Joint Institute of Engineering

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

Authors: Tianyan Zhou, Weicheng Cai, Ming Li
Speaker: Weicheng Cai

SUN YAT-SEN UNIVERSITY | Carnegie Mellon University

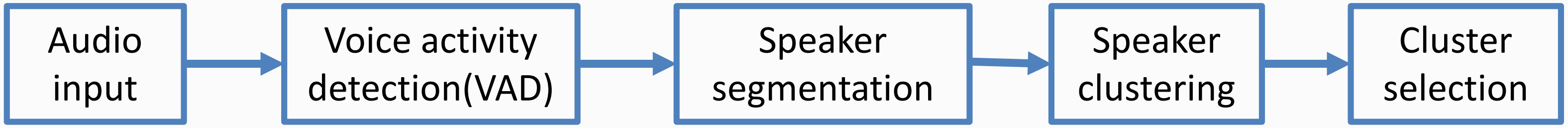
Carnegie Mellon University

Introduction

- Speaker diarization system

Speaker diarization has been used in many applications. The objective of speaker diarization is to determine “who spoke when?”. 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.

- General process



Audio input → Voice activity detection(VAD) → Speaker segmentation → Speaker clustering → Cluster selection

Introduction

- Autism spectrum disorder (ASD)

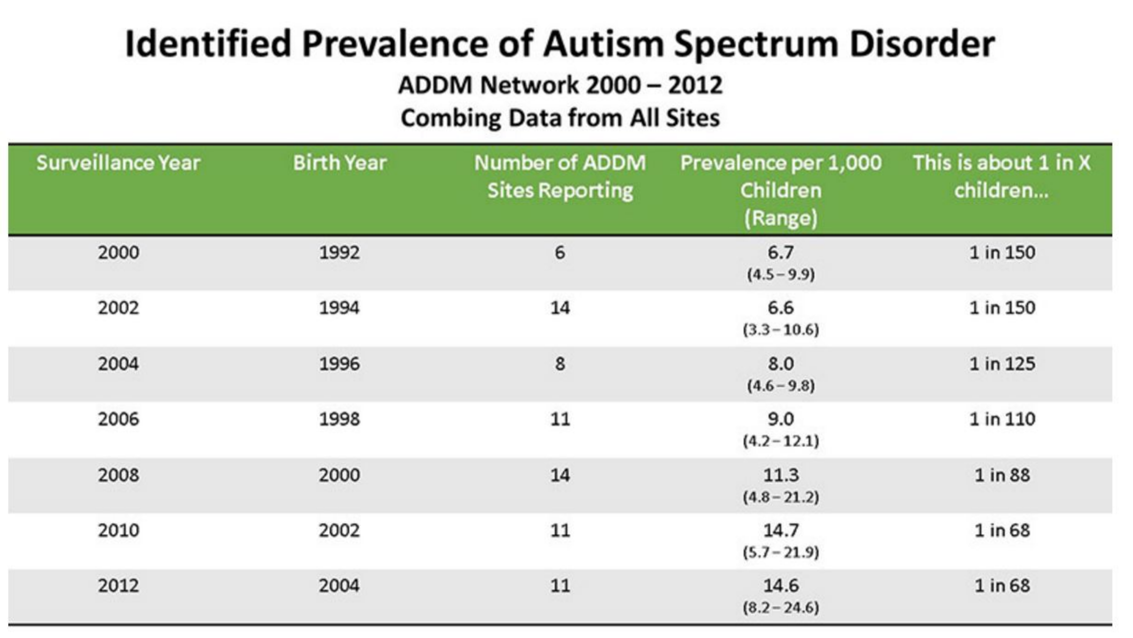
ASD is the name for a group of developmental disorders. ASD includes a wide range, “a spectrum,” of symptoms, skills, and levels of disability[1].

- Children with ASD often have these characteristics[1]:

→ Ongoing social problems including difficulty communicating and interacting with others.
→ Repetitive behaviors as well as limited interests or activities.
→ Symptoms that typically are recognized in the first two years of life.
→ Symptoms that hurt the individual’s ability to function socially,
at school or work, or other areas of life.

[1]. National Institute of Mental Health (https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml).

Introduction



Identified Prevalence of Autism Spectrum Disorder

ADDM Network 2000 – 2012

Combing Data from All Sites

Surveillance Year	Birth Year	Number of ADDM Sites Reporting	Prevalence per 1,000 Children (Range)	This is about 1 in X children...
2000	1992	6	6.7 (4.5 – 9.9)	1 in 150
2002	1994	14	6.6 (3.3 – 10.6)	1 in 150
2004	1996	8	8.0 (4.6 – 9.8)	1 in 125
2006	1998	11	9.0 (4.2 – 12.1)	1 in 110
2008	2000	14	11.3 (4.8 – 21.2)	1 in 88
2010	2002	11	14.7 (5.7 – 21.9)	1 in 68
2012	2004	11	14.6 (8.2 – 24.6)	1 in 68

[2]. Center for Disease Control and Prevention (http://www.cdc.gov/features/dsautismdata/).

- Research background

About 1 in 68 children (or 14.7 per 1,000 eight year olds) were identified with ASD[2].

This new estimate is roughly 30% higher than the estimate for 2008 (1 in 88), roughly 60% higher than the estimate for 2006 (1 in 110), and roughly 120% higher than the estimates for 2002 and 2000 (1 in 150)[2].

Introduction

Introduction

- Research background

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.

- Research objective

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.

Methods

- Voice activity detection

  We simply use an energy-based VAD method due to its efficiency.

- Speaker segmentation

  We employ the LSP feature and BIC metric to detect speaker changes.

- Speaker clustering

  → Introducing discriminative features (pitch, energy, phoneme duration).
  → Revised distance measure.
  → Agglomerative hierarchical clustering (AHC) plus early stop.

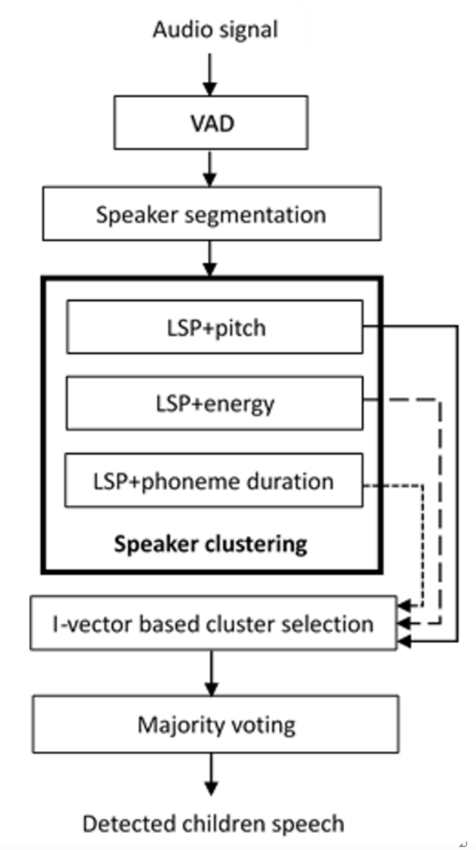


Figure 1: System overview..

Methods

- Discriminative features: pitch

Pitch is the fundamental frequency which exhibits speaker variance. Normally, children have higher pitch than adults

We manually labeled a 16 minutes long child-therapist interaction data with child and non-child tags and calculated the pitch distribution separately.

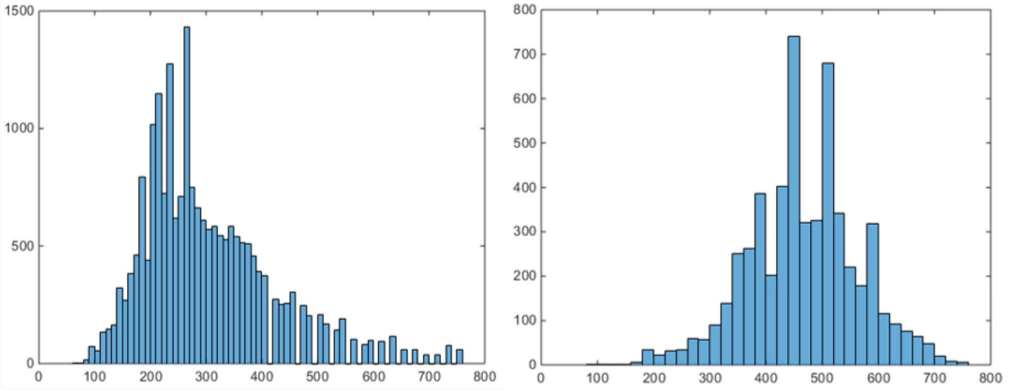


Figure 2: (a) non-child F0-histogram; (b) child F0-histogram.

Methods

- Discriminative features: energy

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.

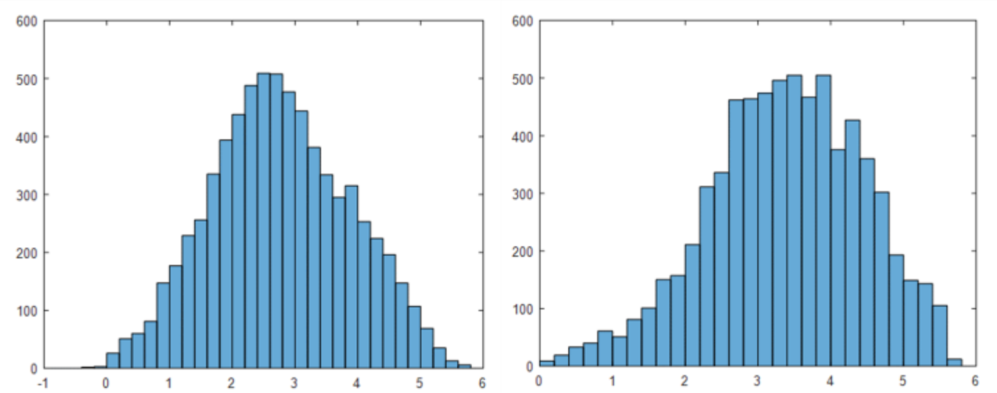


Figure 3: (a) non-child energy-histogram; (b) child energy-histogram.

Methods

- Discriminative features: phoneme duration

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.

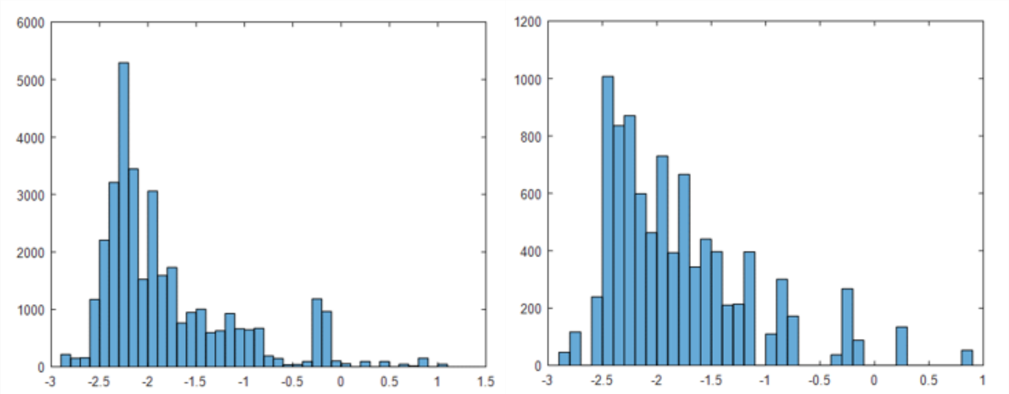


Figure 4: (a) non-child phoneme duration-histogram; (b) child phoneme duration-histogram.

Methods

- Revised distance measure

We revise the classical GLR metric to a weighted summation version.

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

- AHC plus early stop

→ Agglomeration iterations determine the number and size of clusters.
→ Adjusting the agglomeration iterations by changing the merge ratio, M.

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

Methods

- Cluster selection

Picking out target clusters from merged alternative clusters.

Model training: 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.

I-vector based cluster selection: based on 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.

Experimental results

- 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 randomly select 6 audio segments during the child-therapist interactions 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).

Experimental results

- 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.

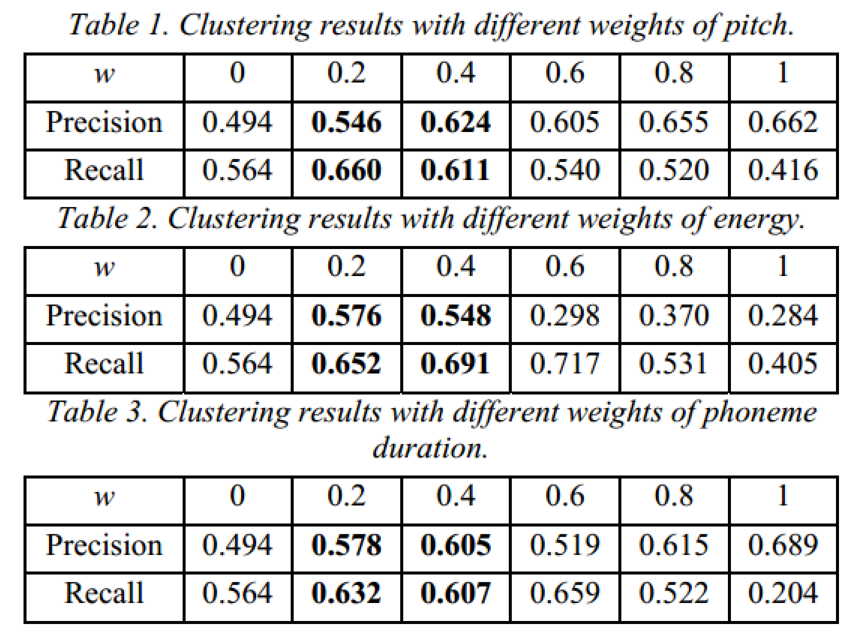


Table 1. Clustering results with different weights of pitch.

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

Experimental results

- System performance

We evaluate the system performance by adjusting M(merge ratio) and cosine similarity threshold. As shown in figure 5.

Figure 6 shows the performance of our system compared to the GLR-AHC baseline system on the recall measurement.

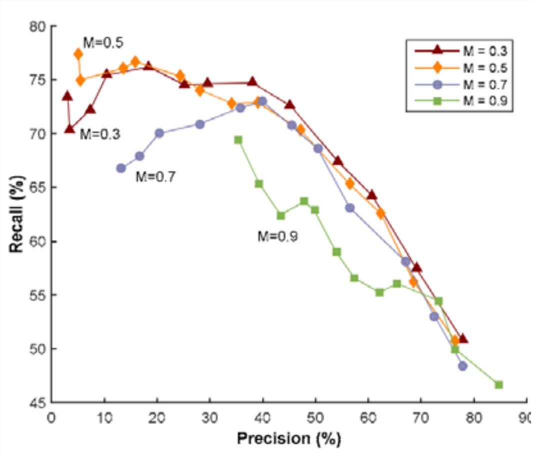


Figure 5: System performance by tuning distance threshold.

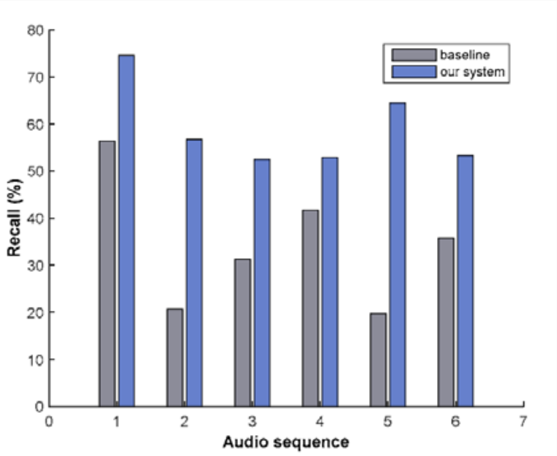


Figure 6: Our system versus GLR+AHC baseline system.

Future work

- System improvement

Automatically estimating the contribution weight of each feature in the GLR distance calculation and more effectively using the child voice’s prior knowledge.

- Follow-up work

After extracting children's speech from the audio data, we intend to focus on the acoustic patterns and quantify some objective and effective indicators for clinicians.

Performing a correlation analysis over our indicators and results of Autism Diagnostic Observation Scale (ADOS) scores and those with high correlation values can be used as items of quantitative diagnosis.

Thank you!