Speaker Diarization System for Autism Children's Real-Life Audio Data

Tianyan Zhou[1], *Weicheng Cai*[1], *Xiaoyan Chen*[3], *Xiaobing Zou*[3], *Shilei Zhang*[4], *Ming Li*[1,2]

[1]SYSU-CMU Joint Institute of Engineering, Sun Yat-sen University, China
[2]SYSU-CMU Shunde International Joint Research Institute, China
[3]The Third Affiliated Hospital, Sun Yat-sen University, China
[4]Speech Technology and Solution Group, IBM China Research Laboratory
liming46@mail.sysu.edu.cn

Abstract

In this paper, we introduce several methods to improve the performance of speaker diarization system for autism children's real-life audio data. This system serves as the frontend module for further speech analysis. Our objective is to detect the children's speech from single channel, noisy and daily audio recordings collected by wearable devices in real environment. First, in the conventional generalized likelihood ratio (GLR) distance with agglomerative hierarchical clustering (AHC) framework, besides using the line spectral pair (LSP) based GLR distance, we further propose a weighted summation of multiple GLR distances combining LSP, pitch, energy and phoneme duration information together. Second, since we only want to extract children's speech in high purity for further speech analysis, we utilize a 30 seconds long enrollment utterance from each child to perform supervised child cluster selection using i-vector cosine distance. We find out that performing supervised cluster selection at AHC early stages generates higher purity. We evaluate our methods on a 120 minutes subset data collected from three children during the child-therapist interactions. Experimental results show that our methods significantly outperform the GLR-AHC baseline in terms of child cluster's recall and precision.

Index Terms: speaker diarization, speaker clustering, autism.

1. Introduction

Autism spectrum disorder (ASD) is neurodevelopmental disorder that causes impaired social interaction [1]. Recent studies report that 1 out of 110 children has ASD [2,3]. However, the diagnosis of autism mainly relies on the experience and judgement of the clinicians, which tends to be subjective and makes it difficult to form a precise diagnosis for children under 1.5 years old. Early speech characteristics, such as atypical prosody can be one of the obvious and helpful clues.

We intend to focus on the acoustic patterns of children with autism in real life environment and quantify some objective and effective indicators for clinicians. By collecting the autism children's audio data under a daily and natural circumstance, we can get plenty of valuable data for further analysis. The first step of this analysis task is to perform speaker diarization. The objective of speaker diarization is to determine "who spoke when?" [4]. In other words, given an audio stream, the speaker diarization system needs to separate the speech data into several clusters, each cluster only contains one speaker. In our case, we need to detect children's voices from noisy, single channel recordings.

Speaker diarization has been used in many applications, such as broadcasting, telephone conversations, debates, meeting environments, etc. [5,6]. It is of great benefit for further analysis when the system can automatically detect the number of speakers as well as the starting and stopping timestamps for each of them. Several existing methods adopt cepstral features to calculate the distances between different speech segments using Bayesian information criterion (BIC) [7] and perform hierarchical clustering on top of these segments [8,9]. Recently, the i-vector representation followed by probabilistic linear discriminant analysis (PLDA) modeling has been shown to be successful in speaker diarization [8,10,11]. Since i-vector extraction requires at least several seconds' speech data, an initial BIC clustering stage is proposed as the frontend [11].

Comparing autism children with adults, we believe that children have higher pitch and shorter phoneme duration. Also since the audio recorder is worn by the children, their speech volume or energy may also contain discriminative information. Therefore, on top of the LSP feature, we introduce pitch, energy and phoneme duration these children informative features to the GLR distance calculation.

In addition, since the goal in our application is to accurately detect children's speech, we combine both speaker diarization and speaker verification together in the child cluster selection stage. We utilize a 30 seconds long utterance from each child to extract their enrollment i-vectors. Then after several iterations of GLR-AHC clustering, we perform supervised child cluster selection using i-vector cosine distance. The UBM and i-vector models are trained in an unsupervised manner from an extra 10 children's speech data recorded by the same wearable audio recorder [10]. PLDA is not adopted for i-vector modeling due to the lack of labeled speakers in our training data.

The rest of this article is organized as follows. Section 2 presents the methods of our speaker diarization system. Experiment setup and results are provided in Section 3. Section 4 summarizes the paper and discusses our future work.

2. Methods

Figure 1 shows the architecture of our speaker diarization system. It consists of voice activity detection (VAD), speaker segmentation, speaker clustering and cluster selection these four stages.

2.1. Voice activity detection

At the beginning, we need to reduce the number of non-speech segments using voice activity detection (VAD) [12]. In our

978-1-5090-4294-4/16/$31.00

speaker diarization system, we simply use an energy-based VAD method due to its efficiency.

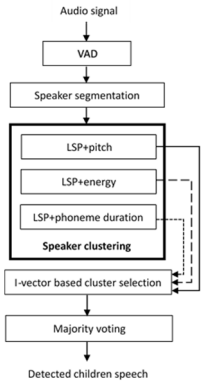


Figure 1: *Overview of our speaker diarization system.*

2.2. Speaker segmentation

After VAD procedure, the audio data is divided into multiple separate speech segments. Speaker segmentation aims at detecting speaker change points within a continuous speech segment and splitting the segment into several fragments according to recognized change points.

There are various features reported in the literature, including Mel frequency cepstral coefficient (MFCC), line spectral pair (LSP) [13], perceptual linear prediction (PLP) cepstral coefficients [14] and combinations of different features. In our system, we simply adopt the LSP feature.

Distance measure is considered as the core issue in the speaker segmentation and clustering field, the most common choices are Bayesian information criterion (BIC) [15] and generalized likelihood ratio (GLR) [16]. Both of them are hypothesis testing methods based on a simple merging or not assumption [17].

We employ the BIC metric in the segmentation stage and adopt the revised GLR distance in the clustering stage. In our system, considering the frequent speaker change points in our collected audio data, we reduce the window size when detecting speaker change points to ensure the fragment purity. By doing this, our final speech fragments achieve an average length of 1 second, in which we expect only one speaker appears.

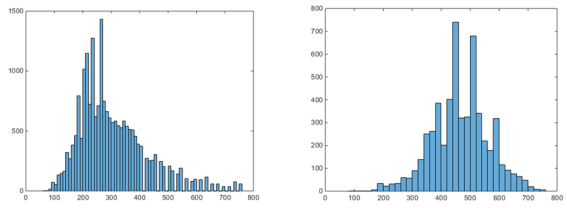
2.3. Speaker clustering

Hierarchical clustering aims at grouping the fragments from the same speaker. It contains two categories according to the initial number of clusters. The first one is agglomerative hierarchical clustering (AHC), and the second one is divisive hierarchical clustering (DHC). A study of analyzing the relative merits of the mentioned strategies can be found in [18]. In our system, we use the AHC approach to perform clustering.

2.3.1. Feature introduced

Pitch is the fundamental frequency which exhibits speaker variance. Normally, females have higher pitch than males and children have higher pitch than adults. As our system is designed for capturing children's audio data, pitch may be an effective feature. In order to prove this, we manually labeled a 16 minutes long child-therapist interaction data with child and non-child tags and calculated the pitch distribution separately. Two fundamental frequency histograms are showed in Figure 2. Only one child but several adults appear in the audio. As it can be observed from Figure 2, the child has a relatively higher pitch than adults, which supports pitch to become a useful feature.

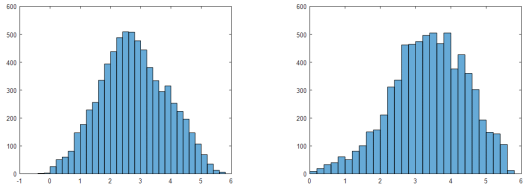
At the clustering stage, we extract the pitch frame by frame. Every small fragment containing frames with non-zero pitch values can be represented by a one dimensional pitch vector. These vectors will be represented by a Gaussian model and used in the mixed GLR distances calculation.



(a) (b)

Figure 2: (a) non-child F0-histogram; (b) child F0-histogram.

The second proposed feature is the energy. In this case, considering the fact that our wearable recording devices are carried by children, the speech from children may have higher energy than adults because they are closer to the microphone. On the other hand, according to our observation, many children with autism are not willing to speak loudly. In the majority situation, they tend to repeat the adults' speech softly or just soliloquize. Therefore, the speech of children have lower energy than the adults. Either the energy of children's speech is higher or lower than adults, they are different and can be extracted as a feature. Figure 3 shows the histogram of energy in logarithm from a recording with one child and several adults. The testing audio is the same as the previous one. From Figure 3, we can see that the energy in the speech of child is slightly higher than adults'.

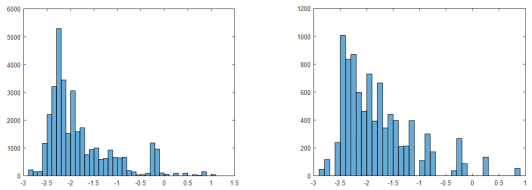


(a) (b)

Figure 3: (a) non-child energy-histogram; (b) child energy-histogram.

At the clustering stage, we also extract energy frame by frame and a similar one-dimensional vector is generated for every small fragment. These energy vectors will also be included as a part of the distance measure.

Phoneme duration is another new feature proposed in our system. Children with autism are reported to exhibit impaired language abilities. Specifically, many of them cannot speak a complete sentence but only several short phrases. Compared with other individuals that appeared in the audio, they tend to have a lower speaking speed, i.e., longer phoneme duration on a certain phoneme. Accordingly, we perform phoneme decoding on the audio stream and calculate the duration for each phoneme spoken. Within a small fragment, a series of phoneme duration values can be calculated as its feature vector and applied to the distance measure. Figure 4 shows the phoneme duration distribution of children's speech and adults' speech. High proportion of long phoneme duration reflects lower speaking speed, which may be also effective for clustering.



(a) (b)

Figure 4: (a) non-child phoneme duration-histogram; (b) child phoneme duration-histogram.

2.3.2. Revised distance measure

Since we have introduced several new features to our clustering stage, we revise the classical GLR metric to a weighted summation version. It can be written as follows.

$$distance = (1 - w) * GLR_{LSP}(C_x, C_y) + w * GLR_{others}(C_x, C_y) \quad (1)$$

where C_x, C_y stands for two candidate clusters and w is a parameter to adjust the weight of the two GLR metrics. The first part, $GLR_{LSP}(C_x, C_y)$, is a GLR metric based on the conventional LSP feature and the second part,$GLR_{others}(C_x, C_y)$, is a GLR metric based on new features.

2.4. Cluster selection

After the clustering stage, we have merged thousands of speech fragments into a limited number of clusters related to different speakers. The final task is to pick out target cluster.

2.4.1. Model training

We intend to use the state of the art i-vector technique to select the child cluster. Consequently, we need to train an UBM model and an i-vector total variability matrix model using our collected autism children' audio data in order to reduce the effect of channel variability. Here, we use an extra 10 children's data for the UBM and i-vector model training. For every child, we have collected 10 to 15 days audio data, each of them remains a length of around 500 minutes after passing through VAD. Right now, we only use the audio data with the child-therapist interactions (two sections per day) for every child, according to a schedule provided by the hospital. The reason we choose the child-therapist interactions is that they contain the most children's speech during the whole day activity and with relatively less ambient noise since the conversation participants are usually 3 to 5. The duration of each interaction session is about 20 minutes.

However, our data contains frequent speaker changes, large overlapping regions and strong environmental noises, which prevents us to simply divide the whole segment equally with a fixed duration for segmentation. As a result, for every child-therapist interaction section, we process it through our VAD stage, speaker segmentation stage and AHC stage. Specifically, at the AHC stage, we set an appropriate times of agglomeration to make an average segment duration of 5 seconds, while the maximum duration and minimum duration range from 12 seconds to 0.5 second. We discard those segments with duration less than 2 seconds. After aforementioned operations, we use all selected segments to train an UBM model and an i-vector total variability matrix for the target cluster selection.

We manually selected and labeled a total duration of 30 seconds speech (from UBM and i-vector training data, no overlapping with our testing data) for each child's enrollment. At the cluster selection stage, we calculate the i-vector cosine similarity score between each cluster and the target child's enrollment, then select target clusters against a pre-defined score threshold.

2.4.2. Majority voting

In order to utilize the complementarity among all the features, we employ a majority voting [19], [20] strategy to select the target child cluster. Specifically, we apply the three features combining with the LSP feature to the clustering stage separately, i.e., pitch plus LSP, energy plus LSP and phoneme duration plus LSP. Each of them will pick out a target cluster, thus we get three candidate clusters. Then we choose those small fragments identified by majority of the system. This approach may decrease the precision but can increase the purity (recall) of the kid cluster, which is more critical for our further analysis.

2.4.3. I-vector based cluster selection

We have examined two strategies to combine the AHC clustering with the i-vector based cluster selection. The first one is to use i-vector based cluster selection when the AHC algorithm stops. As a result, every cluster is supposed to represent one speaker, we need to choose one of them as our final target speaker cluster. The second one is introducing cluster selection at the early stages of AHC. At first, we intend to use i-vector instead of LSP as a feature at the AHC module, but the segment length after speaker segmentation is too short to extract a reliable i-vector. Consequently, we still utilize LSP based AHC to cluster these segments, but employ an early stop strategy. After several rounds of merging, we calculate i-vectors for every segments and compare with the enrollment directly using i-vector cosine similarity. We can further set a threshold to balance the resulted target cluster's precision and purity (recall).

Our evaluation results show that the second strategy significantly increases cluster purity with a tolerable loss of target segment recall.

3. Experimental results

We evaluate the performance of our system on the 120 minutes subset data collected from three children during the child-therapist interactions. The experiments is conducted using both the conventional GLR-AHC system and our system.

3.1. Description of data

Our audio database is collected from children who have been diagnosed as autism and stayed in hospital for a one month rehabilitation treatment. An audio recording wrist-band is worn by each child at the daytime to record speech data in real environment. We have gathered 10 to 15 days audio data for each child depending on their hospital stay period. The data collection is approved by the children' family, the hospital and the university. There are totally 40 children in our database.

We randomly select 6 audio files with a total length of 120 minutes as our diarization evaluation data. They are selected from 3 children, two boys (K1, K3) and one girl (K2). These three children are mutual exclusive with those 10 children appearing in the UBM and i-vector modeling training data. The manually selected 30 seconds long enrollment data for each of

these three children are from their recordings at other days, no overlap with the evaluation data.

The girl is 4 years old, while the boys are 7 and 12 years old. The daily activity of the children contains music class, exercise class, playing games and so on. We select the segments during the child-therapist interactions as it contains the majority of children's speech. Different from other structured scenario, our audio data includes much ambient noise such as sound of toys or non-verbal noise such as laughing, crying, and breathing. Therefore, it is indeed a difficult task to detect autism children's speech from this kind of single channel audio data.

3.2. Performance evaluation

3.2.1. Introducing new features

We first evaluate the proposed three features one by one in the linear weighted summation fusion with LSP feature for GLR distance calculation. When the weight equals to 0, only LSP feature is applied. As you can observe from Table 1 to Table 3, the proposed multiple feature based weighted GLR distance calculation outperforms the single LSP feature baseline dramatically in terms of both precision and recall. We use two measurements to evaluate our system, namely precision and recall, defined as follows,

$$precision = \frac{d_{child\ speech}}{d_{child\ annotation}} \quad (2)$$

$$recall = \frac{d_{child\ speech}}{d_{target\ cluster}} \quad (3)$$

where $d_{child\ speech}$ stands for the duration of target child's speech, $d_{child\ annotation}$ stands for the duration of child's speech refer to annotation, and $d_{target\ cluster}$ stands for the duration of final target cluster. Generally speaking, precision is a ratio of child's speech in the final cluster to the child's entire speech, and recall is also a ratio of child's speech in the final cluster to the entire length of the final cluster.

Table 1. Clustering results with different weights of pitch.

w	0	0.2	0.4	0.6	0.8	1
Precision	0.494	**0.546**	**0.624**	0.605	0.655	0.662
Recall	0.564	**0.660**	**0.611**	0.540	0.520	0.416

Table 2. Clustering results with different weights of energy.

w	0	0.2	0.4	0.6	0.8	1
Precision	0.494	**0.576**	**0.548**	0.298	0.370	0.284
Recall	0.564	**0.652**	**0.691**	0.717	0.531	0.405

Table 3. Clustering results with different weights of phoneme duration.

w	0	0.2	0.4	0.6	0.8	1
Precision	0.494	**0.578**	**0.605**	0.519	0.615	0.689
Recall	0.564	**0.632**	**0.607**	0.659	0.522	0.204

3.2.2. Early stop

Considering the frequent speaker changes in our collected audio data, we reduce the detect window size at the segmentation stage to ensure the segment purity. By doing this, our final speech fragments have an average duration of around 1 second, in which we expect only one speaker appears. However, these fragments are too short to extract i-vectors directly. We intend to perform AHC for several times in order to obtain longer duration fragments, then use the i-vector feature and cosine distance measure to select the high similarity fragments to form the final target cluster. We define a merge ratio, M, as follows,

$$Agglomeration\ iterations = floor(M * N_{cluster}) \quad (4)$$

where $N_{cluster}$ is the number of fragments after speaker segmentation. We evaluate the system performance by adjusting M and cosine similarity threshold. As shown in figure 6, we can discover following phenomena.

First, when the cosine similarity threshold increases, the purity is also increases, but the accuracy decreases, which reflects that the target cluster size becomes smaller and target speaker's purity increases. Second, big merge ratio generates high accuracy, while small merge ratio brings high purity. How to balance the two indices and get better performance is basically depends on each specific application.

Figure 6 shows the performance of our system compared to the GLR-AHC baseline system on the recall measurement. We can see that our system significantly improved the recall especially for older children (audio sequence 1, 2, 5, 6). We also test the early stop strategy with the fusion of pitch, energy and phoneme duration features, but system performance is similar with LSP single feature system. Maybe those new features are not reliable when speech segments are quite short.

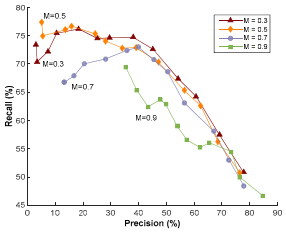


Figure 5: System performance by tuning distance threshold.

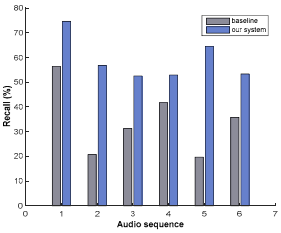


Figure 6: Our system versus GLR+AHC baseline system.

4. Conclusions and future work

We propose a speaker diarization system designed for autism children' real-life audio data. At the clustering stage, we introduced pitch, energy and phoneme duration features to better distinguish the speech between children and adults. With a few seconds of the children' speech, we further use the speaker verification technique to pick out the target child cluster. Future work includes automatically estimating the contribution weight of each feature in the GLR distance calculation and more effectively using the child voice's prior knowledge.

5. Acknowledgement

This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), the Fundamental Research Funds for the Central Universities (15lgjc10) and National Key Research and Development Program (2016YFC0103905).

6. References

[1] *Diagnostic and Statistical Manual of Mental Disorder, Ed. 4 text revision*, American Psychiatric Assoc., Washington D.C., 2000.

[2] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous speech acoustic-prosodic features of children with autism and the interacting psychologist," in *Proceedings of INTERSPEECH*, 2012.

[3] National Center on Birth Defects and Developmental Disabilities, (2010, March) Autism Spectrum Disorders (ASDs). [online]. http://www.cdc.gov/ncbddd/autism/.

[4] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of the IEEE International Conference on Audio and Speech Signal Processing*, 2005.

[5] X. Anguera, et al. "Speaker Diarization: A Review of Recent Research." *IEEE Transactions on Audio Speech & Language Processing* 20.2(2010):356 - 370.

[6] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio Speech & Language Processing*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[7] Y. Huang, O. Vinyals, G. Friedland, C. Muller, N. Mirghafori and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization", in *Proceedings of IEEE Workshop Autom. Speech Recognition Understanding*, pp. 693-698.

[8] A. E. Bulut, H. Demir, Y. Z. Işik and H. Erdogan, "PLDA-based diarization of telephone conversations," in *Proceedings of ICASSP, 2015.*

[9] K.J. Han, K. Samuel and S. Narayanan. "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization." *IEEE Transactions on Audio, Speech, and Language Processing,* vol 16, issue 8, pp. 1590-1601.

[10] S. Madikeri, I. Himawan, P. Motlicek and M. Ferras, "Integrating Online I-vector extractor with Information Bottleneck based Speaker Diarization system," in *Proceedings of INTERSPEECH*, 2015.

[11] B. Desplanques, K. Demuynck, J. Martens, "Factor Analysis for Speaker Segmentation and Improved Speaker Diarization," in *Proceedings of INTERSPEECH*, 2015.

[12] J. Ramirez, J. M. Girriz, and J. C. Segura, M. Grimm and K. Kroschel, Eds., "Voice activity detection. Fundamentals and speech recognition system robustness," in *Proceedings of Robust Speech Recognit. Understand.*, Vienna, Austria, Jun. 2007, p. 460.

[13] N. P. Jawarkar, R. S. Holambe and T. K. Basu, "Audio Segmentation using Line Spectral Pairs," in *Proceedings of National Conference NCIPET*, pp.1–5, Dec. 2011.

[14] S. E. Tranter, K. Yu, G. Evermann, P.C. Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, pp. 433–477, May 2004.

[15] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, Feb. 1998, pp. 127–132.

[16] P. Delacourt and C. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication.*, pp. 111–126, 2000.

[17] K. J. Han, S. Kim, and S. S. Narayanan, "Robust speaker clustering strategies to data source variation for improved speaker diarization," in *Proceedings of IEEE Autom. Speech Recognition and Understanding Workshop*, ec. 2007, pp. 262–267.

[18] E. Nicholas, B. Simon, W. Dong, F. Corinane, and T. Raphael, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol.20, no. 2, pp. 382–392, Feb. 2012.

[19] H. Hung and G. Friedland, "Towards audio-visual on-line diarization of participants in group meetings," in *Proceedings of Workshop Multi-Camera and Multi-Modal Sensor Fusion Algorithms Applicat. –M2SFA2*, Marseille, France, 2008.

[20] G. Friedland and O. Vinyals, "Live speaker identification in conversations," in *Proceedings of 16th ACM Multimedia*, New York, 2008, pp. 1017–1018.