Interspeech 2018
2-6 September 2018, Hyderabad

Talker diarization in the wild: The case of child-centered daylong audio-recordings

Alejandrina Cristia[1], *Shobhana Ganesh*[2], *Marisa Casillas*[3], *Sriram Ganapathy*[2]

[1]LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, 75005, France. [2]Learning and Extraction of Acoustic Patterns (LEAP) lab, Electrical Engineering, Indian Institute of Science, Bangalore, 560012, India.
[3]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

alecristia@gmail.com,shobhana224@gmail.com, marisa.casillas@mpi.nl,sriramg@iisc.ac.in

Abstract

Speaker diarization (answering 'who spoke when') is a widely researched subject within speech technology. Numerous experiments have been run on datasets built from broadcast news, meeting data, and call centers—the task sometimes appears close to being solved. Much less work has begun to tackle the hardest diarization task of all: spontaneous conversations in real-world settings. Such diarization would be particularly useful for studies of language acquisition, where researchers investigate the speech children produce and hear in their daily lives. In this paper, we study audio gathered with a recorder worn by small children as they went about their normal days. As a result, each child was exposed to different acoustic environments with a multitude of background noises and a varying number of adults and peers. The inconsistency of speech and noise within and across samples poses a challenging task for speaker diarization systems, which we tackled via retraining and data augmentation techniques. We further studied sources of structured variation across raw audio files, including the impact of speaker type distribution, proportion of speech from children, and child age on diarization performance. We discuss the extent to which these findings might generalize to other samples of speech in the wild.

Index Terms: speaker diarization, language acquisition, spontaneous speech, i-vectors

1. Introduction

At a glance, the problem of automatic, unsupervised speaker diarization (deciding who is talking at a given time) appears to be a solved task. For instance, in a 2012 review on meeting recording diarization [1], the top-performing system achieved a 4% Diarization Error Rate (DER) for single distant microphone, multi-talker settings. Such accurate diarization is also needed for analyses of the messier acoustic environments characteristic of our everyday language use. In particular, audio recordings gathered with personal devices worn by small children have enormous potential for shedding light on how children learn language.

While it is obvious that typically developing children come to speak their ambient language(s) effortlessly, it is less clear how exactly this process comes about. Surely children learn about language from what they hear, but what exactly is *available* in their speech environment for them to learn from? By recording children's at-home language environments, we can inspect what children say and hear on an everyday basis [2]. In turn, we can better hypothesize about the learning mechanisms that process this linguistic 'input' into full-fledged linguistic knowledge. It is imperative to include geographically, culturally, and linguistically diverse populations in this process so that we capture the whole range of early language experiences that language-learning children encounter [3]. Findings in this domain also indirectly further efforts on language documentation, preservation, and revitalization, as well as inform clinical applications.

Advances in recording technology have broadened our view of children's speech environments—we can now gather recordings that last whole days or weeks—and we can better appreciate the diversity of activities and interactive environments that make up children's daily linguistic experience. An enormous challenge now is how to extract useful information from these recordings, which quickly accumulate to hundreds or thousands of hours, and can therefore no longer simply be manually diarized and annotated for speech properties.

That said, the use of diarization with child audio remains relatively rare. Over the years diarization systems have typically focused on broadcast news and telephone conversations between adults with reasonably clean audio. Advances in systems have also been evaluated based on these datasets [4], and only some diarization studies have been performed on children's speech (e.g, [5, 6])

Daylong child language recordings present a stimulating next challenge for diarization systems. A typical day includes variable background noise conditions as the child moves between various reverberant and dampened spaces inside and outside of their home (see Fig. 1). The recording devices used are typically equipped with one or two omnidirectional microphones. Among the many voices that may be captured over the day, most come from relatives who often sound similar to each other; much more similar than two speakers in an average business meeting or clinical recording.

There is increasing interest in solving talker diarization in these difficult daylong recordings. Using their patented recording device, researchers associated with the LENA^{TM} Foundation have gathered an extensive dataset of daylong recordings varying in child age and socio-economic status, and have developed a set of algorithms to parse the audio (e.g., [7, 8, 9, 10]). Their proprietary software extracts the recordings and processes them as follows: It first extracts 36 mel frequency cepstral coefficients and their deltas in 25 ms windows every 10ms. It then analyzes these features with an iterative system that performs joint vocalization activity detection and talker diarization to break the stream into uniform segments that are minimally .6 seconds long. This process is performed with a Minimum Duration Gaussian Mixture Model (MD-GMM) combined with dynamic programming to find the sequences with maximum likelihood.

10.21437/Interspeech.2018-2078

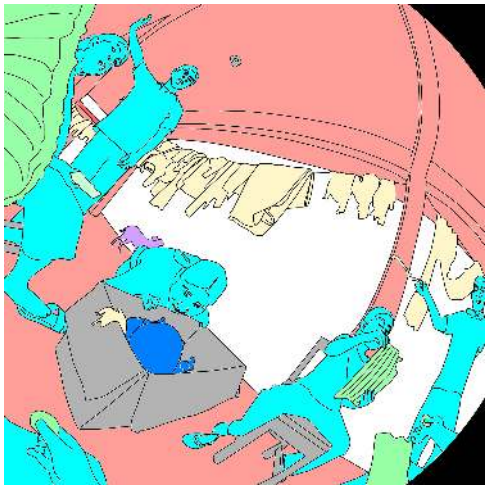

Figure 1: *Representative moment from a Mayan daylong recording, captured via a mother's chest-worn camera with a fish-eye lens. The infant (dark blue) and seven of her family members (light blue) take an afternoon break on their patio, which is made of poured concrete (pink) and partly covered with hanging clothes (yellow).*

The MD-GMM model had been trained on over 150 hours of recordings (30 minutes extracted from 309 daylong recordings, gathered from as many American English-learning children) that were segmented by professional transcribers, eventually resulting in eight categories: Key Child, Other Child, Adult Male, Adult Female, TV/other electronic sound, Noise, Silence, and Overlapping sound (overlap of any two categories, e.g., Key Child + Noise). In evaluating their system [11], they find 71–86% agreement in terms of broad categories of "adult", "child", "TV", and "other", the latter including all overlap regions. Subsequent independent research has largely confirmed these high levels of accuracy (e.g., [12, 13]).

Although their results are *prima facie* promising, further independent work is needed to improve some aspects of the current system. First and foremost, the LENA^{TM} software is proprietary and can neither be modified nor interrogated beyond the descriptions found in published work. Moreover, it cannot be applied to recordings that have not been collected with their recording device (the LENA^{TM} DLP). Second, their approach effectively removes regions with background noise, as well as any lively spontaneous conversation that may contain a great deal of overlap, thus biasing the initial sample towards "easy" regions. In an attempt to keep their data comparable, other researchers have tended to use the same sampling strategy, and we thus have no reliable, generalizable estimate of global diarization performance in daylong recordings. Finally, the LENA^{TM} Foundation and others working in their wake have made the analytic choice of collapsing across all female adult speakers, all male adult speakers, and all child speakers for most accuracy reports. This means that the LENA^{TM}-based reports are not penalized for confusing the mother with other females, or the target child with other children. Yet it would be very informative to provide accuracy estimates that take into account the real identity of these different conversational partners.

1.1. Main goal

Child-centered audio recordings probably constitute the most difficult, yet interesting, challenge facing current speaker diarization systems. Previous reports of speaker diarization performance were based on a single system and may suffer from some biases. Therefore, our main goals were to assess performance of current off-the-shelf diarization systems in such difficult settings and to explore salient avenues to improve the performance, e.g., via retraining.

2. Methods

2.1. Corpus

The third author has collected a large corpus of daylong recordings from children who are growing up in traditional, non-Western, preindustrial societies [14, 15]. The present paper focuses on 10 hours from that corpus which have been carefully annotated jointly by the third author (who is linguistically trained) and a native speaker of the language who personally knew the recorded families. Therefore, these data can be treated as a high standard against which to compare the performance of automated diarization tools.

The ∼10-hour at-home recordings come from 10 Tseltal Mayan children between the ages of 2 and 36 months who live with 3–11 other people (0–5 of whom are siblings). So both (a) the rate and type of vocalizations made by the children and (b) the number and types of other speakers present varies greatly across children. For each child, there is one hour of annotated audio, divided into 19 clips sliced out of the original recordings. The audio scenes vary dramatically, even within a single child's clips, as the child moves from one activity to another over the course of the day. The 19 annotated clips from each of the 10 recordings were selected in multiple ways: random sampling (9 x 5 minutes), or hand-selected moments of high talk or interaction by the child (9 x 1 minutes + 1 x 6 minutes).

2.2. Processing and analyses

Our dataset included 203 clips, 8 of which did not have any speech and were thus discarded. We randomly split the families into training (N families = 5; N clips 17–21 per family, ages 2–36 months) and generalization (N families = 5; N clips 18–21 per family, ages 4–32 months) data sets to ensure that the training set covered the whole range of child ages.

The recording device had two omnidirectional microphones, one slightly closer to the child's mouth (∼20 vs. 22 centimeters). We extracted the channel closer to the child's mouth from the audio and extracted the speech intervals as the segments where one or more people were speaking from the annotations. These two information sources were the input given to the diarization system.

We used the i-vector based system using a PLDA scoring metric [16], followed by clustering, to compute the diarization error rate for each of the audio files. We used the Kaldi pipeline to run the system [17]. Mel Cepstral Frequency Coefficient (MFCC) features were extracted from the audio files with a window size of 25 ms and a stride of 10 ms. These features were used to train the background model consisting of a 2048 mixture component GMM. Following this the T-matrix was trained and i-vectors were obtained. These i-vectors were obtained at every 150 ms with a 75 ms stride for evaluation and for the data used in training, we extract 300 ms i-vectors for every 10 s. The i-vectors are of dimension 128. A PLDA scoring is applied on these i-vectors to compute the similarity between each pair. The i-vectors obtained from the training dataset (depending on the training regime, as explained below) was used to train the PLDA system after which the evaluation data i-vectors (always

the Tseltal data) were scored. These scores were then clustered using agglomerative hierarchical clustering to group all scores belonging to individual speakers together. The stopping criteria in the clustering stage is based on a threshold which is obtained using unsupervised calibration by fitting a two mixture GMM on the PLDA scores. This parameter proved to be an important factor during experimentation wherein we observed that the DER values changed significantly on changing the calibration score.

We explored a variety of training regimes during our experiments. The first dataset we chose for our analysis was LibriSpeech [18], which is a collection of stories read aloud by a number of different speakers in English. This corpus is clean and free of any background noise or disturbances. Using this as our baseline, we experimented with training on different datasets, as follows.

Keeping in mind that the Tseltal dataset analyzed in this paper is a noisy corpus—with many background sounds, animal sounds, adult speech (dominated by females), and sibling speech in addition to the child's own speech—we added the AMI corpus [19] to the training set. The AMI corpus contains recordings of meetings consisting of spontaneous speech with a natural room reverberation and overlapping speech. Since the AMI corpus is male dominated, to balance the gender ratio we also used the Switchboard cellular dataset [20], a speech corpus on mobile phone conversations. We picked out all conversations having at least one female talker taking place in either an outdoor or indoor setting. To account for the child speaking, we used the Paidologos dataset, which are laboratory recordings of words in isolation spoken by children in English, Japanese, Greek, and Cantonese [21, 22, 23, 24], available from the CHILDES repository [25]. To simulate the Tseltal environment, we augmented this dataset with babble, reverb, and noise [26].

3. Results

The average DER collapsing across all clips and all systems was 48.2%, with a range between 0 and 86.2%. Anguera and colleagues' (2012) review [1] show an average DER between 4 and 32% for a range of systems applied to a varied set of meeting recordings. Thus, the first conclusions may be that these family-based recordings are indeed more challenging than the meeting data that has been the focus of diarization attention in the recent past. Overall, systems underestimated the number of speakers when the audio files had a large number of annotated talkers.

Manual inspection suggested that DER changed as a function of training regime at the clip level. For example, on one file the DER was 48.9% when trained on LibriSpeech alone, while the DER dropped to 19.2% when trained on the combination of AMI, augmented Paidologos, and Switchboard cellular. However, statistical inspection of performance across the different training systems suggested that the impact of the training regime on performance was not statistically significant (all p's $> .05$), and only the threshold manipulation helped (all p's $< .05$), with gains against all other systems of about 6% DER for the .7 threshold, and of about 8% DER for the .8 threshold. These systems may be outperforming the others for the wrong reason: the best-performing system estimates that there is only one speaker for all clips.

We additionally observed that certain clips exhibited a very consistent performance (i.e., a very low DER or very high DER) regardless of the training dataset. This suggested that gaining

Figure 2: *Illustration of speaker identity in the gold annotation versus two of the systems' output in a file yielding average scores. Notice particularly the complexity in speaker overlap and turn-taking in the gold (top row of annotation).*

an overall insight on performance over all families and clips is not as productive as analyses on each clip separately. We thus explored variability in clip performance. We predicted that the following would lead to better performance (lower DER):

- in recordings with older children, they and their same-age peers will have more recognizable speech (closer to that which the speech diarization systems have been trained on), thus leading to higher performance with child age;
- longer turns, which would be easier to classify;
- fewer speakers, which reduces the chance of confusion errors;
- a higher proportion of adults, which is a better fit to data used in the past;
- clips with more diverse speaker profiles, which are easier to classify (i.e., clip with a child + female adult + male adult, versus a clip with 3 children);
- clips with less speech (and thus random, over selected, clips), since less speech means fewer opportunities for error.

We also hypothesize that further study on the scoring technique as well as the calibration methodology from the system perspective could lead to an improvement in DER with the current dataset. Therefore, we fit a linear regression predicting DER from the best-performing system (trained on AMI, the augmented Paidologos data, and the Switchboard; threshold set at .8) from the child's age, the average turn duration in the clip, the number of people who spoke, their diversity (on a three-point scale, counting the presence of female adults, male adults, and children separately), the proportion of speakers who were adults as compared to children, controlling for the family's ID. Results should be taken with a grain of salt since the data violated equality of variance. This model was overall significant: $F(14,177) = 19.72$, $p < .001$; and it explained a substantial proportion of variance: $R^2 = .58$. The overall number of speakers, child age, average turn duration, whether the clip was randomly or purposefully selected, and the family ID significantly predicted diarization performance. Figure 3 shows that, unsurprisingly, the system performs worse when more talkers were present in the clip [$\beta = 7.84$ (0.81)]. Also as predicted, longer turns led to lower error rates [$\beta =$ -8.62 (4.34)]. However, counter to predictions, age was positively associated with DER: clips from older children had *higher* error rates [$\beta =$ -0.54 (0.22)]. Also, clips that had been selected to have a greater amount of speech or higher talker change rate in fact had *lower* DER, i.e. higher performance [$\beta =$ -9.12 (2.4)].

4. Discussion

Our findings confirm that daylong recordings of children's natural language environments are incredibly challenging for cur-

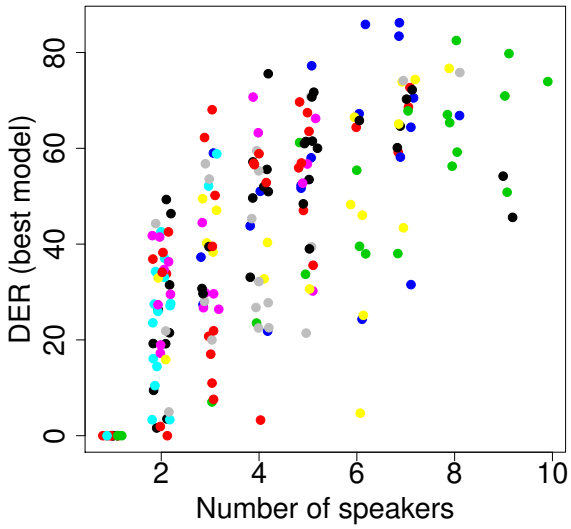

Figure 3: *Diarization error rate as a function of the number of speakers found in the gold human annotation. Each dot is a clip, and a dot's color indicates which of the 10 recordings it came from.*

rent diarization systems. Based on a variety of clips extracted from daylong recordings, we estimate the DER from current systems to be no lower than 40%. This contributes some much needed unbiased estimations regarding the accuracy of automatic diarization systems.

We had expected difficulties due to higher voice similarity, more variable number of participants, shorter turn duration, and more overlap than in previously studied meeting datasets. Although the first two are probably true, at least in this dataset the turn duration and level of overlap is comparable to that reported for meetings achieving much better DERs. Indeed, Anguera and colleagues' (2012) review [1] estimates an average turn duration of 1.4 seconds, with 7 to 16% overlap for meeting data (on which 4% DER has been documented), whereas turns in our data were an average of 1.1 seconds long with 12% overlap. Our regression analyses suggested that, while shorter turns led to lower performance, overlap itself did not explain significant variance. Beyond these factors, the enormous difference between the meeting accuracy (50% here versus minimally 4% to 25% across a range of systems in meeting data [1]) indirectly suggests that voice similarity and variability or talker number pose a formidable challenge for current diarization systems. While we could not test for voice similarity effects (beyond broad classes of male/female adult and child, which was not significant), our regression did confirm that a larger number of talkers led to lower performance. Additionally, we found that clips selected because they had a lot of child speech or active verbal interaction between the child and others actually led to higher performance than a random selection, contrary to our expectations based on the LENA work. Overall, we believe there is a great deal more work to be carried out to understand which factors are most difficult about daylong recordings, and how to address these roadblocks.

Perhaps the most surprising finding is that training did not help improve performance. This is far from obvious, as the idea that in-domain data helps is almost a truism. Yet pretraining on a corpus containing children's voices, including children's voices augmented with noise, did not significantly change DERs. And while we did manage to build a system that outperformed the others, it did so by collapsing all speakers onto one, which is conceptually unacceptable.

One may wonder whether some top-down information could help raise performance. For instance, it would be easy to provide systems with the number of family members. However, the likelihood of each talking at a given time in the recording is completely unknown. Almost every family provided clips with the entire range of number of talkers for that family, so it is unlikely that one will be able to constrain inferences based on the composition of the family. A semi-supervised system that provides annotators with a first classification (e.g., [27]) may, however, be more useful.

Might our results generalize to other recordings "in the wild"? All of our recordings were made in a rural, traditional setting with (mainly) large families. Therefore, we believe that the task we have tested here is harder than that which will be encountered with recordings from typical Western middle-class households. In an average household in the USA, there are 1–2 parents and 1–2 children, whereas the average household size in the present sample is 7 people, with a range of 4 to 14. Further work should revisit these questions with recordings that are not centered on children. Indeed, a growing field of research is investigating the possibility of using adults' speech as a potential biomarker (e.g., [28]). We believe that adult-centered recordings will be, on average, less challenging than child-centered ones, with difficulty levels increasing for certain neurological conditions affecting speech production (e.g., aphasia).

5. Conclusions

In sum, this paper provides the first systematic assessment of speaker diarization of audio recordings collected as children go about their normal day. We find that performance is much lower than that found in previously "difficult" data, notably multi-talker meetings. A main cause for errors is found in marked misestimations of talker number, with increased difficulty when more talkers are present, even after controlling for turn duration.

6. Acknowledgements

AC and SrG initiated the project, MC provided the development and test data, ShG conducted the experiments under SrG's supervision, AC provided the first draft with input from MC and ShG, AC carried out the statistical analyses; all authors approved the manuscript. Funding: TransAtlantic Platform "Digging into Data" collaboration grant (ANR-16-DATA-0004 ACLEW: Analyzing Child Language Experiences Around The World); Agence Nationale de la Recherche (ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC); and J. S. McDonnell Foundation to AC; NWO Veni Innovational Research Scheme (275-89-033) to MC. Some calculations used the Extreme Science and Engineering Discovery Environment (XSEDE), supported by National Science Foundation grant number OCI-1053575. Specifically, they used the Bridges system, supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). We benefited from the 2017 Jelinek Workshop, which was supported by Amazon, Apple, Facebook, Google, and Microsoft (in alphabetical order).

7. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent

research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] B. MacWhinney, "The CHILDES project part 1: The CHAT transcription format," *Department of Psychology*, 2009. [Online]. Available: https://talkbank.org/manuals/CHAT.pdf

[3] S. Stoll and E. Lieven, "Studying language acquisition cross-linguistically," *South and Southeast Asian psycholinguistics*, pp. 19–35, 2014.

[4] A. Joshi, M. Kumar, and P. K. Das, "Speaker diarization: A review," in *2016 International Conference on Signal Processing and Communication (ICSC)*, Dec 2016, pp. 191–196.

[5] M. Najafian and J. H. L. Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 114–120.

[6] S. Schuster, S. Pancoast, M. Ganjoo, M. C. Frank, and D. Jurafsky, "Speaker-independent detection of child-directed speech," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 366–371.

[7] J. Gilkerson and J. A. Richards, "The lena natural language study," 2008.

[8] D. Xu, U. Yapanel, S. Gray, and C. T. Baer, "The LENA^{TM} language environment analysis system: the interpreted time segments (its) file," 2008.

[9] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen, "Signal processing for young child speech language development," in *First Workshop on Child, Computer and Interaction*, 2008.

[10] D. Xu, J. Gilkerson, J. Richards, U. Yapanel, and S. Gray, "Child vocalization composition as discriminant information for automatic autism detection," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 2518–2522.

[11] D. Xu, U. Yapanel, and S. Gray, "Reliability of the LENA^{TM} language environment analysis system in young children's natural home environment," 2008.

[12] E.-S. Ko, A. Seidl, A. Cristia, M. Reimchen, and M. Soderstrom, "Entrainment of prosody in the interaction of mothers with their young children," *Journal of Child Language*, vol. 43, no. 2, pp. 284–309, 2016.

[13] M. VanDam and N. H. Silbert, "Fidelity of automatic speech processing for adult and child talker classifications," *PloS one*, vol. 11, no. 8, p. e0160588, 2016.

[14] Casillas, Marisa and Brown, Penelope and Levinson, Stephen C., "Casillas-X-cultural," https://hdl.handle.net/1839/9E3EF620-690E- 4BC1-8A10-B6815AF84DAB@view, 2017.

[15] ——, "Casillas HomeBank Corpus," https://homebank.talkbank.org/access/Secure/Casillas.html, 2017.

[16] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 413–417.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[19] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[21] H. Chung, E. J. Kong, J. Edwards, G. Weismer, M. Fourakis, and Y. Hwang, "Cross-linguistic studies of children's and adults' vowel spaces," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 442–454, 2012.

[22] J. Edwards and M. E. Beckman, "Methodological questions in studying consonant acquisition," *Clinical linguistics & phonetics*, vol. 22, no. 12, pp. 937–956, 2008.

[23] E. J. Kong, M. E. Beckman, and J. Edwards, "Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese," *Journal of phonetics*, vol. 40, no. 6, pp. 725–744, 2012.

[24] F. Li, "Language-specific developmental differences in speech production: A cross-language acoustic study," *Child development*, vol. 83, no. 4, pp. 1303–1315, 2012.

[25] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk (Third Edition)*. Lawrence Erlbaum Associates, 2000.

[26] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[27] C. Yu and J. H. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.

[28] C. Kohlschein, M. Schmitt, B. Schüller, S. Jeschke, and C. J. Werner, "A machine learning based system for the automatic evaluation of aphasia speech," in *e-Health Networking, Applications and Services (Healthcom), 2017 IEEE 19th International Conference on*. IEEE, 2017, pp. 1–6.