



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Ultrasound tongue imaging for diarization and alignment of child speech therapy sessions

Citation for published version:

Ribeiro, M, Eshky, A, Richmond, K & Renals, S 2019, Ultrasound tongue imaging for diarization and alignment of child speech therapy sessions. in INTERSPEECH 2019: Proceedings of the 20th Annual Conference of the International Speech Communication Association (ISCA). International Speech Communication Association, Graz, Austria, pp. 16-20, Interspeech 2019, Graz, Austria, 15/09/19. <https://doi.org/10.21437/Interspeech.2019-2612>

Digital Object Identifier (DOI):

[10.21437/Interspeech.2019-2612](https://doi.org/10.21437/Interspeech.2019-2612)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

INTER_SPEECH 2019: Proceedings of the 20th Annual Conference of the International Speech Communication Association (ISCA)

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Ultrasound tongue imaging for diarization and alignment of child speech therapy sessions

Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond, Steve Renals

The Centre for Speech Technology Research, University of Edinburgh, UK

{sam.ribeiro, aeshky, korin.richmond, s.renals}@ed.ac.uk

Abstract

We investigate the automatic processing of child speech therapy sessions using ultrasound visual biofeedback, with a specific focus on complementing acoustic features with ultrasound images of the tongue for the tasks of speaker diarization and time-alignment of target words. For speaker diarization, we propose an ultrasound-based time-domain signal which we call estimated tongue activity. For word-alignment, we augment an acoustic model with low-dimensional representations of ultrasound images of the tongue, learned by a convolutional neural network. We conduct our experiments using the Ultrasuite repository of ultrasound and speech recordings for child speech therapy sessions. For both tasks, we observe that systems augmented with ultrasound data outperform corresponding systems using only the audio signal.

Index Terms: speech recognition, speaker diarization, ultrasound tongue imaging, speech therapy, child speech, Ultrasuite

1. Introduction

Developmental speech sound disorders (SSDs) are a common communication impairment in childhood [1]. Children with SSDs consistently exhibit difficulties in the production of specific speech sounds in their native language. Speech disorders have the potential to negatively affect the lives and the development of children. For example, self-awareness of disordered speech may lead to low-confidence in social situations or introduce communication barriers that lead to increased difficulty in learning and decreased literacy levels [2, 3, 4].

Clinical intervention is available for these children. Auditory-based techniques are often efficient for pre-schoolers, but for older children with persistent disorders such methods may be unsuccessful [5]. For these cases, there is growing evidence that visual biofeedback (VBF) is helpful [6, 7, 8, 9]. VBF allows the visualization of the vocal tract during the speech production process, enabling patients to correct inaccurate articulations in real-time. Ultrasound VBF is a clinically safe and non-invasive method that uses standard medical ultrasound to visualize tongue movements [5, 9, 10, 11]. Besides being useful for patients, ultrasound tongue imaging (UTI) helps therapists in the assessment and diagnosis of SSDs, as it provides information not available in the acoustic signal (e.g., the presence of double articulations or undifferentiated lingual gestures).

The automatic processing of speech therapy data can be helpful to patients and therapists. Children can use screening tools at an early age to determine whether they need to be assessed by a therapist [12, 13]. Instrumented methods such as spectrogram or ultrasound analysis can be used to assist the therapist in the assessment, diagnosis, or quantification of treat-

ment efficacy. However, current practice relies on manual annotation by the therapist or other trained professionals. Examples include the identification of boundaries for target words or phones, using the spectrogram or the ultrasound signal. This is a laborious process that could be alleviated using automated methods.

There are several challenges associated with the automatic processing of child speech therapy sessions. They contain dialogue between therapist and child, extraneous child speech, multiple attempts at pronouncing target words, or mispronunciations due to SSDs. Although there is knowledge of target words, full transcriptions of the recorded speech are not available. Additionally, there are various other challenges associated with child speech processing [14], disordered speech processing [15], and ultrasound image processing [16].

In this work, we are concerned with the speaker diarization and the time-alignment of target words from child speech therapy sessions using U-VBF. Robust methods to solve these tasks can alleviate the manual workload of speech therapists, while having the added benefit of preparing data for further processing, such as the development of screening tools. It is the goal of this paper to investigate methods that *complement the audio signal with ultrasound images of the tongue*. For diarization, we propose a time-domain signal which we call *estimated tongue activity*. For word-alignment, we augment an acoustic model with low-dimensional representations of ultrasound tongue images learned by a convolutional neural network. Section 2 introduces our dataset, while Sections 3 and 4 describe our methods for speaker diarization and word alignment respectively. Section 5 provides a discussion and conclusion.

2. The UltraSuite repository

UltraSuite¹ [17] is a repository of synchronized ultrasound and audio data from child speech therapy sessions. The repository currently contains three data collections. One includes recordings from 58 typically developing children: Ultrax Typically Developing (UXTD). Additional collections include data from children with speech sound disorders recorded over the course of assessment and therapy sessions: Ultrax Speech Sound Disorders (UXSSD, 8 children) and Ultraphonix (UPX, 20 children). Each waveform is accompanied by a prompt and corresponding ultrasound recording. The prompt includes a list of target words to be elicited by the child and does not correspond to a direct transcription of the utterance. Recordings may include therapist intervention to encourage or guide the patient or extraneous speech by the child (e.g. false starts, multiple attempts at a target word, or unrelated speech). Ultrasound was recorded using an Ultrasonix SonixRP machine using Articulate Assistant Advanced (AAA) software at ~ 121 fps with a 135°

Supported by the EPSRC Healthcare Partnerships Programme grant number EP/P02338X/1 (Ultrax2020).

¹<http://www.ultrax-speech.org/ultrasuite>

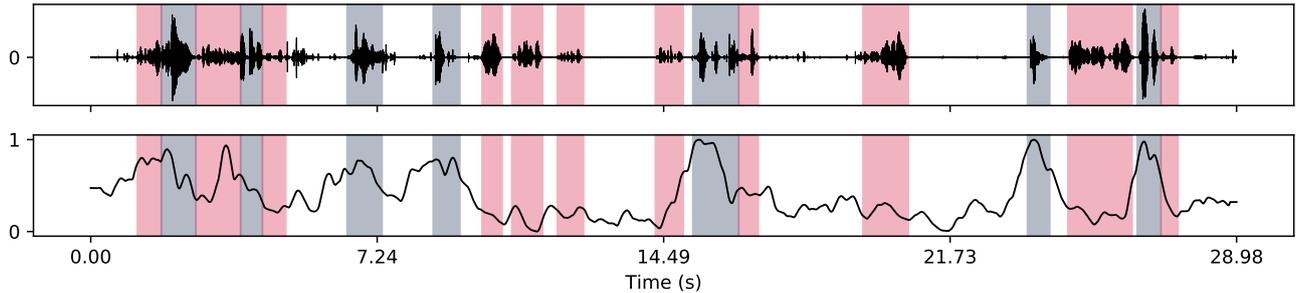


Figure 1: Sample waveform (top) with corresponding unity-based normalized estimated tongue activity (bottom). Highlighted segments denote speech from the child (blue) and the therapist (red).

field of view. A single B-Mode ultrasound frame has 412 echo returns for each of 63 scan lines, giving a 63x412 “raw” ultrasound frame that captures a midsagittal view of the tongue.

3. Speaker diarization

Speaker diarization aims to identify “who spoke when” [18]. For our scenario, the goal is to identify speech segments belonging to the child or the therapist. The difficulty of this task is increased due to very short speech segments with very fast turns between speakers [19]. In this section, we compare baselines based on voice activity detection (VAD), diarization methods using i-vectors [20], and hidden Markov model (HMM) / Gaussian mixture model (GMM) systems bootstrapped by a small set of labeled data [21]. One of our key contributions is the inclusion of the ultrasound signal via *estimated tongue activity*. For evaluation, we use a set of 120 utterances (60 from typically developing and 60 from speech disordered speakers), which have been manually annotated. We measure precision and recall in terms of retrieved child speech and diarization error rate (DER) in terms of silence, child, and therapist speech. Evaluation metrics are computed in terms of time in seconds using a 100ms collar with *pyannote.metrics* [22].

3.1. Estimated tongue activity (ETA)

The ultrasound signal visualizes the vocal tract of the child, but it does not capture information regarding other speakers in the room. We propose a time domain signal which we call *estimated tongue activity* (ETA). This is computed given a sliding time window over the ultrasound frames. The variance of each echo return is taken over time, and the overall mean variance of all echo returns is computed. Higher variance indicates that the content of the ultrasound frames changes rapidly over time. Figure 1 illustrates the ETA signal, considering a sliding window of approximately 160 milliseconds. Child speech (highlighted in blue) generally occurs with high variance, while therapist speech (in red) occurs with low variance.

3.2. Experiments and results

Baselines. We consider threshold-based baselines using frame-level energy-based voice activity detection (VAD), available in Kaldi [23], and the proposed estimated tongue activity (ETA). In these systems, we use a threshold of 7 to perform VAD and a threshold of 0.5 to perform tongue activity detection using ETA. For the method using VAD, we simply term all activity above the threshold as child speech. For the method combining both signals (VAD + ETA), all frames below the VAD threshold are taken as silence and frames that are above the threshold are

Table 1: Results for speaker diarization for typically developing speech (UXTD) and disordered speech (UXSSD). Precision, recall, and F_1 score are measured in terms of retrieved child speech. Diarization Error Rate (DER) measures silence and speech, while Confusion Error (Conf) measures child and therapist speech error. Highlighted results indicate best systems.

Systems	Prec	Rec	F_1	DER	Conf
Typically developing speech					
VAD	0.89	0.93	0.91	18.7	1.5
VAD+ETA	0.92	0.66	0.77	37.1	19.9
PLDA i-vectors (mfcc)	0.89	0.80	0.85	23.3	6.2
+ f_0	0.89	0.70	0.78	26.8	9.8
HMM-GMM (mfcc)	0.86	0.72	0.79	38.0	0.8
+ f_0	0.83	0.81	0.87	26.0	0.2
+ f_0 + ETA	0.93	0.90	0.92	17.1	0.4
+ f_0 + ETA + semi-sup	0.98	0.84	0.90	18.5	0.0
Disordered speech					
VAD	0.55	0.90	0.68	47.6	29.7
VAD+ETA	0.81	0.76	0.79	32.0	14.1
PLDA i-vectors (mfcc)	0.67	0.71	0.69	32.4	14.6
+ f_0	0.69	0.75	0.72	34.4	16.6
HMM-GMM (mfcc)	0.57	0.71	0.63	52.5	12.5
+ f_0	0.82	0.81	0.82	29.1	8.6
+ f_0 + ETA	0.81	0.89	0.85	24.0	8.4
+ f_0 + ETA + semi-sup	0.95	0.81	0.87	28.2	0.2

further processed using the ETA signal. Those frames above the ETA threshold are identified as child speech, while those below are assumed to be therapist speech.

i-vectors. This baseline follows the method presented by Sell and Garcia-Romero [20], available in Kaldi [23]. We use the PF-STAR children’s speech corpus [24] (7.4 hours, 86 speakers) to train a Universal Background Model (GMM with 512 components) and extract 64 dimensional i-vectors over 1.5 second windows. We use in-domain heldout data from UltraSuite to normalize the training data for the probabilistic linear discriminant analysis (PLDA) scoring and perform agglomerative clustering assuming 2 speakers in each recorded session.

HMMs. Our proposed HMM approaches require less data than typical i-vector clustering approaches. We use the transcriptions available with the UXTD dataset, which contain tags denoting whether a word is spoken by the therapist or the child [17]. Each word is reduced to “child” or “therapist” tokens, corresponding to speaker turns in each utterance. These are modelled with 5 state ergodic HMMs with 1000 Gaussian com-

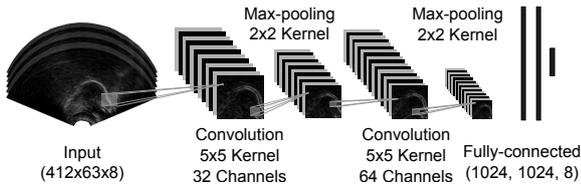


Figure 2: Convolutional neural network architecture for ultrasound embedding extraction.

ponents. We allow HMMs for child, therapist, and optional silence and noise. Turn-taking sequences are derived from a transcription and there is no initial alignment with respect to the features. We evaluate this method using a combination of 20 MFCCs, 3 f_0 features [25], and the ETA signal. Because we bootstrap HMM/GMMs with a small set of annotated data (1.75 hours), we further investigate a *semi-supervised* approach incorporating additional unlabeled disordered speech data, in which a model trained on typically development data is used to decode the UPX dataset (11.05 hours), with the hypothesized labels then used to retrain the HMM/GMM.

For all systems, a post-processing step merges identical labels separated by silence shorter than 100ms and removes labels with duration less than 50ms.

Results are presented in Table 1. The simple baseline approaches perform well for the typically developing dataset, which has minimal adult intervention, and therefore less need for diarization [17]. The loss in recall with ETA thresholding is caused by losing segments with reduced tongue activity (e.g. words such as “apa”). In the case of the disordered speech dataset, therapist intervention is more frequent, and ETA is a useful addition to VAD.

The i-vector system does not perform as well as expected. This might be due to size of the database used to train the UBM, but also to the challenges of fast speaker turns. Similar results were found using a large adult database (NIST SRE datasets, similar to [20]). We hypothesize that the i-vectors are not robust enough, as we enforce a very small window due to the rapid turns between speakers. High error rates have previously been found for a similar scenario [19]. Allowing a longer window would lead to equally poor i-vectors, as they would be computed over a mixture of adult and child speech. Surprisingly, including f_0 was not helpful using this method.

The proposed HMM-GMM systems achieved better results, especially when using additional features. The use of f_0 leads to good improvements, which are complemented by the inclusion of the ETA signal. Overall, a basic HMM-GMM using a combination of all features achieves good performance on both datasets. The semi-supervised approach leads to very low speaker confusion, but DER is increased due to missed cases. Although the model becomes stronger at differentiating speakers, it does so at the expense of the silence and noise models. This can also be seen from the lower recall scores.

4. Word alignment

Word alignment involves time-aligning target words, given in the prompt, with the speech signal. There are various challenges associated with this task when applied to speech therapy sessions. This is noticeable when force-aligning the audio with the expected prompt, using baseline standard methods. In this section, we complement the audio signal with ultrasound images of the tongue for acoustic modeling.

Table 2: Results for word alignment (precision, recall, F_1) and recognition (WER) for typically developing speech (UXTD) and disordered speech (UXSSD). Baseline systems use GMMs trained on Ultrasuite and PF-Star data. USuite systems use neural network acoustic models. Highlighted results indicate best performing systems.

Systems	Pre	Rec	F_1	WER
Typically developing speech				
Baseline (no diarization)	0.88	0.90	0.89	36.29
Baseline	0.89	0.91	0.90	37.18
USuite	0.86	0.92	0.89	31.35
USuite+PFstar	0.85	0.92	0.89	30.96
USuite+UTI	0.86	0.92	0.90	31.22
System combination ($\alpha = 0.6$)	0.86	0.93	0.89	28.68
System combination ($\alpha = 0.7$)	0.86	0.93	0.89	29.44
Disordered speech				
Baseline (no diarization)	0.53	0.46	0.50	71.19
Baseline	0.77	0.68	0.72	69.24
USuite	0.75	0.70	0.73	63.21
USuite+PFstar	0.76	0.72	0.74	59.35
USuite+UTI	0.75	0.90	0.72	62.73
System combination ($\alpha = 0.6$)	0.77	0.73	0.75	59.35
System combination ($\alpha = 0.7$)	0.77	0.72	0.75	58.14

4.1. Ultrasound embeddings

We augment the audio features with ultrasound embeddings extracted from a convolutional neural network (CNN). The network uses the UXTD dataset as training data and follows the architecture and findings reported by [26], illustrated in Figure 2. Input consists of 8 ultrasound frames, grouped as multiple channels. These are: the current frame, 3 left and right neighboring frames, and the speaker mean. Two sets of convolution and max-pooling layers are followed by three fully-connected layers. The final layer of the network is 8 dimensional, and we let this representation be the ultrasound embedding for the current timestamp. Output classes roughly correspond to place of articulation, defined over the entire phoneset.

4.2. Acoustic model training and evaluation

We train the acoustic models with the Kaldi speech recognition toolkit [23]. GMMs are initialized on pooled data from UXTD (1.75 hours), UPX (11.05 hours), and the training set of the PF-STAR children’s speech corpus (7.4 hours) [24]. The inclusion of the PF-STAR corpus is beneficial for model initialization as it contains additional manually transcribed data for 86 speakers.² For UXTD and UPX data, we remove speech frames associated with the therapist using the system with the best DER from Section 3. This way, we avoid aligning target words with time segments belonging to the speech therapist. After monophone and triphone training, Mel Frequency Cepstral Coefficients (MFCCs) are processed with Linear Discriminant Analysis (LDA) and a Maximum Likelihood Linear Transform (MLLT). This is followed by Speaker Adaptive Training (SAT) with feature-space MLLR (fMLLR) [27, 28]. This HMM-GMM system is denoted **Baseline** in Table 2. Alignment

²We have attempted to initialize models using only in-domain data, but including the PF-STAR corpus lead to a stronger initialization at this stage. We therefore opt to treat that as our baseline and we conduct our analyses over the neural network acoustic models.

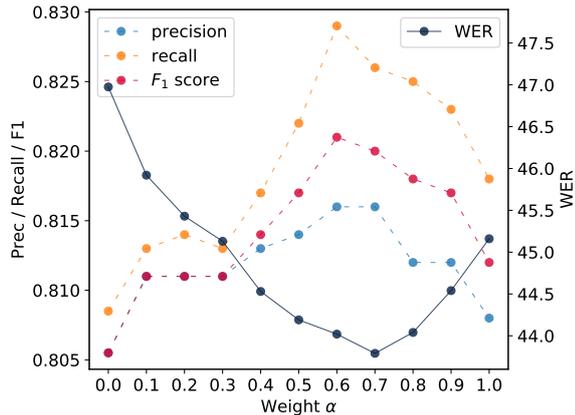


Figure 3: *Combination of models augmented with out-of-domain acoustic data (USuite+PFstar) with weight α and augmented with ultrasound embeddings (USuite+UTI) with weight $(1 - \alpha)$. Results are averaged across typically developing and speech disordered evaluation sets.*

and fMLLR features from this system are then used to train feedforward neural networks, following Kaldi’s *nnet1* recipe.

For evaluation, we use a set of 398 utterances (199 from typically developing speakers, and 199 from speakers with speech disorders), manually annotated with word boundaries from target words given by the prompt. Precision and recall are measured in terms of correctly retrieved time with respect to the manual gold standard, given matching word labels [22]. Additionally, in order to further quantify the robustness of the acoustic models, we consider an experimental scenario where we decode within oracle word boundaries. For this case, we use a uniform unigram language model, built from the joint vocabulary of the UXTD/UXSSD datasets (1161 words). Results are reported in terms of word error rate (WER).

4.3. Experiments and results

We compare various neural network acoustic models. The first uses only acoustic in-domain data from speech therapy sessions (UXTD and UPX training data), termed **USuite**. This training set is also augmented with out-of-domain acoustic data from the PF-STAR children’s speech corpus (**USuite+PFstar**). We augment the in-domain system with the ultrasound embeddings described in Section 4.1 using a context of 4 frames. Acoustic and ultrasound features are concatenated for the neural network (**USuite+UTI**). Ideally we would include all data sources, but ultrasound does not exist in the PF-STAR corpus. Therefore, we combine the two systems by interpolating their posterior probabilities, with weight α for USuite+PFstar and $(1 - \alpha)$ for USuite+UTI (**System combination**).

Results are presented in Table 2. Diarization is not beneficial for the typically developing dataset, as it has a small amount of adult speech. This is not the case for disordered speech, where diarization leads to more accurate alignment and recognition. Although baseline GMMs perform reasonably well in terms of alignment, they do not outperform neural network acoustic models in terms of recognition. Augmenting in-domain data with ultrasound embeddings leads to improvements on both datasets, but not as much as using additional out-of-domain acoustic data. The best system is a combination of the systems augmented with out-of-domain data and ultra-

Table 3: *Results for combination of models with $\alpha = 0.7$ for disordered speech (UXSSD), grouped by assessment session at various stages of therapy. Post-therapy is recorded immediately after therapy and Maintenance several months after.*

Sessions	Pre	Rec	F ₁	WER
Baseline	0.77	0.72	0.75	62.01
Mid-therapy	0.82	0.76	0.80	60.90
Post-therapy	0.80	0.74	0.77	47.89
Maintenance	0.71	0.69	0.70	60.35
Global	0.77	0.72	0.75	58.14

sound embeddings. Although we observe small improvements in terms of alignment metrics, WER shows that combined models are more robust than either of them separately. Figure 3 shows results for model combination across the range of α values. Best results occur when there is a slight bias towards the system augmented with out-of-domain training data. These results also show that ultrasound and audio data can complement each other well.

5. Discussion and conclusions

Speech therapy using U-VBF is beneficial for patients, who may visualize their articulators during speech production, and therapists, who may use ultrasound for more accurate diagnoses and treatments. The automatic processing of speech therapy sessions provides additional benefits by alleviating some of the time-consuming manual tasks undertaken by therapists, as well as preparing data for further processing. In this paper we proposed methods that use ultrasound tongue imaging to complement the audio signal for speaker diarization and the time-alignment of target words. For both tasks, we have observed improvements in models augmented with ultrasound when compared with models using only audio recordings.

For speaker diarization, we have proposed a measure of Estimated Tongue Activity (ETA) which we have used for tongue activity detection. In the future, this signal could be useful to measure patients’ overall tongue activity during therapy sessions. For example, to investigate whether children shadow therapists or whether they exhibit a larger amount of tongue movement after therapy. One of the main issues for diarization, however, is the lack of annotated training data. These methods are designed to be used by therapists during and after assessment sessions, and they could learn from user input. For example, a single session typically contains the same two speakers. Adaptation based on a small amount of data annotated by the therapist could help with speaker diarization.

In terms of alignment, we showed that a combination of models augmented with out-of-domain acoustic data and in-domain ultrasound data led to more robust acoustic models. Table 3 shows results by assessment session at various stages of therapy. These results should not be directly compared, as they contain different target words, but they do illustrate the temporal dependency of therapy sessions, which could be used for longitudinal online learning [29]. Additionally, there are various issues that were not addressed by this paper. Future work may consider learning speaker- or session-specific pronunciations [30], as baseline sessions are expected to have different pronunciations than post-therapy or maintenance sessions. Furthermore, insertions and deletions with respect to the prompt could be handled by relaxing the constraints of the linear transducer used for alignment [31].

6. References

- [1] Y. Wren, L. L. Miller, T. J. Peters, A. Emond, and S. Roulstone, "Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 4, pp. 647–673, 2016.
- [2] C. J. Johnson, J. H. Beitchman, and E. Brownlie, "Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes," *American Journal of Speech-Language Pathology*, 2010.
- [3] B. A. Lewis, A. A. Avrich, L. A. Freebairn, A. J. Hansen, L. E. Sucheston, I. Kuo, H. G. Taylor, S. K. Iyengar, and C. M. Stein, "Literacy outcomes of children with early childhood speech sound disorders: Impact of endophenotypes," *Journal of Speech, Language, and Hearing Research*, 2011.
- [4] J. McCormack, L. J. Harrison, S. McLeod, and L. McAllister, "A nationally representative study of the association between communication impairment at 4–5 years and children's life activities at 7–9 years," *Journal of Speech, Language, and Hearing Research*, 2011.
- [5] J. Cleland, J. M. Scobbie, Z. Roxburgh, C. Heyde, and A. Wrench, "Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback," *Journal of Speech, Language, and Hearing Research*, pp. 1–18, 2019.
- [6] W. F. Katz, M. R. McNeil, and D. M. Garst, "Treating apraxia of speech (AOS) with EMA-supplied visual augmented feedback," *Aphasiology*, vol. 24, no. 6-8, pp. 826–837, 2010.
- [7] F. E. Gibbon and S. E. Wood, "Visual feedback therapy with electropalatography," in *Interventions in speech sound disorders*. Paul H. Brookes Publishing Co., Inc., 2010.
- [8] T. M. Byun and E. R. Hitchcock, "Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation," *American Journal of Speech-Language Pathology*, 2012.
- [9] J. Cleland, J. M. Scobbie, and A. A. Wrench, "Using ultrasound visual biofeedback to treat persistent primary speech sound disorders," *Clinical Linguistics & Phonetics*, vol. 29, no. 8-10, pp. 575–597, 2015.
- [10] B. Bernhardt, B. Gick, P. Bacsfalvi, and M. Adler-Bock, "Ultrasound in speech therapy with adolescents and adults," *Clinical Linguistics & Phonetics*, vol. 19, no. 6-7, pp. 605–617, 2005.
- [11] R. Furniss and T. Wenger, "Seeing the big picture. the use of ultrasound in treating functional speech disorders in school-aged children in a community health setting," *Journal of Clinical Practice in Speech-Language Pathology*, vol. 20, no. 2, pp. 76–82, 2018.
- [12] R. Sadeghian and S. A. Zahorian, "Towards an automated screening tool for pediatric speech delay," in *Sixteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2015.
- [13] L. Ward, A. Stefani, D. Smith, A. Duenser, J. Freyne, B. Dodd, and A. Morgan, "Automated screening of speech development issues in children by identifying phonological error patterns," in *Seventeenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2016.
- [14] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Workshop on Speech and Language Technology in Education*, 2007.
- [15] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Thirteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.
- [16] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.
- [17] A. Eshky, M. S. Ribeiro, J. Cleland, K. Richmond, Z. Roxburgh, J. M. Scobbie, and A. A. Wrench, "UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions," in *Nineteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2018.
- [18] 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, 2012.
- [19] A. Cristia, S. Ganesh, M. Casillas, and S. Ganapathy, "Talker diarization in the wild: The case of child-centered daylong audio-recordings," in *Nineteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2018.
- [20] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 413–417.
- [21] M. Najafian and J. H. Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.
- [22] H. Bredin, "pyannote. metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Eighteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2017.
- [23] 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 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.
- [24] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF.STAR children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [25] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [26] M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals, "Speaker-independent classification of phonetic segments from raw ultrasound in child speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [27] S. P. Rath, D. Povey, K. Veselý, and J. Cernocký, "Improved feature processing for deep neural networks," in *Fourteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2013.
- [28] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Seventeenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2016.
- [29] P. Karanasou, M. J. Gales, P. Lanchantin, X. Liu, Y. Qian, L. Wang, P. C. Woodland, and C. Zhang, "Speaker diarisation and longitudinal linking in multi-genre broadcast data," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.
- [30] H. Christensen, P. D. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech," in *Fourteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2013.
- [31] P. Bell and S. Renals, "A system for automatic alignment of broadcast media captions using weighted finite-state transducers," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.