Child vs Adult Speaker Diarization of naturalistic audio recordings in preschool environment using Deep Neural Networks†

Prasanna V. Kothalkar, John H.L. Hansen
Center for Robust Speech Systems (CRSS)
University of Texas at Dallas, TX, USA

Dwight Irvin, Jay Buzhardt
Juniper Garden's Children Project
University of Kansas, KS, USA

Beth Rous
Department of Educational Leadership Studies
University of Kentucky, KY, USA

Abstract

Speech and language development in children is crucial for ensuring optimal outcomes in their long term development and life-long educational journey. A child's vocabulary size at the time of kindergarten entry is an early indicator of learning to read and potential long-term success in school. The preschool classroom is thus a promising venue for monitoring growth in young children by measuring their interactions with teachers and classmates. Automatic Speech Recognition (ASR) technologies provide the ability for 'Early Childhood' researchers for automatically analyzing naturalistic recordings in these settings. For this purpose, data are collected in a high-quality childcare center in the United States using Language Environment Analysis (LENA) devices worn by the preschool children. A preliminary task for ASR of daylong audio recordings would involve diarization, i.e., segmenting speech into smaller parts for identifying 'who spoke when.' This study investigates a Deep Learning-based diarization system for classroom interactions of 3-5-year-old children. However, the focus is on 'speaker group' diarization, which includes classifying speech segments as being from adults or children from across multiple classrooms. SincNet based diarization systems achieve utterance level Diarization Error Rate of 19.1%. Utterance level speaker group confusion matrices also show promising, balanced results. These diarization systems have potential applications in developing metrics for adult-to-child or child-to-child rapid conversational turns in a naturalistic noisy early childhood setting. Such technical advancements will also help teachers better and more efficiently quantify and understand their interactions with children, make changes as needed, and monitor the impact of those changes.

† This work was supported by the National Science Foundation (NSF) under Grant Award #1918032 (PI: Hansen), NSF Grant Award #1918012 (PI: Irvin), Kentucky Race to the Top Early Learning Challenge Grant Award #S412A130045 (PI: Rous)

Introduction

The diversity of language background, socio-economic conditions, development level, or potential communication disorders represents a challenge in the assessment of child speech and language skills[1]. The language environment of young children plays an important role in the development of speech, language, vocabulary, and thus, thinking and learning ability, and has an impact on the lifelong outcomes for the child. The quality and number of interactions in a rich language environment help support essential language development outcomes in early childhood[2]. Thus, early childhood researchers are focusing on analyzing classroom interactions of preschool children to monitor and provide proactive support to them. Due to the vast amount of daylong recordings to be analyzed, the usage of automated speech processing and machine learning techniques would be highly beneficial.

The preliminary task of analyzing such data environments involves Speaker Diarization, i.e., segmenting and tagging 'who spoke when', followed by Speech Recognition, Keyword Spotting etc. In this study, we perform Speaker group Diarization on child-adult and child-child interactions of preschool children in naturalistic active learning environments. The audio data were collected using LENA devices[3,4] worn by the children (Figure 1) in different classrooms at different times. The recordings continue as subjects move around during a school day and are paused during nap time.

Specifically, a SincNet-based Speaker Diarization system that uses oracle segmentation is presented. Additionally, we analyze classifications of different speaker groups, which provides insights into how the results can be improved.

Modeling Speaker Characteristics

i-Vectors[5,6] are fixed length vectors that characterize speaker identity from arbitrary length sequential data (i.e. speech samples) and are traditional features for speaker recognition[6]. They have also been used for language recognition[7], accent recognition[8], emotion recognition[9] etc.

DNNs[10,11,12] can be used to directly capture language or speaker characteristics. They have provided improved results over i-Vectors using Mel-Frequency Cepstral Coefficients[5] or Filterbank Coefficients[5] as features. But these standard features smooth the speech spectrum, discarding crucial narrow-band speaker characteristics such as pitch and formants. The current standard framework consists of a discriminatively trained DNN that maps variable-length speech segments to embeddings called x-vectors[12]. x-Vectors are deep speaker embeddings based on time-delay recurrent neural network architecture. It has provided excellent results for speaker recognition[12], diarization[13] and language recognition[14] with innovations being actively researched.

Previous work on child speech utilized i-Vectors[15,16] and x-Vectors[17] as features for speaker classification. SincNet-based speaker identification model[18] has been used in university classroom settings[19] as well as for adult vs. child speech classification[20] with good results.

Previous work on this dataset[15] used much lesser data and fixed segments of length 1.5 seconds with a Support Vector Machine (SVM) backend for classification. A recent work[16] with more data transcribed for the dataset, used Deep Neural Network (DNN) modeling with i-Vectors as features provided promising results.

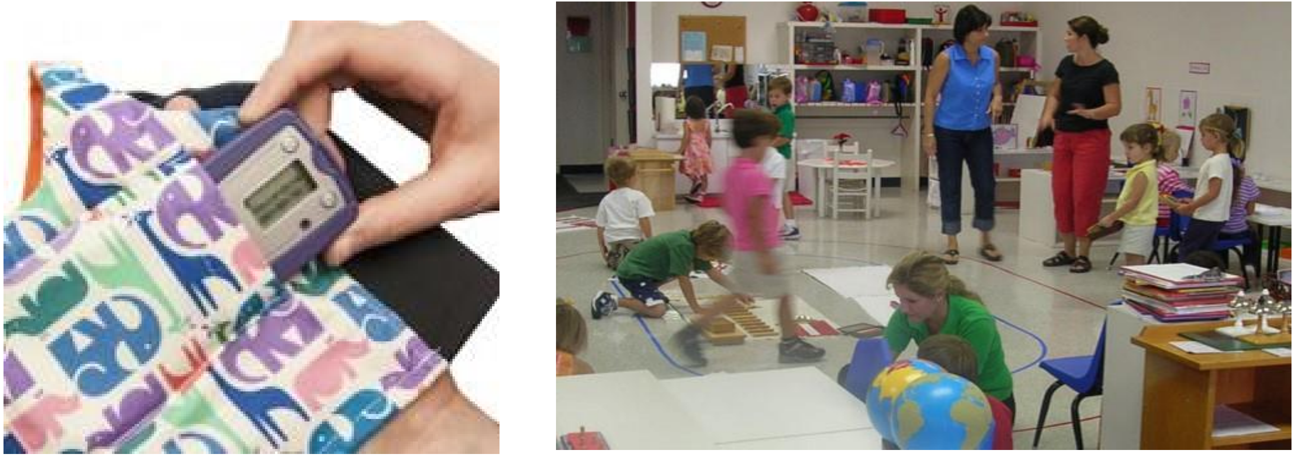

Figure 1. LENA device in jacket pouch (left) and an illustrative preschool classroom (right)

Dataset Details

Speaker group /Dataset	Primary Child	Secondary Children	Adults	All speaker groups
Training set	2:37:39	3:32:19	7:45:21	13:55:19
Development set	1:34:55	1:37:30	2:44:02	5:56:27
Test set	0:55:05	0:57:52	1:24:23	3:17:20
Overall set	5:07:39	6:07:41	11:53:46	

Table 1. Dataset split in terms of duration (hh:mm:ss) for training, development and test sets as well as speaker groups of Primary Child, Secondary Children and Adults

The dataset in this paper consists of spontaneous conversational speech recorded with the help of LENA recording units attached to the subjects in a high quality child care center in the United States. The 48 recording sessions have children who are 3 to 5 years old. About 15 hours of child speech was manually transcribed by the CRSS transcription team at UT Dallas. Another 28 hours of adult speech from 4 teachers/caregivers were manually transcribed as well. A total of 79 hours of speech and non-speech child and adult data was tagged by our transcribers. Out of these, 27 randomly selected sessions were divided into training, development, and test sets for initial investigation of a SincNet-based diarization system. Transcribed segments of less than 1 second of duration are not selected, as any audio of lesser duration mostly has pseudo-words like 'hmm','aah' etc. This allots 22,698 utterances in training set, 9,628 utterances in development set and 5306 utterances for testing set.

Method

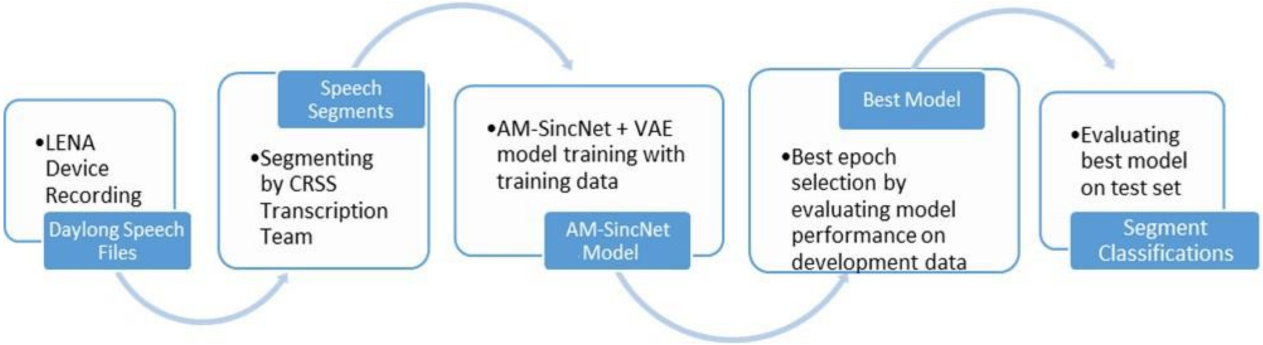
•LENA Device Recording
Daylong Speech Files
Speech Segments
•Segmenting by CRSS Transcription Team
•AM-SincNet + VAE model training with training data
AM-SincNet Model
Best Model
•Best epoch selection by evaluating model performance on development data
•Evaluating best model on test set
Segment Classifications

Figure 2. System block diagram for SincNet-based diarization system that classifies speech segments from daylong recordings using oracle segmentation

The system block diagram (Figure 2) presents the approach for learning AM-SincNet models using training data and evaluating them on testing set, based on best performance on development set.

The classification task is to map quick conversational turns from the LENA audio data as being from Primary Child (PC), Secondary Children (SC), or Adults (AD). Here, PC carries the LENA recording device on his person, while SC/AD are the other children/adults that are recorded by the PC's LENA device. In some sessions, an adult carries the LENA device, and all children in such sessions are marked as SC. Transcripts provided by CRSS transcription team are used as oracle segmentation for speech activity detection. Thus, final evaluation involves comparing the predicted classes for the segmented audio files to their actual class value based on human transcription.

AM-SincNets for child vs adult speaker characteristics
Convolutional Neural Networks[21] (CNNs) have been outperforming i-Vector-based and Deep Learning-based systems. They are used along with raw waveforms[22] to learn low-level speech representations, capturing important speaker characteristics such as pitch and formants.

Parameterized sinc functions that represent band-pass filters have been used to discover more meaningful filters in the first convolutional layer. This interpretable architecture[18,23] provides better results than conventional CNNs and has faster convergence with fewer parameters. The cumulative frequency response characteristics learned by SincNet correspond to pitch and formant of an adult male or female. These are better than those of a CNN, due to the interpretable filters and are proposed for better classification of adult vs child audio. The hypothesis is that the architecture has potential discriminative capability between adults and children by automatically learning filter frequencies corresponding to the same.

Additive Margin SincNet[24] (AM-SincNet) replaces the Softmax layer of the SincNet with Additive Margin Softmax which introduces an additive margin to its decision boundary. The advantages include better inter-class separability and intra-class clustering, making it ideal for classification and verification tasks. The experimental setup in AM-SincNet architecture block is the same as the setup in the original paper[18]. Experiments are conducted for margin values varying from m=0.05 to m=0.9

with step size 0.1. The best results are obtained for m=0.15, so only these are explored further.

Results and Discussion

The best performing model on the development set in terms of frame-level DER, is evaluated on the test set. Diarization error rate is the metric for evaluating the diarization performance.

Diarization Error Rate

Diarization error rate (DER) can be defined as the sum of errors due to incorrect speaker (E_{spkr}), missed speech (E_{MISS}), false alarm speech (E_{FA}) and overlapping speakers (E_{ovl}).

$$DER= E_{spkr} + E_{MISS} + E_{FA} + E_{ovl}$$

Since oracle speech segments are used, only first type of errors are reported. DER on the development set using the proposed system is 27.5% while DER on the test set using the proposed system is 19.1%.

Confusion Matrix for PC, SC and AD

Predicted /Actual	Primary Child	Secondary Children	Adults
Primary Child	78.41%	11.12%	10.50%
Secondary Children	20.82%	81.81%	8.38%
Adults	0.77%	7.07%	81.12%

Table 2. Confusion matrix for AM-SincNet model for Primary Child, Secondary Children and Adults as the speaker groups

Confusion matrices for the three speaker groups are compiled (Table 2) for a better understanding of the error rates for the AM-SincNet model with margin=0.15. Here the horizontal rows represent actual classification of the utterance-level speech segments, while the vertical columns represent the predicted classification of the utterance-level segments. Thus, the diagonal of this matrix contains the percentage of correct predictions by the corresponding model.

In table 2, the AM-SincNet model provides accurate predictions of 78.41% for PC, 81.81% for SC and 81.12% for AD speaker group. Thus, PC speaker group performs slightly worse than SC and AD speaker groups. This could be attributed to the disproportionate data distribution for PC class (22.1%) Vs. SC (26.5%) and AD (51.4%) classes. The data imbalance ensures low false positives (FP) for AD than for any other class. The highest FP is for SC class being predicted as PC, which is likely when the SC are close to PC and interacting with him/her. PC and SC are predicted as AD around 8-11% of the time. These near-balanced results with good performance for PC can be further improved with better modeling and algorithmic advancements.

Application for predicting talk time

A practical application for our system involves the prediction of talk time for the different speaker groups by summing the individual segment classifications. We can compare the actual and the predicted talk times using figures 3 and 4. Corresponding to analyses from confusion matrices, it can be seen that PC and SC classes are predicted lesser than their actual duration while AD is predicted to have higher talk time than is the reality. With better precision, these details can be useful to teachers in planning their classes for improved participation and more equitable talk time among the participants.

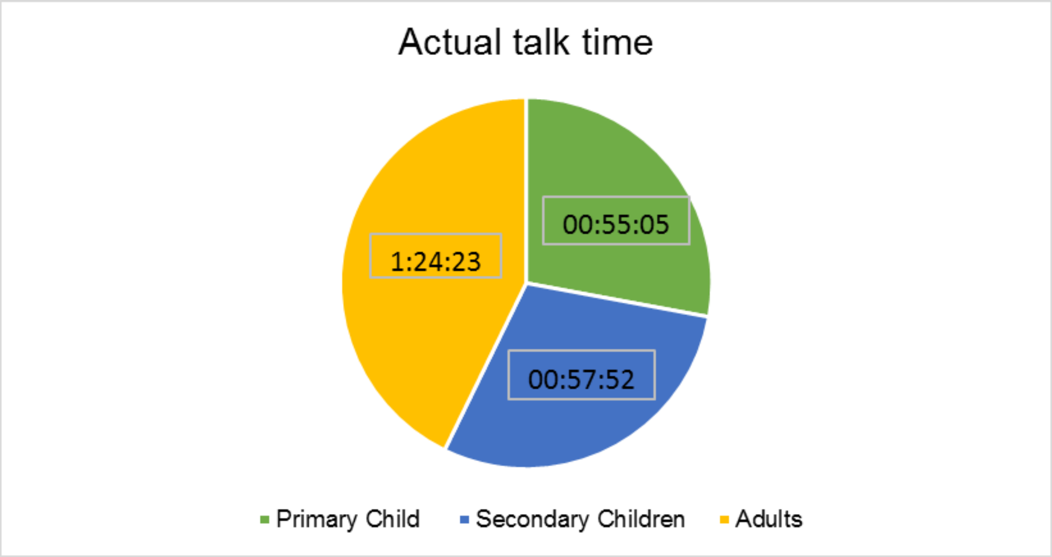

Figure 3. Actual talk time for speaker groups of Primary Child, Secondary Children and Adults

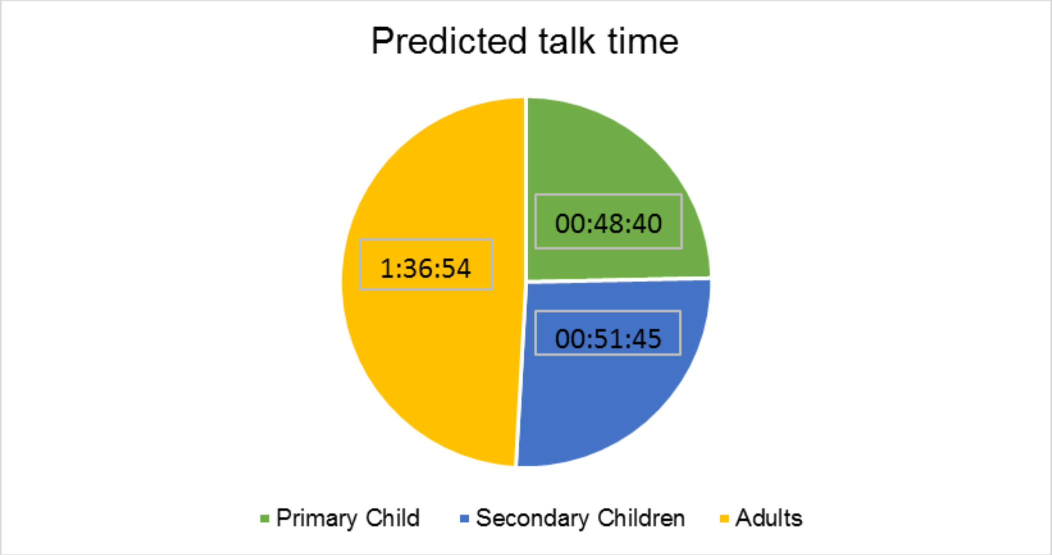

Figure 4. Predicted talk time for speaker groups of Primary Child, Secondary Children and Adults using AM-SincNet model

Conclusion and Future Work

This study investigated AM-SincNet architecture for child vs adult speaker group diarization and provided promising results for classifying oracle segmentation. Further investigation revealed AD speaker group has the lowest false positive rate, probably due to higher data proportion. SC is predicted as PC- almost twice as much Vs. PC predicted as SC. This could also be due to slightly higher proportion of SC data Vs. PC data overall. Thus, SincNet embedding distinguished group characteristics of the input signal with impressive DER of 19.1% on the test set. The classified speech segments can be summed to measure the total talk time for individual speaker groups. The predicted talk time matches the actual talk time with a difference of around 12 minutes between the groups. For future work, advanced modeling in terms of architecture and more balanced data could help in improving the performance. Better performance will be very helpful as diarization is preprocessing step for further systems like Keyword Spotting, Automatic Speech Recognition, and Word Counting.

References

1. National Academies of Sciences, Engineering, and Medicine. (2016). Speech and language disorders in children: Implications for the Social Security Administration's Supplemental Security Income program.
2. Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.
3. URL: https://www.lenafoundation.org
4. Ziaei, A., Sangwan, A., & Hansen, J. H. (2013, May). Prof-Life-Log: Personal interaction analysis for naturalistic audio streams. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7770-7774). IEEE.
5. Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, *32*(6), 74-99.
6. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker

verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788-798.

7. Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association*.
8. Bahari, M. H., Saeidi, R., & Van Leeuwen, D. (2013, May). Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7344-7348). IEEE.
9. Xia, R., & Liu, Y. (2012). Using i-vector space model for emotion recognition. In *Thirteenth Annual Conference of the International Speech Communication Association*.
10. Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., & Khudanpur, S. (2016, December). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 165-170). IEEE.
11. McLaren, M., Lei, Y., & Ferrer, L. (2015, April). Advances in deep neural network approaches to speaker recognition. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4814-4818). IEEE.
12. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5329-5333). IEEE.
13. Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., ... & Khudanpur, S. (2018, September). Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In *Interspeech* (pp. 2808-2812).
14. Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., & Khudanpur, S. (2018, June). Spoken language recognition using x-vectors. In *Odyssey* (pp. 105-111).
15. Najafian, M., Irvin, D., Luo, Y., Rous, B. S., & Hansen, J. H. (2016). Automatic measurement and analysis of the child verbal communication using classroom acoustics within a child care center. In *WOCCI* (pp. 56-61).
16. Kothalkar, P. V., Irvin, D., Luo, Y., Rojas, J., Nash, J., Rous, B. S., & Hansen, J. H. (2019). Tagging child-adult interactions in naturalistic, noisy, daylong school environments using i-vector based diarization system. In *SLaTE* (pp. 89-93).
17. Xie, J., Garcia-Perera, L. P., Povey, D., & Khudanpur, S. (2019). Multi-PLDA Diarization on Children's Speech. In *Interspeech* (pp. 376-380).
18. Ravanelli, M., & Bengio, Y. (2018, December). Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1021-1028). IEEE.
19. Dubey, H., Sangwan, A., & Hansen, J. H. (2019, May). Transfer learning using raw waveform sincnet for robust speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6296-6300). IEEE.
20. Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*.
21. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5200-5204). IEEE.
22. Jung, J. W., Heo, H. S., Yang, I. H., Shim, H. J., & Yu, H. J. (2018, April). A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5349-5353). IEEE.
23. Ravanelli, M., & Bengio, Y. (2018). Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725*.
24. Nunes, J. A. C., Macêdo, D., & Zanchettin, C. (2019, July). Additive margin sincnet for speaker recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-5). IEEE.

PRASANNA V. KOTHALKAR

Prasanna V. Kothalkar, received the B.S. degree in Computer Engineering from Mumbai University, Mumbai, India in 2010, M.S. degree in Computer Science from Univ. of Texas at Dallas (UT Dallas) in 2014. He received the merit-based Erik Jonsson Scholarship for incoming graduate students during his masters. He has interned at technology companies for research positions in the areas of Speech Processing and Machine Learning. Currently he is pursuing his Ph.D. degree as a Research Assistant in the Erik Jonsson School of Engineering and Computer Science, UT Dallas under supervision of Dr. John H. L. Hansen. His research interests focus on Speech Recognition and Diarization, Machine Learning and Deep Learning.

JOHN H.L. HANSEN
John H.L. Hansen, received Ph.D. & M.S. degrees from Georgia Institute of Technology, and B.S.E.E. degree from Rutgers Univ. At Univ. of Texas at Dallas (UT Dallas), he is Associate Dean for Research, Professor of Electrical & Computer Engineering, Distinguished Univ. Chair in Telecommunications Engineering, and holds a joint appointment in School of Behavioral & Brain Sciences (Speech & Hearing). At UT Dallas, he established Center for Robust Speech Systems (CRSS). He is an ISCA Fellow, IEEE Fellow, past Member and TC-Chair of IEEE Signal Proc. Society, Speech & Language Proc. Tech. Comm.(SLTC), and Technical Advisor to U.S. Delegate for NATO (IST/TG-01). He currently serves as ISCA President. He has supervised 92 PhD/MS thesis candidates, was recipient of 2020 UT-Dallas Provost's Award for Grad. Research Mentoring, 2005 Univ. Colorado Teacher Recognition Award, and author/co-author of +750 journal/conference papers in the field of speech/language/hearing processing & technology. He served as General Chair for Interspeech-2002, Co-Organizer and Tech. Chair for IEEE ICASSP-2010, and Co-General Chair and Organizer for IEEE Workshop on Spoken Language Technology (SLT-2014) (Lake Tahoe, NV). He is serving as Co-Chair for ISCA INTERSPEECH-2022, and Tech. Chair for IEEE ICASSP-2024.

DWIGHT IRVIN
As an assistant research professor at Juniper Gardens Children's Project (JGCP) at the University of Kansas, Dr. Dwight Irvin's research interests center on developing/refining measurement approaches to: 1) capture movement and location in young children at-risk for or with intellectual and developmental disabilities; 2) better understand and enhance the language environments that these children experience in school, home, and community settings. He directs and co-directs several research projects from federal agencies. For example, he leads an Institute for Education Sciences project titled, Validity Studies of the Classroom Code for Interactive Recording of Children's Learning Environments; this project focused on validating the use of the CIRCLE, an observational instrument designed to close the literacy gap for all learners. He also directs a National Science Foundation project focused on using existing wearable technology and advanced speech recognition/diarization algorithms to monitor student engagement over time in science activity areas in classroom and community-based settings.

JAY F. BUZHARDT
As an associate research professor at the Juniper Gardens Children's Project, Dr. Buzhardt's interests focus on investigating factors that impact the implementation, usability and effectiveness of technology-based intervention, assessment, and training. He has directed and co-directed several federally-funded research projects from NIH, IES, OSEP, NIDRR, and local foundations, including the following: He is co-developer and investigator of the OASIS (Online and Applied System for Intervention Skills) Training to train parents how to implement evidence-based practices with their young children with autism. He has been the lead scientist in the development and experimental evaluation of the web-based progress monitoring and decision-making tools for Infant and Toddler IGDIs. He is also co-developer and investigator of the Distance Mentorship Program, an online system to promote distance collaboration and training for school teams in rural areas who serve learners deaf-blindness.

BETH S. ROUS
Beth S. Rous is a professor in the Department of Educational Leadership Studies. She teaches courses in research methods. Her research agenda focuses on education policy, leadership and practices to support the design, implementation and large-scale implementation of programs. Her work has centered on cross-sector programs (education, health, and human services) designed to enhance the quality of services for vulnerable populations of children from birth through the early grades. Her research in this area has included accountability and standards, inclusion, online and hybrid learning, professional development, quality initiatives, and transition/school readiness. Beth also serves as a research and policy associate at the Human Development Institute, serving as the founding Director for the Kentucky Partnership for Early Childhood Services from 1996 through 2017. To date, Dr. Rous has generated over $97 million in grants and contracts to support her work. She regularly serves as a technical advisor/consultant at the national level for both the U.S Department of Education and Health and Human Services (e.g., National Study of IDEA Implementation; Pre-Elementary Education Longitudinal Study; Head Start Family Child Experiences Survey – FACES).