# Measuring Frequency of Child-directed WH-Question Words for Alternate Preschool Locations using Speech Recognition and Location Tracking Technologies

Prasanna V. Kothalkar*
Sathvik Datla
Satwik Dutta
John H. L. Hansen
prasanna.kothalkar@utdallas.edu
sathvik.datla@utdallas.edu
satwik.dutta@utdallas.edu
john.hansen@utdallas.edu
Center for Robust Speech Systems (CRSS), University of Texas at Dallas.
Richardson, Texas, USA

Yagmur Seven
Dwight Irvin
Jay Buzhardt
yagmur.seven@ku.edu
dwirvin@ku.edu
jaybuz@ku.edu
Juniper Garden's Children's Project (JGCP), University of Kansas.
Kansas City, Kansas, USA

## ABSTRACT

Speech and language development in children are crucial for ensuring effective skills in their long-term learning ability. A child's vocabulary size at the time of entry into kindergarten is an early indicator of their learning ability to read and potential long-term success in school. The preschool classroom is thus a promising venue for assessing growth in young children by measuring their interactions with teachers as well as classmates. However, to date limited studies have explored such naturalistic audio communications. Automatic Speech Recognition (ASR) technologies provide an opportunity for 'Early Childhood' researchers to obtain knowledge through automatic analysis of naturalistic classroom recordings in measuring such interactions. For this purpose, 225 hours of audio recordings across 48 daylong sessions are collected in a childcare learning center in the United States using Language Environment Analysis (LENA) devices worn by the preschool children. Approximately 29 hours of adult speech and 26 hours of child speech is segmented using manual transcriptions provided by CRSS transcription team. Traditional as well as End-to-End ASR models are trained on adult/child speech data subset. Factorized Time Delay Neural Network provides a best Word-Error-Rate (WER) of 35.05% on the adult subset of the test set. End-to-End transformer models achieve 63.5% WER on the child subset of the test data. Next, bar plots demonstrating the frequency of WH-question words in Science vs. Reading activity areas of the preschool are presented for sessions in the test set. It is suggested that learning spaces could be configured to encourage greater adult-child conversational engagement given such speech/audio assessment strategies.


*ICMI '21 Companion, October 18–22, 2021, MontrÃĭal, QC, Canada*

ACM ISBN 978-1-4503-8471-1/21/10…$15.00
https://doi.org/10.1145/3461615.3485440

## CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Applied computing** → **Interactive learning environments**;

## KEYWORDS

early childhood, education, neural networks, speech recognition

**ACM Reference Format:**
Prasanna V. Kothalkar, Sathvik Datla, Satwik Dutta, John H. L. Hansen, Yagmur Seven, Dwight Irvin, and Jay Buzhardt. 2021. Measuring Frequency of Child-directed WH-Question Words for Alternate Preschool Locations using Speech Recognition and Location Tracking Technologies. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, MontrÃĭal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3461615.3485440

## 1 INTRODUCTION

The diversity of language background, socio-economic conditions, development level, or potential communication disorders represents a challenge in assessment of child speech and language skills [17, 22]. The quality and amount of interaction in rich language environments help in meeting essential language development outcomes in early childhood [4]. Thus, early childhood researchers are focusing on analyzing classroom interactions of preschool children with adults (e.g., word counts, turn-taking, word diversity) to monitor and provide proactive support to them. In such settings, teachers prompt exploration by asking questions that engage the children's curiosity and experimentation, particularly in science-focused activities[3]. Thus, tracking sentences with these WH-question words or WH-words [18, 26] (what, where, when, who, why, how) can help teachers review their interactions with the children. Further, the WH-words representing the questions, can be analyzed in terms of frequency of occurrence based on location.

For this purpose, the authors have collected multi-session dataset in a real preschool during their daily activities. A typical preschool is composed of separate areas for specific activities to be conducted during alternate times of the day as seen in figure 1. Due to the extensive amount of daylong recordings to be analyzed, it is not

**Figure 1: Illustrative example of floor plan for child learning spaces within preschool classrooms. (i.e. learning stations: Books/Reading, Science etc.)**

feasible for humans to manually perform such analysis. Recently, speech recognition[10, 20] and machine learning [6, 11] techniques have been utilized for automated processing and analysis of child-centered data. Previously, diarization[8, 9, 12] and children's speech recognition[10] has been attempted on this dataset.

In this study we are focusing on recognizing location-specific WH-word frequency for adults using ASR. Adult speech trained models are also evaluated on children's speech for potential applications to child vocabulary and WH-word frequency measurement.

Next, an overview of this study is presented. In section 2, we look at the Speech and Location data specifics. Section 3 explains the Methodology including details of the acoustic and language models. Section 4 presents the experimental Results with a discussion about them. In section 5, Conclusion and Future Work is mentioned.

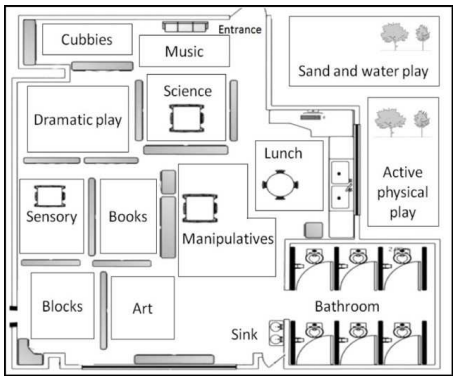


**Figure 2: Data collection platforms used in this study: i) Ubisense: RFID system location tracking, ii) LENA: Speech and language capture recorder.**

| Set | Adult Speech Duration | Child Speech Duration | Total Audio Duration |
|---|---|---|---|
| Train. | 18:52:17 | 16:38:26 | 106:11:19 |
| Dev. | 4:07:32 | 4:22:07 | 29:58:19 |
| Test. | 6:04:28 | 4:45:36 | 64:38:55 |
| Overall | 29:04:18 | 25:46:09 | 224:48:33 |

**Table 1: Setwise Database Details where the durations are in hours:minutes:seconds format.**

| WH-Word /Data | WHAT | WHEN | WHERE | HOW | WHY | WHO | Total |
|---|---|---|---|---|---|---|---|
| Adult Speech | 27/31 (755) | 2/7 (163) | 2/11 (152) | 6/27 (173) | 2/7 (97) | 2/5 (141) | 41/88 (1481) |
| Child Speech | 16/23 (412) | 0/6 (78) | 15/4 (97) | 5/11 (104) | 3/13 (107) | 1/3 (35) | 40/60 (833) |

**Table 2: Total count of WH-keywords as measured in manual transcripts of the Testing subset for both adult and child speech data in Science, Reading and all activity locations represented as Total $\text{Count}_{\text{Science}}$/Total $\text{Count}_{\text{Reading}}$ (Total $\text{Count}_{\text{All locations}}$).**

## 2 SPEECH AND LOCATION DATA SPECIFICS

The details for Kentucky Preschool child-adult naturalistic audio dataset used in this study, are mentioned below.

### 2.1 Speech data

*2.1.1 Kentucky Preschool Corpus.* The dataset in this paper consists of spontaneous conversational speech recorded with the help of lightweight compact digital audio recorder (LENA device[1, 27]) attached to children who are 3 to 5 year olds. The violet-colored LENA device is shown in figure 2 are lightweight causing minimal self awareness allowing voice capture of naturalistic conversations. Out of the 48 recording sessions recorded at a daycare center in the United States, some also have the devices worn by an adult. The recordings continue as subjects move around during a school day and are paused during nap time.

A total of 225 hours of speech and non-speech data was tagged by the CRSS transcription team at UT Dallas resulting in around 29 hours of adult speech data and 26 hours of child speech data. This was divided into training (Train.), development (Dev.) and testing (Test.) sets as seen in table 1, for initial investigation of a Deep Learning-based ASR system.

### 2.2 Location data

For the purpose a real-time classroom location tracking, a location data collection system (Ubisense device [25]) was worn by all participating children. Ubisense relies on RF receivers and transmitters, which communicate using ultra wide band radio frequencies to report the child's location every second. The device is shown in on the left half of figure 2 . These communications are logged by a separate computer running the Ubisense Location Engine software packages [25]. With proper calibration, the accuracy of Ubisense is +/- 15 cm under ideal measurement conditions, and +/- 30 cm in challenging measurement conditions [13]. Ubisense has been widely adopted in commercial and research endeavors [2, 16].

### 2.3 WH-word location-based manual transcription

The frequency of occurrence of WH-words in science, reading as well as all locations, is specified for both adult as well as children's

speech in table 2. It is interesting to observe that number of WH-words in science and reading locations comprised 8.7% of WH-word occurrences in all locations for adults, but 12.0% of WH-word occurrences in all locations for children. Thus, children spoke a slightly higher percentage of their WH-words in science and reading locations Vs. all locations in comparison with adults, however the context of these occurrences is not investigated in this study.

# 3 METHODOLOGY

In this section, we present the Acoustic Model (AM) and the Language Model (LM) used to train the ASR system.

## 3.1 Acoustic models

Current AMs in both End-to-End as well as Hybrid ASR with strong performance on standard Large Vocabulary Continuous Speech Recognition (LVCSR) datasets are utilized for Kentucky data. End-to-End systems have widely used Sequence-to-Sequence (S2S) models for their natural application of speech to text conversion as they perform source sequence conversion to target sequence. These include conventional Recurrent Neural Network (RNN) and Transformers and here Transformers have outperformed RNN on several benchmark datasets [5]. The standard hybrid model consists of n-gram language model and a Time Delay Neural Network (TDNN).

*3.1.1 CTC-Attention.* Connectionist Temporal Classification, also refferred as CTC, is a type of neural network output for a scoring function, in training S2S models. It relieves the requirement of one-to-one mapping with the help of a 'blank' symbol, indicating no label is seen. Using this, it can compute the probability of all possible alignment paths taken during training for calculating the probability of the entire target transcription. The attention based models in contrast incorporates contextual information using both input frames and history of the target label for the inference process. The role of attention mechanism here is to find alignment between input acoustic frames and text output. Attention-based approaches are harder to train and susceptible to noise. Mostly a combination of the two approaches in the form of multitask learning [7] is used.

*3.1.2 Transformer.* Transformers[5] were initially proposed for Machine Translation task[21] and have been adapted for Speech Recognition. It learns sequential information via Self-Attention mechanism instead of recurrent connection employed in RNN. It consists of multiple Dot-Attention layers. A Multi-Head Attention (MHA) block allows the model to deal with multiple attentions in parallel. It is used in both the Self-Attention Encoder and Self-Attention Decoder networks. The MHA block between these two networks is referred to as Encoder-Decoder Attention. It is trained using CTC-attention based approach.

*3.1.3 TDNN-F model.* TDNNs are 1-dimensional Convolutional Neural Networks that performed well for ASR. Factored form of TDNNs known as TDNN-F[14], were introduced for improved performance with lower footprint. TDNN-F models have same structure as TDNNs but have their layers compressed by Singular Value Decomposition while ensuring one of the factors of each matrix are constrained to be semi-orthogonal. TDNN-F model with skip connections provide best results with Mel-Frequency Cepstral Coefficient (MFCC) features.

## 3.2 Language models

SRILM toolkit[19] is used for building and evaluating statistical LMs based on backoff models, using standard smoothing algorithms. SRILM toolkit along with CMU dictionary was used for developing the 4-gram LM for TDNN-F AM using text from the training data. RNN LM trained with the train subset was used with the various End-to-End AMs for model rescoring.

## 3.3 Experimental Details

The Kaldi[15] and Espnet[23, 24] recipes perform speech perturbation before training DNN models, so the final data for training was thrice the training data listed in table 1. This resulted in approximately 54 hours of Kentucky adult data and 50 hours of Kentucky child data for training the respective models.

*3.3.1 End-to-End models for testing on child/adult test data.* End-to-End Transformer model utilized 80-dimensional Mel-Filterbank (FBANK) as input features to the model. They were trained for 200 epochs with patience of 15 for early stopping. The best performing model on development set was evaluated on the testing set.

*3.3.2 Hybrid models for testing on adult test data.* Hybrid TDNN-F model utilized 39-dimensional MFCC as input for the model. They were trained for 4 epochs consisting of 168 iterations in all. The models trained on the last 20 iterations were combined for evalution on the development and test set.

*3.3.3 Evaluation metrics.* ASR system performance is measured in terms of Word-Error-Rate(%) which is defined as the percentage of substitutions, insertions and deletions in the recognized text transcript compared to the ground truth transcription.

Next we define the terms of Precision, Recall and F-score for our task of predicting WH-words correctly by the ASR system vs. the actual transcript. Precision is defined as the fraction of correct predictions of a WH-word with all the predictions of that WH-word as given by the ASR system. Recall is defined as the fraction of relevant WH-words in the ground truth transcript, that are predicted correctly by the ASR system. F-score is defined as the harmonic mean of Precision and Recall.

# 4 RESULTS AND DISCUSSION

| ASR Model | Features | Train Data | Test Data | Dev-set WER (%) | Test-set WER (%) |
|---|---|---|---|---|---|
| Transformer | FBANK | Adult | Adult | 47.3% | 53.9% |
| TDNN-F | MFCC | Adult | Adult | **28.82%** | **35.05%** |
| Transformer | FBANK | Child | Child | 57.7% | 68.3% |
| Transformer | FBANK | Adult+Child | Child | **51.7%** | **63.5%** |

**Table 3: Automatic Speech Recognition system results on development and test sets of adult speech data.**

## 4.1 Adult data ASR performance

As expected, Transformers perform much worse than TDNN-F model on the Kentucky adult data as seen in table 3. It provides us a baseline performance for an end-to-end system. The Hybrid

| WH-Word / Location | WHAT | WHEN | WHERE | HOW | WHY | WHO |
|---|---|---|---|---|---|---|
| Science | 65.2% | **100.0%** | 50.0% | 83.3% | **100.0%** | 66.7% |
| Reading | **85.3%** | 92.3% | **81.8%** | **94.1%** | 92.3% | **75.0%** |
| All | 72.2% | 69.0% | 71.0% | 72.8% | 52.4% | 73.2% |

**Table 4: F1-scores for correctly predicting WH-keywords using Automatic Speech Recognition system for both adult speech data in Science, Reading and all activity locations.**

| WH-Word / Location | WHAT | WHEN | WHERE | HOW | WHY | WHO |
|---|---|---|---|---|---|---|
| Science | 41.5% | 0% | **53.3%** | 28.6% | 50.0% | 0% |
| Reading | **56.0%** | **60%** | 40.0% | **58.9%** | **66.6%** | 0% |
| All | 47.3% | 39.1% | 30.2% | 44.5% | 56.0% | **39.8%** |

**Table 5: F1-scores for correctly predicting WH-keywords using Automatic Speech Recognition system for child speech data in Science, Reading and all activity locations.**

TDNN-F model is able to achieve decent performance based on the strength of it's language model in combination with a powerful acoustic model.

## 4.2 Child data ASR performance

For the children's speech, enough triphones are not present in this combination of training dataset. Hence, only End-to-End models are trained for evaluating children's speech and results are presented in table 3. Training models on combined adult and children's speech component of the training data, improves the performance of the model on development set (6% absolute) and test set (4.8% absolute).

## 4.3 F1-scores for WH-word recognition in Science and Reading activity locations

Most of the F1-scores for recognition of WH-words are better or equivalent in Reading zones compared to Science zones as seen in tables 4 and 5. A contributing factor could be that the count of WH-words in reading zones is more than twice the count of WH-words in science zones, as seen in table 2. The highest occurring WH-word 'What' in both adult as well as children's speech and across all locations, has better F1-scores for Reading, followed by all locations and lastly Science location. This maybe due to the environmental acoustic conditions created by the activities conducted in these locations, resulting in better performing ASR system in Reading locations.

## 4.4 WH-word frequency bar plots

There are 41 occurrences of WH-words in Science areas and 88 occurrences of WH-words in Reading areas for the adult speech test subset as seen in table 2. The ASR system predicts 26 WH-words in Science areas and 77 WH-words in Reading areas correctly. Figure 3 shows the bar plots for the frequency of occurrence of WH-words in Science vs. Reading areas using manual transcription (actual) as well as ASR prediction (predicted) i.e. Recall for the actual occurrences. Here, 'What' has the highest frequency in both the areas, followed

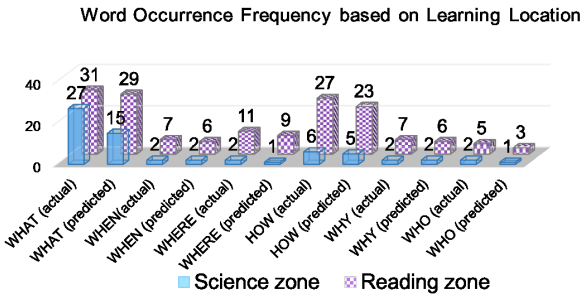


**Figure 3: Distribution bar plot of actual and predicted adult word frequency for Science vs. Reading based activity area(e.g. defined locations in child learning space; see Fig. 1); for sessions in the test set.**

by 'How'. But the actual count of 'How' keyword is much less in Science location compared to Reading location, which indicates except for the most frequent WH-word 'What', other WH-words were far fewer in science zones for the test subset.

The ASR system has closely predicted the actual WH-word frequencies (i.e. Recall) in Reading locations, even for low frequency WH-words like 'When', 'Where', 'Why' and 'Who'. Except for the high frequency word 'What', the ASR system performs equally well in terms of Recall for 'Science' as well as 'Reading' locations.

# 5 CONCLUSION AND FUTURE WORK

In preschool classroom, teachers prompt exploration by asking questions that engage the children's attention. Thus, tracking sentences with these WH-words can help teachers review their interactions with the children. To achieve this goal in an automated fashion, ASR systems with good performance on benchmark datasets are evaluated for a naturalistic audio dataset collected at a preschool daycare center in the United States. TDNN-F Hybrid ASR model achieved the best WER performance overall while Transformer-based ASR model achieved best performance for End-to-End ASR systems on adult speech. End-to-End speech recognition for children's speech showed potential for future applications by utilizing adult speech in the training. Bar plots for WH-word frequency are plotted for Science vs. Reading activity locations and followed the pattern of the manual transcripts closely, which was confirmed by the F-scores. Utilization of more advanced models and training strategies will help in improving the ASR system performance. Introducing a diarization pre-processing component in the pipeline, will help in transitioning this research system for real-world application of providing feedback to educators.


# ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) under Grant Awards #1918032, #1918012. The authors would like to thank Dr. Beth S. Rous, Dr. Dwight Irvin and Ying Luo at University of Ketucky for establishing the protocol and collecting the dataset. Thanks to the CRSS transcription team for their dedicated transcription effort and Satwik Dutta at CRSS and Kyle Consolver at JGCP for managing the Institutional Review Board documentation.

# REFERENCES


[1] [n. d.]. https://www.lenafoundation.org. ([n. d.]).

[2] Damien Connaghan, Sarah Hughes, Gregory May, Philip Kelly, Ciarán Ó Conaire, Noel E O'Connor, Donal O'Gorman, Alan F Smeaton, and Niall Moyna. 2009. A sensing platform for physiological and contextual feedback to tennis athletes. In *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*. IEEE, 224–229.

[3] Catherine Eberbach and Kevin Crowley. 2017. From seeing to observing: How parents and children learn to see science in a botanical garden. *Journal of the Learning Sciences* 26, 4 (2017), 608–642.

[4] Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

[5] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs RNN in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 449–456.

[6] Heysem Kaya, Oxana Verkholyak, Maxim Markitantov, and Alexey Karpov. 2020. Combining clustering and functionals based acoustic feature representations for classification of baby sounds. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 509–513.

[7] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4835–4839.

[8] Prasanna Vasant Kothalkar, John HL Hansen, Jay Buzhardt, Dwight Irvin, and Beth S Rous. 2021. Child vs Adult Speaker Diarization of naturalistic audio recordings in preschool environment using Deep Neural Networks. In *ASEE 2021 Gulf-Southwest Annual Conference*.

[9] Prasanna V Kothalkar, Dwight Irvin, Ying Luo, Joanne Rojas, John Nash, Beth Rous, and John HL Hansen. 2019. Tagging child-adult interactions in naturalistic, noisy, daylong school environments using i-vector based diarization system. In *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*. 89–93.

[10] Rasa Lileikyte, Dwight Irvin, and John HL Hansen. 2020. Assessing Child Communication Engagement via Speech Recognition in Naturalistic Active Learning Spaces. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*. 396–401.

[11] Elena E Lyakso and Olga V Frolova. 2020. Early Development Indicators Predict Speech Features of Autistic Children. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 514–521.

[12] Maryam Najafian, Dwight Irvin, Ying Luo, Beth S Rous, and John HL Hansen. 2016. Automatic measurement and analysis of the child verbal communication using classroom acoustics within a child care center.. In *WOCCI*. 56–61.

[13] Terry Phebey. 2010. The Ubisense assembly control solution for BMW. In *RTLS in Manufacturing Workshop-RFID Journal Europe Live*.

[14] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks.. In *Interspeech*. 3743–3747.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[16] Timothy H Riehle, P Lichter, and Nicholas A Giudice. 2008. An indoor navigation system to support the visually impaired. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 4435–4438.

[17] Sara Rosenbaum and Patti Simon. 2016. *Speech and Language Disorders in Children: Implications for the Social Security Administration's Supplemental Security Income Program*. ERIC.

[18] Meredith L Rowe, Kathryn A Leech, and Natasha Cabrera. 2017. Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development. *Cognitive science* 41 (2017), 162–179.

[19] Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of IEEE automatic speech recognition and understanding workshop*, Vol. 5.

[20] Anika van der Klis, Frans Adriaans, Mengru Han, and René Kager. 2020. Automatic Recognition of Target Words in Infant-Directed Speech. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 522–522.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[22] Lynne Vernon-Feagans, Mary Bratsch-Hines, Elizabeth Reynolds, and Michael Willoughby. 2020. How early maternal language input varies by race and education and predicts later child language. *Child development* 91, 4 (2020), 1098–1115.

[23] Shinji Watanabe, Florian Boyer, Xuankai Chang, Pengcheng Guo, Tomoki Hayashi, Yosuke Higuchi, Takaaki Hori, Wen-Chin Huang, Hirofumi Inaguma, Naoyuki Kamo, et al. 2020. The 2020 ESPnet update: new features, broadened applications, performance improvements, and future plans. *arXiv preprint arXiv:2012.13006* (2020).

[24] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015* (2018).

[25] Marek Woźniak, Waldemar Odziemczyk, and Kamil Nagórski. 2013. Investigation of practical and theoretical accuracy of wireless indoor positioning system Ubisense. *Reports on Geodesy and Geoinformatics* 95, 1 (2013), 36–48.

[26] Paul J Yoder, Betty Davies, Kerri Bishop, and Leslie Munson. 1994. Effect of adult continuing wh-questions on conversational participation in children with developmental disabilities. *Journal of Speech, Language, and Hearing Research* 37, 1 (1994), 193–204.

[27] Ali Ziaei, Abhijeet Sangwan, and John HL Hansen. 2013. Prof-Life-Log: Personal interaction analysis for naturalistic audio streams. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7770–7774.