and then carefully checked by one of 10 verifiers. 10 annotators and 10 verifiers are different, i.e. 20 different people in total. If there are no different opinions about the annotations in one audio file between annotator and verifier, then the annotations are considered as correct; otherwise, the controversial annotations will be checked by other 3 verifiers and the opinions agreed by at least 2 verifiers are used as the final annotations. The annotation information includes the endpoints of speech segments and silence segments, the endpoints of turns of each speaker, the identities of speakers and speaker roles, and so on. The annotation information is used as the ground truth for the performance evaluation of speaker role classification methods. All audio data are saved as mono-channel WAV format with a sample frequency of 16 kHz with 16 bits quantization. The details of experimental data are listed in Table 1.

As can be seen from Table 1, 4 different types of conversational speech recordings are used as experimental data, which have different roles, speaker numbers per role, speaker numbers per file, and speaking time per role. For example,

the speech documents of Press conference are 70 hours in length in total and have 4 roles (i.e. *Guest*, *Presider*, *Translator*, and *Questioner*). The average speaker numbers for the role of *Guest*, *Presider*, *Translator* and *Questioner* are 1, 1–2, 1–2, and 5–16, respectively, in the speech documents of Press conference. In addition, there are 8–20 speakers per file in this kind of speech documents.

Speaker role clustering is similar to speaker clustering in the view of merging the data belonging to the same cluster. The former is to merge the turns belonging to the same role into one cluster; whereas the latter is to merge the speech segments belonging to the same speaker into one cluster. Speaker clustering is a basis for speaker role clustering. Hence, the performance metrics for speaker clustering, i.e. average cluster purity and average speaker purity (Kotti et al., 2008), can be used for measuring the performance of speaker role clustering after appropriate revisions. Here, two metrics are used to evaluate the performance of speaker role clustering, i.e. average cluster purity (*ACP*) and average role purity (*ARP*).

Given n_{cr} is the total length of turns in cluster c uttered by speakers of role r; N_R is the total number of speaker roles; N_C is the total number of clusters; N_F is the total length of all turns; $n_{\cdot r}$ is the total length of turns uttered by speakers of role r; $n_{c\cdot}$ is the total length of turns in cluster c. Here, the basic unit of length is 20 milliseconds (i.e. the length of one speech frame). The relationships between n_{cr} and $n_{c\cdot}$, between n_{cr} and $n_{\cdot r}$, between n_{cr} and N_F, are given in Eqs. (26)–(28), respectively.

$$n_{c\cdot} = \sum_{r=1}^{N_R} n_{cr} \tag{26}$$

$$n_{\cdot r} = \sum_{c=1}^{N_C} n_{cr} \tag{27}$$

$$N_F = \sum_{c=1}^{N_C} \sum_{r=1}^{N_R} n_{cr} \tag{28}$$

The purity of cluster c, $\pi_{c\cdot}$ is given as shown below:

$$\pi_{c\cdot} = \sum_{r=1}^{N_R} \frac{n_{cr}^2}{n_{c\cdot}^2} \tag{29}$$

and the *ACP* is defined as

$$ACP = \frac{1}{N_F} \sum_{c=1}^{N_C} \pi_{c\cdot} n_{c\cdot} \tag{30}$$

The role purity for role r, $\pi_{\cdot r}$ is given as shown below:

$$\pi_{\cdot r} = \sum_{c=1}^{N_C} \frac{n_{cr}^2}{n_{\cdot r}^2} \tag{31}$$

and the *ARP* is defined as

$$ARP = \frac{1}{N_F} \sum_{r=1}^{N_R} \pi_{\cdot r} n_{\cdot r} \tag{32}$$

Finally, F score is adopted to characterize the overall performance of the methods, which is equal to the harmonic mean of *ACP* and *ARP*, as defined below:

$$F = \frac{2 \times ACP \times ARP}{ACP + ARP} \tag{33}$$

ACP, *ARP* and *F* range from 0 to 1. The higher *F* score is, the better performance the method is.

The predominant method for speaker clustering is Agglomerative Hierarchical Clustering (AHC) using Bayesian Information Criterion (BIC) as stopping criterion (i.e. AHC + BIC) (Kotti et al., 2008). The stopping criterion, i.e. ΔBIC, is used to compare the likelihood of two clusters (i.e. $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$) with that of the merged cluster $\boldsymbol{C}_{kl}$ (Kotti et al., 2008):

$$\Delta BIC = I_k \times \ln\left(\left|\det\left(\mathrm{cov}\left(\boldsymbol{F}^k\right)\right)\right|\right) + I_l \times \ln\left(\left|\det\left(\mathrm{cov}\left(\boldsymbol{F}^l\right)\right)\right|\right) + \alpha \times \left(J + 0.5J \times (J+1)\right) \times \ln\left(I_{kl}\right) - I_{kl} \times \ln\left(\left|\det\left(\mathrm{cov}\left(\boldsymbol{F}^{kl}\right)\right)\right|\right) \tag{34}$$

where $\mathrm{cov}(\cdot)$ represents covariance matrix of feature matrix, $\det(\cdot)$ is determinant of square matrix, $\ln(\cdot)$ denotes natural logarithm function, α is a BIC penalty coefficient, i.e. the tuning parameter of the BIC penalty and $\boldsymbol{C}_{kl}$ consists of both $\boldsymbol{C}_k$ and $\boldsymbol{C}_l$. J is the dimension of feature vector. $I_{kl} = I_k + I_l$ is the number of feature vectors in $\boldsymbol{C}_{kl}$, while I_k and I_l are the numbers of feature vectors in $\boldsymbol{C}_k$ and in $\boldsymbol{C}_l$, respectively. $\boldsymbol{F}^k$, $\boldsymbol{F}^l$ and $\boldsymbol{F}^{kl}$ are the corresponding feature matrices of $\boldsymbol{C}_k$, $\boldsymbol{C}_l$ and $\boldsymbol{C}_{kl}$, respectively. For each step of clustering, the pair of clusters with the highest ΔBIC is merged. The procedure of clustering is terminated when the highest ΔBIC becomes less than a fixed threshold (e.g. 0) (Kotti et al., 2008).

The AHC + BIC clustering method is implemented for role clustering and acts as a baseline in this paper. The BIC penalty coefficient α needs to be tuned for different experimental data in order to obtain the best result. Therefore, α is experimentally tuned and is finally set as 3.1 for obtaining the highest *F* score. The features used by the AHC + BIC are the same as that adopted by the proposed method. The conventional supervised method of speaker role classification is also implemented for performance comparison. The features used by the supervised method are the same as that used by the proposed method, and the classifier is Gaussian mixture model. The number of Gaussian mixtures for each model is experimentally set as 16. The features are extracted from audio data according to the definitions in Subsection 2.2 and the dimension of feature vector, *J*, is 20.

All experiments are done under the platform using C/C++ computer language on three LENOVO computers (AMD A10-7800 CPU 3.50 GHz, 8 GB RAM) with Windows 8. Speaker diarization is first implemented on all experimental data, and the results of speaker diarization are saved on the hard discs for later usage. The outputs of speaker diarization have been obtained already in our previous work, and are used as the inputs of speaker role discovery (including features extraction and clustering of speaker roles) in this work. The computational loads of features extraction of speaker roles and roles clustering are relatively light in comparison with that of speaker diarization. In addition, we process the experimental audio files one by one. Hence, the processing time for the experimental data using the proposed method is not a significant problem.

3.2. Experimental results

We first present the contributions of different feature subsets for speaker role classification when they are used in the proposed method. Next, we describe the clustering details of different roles for the proposed method using all feature subsets. Then, the effects of speaker diarization errors and feature dimensions on the performance of the proposed method are discussed. Finally, the proposed method is compared with both the previous unsupervised method (i.e. the AHC + BIC method) and the conventional supervised method (i.e. the GMM based method), evaluated on the turns generated by both manual annotations and automatic speaker diarization in different types of multi-participant conversational speech documents.

3.2.1. Contribution comparison of different feature subsets

To compare the contributions of feature subsets extracted from turns and speech segments, one experiment is carried out in which the features are divided into three subsets, i.e. F_{Turn} (the feature subsets extracted from turns), F_{Seg} (the feature subsets extracted from speech segments) and $F_{T/S}$ (the feature subsets extracted from both turns and speech segments). All data listed in Table 1 are used in this experiment. According to the feature definitions in Subsection 2.2, F_{Turn} include 10 components: NT, LIAT, TIAT, MIAT, DLT, TDT, MDT, STFT, ETLT and RU. F_{Seg} consist of 4 elements: NSS, DLSS, TDSS and MDSS; while $F_{T/S}$ are composed of 6 parts: RTSS, RTT, RSST, NSST, TDSST and MDSST. The contribution comparisons of different combinations of feature subsets for the proposed method evalu-

Table 2
The contribution comparisons of different combinations of feature subsets for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of Press conference.

Feature subsets	*ACP (%)*	*ARP (%)*	*F (%)*	*#C*
F_{Turn}	65.37	66.59	65.97	4
F_{Seg}	56.11	57.13	56.62	3
$F_{T/S}$	59.62	60.53	60.07	3
$F_{Turn} + F_{Seg}$	68.24	69.27	68.75	4
$F_{Turn} + F_{T/S}$	69.21	70.05	69.63	4
$F_{Seg} + F_{T/S}$	66.74	67.95	67.34	4
$F_{Turn} + F_{Seg} + F_{T/S}$	70.48	71.53	71.01	4

[#] *#C*: number of clusters.

Table 3
The contribution comparisons of different combinations of feature subsets for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of Lectures.

Feature subsets	*ACP (%)*	*ARP (%)*	*F (%)*	*#C*
F_{Turn}	66.73	67.98	67.35	3
F_{Seg}	56.95	57.43	57.19	2
$F_{T/S}$	61.36	62.31	61.83	2
$F_{Turn} + F_{Seg}$	69.70	70.83	70.26	3
$F_{Turn} + F_{T/S}$	70.59	71.65	71.12	3
$F_{Seg} + F_{T/S}$	68.58	69.76	69.16	3
$F_{Turn} + F_{Seg} + F_{T/S}$	72.40	73.27	72.83	3

[#] *#C*: number of clusters.

Table 4
The contribution comparisons of different combinations of feature subsets for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of Summits.

Feature subsets	*ACP (%)*	*ARP (%)*	*F (%)*	*#C*
F_{Turn}	60.78	62.46	61.61	3
F_{Seg}	51.22	52.52	51.86	2
$F_{T/S}$	55.81	56.47	56.14	2
$F_{Turn} + F_{Seg}$	63.49	64.52	64.00	3
$F_{Turn} + F_{T/S}$	64.80	65.91	65.35	3
$F_{Seg} + F_{T/S}$	62.82	63.96	63.38	3
$F_{Turn} + F_{Seg} + F_{T/S}$	66.78	67.33	67.05	3

[#] *#C*: number of clusters.

Table 5
The contribution comparisons of different combinations of feature subsets for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of TV news.

Feature subsets	*ACP (%)*	*ARP (%)*	*F (%)*	*#C*
F_{Turn}	62.98	64.22	63.59	3
F_{Seg}	53.31	54.72	54.01	2
$F_{T/S}$	57.71	58.51	58.11	2
$F_{Turn} + F_{Seg}$	66.11	67.03	66.57	3
$F_{Turn} + F_{T/S}$	66.81	67.97	67.39	3
$F_{Seg} + F_{T/S}$	64.83	66.11	65.46	3
$F_{Turn} + F_{Seg} + F_{T/S}$	68.82	69.16	68.99	3

[#] *#C*: number of clusters.

ated on the outputs of automatic speaker diarization in the speech documents of Press conference, Lectures, Summits and TV news, are given in Tables 2, 3, 4 and 5, respectively. As can be seen from these four tables, with the increase of numbers of feature subsets, the *F* scores steadily increase and the final numbers of clusters become being close and even equal to the number of speaker roles of ground truth for all types of speech documents. As far as the

Table 6
The clustering details of different roles for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of Press conference.

	Guest	Presider	Translator	Questioner
π_r (%)	76.27	70.02	73.68	65.21

Table 7
The clustering details of different roles for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of Lectures.

	Orator	Presider	Questioner
π_r (%)	78.35	72.64	68.28

Table 8
The clustering details of different roles for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of Summits.

	Guest	Presider	Translator	Questioner
π_r (%)	70.36	65.81	68.43	60.12

Table 9
The clustering details of different roles for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of TV news.

	Anchor	Reporter	Interviewee
π_r (%)	73.42	67.25	62.58

individual feature subsets are concerned, F_{Turn} is the best one and F_{Seg} is the worst one, among three individual feature subsets for speaker role classification. The reason may be that F_{Turn} represents more characteristics of different speaker roles in comparison with F_{Seg}, and the information contained in F_{Seg} is a part of F_{Turn} if the turns consist of only one speech segment. As for the combinations of two individual feature subsets, $F_{Turn} + F_{T/S}$ is the most effective one among three combinations of two individual feature subsets for classifying speaker roles from all types of speech documents. Finally, the combination of three individual feature subsets, i.e. $F_{Turn} + F_{Seg} + F_{T/S}$, obtain the best results in all different combinations of feature subsets, which indicates that these three individual feature subsets are complementary for speaker role classification.

3.2.2. Clustering details of different roles

The clustering details of different roles for the proposed method evaluated on the outputs of automatic speaker diarization in the speech documents of Press conference, Lectures, Summits and TV news, are given in Tables 6, 7, 8 and 9, respectively. All data listed in Table 1 are used in this experiment. All feature subsets, i.e. $F_{Turn} + F_{Seg} + F_{T/S}$, are used in the experiments. The role purity for role r, i.e. π_r, defined in Eq. (31), is used to characterize the clustering differences of different roles. As can be seen from these four tables, the role purities for different roles fluctuate in all types of speech documents. The role purities of *Guest* in the speech documents of Press conference and Summits are much higher than that of other roles, especially higher than that of *Questioner*. Similarly, the role purities of *Orator* in the speech documents of Lectures and the role purities of *Anchor* in the speech documents of TV news are much higher than that of other roles, especially higher than that of *Questioner* and that of *Interviewee*, respectively. The reason may be that the corresponding speakers of *Guest*, *Orator* and *Anchor* are the most active and regularly utter during the procedure of conversations and thus they can be effectively characterized by the features defined in Subsection 2.2. On the contrary, the corresponding speakers of *Questioner* and *Interviewee* seldom utter with short duration and few turns.

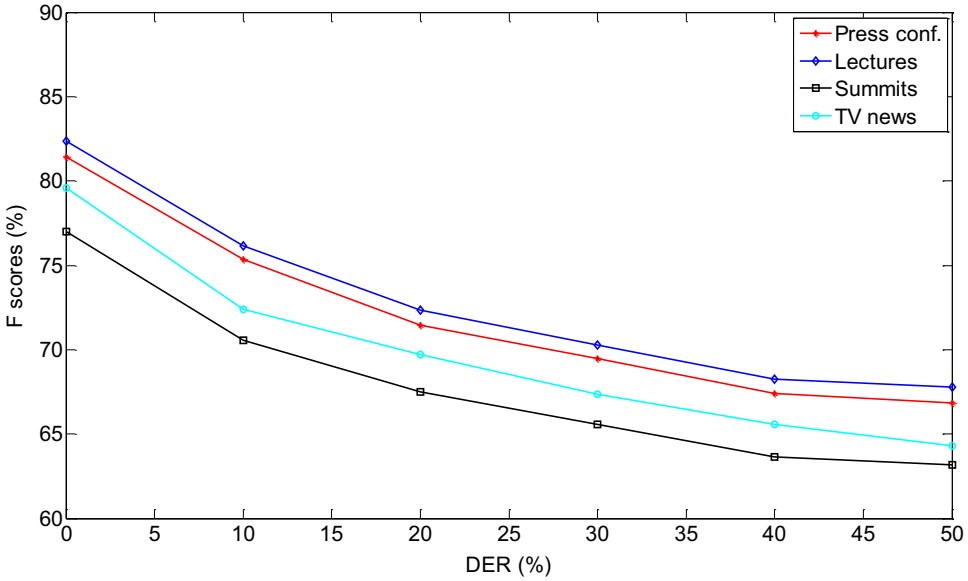

Fig. 4. The relationship between *F* scores of the proposed method and DERs (Diarization Error Rates) of speaker diarization for different types of data.

3.2.3. Effects of speaker diarization errors and feature dimensions

The outputs of speaker diarization are the basis to extract features of speaker roles for role clustering. Hence, there should be a relationship between speaker diarization quality and speaker role discovery. The outputs of speaker diarization with different DERs (Diarization Error Rates) (NIST, 2006) from 0 to 50% are first artificially created by controlling experimental conditions of speaker diarization. Then, the proposed method is performed based on these outputs of speaker diarization evaluated on the experimental data listed in Table 1. The relationship between *F* scores of the proposed method and DERs of speaker diarization for different types of data (Press conferences, Lectures, Summits, TV news) is shown in Fig. 4. DERs (i.e. diarization quality) have an obvious impact on *F* scores (i.e. performance of the proposed method), especially lower DERs. However, with the increase of DERs, this impact becomes weak.

To further know the effect of feature dimensions on the performance of the proposed method, we apply a principal component analysis (Duda et al., 2000) on the initial 20 features extracted from the outputs of automatic speaker diarization evaluated on the experimental data listed in Table 1. The relationship between *F* scores of the proposed method and feature dimensions for different types of experimental data (Press conferences, Lectures, Summits, TV news) is shown in Fig. 5. With the increase of feature dimensions from 1 to 18, *F* scores steadily increase for all types of experimental data. When the feature dimensions are equal to 19, 18, 19 and 18 for Lectures, Press conferences, TV news and Summits, respectively, the proposed method obtains the highest *F* scores. However, the differences between the highest *F* scores and that obtained by the initial 20 features are very minor for all types of experimental data.

3.2.4. Performance comparison of different methods

To compare the proposed method with both the previous unsupervised method and the conventional supervised method, another experiment is carried out in which the experimental data listed in Table 1 are equally divided into two parts. The first part is used to train GMMs in the supervised method, while the second part is used as test data for three different methods. It should be noted that the proposed method and the AHC + BIC method are all unsupervised methods and thus they do not need training data (i.e. the data of the first part). In order to use the same metrics for performance comparison, *ACP*, *ARP* and *F* scores are also used for the supervised method. The number of speaker roles and the number of clusters are known for the supervised method, whereas they are all unknown for the unsupervised methods in advance. The feature dimension is not reduced in these experiments, and is equal to 20.

Under the same experimental conditions, *ACP*, *ARP* and *F* scores of the proposed method, the AHC + BIC method and the supervised method evaluated on the outputs of manual annotations and automatic speaker diarization are given

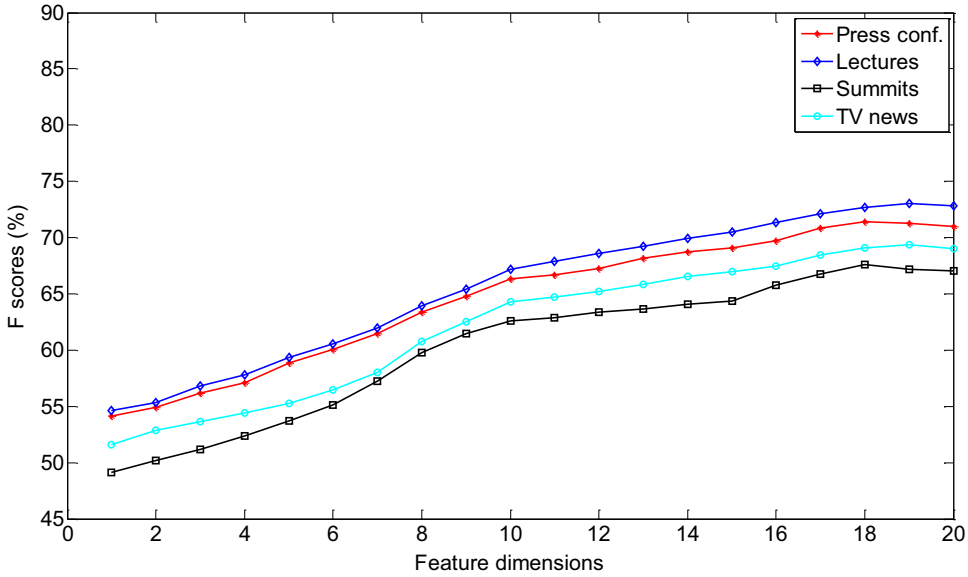

Fig. 5. The relationship between F scores of the proposed method and feature dimensions for different types of data.

Table 10
The performance comparison of different methods evaluated on the outputs of manual annotations.

Data type	The proposed				AHC + BIC				The supervised			
	ACP (%)	*ARP* (%)	*F* (%)	#*C*	*ACP* (%)	*ARP* (%)	*F* (%)	#*C*	*ACP* (%)	*ARP* (%)	*F* (%)	#*C*
Press conf.	80.84	82.01	81.42	4	77.15	77.86	77.50	4	84.92	85.35	85.13	4
Lectures	81.83	82.91	82.37	3	78.91	80.45	79.67	3	86.57	87.42	86.99	3
Summits	76.11	77.86	76.98	4	72.41	74.27	73.33	3	83.02	85.35	84.17	4
TV news	79.23	79.92	79.57	3	76.16	77.76	76.95	3	85.19	87.21	86.19	3
Mean	79.50	80.68	80.09	3.5	76.16	77.59	76.86	3.25	84.93	86.33	85.62	3.5

[#] #*C*: number of clusters.

Table 11
The performance comparison of different methods evaluated on the outputs of automatic speaker diarization.

Data type	The proposed				AHC + BIC				The supervised			
	ACP (%)	*ARP* (%)	*F* (%)	#*C*	*ACP* (%)	*ARP* (%)	*F* (%)	#*C*	*ACP* (%)	*ARP* (%)	*F* (%)	#*C*
Press conf.	70.97	71.89	71.43	4	66.27	68.23	67.24	3	76.52	78.02	77.26	4
Lectures	72.88	73.76	72.32	3	69.05	70.21	69.63	3	79.91	80.47	80.19	3
Summits	67.23	67.82	67.52	3	62.86	65.10	63.96	3	77.18	78.25	77.71	4
TV news	69.23	70.15	69.69	3	66.45	68.47	67.45	3	75.62	76.53	76.07	3
Mean	70.08	70.91	70.49	3.25	66.16	68.00	67.07	3	77.31	78.32	77.81	3.5

[#] #*C*: number of clusters.

in Tables 10 and 11, respectively. In Table 10, the proposed method averagely achieves *ACP* of 79.50%, *ARP* of 80.68%, and *F* of 80.09%. In Table 11, the proposed method averagely achieves *ACP* of 70.08%, *ARP* of 70.91%, and *F* of 70.49%. The proposed method obtains higher *F* scores for the speech documents of Lectures and lower *F* scores for the speech documents of Summits. Although the *F* scores obtained by the proposed method are different for different types of data, the gaps of *F* scores are not significant, i.e. with maximum of 5.39% (82.37%–76.98%) in Table 10 and maximum of 4.8% (72.32%-67.52%) in Table 11. That is, the proposed method is effective and universal for classifying speaker roles in different types of multi-participant conversational speech documents.

The performances of the three different methods all degrade when they are evaluated on the turns generated by automatic speaker diarization. That is, the pre-processing step of speaker diarization introduces *F* score decreases by

9.6% (80.09%–70.49%) for the proposed method, 9.79% (76.86%–67.07%) for the AHC + BIC method, and 7.81% (85.62%–77.81%) for the supervised method. Compared with the AHC + BIC method, the proposed method obtains improvement of *F* scores by 3.23% (80.09%–76.86%) and 3.42% (70.49%–67.07%) under two different experimental conditions, respectively. The results show that the proposed method can be used for speaker role classification from the turns generated by both manual annotations and automatic speaker diarization with better performance in comparison with the AHC + BIC method. What's more, the proposed method does not need to manually tune any thresholds, whereas the AHC + BIC method needs to properly set the BIC penalty α. Compared with the supervised method, the unsupervised methods (including the proposed method and the AHC + BIC method) obtain poor performance in terms of *F* scores. As shown in Tables 10 and 11, the gaps of average *F* scores between the proposed method and the supervised method are 5.53% (85.62%–80.09%) and 7.32% (77.81%–70.49%) evaluated on the outputs of manual annotations and the outputs of automatic speaker diarization, respectively. The improvements obtained by the supervised method are mainly due to the usages of pre-trained classifiers and the known number of roles and clusters. Although the proposed method is inferior to the supervised method in terms of *F* scores, but its advantage over the supervised method is that it can obtain the number of speaker roles and the utterances of the corresponding roles without training any classifier and preknowing any information about the experimental data. This advantage is very useful for processing huge mass of speech documents with different types in practice.

When evaluated on the turns generated by the manual annotations, the final numbers of roles (as given in Table 10) obtained by the proposed method are same as the number of roles of ground truth for different types of data (as shown in Table 1). But, the final number of roles for Summits obtained by the AHC + BIC method is incorrect (3 in Table 10 vs 4 in Table 1). Similarly, the proposed method has hypothesized the correct number of roles from the turns generated by automatic speaker diarization for Press conferences, Lectures and TV news, except Summits (3 in Table 11 vs 4 in Table 1), whereas the final number of roles for both Press conferences and Summits estimated by the AHC + BIC method are all wrong (3 and 3 in Table 11 vs 4 and 4 in Table 1, respectively). Hence, the ability for correctly estimating the number of roles of the proposed method is stronger than that of the AHC + BIC method. In Table 10, the number of roles estimated by the proposed method are equal to the true number of roles for all types of data, the corresponding *ACP*, *ARP* and *F* scores will be same whether the number of roles are known or unknown a priori. However, in Table 11, the number of roles estimated by the proposed method for Summits is 3, whereas the true number of roles for Summits is 4. In order to obtain the error compared to when the number of roles is known a priori, the number of roles is artificially set as 4 for processing the speech documents of Summits. As a result, *ACP*, *ARP* and *F* scores increase from 67.23%, 67.82% and 67.52% (as given in Table 11), to 68.57%, 69.13% and 68.85%, respectively. The performance is improved for processing Summits when the number of roles is known a priori, but the errors between 3 clusters (role number unknown a priori) and 4 clusters (role number known a priori) are not significant. This is partly due to the unbalance of speaking times of different types of speakers in Summits. The speaking time of *Questioner* in Summits is shorter than that of other three types of speakers as shown in Table 1.

4. Conclusions

In this paper, we proposed an unsupervised method for classifying speaker roles in multi-participant conversational speech by using the features extracted from the outputs of speaker diarization and by using a role clustering algorithm. This method obtained the number of speaker roles in a speech document and merged the turns belonging to the same role into one cluster without preknowing any information concerning the speech documents, such as the type of speech documents, the name and number of speaker roles, the number of speakers per role, and the language spoken by speakers.

The contributions of feature subsets extracted from turns, speech segments, both turns and speech segments, were compared for the proposed method on the outputs of automatic speaker diarization in different types of speech documents. It was found that the feature subsets extracted from turns achieved the best results in comparison with other two individual feature subsets. In addition, the combinations of these three feature subsets are superior to these three individual feature subsets and were also better than the combinations of two individual feature subsets for speaker role classification. The role purities for different roles varied in different speech documents. The role purities of *Guest* in the speech documents of Press conference and Summits were much higher than that of other roles, especially higher than that of *Questioner*. Similarly, the role purities of *Orator* and *Anchor* in the speech documents of Lectures and TV news were much higher than that of *Questioner* and that of *Interviewee*, respectively.

Both speaker diarization errors and feature dimensions have impacts on the performance of the proposed method. The effects of speaker diarization errors on the performance of the proposed method are great when speaker diarization errors are lower. But, this effect gradually becomes weak when the diarization errors increase. With the increase of feature dimensions, the performance of the proposed method is generally improved for all types of experimental data.

Compared with the AHC + BIC method, experimental evaluations showed that the proposed method: (1) achieved higher *F* scores for speaker role clustering on the outputs of both manual annotations and automatic speaker diarization; (2) was less sensitive to the errors of speaker diarization; and (3) was not required to manually tune any thresholds. Compared with the supervised method, although the proposed method obtained lower *F* scores for speaker role clustering on the outputs of both manual annotations and automatic speaker diarization, but it did not need to train any classifier and to preknow any information about the speech documents. That is, the proposed method is more universal than the supervised method for processing various types of multi-participant speech documents.

The classification of speaker roles in an unsupervised way is a solid foundation for audio content analyses and higher level semantic information extraction in huge mass of speech documents. In the next work, we plan to further study how to effectively extract more speaker information from speech documents and integrate these methods into the speaker analyses and retrieval system. Two directions about speaker role discovery need to be further investigated for performance improvement. The first direction is to extract more effective features for characterizing different speaker roles. Although the features used in this work are effective and relatively easy to extract from the outputs of speaker diarization, other features are still needed to further improve the performance of speaker role discovery. The second one is to study the global optimization clustering method instead of using hierarchical clustering algorithm. The hierarchical clustering algorithm suffers from longer computational time and error propagation, since it iteratively merges two clusters until the end of the clustering process.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (61101160, 61271314, 61301300, 61401161, 61571192), the Fundamental Research Funds for the Central Universities, South China University of Technology, China (2015ZZ102, 2015ZM145, 2015ZM143), Project of the Pearl River Young Talents of Science and Technology in Guangzhou, China (2013J2200070), Science and Technology Planning Project of Guangdong Province (2014A050503022, 2015A010103006, 2015A030313600, 2015A010103003) and the Foundation of China Scholarship Council (201208440078).

References

Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O., 2012. Speaker diarization: a review of recent research. IEEE Trans. Audio Speech Lang. Process. 20, 356–370.

Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S., 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In: 17th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 679–684.

Bigot, B., Pinquier, J., Ferrane, I., Andre-Obrecht, R., 2010. Looking for relevant features for speaker role recognition. In: Interspeech, pp. 1057–1060.

Bigot, B., Ferrané, I., Pinquier, J., André-Obrecht, R., 2012. Detecting individual role using features extracted from speaker diarization results. Multimed. Tools Appl. Archiv. 60, 347–369.

Bigot, B., Fredouille, C., Charlet, D., 2013. Speaker role recognition on TV broadcast documents. In: Slam@Interspeech, pp. 66–71.

Damnati, G., Charlet, D., 2011. Robust speaker turn role labeling of TV broadcast news shows. In: IEEE ICASSP, pp. 5684–5687.

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification, second ed. Wiley-Interscience.

Dufour, R., Estève, Y., Deléglise, P., 2014. Characterizing and detecting spontaneous speech: application to speaker role recognition. Speech Commun. 56, 1–18.

Hain, T., Burget, L., Dines, J., Garner, P.N., Grézl, F., Hannani, A.E., et al., 2012. Transcribing meetings with the AMIDA systems. IEEE Trans. Audio Speech Lang. Process. 20, 486–498.

Hutchinson, B., Zhang, B., Ostendorf, M., 2010. Unsupervised broadcast conversation speaker role labeling. In: IEEE ICASSP, pp. 5322–5325.

Kotti, M., Moschou, V., Kotropoulos, C., 2008. Speaker segmentation and clustering. Sign. Process. 88, 1091–1124.

Laurent, A., Camelin, N., Raymond, C., 2014. Boosting bonsai trees for efficient features combination: application to speaker role identification. In: Interspeech, pp. 76–80.

Li, Y., Jin, H., Li, W., He, Q., Zhu, Z., Feng, X., 2014. Fast speaker clustering using distance of feature matrix mean and adaptive convergence threshold. IET Sign. Process. 8, 844–851.

Li, Y.-X., He, Q.-H., Li, W., Wang, Z.-F., 2010. Two-level approach for detecting non-lexical audio events in spontaneous speech. In: International Conference on Audio, Language and Image Processing, pp. 771–777.

Li, Y.X., He, Q.H., Kwong, S., Li, T., Yang, J.C., 2009. Characteristics-based effective applause detection for meeting speech. Sign. Process. 89, 1625–1633.

Liu, Y., 2006. Initial study on automatic identification of speaker role in broadcast news speech. In: Human Language Technology Conference of the NAACL, pp. 81–84.

Moattar, M.H., Homayounpour, M.M., 2012. A review on speaker diarization systems and approaches. Speech Commun. 54, 1065–1103.

NIST, 2006. 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>. Accessed on 10 May 2015.

Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tur, D., Harper M., Hillard D., et al., 2008. Speech segmentation and spoken document processing. IEEE Sign. Process. Mag. 25, 59–69.

Salamin, H., Vinciarelli, A., 2012. Automatic role recognition in multiparty conversations: an approach based on turn organization, prosody, and conditional random fields. IEEE Trans. Multimed. 14, 338–345.

Sapru, A., Valente, F., 2012. Automatic speaker role labeling in AMI meetings: recognition of formal and social roles. In: IEEE ICASSP, pp. 5057–5060.

Sinclair, M., King, S., 2013. Where are the challenges in speaker diarization? In: IEEE ICASSP, pp. 7741–7745.

Sun, H., Ma, B., Khine, S.Z.K., Li, H., 2010. Speaker diarization system for RT07 and RT09 meeting room audio. In: IEEE ICASSP. IEEE, pp. 4982–4985.

Vinciarelli, A., 2007. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. IEEE Trans. Multimed. 9, 1215–1226.

Wang, W., Yaman, S., Precodal, K., Richey, C., 2011. Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In: IEEE ICASSP, pp. 5556–5559.

Yella, S.H., Bourlard, H., 2014. Information bottleneck based speaker diarization of meetings using non-speech as side information. In: IEEE ICASSP, pp. 96–100.