

Analysis of Speech Properties of Neurotypicals and Individuals Diagnosed with Autism and Down Syndrome

Mohammed E. Hoque
MIT Media Lab
20 Ames Street
Cambridge, MA 02139
mehoque@mit.edu

ABSTRACT

Many individuals diagnosed with autism and Down syndrome have difficulties producing intelligible speech. Systematic analysis of their voice parameters could lead to better understanding of the specific challenges they face in achieving proper speech production. In this study, 100 minutes of speech data from natural conversations between neurotypicals and individuals diagnosed with autism/Down-syndrome was used. Analyzing their voice parameters indicated new findings across a variety of speech parameters. An immediate extension of this work would be to customize this technology allowing participants to visualize and control their speech parameters in real time and get live feedback.

General Terms

Human Factors

Keywords

Autism, Down syndrome, accessibility, speech production, speech processing

1. INTRODUCTION

Consider a scenario where an individual with speech disorder can visualize and manipulate his/her speech properties through a computer application and get live feedback. Evidently, creation of this application requires understanding of the limitations that an individual may have with a particular speech disorder. Many people diagnosed with Autism Spectrum Disorder (ASD) or Down syndrome (DS) have difficulties in speech production. Their limitation to verbally express themselves often gets equated with their lack of intelligence and “disinterest” in social interaction. However, rapid development of technology, particularly in the area of speech processing has shown promise towards building technologies measuring speech production and processing abilities [1][2]. It has been hypothesized in [3] that real time visualizations of speech properties, which often act as social mirrors, can influence social communication. However, building such computer applications requires understanding and studying divergent speech properties of individuals diagnosed with speech disorders.

In this study, data [4] collected from one-to-one conversation was used to compare properties of speech across three groups: Autism (ASD), Down-syndrome (DS) and Neuro-Typicals (NT).

2. DATA COLLECTION

The study was set up at the Groden Center, a non-profit school in Providence, Rhode Island, which provides early intervention services to individuals diagnosed with autism, behavior and developmental disabilities.

2.1 Human Subjects

There were six participants in this study: two NT, three diagnosed with mild to moderate autism, and one diagnosed with DS.

2.2 Apparatus

The recording system consisted of a MacBook connected to an analog camera via an Analog-to-Digital converter. At each recording session, two participants sat across a desk from each other in conversation with their own recording system in front of them. The NT partner would arbitrarily pick a topic that s/he believed to interest the other conversation partner (ASD or DS). The conversation then continued as question-answer format with the NT partner taking the lead of the conversation.

3. PROCEDURES

The overall flow of the process was to engage two participants into a conversation, separate the sources into different channels and then extract the speech features from them, as shown in Figure 1. In this study, 10 conversations totaling 100 minutes of audio data were used. The conversations mainly took place between NT vs ASD and NT vs DS.

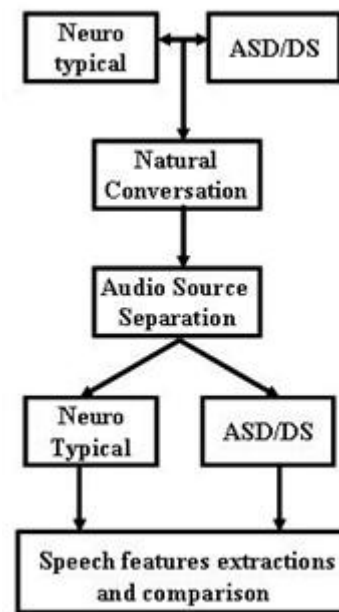


Figure 1. High level diagram.

3.1 Features Extraction

To compare and contrast speech features across ASD, DS and NT, more than 50 features related to segmental and suprasegmental information (information that are not confined to any one segment; rather they evolve in a hierarchy of higher levels of an utterance) were calculated. Computed features were utterance level statistics related to fundamental frequency (F0). Other features were related to duration, pauses, rhythm, voice quality intensity, and formants. The speech processing software *Praat* [5] was used to calculate the prosodic features of speech.

3.2 Feature Mining

The second phase of this study included using a variety of feature mining algorithms to identify the distinguishable speech features for the DS, ASD and NT participants. In this process, optimal feature subsets were first identified and then evaluated using search methods and evaluation techniques. There are two stages to feature selection algorithms. The first stage uses search methods to identify optimal subsets of features and the second stage evaluates the subsets using different measures. Three search techniques were used: best search, greedy stepwise, and ranker. The algorithms that were considered for evaluation of feature sets are Consistency Subset Evaluator and Chi Squared Attribute evaluator. The Waikato Environment for Knowledge Analysis (WEKA) [6] was used to perform the feature mining.

4. RESULTS

The analysis indicates that the average duration per turn was longer in our NT than in our DS and ASD participants. This result is consistent with the pattern of NTs taking the proactive roles in continuing the conversation as opposed to their ASD or DS conversational partner due to the set up of this study. However, the energy parameter in speech yields much higher values for individuals with DS compared to NTs and ASD. This could be due to the DS's habit of being easily and usually excited compared to NT or ASD. It was explicit through the analysis of pause parameters that NTs are capable of using pauses in an utterance/turn more appropriately compared to ASD and DS. It has also been observed that the magnitudes of maximum rising and falling edges in an utterance/turn are much higher in NTs, and then in DS compared to ASD. However, in terms of the number of rising and falling edges in an utterance/turn, the numbers for ASD and NT are very comparable. There are multiple ways to explain this phenomenon. However, one plausible explanation is that individuals with ASDs are capable of being as responsive as NTs. But they often fail to articulate their responses with appropriate intonational parameters.

A complete list of speech features that are similar and dissimilar across the three groups of NT, DS and ASD is shown in Table 1. It's evident from Table 1 that features related to pitch, intensity, formants contain the most dissimilar information whereas voice quality features, speaking rate, patterns of pauses contains the most similar information across speech samples from ASD, DS and NT apart.

5. CONCLUSION

In this study, over 50 features were extracted and combinations of these mined for locating distinctive properties in 100 minutes of speech recorded from one-to-one conversations between NTs and individuals diagnosed with ASD or DS. We found features consistent with the NT's taking the lead in the conversation, as

well as features showing higher DS energy levels, and lower magnitudes of maximum rising and falling edges in ASD. We understand that some of the characteristics of speech properties Table 1. The optimal features sets of speech samples using combination of Ranker search technique and Chi Squared Attribute evaluator, greedy stepwise search technique and CfsSubsetEvaluator, and Bestfirst search technique with ConsistencySubset evaluator.

| Speech features that are similar across three groups (not in any particular order) | Speech features that are dissimilar across three groups (in order of significance) |
|--|--|
| Voice quality features (jitter, shimmer), speaking rate, pauses parameters (# of pauses per turn, maximum duration of pauses), maximum duration of pauses, values associated the second formant | Minimum pitch, mean pitch, maximum pitch, mean intensity, mean intensity, max intensity, values of first and third formants, min intensity, energy, bandwidths of first and third formants |

across NT, ASD and DS are attributed due to the experimental setting, where the NT partner is encouraged to take the lead by asking questions and expect the other partner (DS and ASD) to answer it. Future work will involve designing an experimental setting where subjects are encouraged to take turns while accomplishing a task together or playing a game (card games for example). Such an experimental setting would elicit appropriate emotion with unbiased speech samples without requiring someone to take the lead in the conversation. While it is challenging to recruit a good number of ideal subjects willing to participate in an experiment, it is desired to replicate this study across more subjects to further validate the reported results. However, the outcomes of this study are immediately useful in building speech applications where participants can visualize and manipulate these and other properties of speech with live feedback.

ACKNOWLEDGEMENT

I would like to acknowledge my advisor Prof. Rosalind Picard for her valuable guidance with this work.

6. REFERENCES

- [1] J. Hailpern, Encouraging Speech and Vocalization in Children with Autistic Spectrum Disorder, *ACM SIG ACCESS's Doctoral Consortium/Winter Newsletter*.
- [2] J. Santen, A. Kain, et al., Synthesis of prosody using multi-level sequence units, *Speech Communication* 46(3-4): 365-375, 2005.
- [3] T. Bergstrom and K. Karahalios, Seeing More: Visualizing Audio Cues. *Proceedings of INTERACT 2007*.
- [4] A. Teeters, Use of a Wearable Camera System in Conversation: Toward a Companion Tool for Social-Emotional Learning in Autism, MIT MS Thesis, September 2007.
- [5] P. Boersma and D. Weenink, Praat: doing phonetics by computer (Version 4.6.01) [Computer program]. Retrieved May 16, 2008, from <http://www.praat.org/>
- [6] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," 2 ed., San Francisco: Morgan Kaufmann, 2005.