International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 8, No. 4 (2015), pp. 31-40
http://dx.doi.org/10.14257/ijsip.2015.8.4.03

Estimating Key Speaker in Meeting Speech Based on Multiple Features Optimization

Wei Li Yanxiong Li* and Qianhua He

(School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, Guangdong province, China)
scut_liwei@163.com

Abstract

This paper proposes to estimate key speaker in meeting speech based on multiple features optimization. First, each feature is defined and their differences between key speaker and other speakers are analyzed. Then, a decision function of multiple feature weighting is generated for estimating key speaker in meeting speech, and the genetic algorithm is used to optimize these coefficients of feature weighting. The methods are evaluated on three different meeting speech datasets. Experimental results show that the proposed optimization method obtains average accuracy of 93.3% for estimating key speaker, and gains average accuracy improvement by 9.7% and 4.1% compared with the previous method and the feature weighting method without optimization, respectively.

Keywords: *Meeting speech, Feature optimization, Key speaker*

1. Introduction

Meeting speeches (news conference speech, summit forum speech, interview speech, lecture speech of leaders, *etc.*,) present the tsunami type growth [1-5]. The key speaker in meeting speech refers to the speaker who has the greatest right to speak and is in the leading position during the meeting, such as the national leaders, the person in charge, the special guest and so on. The speech of the key speaker is most likely to be analyzed and processed. Therefore, it is very important to quickly and effectively find out the speech of the key speaker in a mass meeting speech, which is also one of the essential steps for content analysis and semantic understanding of meeting speech.

The topic to estimate the key speaker in meeting speech has been discussed only in recent years. There are not so many relevant researches. By collecting such features as speaking length, turn-talking times, times to obtain speaking right, times to interrupt others' speech, times being introduced, times being asked etc. and using support vector machine as classifier, Rutger Rienks, *et al.*, [6] obtain accuracy of 75% for estimating key speaker. Their research shows that the most effective features are turn-talking times and times to obtain speaking right. Later [7], they processed forty meeting speeches to compare the functions of the dynamic model (dynamic Bayesian network) and the static model (support vector machines, multilayer perception *etc.*,) to estimate the key speaker, the best testing result of which is 70.59%. The method they adopted needed to train complicated model (classifier), and had trouble in choosing the suitable model parameter. The improper selection of the model parameter has a great impact on the performance of the model. Hayley Hung, *et al.*, [8] respectively use audio and video to estimate the key speaker and the experimental result shows that to collect the video feature is time-consuming and its performance is not satisfactory. Jayagopi, *et al.*, [9]attempted to combine the features of audio and video to estimate the key speaker. The adopted features and the feature parameters in reference [6] are the same. However, the obtained testing

* Corresponding author

ISSN: 2005-4254 IJSIP

performance is nearly the same with the performance of the single audio feature. Yukiko, *et al.*, [10]also try to combine the features of audio and video to estimate the key speaker among three speakers. Except collecting the audio feature, they collected other video features like times looking at the other party, times being stared and times looking at each other. The experiment data was obtained in a certain circumstance which needed to install many cameras. Besides, the scope of application of the data was limited. In addition, if the speaker moves and turns around (for example delivering speech) during the speaking process, it becomes very difficult to collect the video feature as eye contact. Hence, research methods based on video feature is greatly restricted when being applied. Based on the segmentation and cluster of the speaker in meeting speech, Hayley Hung, *et al.*, [11, 12] only adopted the total speaking length of every speaker as the unique judgmental basis to define the importance of the speaker. This means that the speaker who has the longest speaking length is the most important, and the one who has the shortest is the least important, and so forth. By evaluating their speech data (communication speech of project members), a preferable performance is obtained. Cao Jie, *et al.*, [13]also collected the audio features like the features in references [6, 11] (total speaking length of every speaker, speech energy *etc.*,) for evaluating the key speaker, whose result shows that the total speaking length of the speaker is the most effective feature.

In conclusion, for evaluating the key speaker, the total speaking length of the speaker is the commonly adopted and known as the most effective feature. The employed classifiers include the complex statistical models (support vector machine, Bayesian network *etc.*,) and the simple threshold decision (for example the longest speaking length the most important). Although the total speaking length of the speaker is fairly effective in the speech data of the above references, it is not always the same in some types of meeting. For example, in the meeting which translators take part in (for example, in the meeting speech of the press conference of premier Wen Jiabao and other meeting speeches which need translators), they need to translate the words of the speakers into English or Chinese or other languages after every speaker making a statement. For this reason, the speaking length of the translator may highly possibly longer than that of the main speaker. For another example, in some meeting speeches, if the host communicates with many guests, its total speaking length may be longer than the special guest. Nevertheless, the translator or the host is not the most important among all the speakers. Therefore, that references [11, 12] only use the total speaking length as the sole feature to estimate the key speaker in this type of meetings is not that effective. Moreover, when using the video features and the complicated classifier, there are such problems as large calculating quantity, video features being difficult to be extracted, narrow scope of application *etc.*, In order to overcome the problems of the current method in estimation of the key speaker, this paper first defines the multi-feature of the speaker and analyzes their differences between the features of the key speaker and those of other speakers. Then, a decision function is generated by extracting four simple and effective audio features and the genetic algorithm is used to optimize these coefficients of feature weighting to obtain the best weighting coefficient. Without training the complicated classifier, this method is effective in estimating the key speaker.

2. Feature Difference of Speakers

Because every speaker plays a different role in the meeting, there exist obvious differences among the speaking features of different speakers. For this reason, to collect the features of speakers that can effectively represent the differences is the key in estimating the key speaker. This chapter analyzes the differences in terms of average speaking rate, total length of speaking time, the maximum length of single speaking time and total times of speaking, which lays the foundation for the following key speaker estimation. We analyzed the experimental data and give the

statistical distribution Figure about the above features of the key speaker and other speakers, as shown in Figure 1.

2.1. Average Speaking Rate

Average speaking rate refers to the ratio between the total words of the speaker said at the meeting and the total speaking length, with the unit number of words per second. In news conferences or interviews, the key speaker needs to impromptu answer problems from other people (reporter, host, and other people at the scene) or make impromptu speeches, while the person who asks questions will prepare the text of statement beforehand. Besides, considering the social effects of their speech, the key speaker thinks twice before opens their mouth or deliberately slows down speaking rate to avoid slip of the tongue. In conclusion, the average speaking rate of the key speaker is relatively slower, while the average rates of the other speakers are relatively faster. Figure 1 (a) shows the distribution of average speaking rate, from which we can see that the speaking rate of the key speaker is slower than those of other speakers.

2.2. The Total Length of Speaking Time

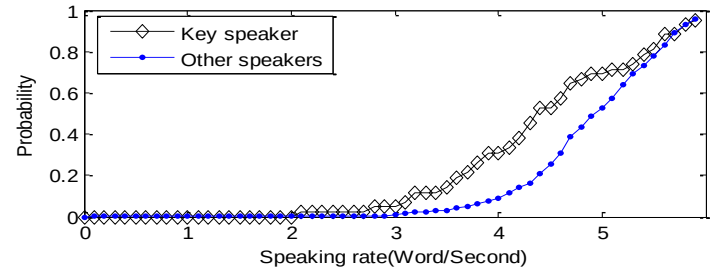
The total length of speaking time refers to overall time for the speaker making speeches (including the silence time during the speaking time), with the unit second. In the meeting speech, the key speaker has a greater right to speak. It can interrupt the speeches of others and need to answer all the other people's questions. Therefore, compared with others, its total length of speaking time is comparatively longer. Figure 1 (b) shows the distribution of total length of speaking time, from which we can see that the total length of speaking time of the key speaker is longer than those of other speakers.

2.3. The Maximum Length of Single Speaking Time

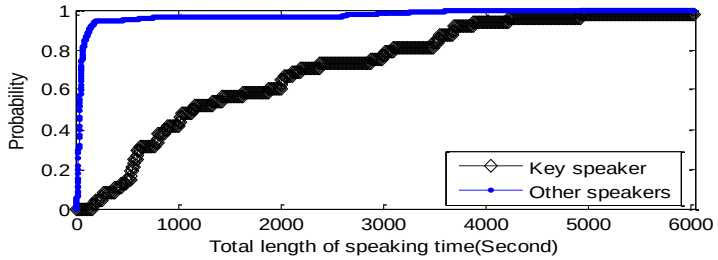
The maximum length of single speaking time refers to the maximum length of time the speaker speaks in a meeting (without being interrupted)(including the silence time during the speaking time), with the unit second. In the meeting speech, the key speaker can control the time and rhythm of one speech and interrupt the speeches of others. Hence, compared with others, its maximum length of single speaking time is comparatively longer. Figure 1 (c) shows the distribution of the maximum length of single speaking time, from which we can see that the maximum length of single speaking time of the key speaker is obviously longer than those of other speakers.

2.4. The Total Times of Speaking

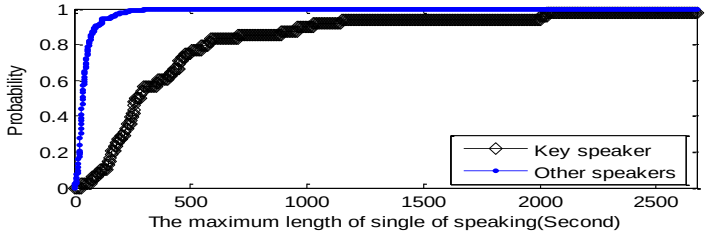
The total times of speaking refers to the total times the speaker speaks continuously in a meeting (without being interrupted). In the meeting speech, the key speaker is the focus among all the speakers. Other speakers will continually ask him/her questions or consult him/her. Consequently, compared with others, it generally has more chance to speak. Thus, its total times of speaking are comparatively more. Figure 1 (d) shows the distribution of the total times of speaking, from which we can see that the total times of speaking of the key speaker is obviously more than those of other speakers.



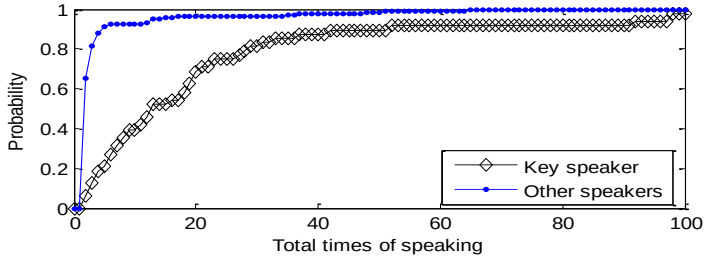
(a) The Distribution of Average Speaking Rate



(b) The Distribution of the Total Length of Speaking Time



(c) The Distribution of the Maximum Length of Single Speaking Time



(d) The Distribution of the Total Times of Speaking

Figure 1. The Statistical Distribution of Each Feature

3. The Methods for Estimating the Key Speaker

The method that combines the multiple features to estimate the key speaker can compensate for those that using only one feature. Moreover, the best feature combination can be obtained by optimizing the weighting coefficients of the multiple features, so as to realize the best estimation of the key speaker in the meeting. On account of the above consideration, this chapter focuses on estimating the key speaker in meeting speech based on multiple features optimization.

3.1. Key Speaker Estimation Method

Supposing there are N speakers S_n $(1 \le n \le N)$ in meeting speech file, the procedure of key speaker estimation method is as following:

Step 1: Speaker segmentation and clustering for meeting speech file to read, obtaining every speaker' speech. The methods of speaker segmentation and clustering are the same as in the references, which are not important of the paper.

Step 2: The following four features are extracted from every speaker's speech : average speaking speed SR_n , total length of speaking time SL_n , the maximum length of speaking time single time SS_n and total speaking times, SN_n ,where n is between 1 and N.

Step 3: The above-mentioned four features are normalized , SC_n （$1 \le n \le N$） stands for one feature of the n-th speaker, $SC_{\min}$ stands for the minimum value of SC_n （$1 \le n \le N$） , $SC_{\max}$ stands for the maximum value of SC_n （$1 \le n \le N$） , SC_n' stands for the feature of normalization, which is as following:

$$SC_n' = (SC_n - SC_{\min}) / (SC_{\max} - SC_{\min}) \quad (1)$$

Step 4: Weighted sum of the four normalization feature: normalization average speaking speed SR'_n , normalization total length of speaking time SL'_n , normalization the maximum length of speaking time single time SS'_n and normalization total speaking times SN'_n , obtaining the degree of importance DI_n of every speaker, which is as following:

$$DI_n = \alpha_R \times SR'_n + \alpha_L \times SL'_n + \alpha_S \times SS'_n + \alpha_N \times SN'_n, \quad 1 \le n \le N \quad (2)$$

where α_R 、 α_L 、 α_S and α_N are the feature weighting coefficients to be optimized, which is between 0 and 1 and the sum of the four coefficients is 1, how to optimize and choose the four weighting coefficients is important in section 2.2.

Step 5: Obtaining the minimum value from DI_n of N speakers, the speaker, whose DI_n is minimum, is the key speaker of the meeting speech file.

$$KS = \arg\min_{1 \le n \le N} [DI_n] \quad (3)$$

3.2. Optimized Method of Feature Weighting Coefficients

Genetic algorithm has been approved to obtain globally optimal solution of non-linear function and parameter estimation [14-15], genetic algorithm is used to optimize feature weight coefficients in our paper, whose specific step is as following:

1) Randomly generated initial population: supposing every population has twenty chromosomes. Real coding mechanism is used to code for every chromosomes, the code length of chromosomes is 4, the range of every gene is between 0 and 1 and the sum of every gene of every gene is 1. The schematic diagram of a chromosome is as Figure 2.

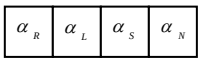


Figure 2. The Schematic Diagram of a Chromosomes

where every coefficient must satisfy：

$$\begin{cases} 0 < \alpha_R, \alpha_L, \alpha_S, \alpha_N < 1 \\ \alpha_R + \alpha_L + \alpha_S + \alpha_N = 1 \end{cases} \quad (4)$$

2) Determining the fitness function: make the accuracy of key speaker estimation be fitness value. The more accuracy, the more fitness.

$$f_i = \frac{N_c}{N_a} \quad (5)$$

where f_i stands for the fitness of the i-th chromosome, N_c stands for the right times of estimation, N_a stands for the total times of estimation

3) Selecting operation: the roulette wheel method is used, selecting probability of some chromosomes P_s is

$$P_s = \frac{f_i}{\sum_{i=1}^{N} f_i} \quad (6)$$

where f_i stands for the fitness, N stands for the chromosomes number in population.

4) Crossover operation: a random probability P_1 is generated by uniform random number generator, comparing P_1 with crossover probability P_c, when P_1 is bigger than P_c, crossover will not operate, otherwise, two chromosomes are randomly selected, then cross random selection of crossover position. Crossover method is used as crossover operator, crossover operator for the i-th chromosome and the j-th chromosome at the k-th position is as following:

$$\begin{cases} x_{ik}^{t+1} = x_{ik}^{t}(1-\phi) + x_{jk}^{t}\phi \\ x_{jk}^{t+1} = x_{jk}^{t}(1-\phi) + x_{ik}^{t}\phi \end{cases} \quad (7)$$

where x_{ik}^{t}、x_{ik}^{t+1} stand for the k-th position of the i-th chromosome before crosser and after crossover, x_{ik}^{t}、x_{ik}^{t+1} stand for the k-th position of the j-th chromosome before crosser and after crossover. ϕ stands for a random between 0 and 1,which is not crossover probability.

5) Mutation operation: a random probability P_2 is generated by uniform random number generator, comparing P_2 with crossover probability P_m, when P_2 is less than P_m, mutation will go on. Mutation operation is employed by non-uniform mutation, selecting the k gene x_{ik} of i-th chromosome to mutate, the operation method is as following:

$$f(g) = [r(1 - g/G_{\max})]^2 \quad (8)$$

$$x_{ik} = \begin{cases} x_{ik} + (x_{\max} - x_{ik}) * f(g) & r > 0.5 \\ x_{ik} + (x_{\min} - x_{ik}) * f(g) & r \le 0.5 \end{cases} \quad (9)$$

where $x_{\max}$ and $x_{\min}$ are upper bound and lower bound of gene x_{ik}, g is the current number of iterations, $G_{\max}$ is the maximum number of iterations.

6) A new round of iteration to the new generation population by repeating from 2) to 5), until the maximum number of iterations is reached or the fitness value will not change.

4. Experimental Result and Analysis

This chapter first introduces the status of experimental data, the experimental parameters setting, performance evaluation index, and then presents the experimental results and analysis.

4.1. Experiment Setting

The experimental data comes from different types of meeting speech, totally 54 hours, 152 voice recordings, including 3 data subsets: press conferences of Premier Li Keqiang, Premier Wen Jiabao and Premier Zhu Rongji, speeches, forum and interviews of President Obama, as shown in Table 1. The experimental data are annotated manually by six undergraduates and a postgraduate. The experimental data are annotated manually by six undergraduates and a postgraduate. As for the annotation of the key speaker in every meeting, the real key speaker must be the one that is annotated by the seven people. The manually annotated key speaker is used as the reference for the evaluation of the algorithm performance. The key speaker in this paper refers to the one who has the greatest right to speak and the most influence, such as in the data President Obama, Premier Li Keqiang, Premier Wen

Jiabao, Premier Zhu Rongji and some special guests in the interviews. Data format is converted to WAV files of 16kHz sampling frequency and 16bit quantization.

Table 1. The Introduction to Experimental Data

Types of meeting speech	Record number of the speeches	Key speaker	Language	Number of people	Role of the speaker
Premier answering reporters' questions	44	Premier	Chinese, English	8~20	guest, host, translator, askers
Speeches of the president	36	President	English	2~9	speaker, host, translator, askers
Forum, interview	72	Special guest	Chinese, English	3~10	guest, host, translator(for some meetings), askers

Table 2. The Training and Test Datas for Genetic Algorithm

Data application	Types of meeting speech	Record number of the speeches
Training	Premier answering reporters' questions	27
	Speeches of the president	22
	Forum, interview	44
Testing	Premier answering reporters' questions	17
	Speeches of the president	14
	Forum, interview	28

The speech of every speaker in every meeting was obtained through the segmentation and cluster system that we made in the earlier stage [16]; the suggested method in reference [17] is used to estimate the average speaking rate, calculate the total length of speaking time, find out the maximum length of single speaking time and count the total times of speaking of every speaker. The importance value of every speaker is gained according to formula (2) and the key speaker of the meeting is found due to formula (3).

The values of the weighting coefficients α_R、α_L、α_S and α_N are obtained through the genetic optimization algorithm, the crossover probability of which is 0.3, the mutation of which is 0.05, and the maximum number of iteration is 100. The training data and the testing data obtained by optimizing feature weighting coefficients through genetic algorithm are shown in Table 2, which are completely different. Estimation accuracy is used as the evaluation index for the performance of the algorithm: the ratio between the number of the correct estimation and the total number of estimation.

4.2. Experimental Result

Average speaking rate (the smaller rate the more important speaker), total length of speaking time (the longer length of speaking time the more important speaker), the maximum length of single speaking time (the bigger the more important speaker) and total times of speaking(the more the more important speaker) are used respectively to estimate the key speaker in the experimental data. This means that only one of the four

coefficients (α_R, α_L, α_S and α_N) will not be zero every time. The experimental result is shown in table 3, in which the total length of speaking time is corresponding to the result of the average accuracy 83.6%, a result obtained in reference [12].

Table 3. The Result of Estimating Key Speaker by using Single Feature

Feature	Types of meeting speech	Accuracy (%)
Average speaking rate	Answering reporters' questions	100
	Speech	50
	Forum, interview	76.7
	Mean value	75.6
Total length of speaking time	Answering reporters' questions	90.9
	Speech	100
	Forum, interview	60
	Mean value	83.6
The maximum length of single speaking	Answering reporters' questions	81.8
	Speech	100
	Forum, interview	66.7
	Mean value	82.8
Total times of speaking	Answering reporters' questions	0
	Speech	100
	Forum, interview	14.3
	Mean value	38.1

From Table 3, we can see that the average speaking rate is effective in estimating the key speaker in such meetings as press conference, forum, and interview, while it is unsatisfactory in speech. The reason may be that in speech the speaker has a relatively fast speaking rate because he/she will make good preparation beforehand to be familiar with the speech content and the questions form others. However, guest in press conferences, forums and interviews need to answer impromptu questions, because of which he/she has a relatively slow speaking rate. The total length of speaking time is effective in estimating the key speaker in such meetings as speech and press conference, especially speech, while a poor performance is got in forums and interviews. The reason could be that the total length of speaking time of the host is longer because there are relatively more guests in forums and interviews and the host needs to communicate with them; while in press conferences the translator may have a longer length of speaking time because he/she needs to translate the words of every speaker into English or Chinese. Even though their performances are close, the result obtained by using the maximum length of single speaking time is better than the one obtained by using total length of speaking time. Compared with other features, the performance obtained by using total times of speaking is the worst, but it is effective in estimating the key speaker in speeches. The reason could be the key speaker in speeches has the most times of speaking, while in press conferences, forums and interview, the speaking times of the host and translators are more than the key speaker. The above conclusion shows that single feature in ineffective

in estimating the key speaker in different types of meeting speeches. Single feature is effective in estimating the key speaker only in some certain kind of meeting but not in another kind of meeting.

This paper carries on a research to estimate the key speaker in the experimental data based on feature optimization and feature weighting coefficient, and the experimental result is shown in Table 4. This experiment gains an average accuracy of 93.3% by combing the four optimized features to estimate the key speaker. Compared with the result 83.6% obtained by using only the feature of total length of speaking time reported in reference [12], it is improved by 9.7%. Compared with result obtained with the method without optimization, it is improved by 4.1%. These show that this method achieves better performance. Table 5 further presents the results of estimating the key speaker in the other three types of meeting with this method. Better results are received and the three results are close. This shows that this method is effective in estimating the key speaker in various types of meeting speech. It is universal.

Table 4. The Result of Estimating Key Speaker by using the Suggested Method

Before or after optimization	$[\alpha_R, \alpha_L, \alpha_S, \alpha_N]$	Average accuracy (%)
Before optimization	[0.25, 0.25, 0.25, 0.25]	89.2
Before optimization	[0.415, 0.336, 0.109, 0.14]	93.3

Table 5. The Result of Estimating Key Speaker in Various Meetings by using the Proposed Method

Types of meeting speech	Accuracy (%)
Answering reporters' questions	94.1
Speech	92.9
Forum, interview	92.9
Mean value	93.3

There is a miscalculation when estimating the key speaker in the press conference of the premier, which is caused by the reason that the total length of speaking time, the maximum length of single speaking and the total times of speaking of the translator may exceed those of the key speaker. When estimating the key speaker in speeches, the miscalculation is mainly caused by the average speaking rate. Because the key speaker sometimes has high speaking rate in the speech, while the people who ask questions have very low speaking rate, which leads to erroneous judgment that the one who speak fast is considered as the unessential speaker. When estimating the key speaker in the forum or the interview, the miscalculation is mainly caused by the total times of speaking. Because the host sometimes has more times of speaking than the key guest, which causes the miscalculation that the host who has the most times of speaking is mistaken as the key speaker.

5. Conclusion

Based on the segmentation and cluster of the speaker, this paper collects four speaking feature coefficients and optimizes the feature weighting coefficients, so as to generate the best decision function of feature weighting for estimating the key speaker. Without training complicated classifier, this method effectively estimates the key speaker. Compared with the mainstream methods reported in reference [12], this experiment gains an average accuracy improvement by 9.7%; compared with feature weighting method without optimization, it gains an average accuracy improvement by 4.1%. With a better estimation result for every type of meeting

speech, this method is proved to be universal. The effective estimation of the key speaker lays a foundation for searching follow-up speakers and fast browsing meeting speeches.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61101160), the Fundamental Research Funds for the Central Universities, South China University of Technology, China (Item No. 2013ZZ0053), Project of the Pearl River Young Talents of Science and Technology in Guangzhou, China (No. 2013J2200070), and the Foundation of China Scholarship Council.

References

[1] T. Hain, L. Burget, J. Dines, *et al.*, "Transcribing meetings with the AMIDA systems", IEEE Transaction on Audio, Speech, and Language Processing, vol. 20, no. 2, **(2012)**, pp. 486-498.
[2] Y.-X. Li, Q.-H. He, S. Kwong, *et al.*, "Characteristics-based effective applause detection for meeting speech", Signal Processing, vol. 89, no. 8, **(2009)**, pp. 1625-1633.
[3] L. Yan-xiong, W. Yong and H. Qian-Hua, "Feature mean distance based speaker clustering for short speech segments", Journal of Electronics & Information Technology, vol. 34, no. 6, **(2012)**, pp. 1404-1407.
[4] S. H. Shum, N. Dehak, R. Dehak, *et al.*, "Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach", IEEE Transactions on Audio, Speech, and Language Processing, Vancouver, vol. 21, no. 10, **(2013)**, pp. 2015-2028.
[5] T.-H. Wen, H.-Y. Lee, P.-H. Su, *et al.*, "Interactive spoken content retrieval by extended query model and continuous state space markov decision processing", International Conference on Acoustics, Speech, and Signal Processing, Vancouver: IEEE Press, **(2013)**, pp. 8510-8514.
[6] R. Rienks and D. Heylen, "Dominance detection in meetings using easily obtainable features", International Workshop on Machine Learning for Multimodal Interaction. Edinburgh: Springer Berlin Heidelberg, **(2005)**, pp. 76-86.
[7] R. Rienks, D. Zhang, D. GaticaPerez, *et al.*, "Detection and application of influence rankings in small group meetings", The 8th International Conference on Multimodal Interfaces, New York: ACM, **(2006)**, pp. 257-264.
[8] H. Hung, D. Jayagopi, C. Yeo, *et al.*, "Using audio and video features to classify the most dominant person in a group meeting", The 15th International Conference on Multimedia, New York: ACM, **(2007)**,pp. 835-838.
[9] D. B. Jayagopi, H. Hung, C. Yeo, *et al.*, "Modeling dominance in group conversations using nonverbal activity cues", IEEE Trans on Audio, Speech and Language Processing, vol. 17, no. 3, **(2009)**,pp. 501-513.
[10] Y. Nakano and Y. Fukuhara, "Estimating conversational dominance in multiparty interaction", ACM International Conference on Multimodal Interaction, New York: ACM, **(2012)**, pp. 77-84.
[11] H. Hung, Y. Huang, G. Friedland, *et al.*, "Estimating the dominant person in multi-party conversations using speaker diarization strategies", IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas: IEEE Press, **(2008)**, pp. 2197 - 2200.
[12] H. Hung, Y. Huang, G. Friedland, *et al.*, "Estimating dominance in multi-party meetings using speaker diarization", IEEE Trans. on Audio, Speech and Language Processing, vol. 19, no. 4, **(2011)**, pp. 847-860.
[13] C. Jie and P. Peng, "Recognize the most dominant person in multi-party meetings using nontraditional features", IEEE International Conference on Intelligent Computing and Intelligent Systems, Xiamen: IEEE Press, **(2010)**, pp. 312-316.
[14] C. Liaw, W. H. Wang, C. T. Tsai, *et al.*, "Evolving a team in a first-person shooter game by using a genetic algorithm", Applied Artificial Intelligence, vol. 27, no. 3, **(2013)**, pp. 199-212.
[15] R. Köker, "A genetic algorithm approach to a neural-network-based inverse kinematics solution of robotic manipulators based on error minimization", Information Sciences, vol. 222, **(2013)**, pp. 528-543.
[16] C. Fen, "Research on unsupervised speaker clustering and its implementation", Guangzhou: School of Electronic and Information Engineering, South China University of Technology, **(2012)**.
[17] L. Yan-Xiong, X. Xin and H. Qian-Hua, "Estimating speaking rate for multi-speaker based on speaker segmentation and clustering", China, ZL201110403577[P].2013-07-24.

