SPEAKER-INDEPENDENT CLASSIFICATION OF PHONETIC SEGMENTS FROM RAW ULTRASOUND IN CHILD SPEECH

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

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

ABSTRACT

Ultrasound tongue imaging (UTI) provides a convenient way to visualize the vocal tract during speech production. UTI is increasingly being used for speech therapy, making it important to develop automatic methods to assist various time-consuming manual tasks currently performed by speech therapists. A key challenge is to generalize the automatic processing of ultrasound tongue images to previously unseen speakers. In this work, we investigate the classification of phonetic segments (tongue shapes) from raw ultrasound recordings under several training scenarios: speaker-dependent, multi-speaker, speaker-independent, and speaker-adapted. We observe that models underperform when applied to data from speakers not seen at training time. However, when provided with minimal additional speaker information, such as the mean ultrasound frame, the models generalize better to unseen speakers.

Index Terms— ultrasound, ultrasound tongue imaging, speaker-independent, speech therapy, child speech

1. INTRODUCTION

Ultrasound tongue imaging (UTI) uses standard medical ultrasound to visualize the tongue surface during speech production. It provides a non-invasive, clinically safe, and increasingly inexpensive method to visualize the vocal tract. Articulatory visual biofeedback of the speech production process, using UTI, can be valuable for speech therapy [1, 2, 3] or language learning [4, 5]. Ultrasound visual biofeedback combines auditory information with visual information of the tongue position, allowing users, for example, to correct inaccurate articulations in real-time during therapy or learning. In the context of speech therapy, automatic processing of ultrasound images was used for tongue contour extraction [6] and the animation of a tongue model [7].

More broadly, speech recognition and synthesis from articulatory signals [8] captured using UTI can be used with silent speech interfaces in order to help restore spoken communication for users with speech or motor impairments, or to allow silent spoken communication in situations where audible speech is undesirable [9, 10, 11, 12, 13]. Similarly, ultrasound images of the tongue have been used for direct estimation of acoustic parameters for speech synthesis [14, 15, 16].

Speech and language therapists (SLTs) have found UTI to be very useful in speech therapy. In this work we explore the automatic processing of ultrasound tongue images in order to assist SLTs, who currently largely rely on manual processing when using articulatory imaging in speech therapy. One task that could assist SLTs is the automatic classification of tongue shapes from raw ultrasound. This can facilitate the diagnosis and treatment of speech sound disorders, by allowing SLTs to automatically identify incorrect articulations, or by quantifying patient progress in therapy. In addition to being directly useful for speech therapy, the classification of tongue shapes enables further understanding of phonetic variability in ultrasound tongue images. Much of the previous work in this area has focused on speaker-dependent models. In this work we investigate how automatic processing of ultrasound tongue imaging is affected by speaker variation, and how severe degradations in performance can be avoided when applying systems to data from previously unseen speakers through the use of speaker adaptation and speaker normalization approaches.

Below, we present the main challenges associated with the automatic processing of ultrasound data, together with a review of speaker-independent models applied to UTI. Following this, we present the experiments that we have performed (Section 2), and discuss the results obtained (Section 3). Finally we propose some future work and conclude the paper (Sections 4 and 5).

1.1. Ultrasound Tongue Imaging

There are several challenges associated with the automatic processing of ultrasound tongue images.

Image quality and limitations. UTI output tends to be noisy, with unrelated high-contrast edges, speckle noise, or interruptions of the tongue surface [17, 18]. Additionally, the oral cavity is not entirely visible from the image, missing the lips, the palate, or the pharyngeal wall.

Inter-speaker variation. Age and physiology may affect the output, with children imaging better than adults due to more moisture in the mouth and less tissue fat [17]. However, dry mouths lead to poor imaging, which might occur in speech therapy if a child is nervous during a session. Similarly, the vocal tracts of children across different ages may be more variable than those of adults.

Probe placement. Articulators that are orthogonal to the ultrasound beam direction image well, while those at an angle tend to image poorly. Incorrect or variable probe placement during recordings may lead to high variability between otherwise similar tongue shapes. This may be controlled using helmets [19], although it is unreasonable to expect the speaker to remain still throughout the recording session, especially if working with children. Therefore, probe displacement should be expected to be a factor in image quality and consistency.

Limited data. Although ultrasound imaging is becoming less expensive to acquire, there is still a lack of large publicly available databases to evaluate automatic processing methods. The UltraSuite Repository [20], which we use in this work, helps alleviate this issue, but it still does not compare to standard speech recognition or image classification databases, which contain hundreds of hours of speech or millions of images.

Supported by the EPSRC Healthcare Partnerships grant number EP/P02338X/1 (Ultrax2020 – http://www.ultrax-speech.org).

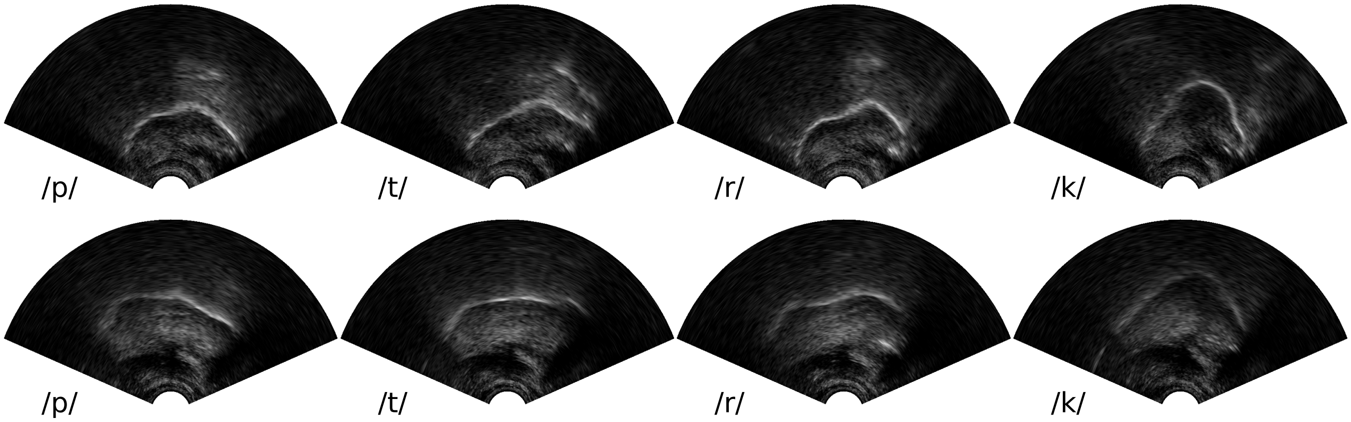

Fig. 1. Ultrasound samples for the four output classes based on place of articulation. The top row contains samples from speaker 12 (male, aged six), and the bottom row from speaker 13 (female, aged eleven). All samples show a midsaggital view of the oral cavity with the tip of the tongue facing right. Each sample is the mid-point frame of a phone uttered in an aCa context (e.g. *apa, ata, ara, aka*). See the UltraSuite repository[2] for details on interpreting ultrasound tongue images.

1.2. Related Work

Earlier work concerned with speech recognition from ultrasound data has mostly been focused on speaker-dependent systems [21, 22, 23, 24]. An exception is the work of Xu et al. [25], which investigates the classification of tongue gestures from ultrasound data using convolutional neural networks. Some results are presented for a speaker-independent system, although the investigation is limited to two speakers generalizing to a third. Fabre et al [6] present a method for automatic tongue contour extraction from ultrasound data. The system is evaluated in a speaker-independent way by training on data from eight speakers and evaluating on a single held out speaker. In both of these studies, a large drop in accuracy was observed when using speaker-independent systems in comparison to speaker-dependent systems. Our investigation differs from previous work in that we focus on child speech while using a larger number of speakers (58 children). Additionally, we use cross-validation to evaluate the performance of speaker-independent systems across all speakers, rather than using a small held out subset.

2. EXPERIMENTAL SETUP

2.1. Ultrasound Data

We use the Ultrax Typically Developing dataset (UXTD) from the publicly available UltraSuite repository[1][20]. This dataset contains synchronized acoustic and ultrasound data from 58 typically developing children, aged 5-12 years old (31 female, 27 male). The data was aligned at the phone-level, according to the methods described in [20, 26]. For this work, we discarded the acoustic data and focused only on the B-Mode ultrasound images capturing a midsaggital view of the tongue. The data was recorded using an Ultrasonix SonixRP machine using Articulate Assistant Advanced (AAA) software at ∼121fps with a 135° field of view. A single ultrasound frame consists of 412 echo returns from each of the 63 scan lines (63x412 raw frames). For this work, we only use UXTD type A (semantically unrelated words, such as *pack, tap, peak, tea, oak, toe*) and type B (non-words designed to elicit the articulation of target phones, such as *apa, eepee, opo*) utterances.

2.2. Data Selection

For this investigation, we define a simplified phonetic segment classification task. We determine four classes corresponding to distinct places of articulation. The first consists of bilabial and labiodental phones (e.g. */p, b, v, f, . . . /*). The second class includes dental, alveolar, and postalveolar phones (e.g. */th, d, t, z, s, sh, . . . /*). The third class consists of velar phones (e.g. */k, g, . . . /*). Finally, the fourth class consists of alveolar approximant */r/*. Figure 1 shows examples of the four classes for two speakers.

For each speaker, we divide all available utterances into disjoint train, development, and test sets. Using the force-aligned phone boundaries, we extract the mid-phone frame for each example across the four classes, which leads to a data imbalance. Therefore, for all utterances in the training set, we randomly sample additional examples within a window of 5 frames around the center phone, to at least 50 training examples per class per speaker. It is not always possible to reach the target of 50 examples, however, if no more data is available to sample from. This process gives a total of ∼10700 training examples with roughly 2000 to 3000 examples per class, with each speaker having an average of 185 examples. Because the amount of data varies per speaker, we compute a "sampling score", which denotes the proportion of sampled examples to the speaker's total training examples. We expect speakers with high sampling scores (less unique data overall) to underperform when compared with speakers with more varied training examples.

2.3. Preprocessing and Model Architectures

For each system, we normalize the training data to zero mean and unit variance. Due to the high dimensionality of the data (63x412 samples per frame), we have opted to investigate two preprocessing techniques: principal components analysis (PCA, often called eigentongues in this context) and a 2-dimensional discrete cosine transform (DCT). In this paper, **Raw** input denotes the mean-variance normalized raw ultrasound frame. **PCA** applies principal components analysis to the normalized training data and preserves the top 1000 components. **DCT** applies the 2D DCT to the normalized raw ultrasound frame and the upper left 40x40 submatrix (1600 coefficients) is flattened and used as input.

[1] http://www.ultrax-speech.org/ultrasuite

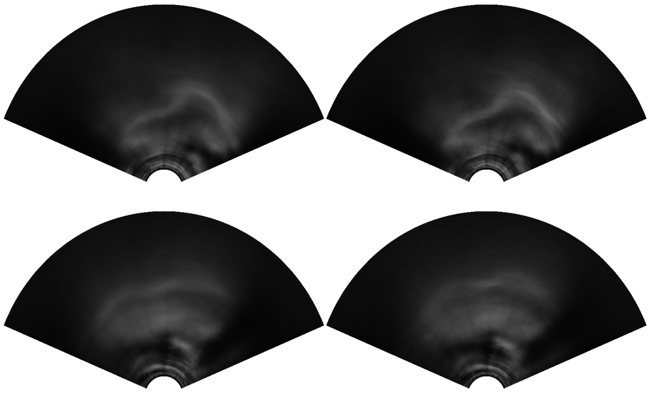

Fig. 2. Ultrasound mean image for speaker 12 (top row) and speaker 13 (bottom row). Means on the left column are taken over the training data, while means on the right are taken over the test data.

The first type of classifier we evaluate in this work are feedforward neural networks (**DNN**s) consisting of 3 hidden layers, each with 512 rectified linear units (ReLUs) with a softmax activation function. The networks are optimized for 40 epochs with a minibatch of 32 samples using stochastic gradient descent. Based on preliminary experiments on the validation set, hyperparameters such learning rate, decay rate, and L2 weight vary depending on the input format (Raw, PCA, or DCT). Generally, Raw inputs work better with smaller learning rates and heavier regularization to prevent overfitting to the high-dimensional data. As a second classifier to evaluate, we use convolutional neural networks (**CNN**s) with 2 convolutional and max pooling layers, followed by 2 fully-connected ReLU layers with 512 nodes. The convolutional layers use 16 filters, 8x8 and 4x4 kernels respectively, and rectified units. The fully-connected layers use dropout with a drop probability of 0.2. Because CNN systems take longer to converge, they are optimized over 200 epochs. For all systems, at the end of every epoch, the model is evaluated on the development set, and the best model across all epochs is kept.

2.4. Training Scenarios and Speaker Means

We train speaker-**dependent** systems separately for each speaker, using all of their training data (an average of 185 examples per speaker). These systems use less data overall than the remaining systems, although we still expect them to perform well, as the data matches in terms of speaker characteristics. Realistically, such systems would not be viable, as it would be unreasonable to collect large amounts of data for every child who is undergoing speech therapy. We further evaluate all trained systems in a **multi-speaker** scenario. In this configuration, the speaker sets for training, development, and testing are equal. That is, we evaluate on speakers that we have seen at training time, although on different utterances. A more realistic configuration is a speaker-**independent** scenario, which assumes that the speaker set available for training and development is disjoint from the speaker set used at test time. This scenario is implemented by leave-one-out cross-validation. Finally, we investigate a speaker **adaptation** scenario, where training data for the target speaker becomes available. This scenario is realistic, for example, if after a session, the therapist were to annotate a small number of training examples. In this work, we use the held-out training data to finetune a pretrained speaker-independent system for an additional 6 epochs in the DNN systems and 20 epochs for the CNN systems. We use all available training data across all training scenarios, and we investigate the effect of the number of samples on one of the top performing systems.

	DNN Raw	DNN PCA	DNN DCT	CNN Raw
Dependent	62.15%	57.78%	68.38%	66.56%
Multi-speaker	69.62%	66.30%	71.91%	74.70%
Independent	54.15%	55.14%	55.36%	59.42%
Adapted	69.26%	68.37%	67.76%	72.67%
with speaker mean				
Multi-speaker	71.61%	67.71%	72.28%	74.81%
Independent	60.52%	55.76%	60.19%	67.00%
Adapted	70.31%	68.02%	69.41%	71.30%

Table 1. Phonetic segment accuracy for the four training scenarios.

This work is primarily concerned with generalizing to unseen speakers. Therefore, we investigate a method to provide models with speaker-specific inputs. A simple approach is to use the speaker mean, which is the pixel-wise mean of all raw frames associated with a given speaker, illustrated in Figure 2. The mean frame might capture an overall area of tongue activity, average out noise, and compensate for probe placement differences across speakers. Speaker means are computed after mean variance normalization. For PCA-based systems, matrix decomposition is applied on the matrix of speaker means for the training data with 50 components being kept, while the 2D DCT is applied normally to each mean frame. In the DNN systems, the speaker mean is appended to the input vector. In the CNN system, the raw speaker mean is given to the network as a second channel. All model configurations are similar to those described earlier, except for the DNN using Raw input. Earlier experiments have shown that a larger number of parameters are needed for good generalization with a large number of inputs, so we use layers of 1024 nodes rather than 512.

3. RESULTS AND DISCUSSION

Results for all systems are presented in Table 1. When comparing preprocessing methods, we observe that PCA underperforms when compared with the 2 dimensional DCT or with the raw input. DCT-based systems achieve good results when compared with similar model architectures, especially when using smaller amounts of data as in the speaker-dependent scenario. When compared with raw input DNNs, the DCT-based systems likely benefit from the reduced dimensionality. In this case, lower dimensional inputs allow the model to generalize better and the truncation of the DCT matrix helps remove noise from the images. Compared with PCA-based systems, it is hypothesized the observed improvements are likely due to the DCT's ability to encode the 2-D structure of the image, which is ignored by PCA. However, the DNN-DCT system does not outperform a CNN with raw input, ranking last across adapted systems.

When comparing training scenarios, as expected, speaker-independent systems underperform, which illustrates the difficulty involved in the generalization to unseen speakers. Multi-speaker systems outperform the corresponding speaker-dependent systems, which shows the usefulness of learning from a larger database, even if variable across speakers. Adapted systems improve over the dependent systems, except when using DCT. It is unclear why DCT-based systems underperform when adapting pre-trained models. Figure 3 shows the effect of the size of the adaptation data when finetuning a pre-trained speaker-independent system. As expected,

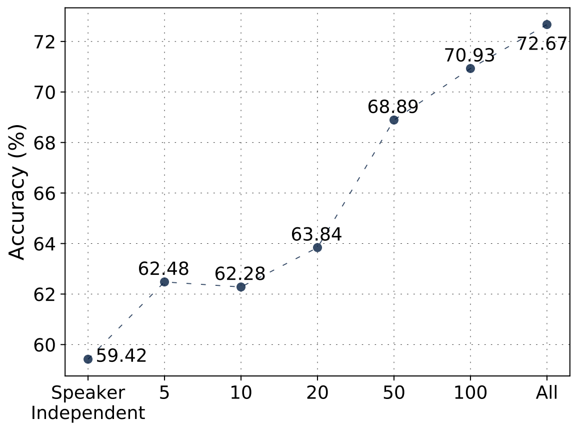

Fig. 3. Accuracy scores for adapted CNN Raw, varying amount of adaptation examples. We separately restrict training and development data to either n or all examples, whichever is smallest.

the more data is available, the better that system performs. It is observed that, for the CNN system, with roughly 50 samples, the model outperforms a similar speaker-dependent system with roughly three times more examples.

Speaker means improve results across all scenarios. It is particularly useful for speaker-independent systems. The ability to generalize to unseen speakers is clear in the CNN system. Using the mean as a second channel in the convolutional network has the advantage of relating each pixel to its corresponding speaker mean value, allowing the model to better generalize to unseen speakers.

Figure 4 shows pair-wise scatterplots for the CNN system. Training scenarios are compared in terms of the effect on individual speakers. It is observed, for example, that the performance of a speaker-adapted system is similar to a multi-speaker system, with most speakers clustered around the identity line (bottom left subplot). Figure 4 also illustrates the variability across speakers for each of the training scenarios. The classification task is easier for some speakers than others. In an attempt to understand this variability, we can look at correlation between accuracy scores and various speaker details. For the CNN systems, we have found some correlation (Pearson's product-moment correlation) between accuracy and age for the dependent ($r = 0.26$), multi-speaker ($r = 0.40$), and adapted ($r = 0.34$) systems. A very small correlation ($r = 0.15$) was found for the independent system. Similarly, some correlation was found between accuracy and sampling score ($r = -0.32$) for the dependent system, but not for the remaining scenarios. No correlation was found between accuracy and gender (point biserial correlation).

4. FUTURE WORK

There are various possible extensions for this work. For example, using all frames assigned to a phone, rather than using only the middle frame. Recurrent architectures are natural candidates for such systems. Additionally, if using these techniques for speech therapy, the audio signal will be available. An extension of these analyses should not be limited to the ultrasound signal, but instead evaluate whether audio and ultrasound can be complementary. Further work should aim to extend the four classes to more a fine-grained place of articulation, possibly based on phonological processes. Similarly, investigating which classes lead to classification errors might help explain some of the observed results. Although we have looked at variables such as age, gender, or amount of data to explain speaker variation, there may be additional factors involved, such as the general quality of the ultrasound image. Image quality could be affected by probe placement, dry mouths, or other factors. Automatically identifying or measuring such cases could be beneficial for speech therapy, for example, by signalling the therapist that the data being collected is sub-optimal.

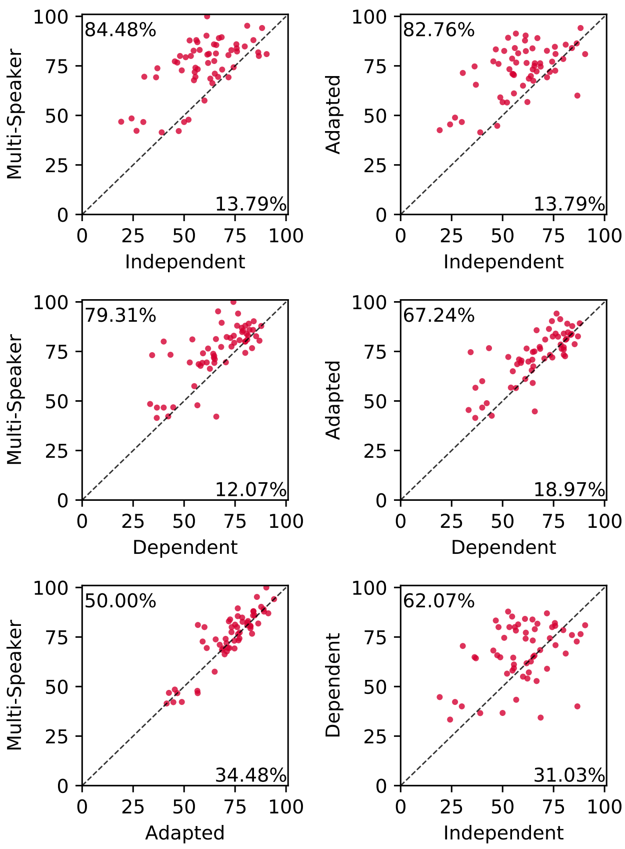

Fig. 4. Pair-wise scatterplots for the CNN system without speaker mean. Each sample is a speaker with axes representing accuracy under a training scenario. Percentages in top left and bottom right corners indicate amount of speakers above or below the dashed identity line, respectively. Speaker accuracies are compared after being rounded to two decimal places.

5. CONCLUSION

In this paper, we have investigated speaker-independent models for the classification of phonetic segments from raw ultrasound data. We have shown that the performance of the models heavily degrades when evaluated on data from unseen speakers. This is a result of the variability in ultrasound images, mostly due to differences across speakers, but also due to shifts in probe placement. Using the mean of all ultrasound frames for a new speaker improves the generalization of the models to unseen data, especially when using convolutional neural networks. We have also shown that adapting a pretrained speaker-independent system using as few as 50 ultrasound frames can outperform a corresponding speaker-dependent system.

6. REFERENCES

[1] Joanne Cleland, James M Scobbie, and Alan 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.

[2] Joanne Cleland, James Scobbie, Zoe Roxburgh, Cornelia Heyde, and Alan Wrench, "Ultraphonix: using ultrasound visual biofeedback to teach children with special speech sound disorders new articulations," in *7th International Conference on Speech Motor Control*, 2017.

[3] Joanne Cleland, James M Scobbie, Zoe Roxburgh, Cornelia Heyde, and Alan Wrench, "Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback," *Journal of Speech, Language and Hearing Research*, 2018 (In Press).

[4] Ian Wilson, Bryan Gick, MG O'Brien, C Shea, and J Archibald, "Ultrasound technology and second language acquisition research," in *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA)*, 2006, pp. 148–152.

[5] Bryan Gick, Barbara Bernhardt, Penelope Bacsfalvi, and Ian Wilson, "Ultrasound imaging applications in second language acquisition," *Phonology and second language acquisition*, vol. 36, pp. 315–328, 2008.

[6] Diandra Fabre, Thomas Hueber, Florent Bocquelet, and Pierre Badin, "Tongue tracking in ultrasound images using eigentongue decomposition and artificial neural networks," in *Proc. Interspeech*, 2015.

[7] Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda, and Pierre Badin, "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract," *Speech Communication*, vol. 93, pp. 63–75, 2017.

[8] Tanja Schultz, Michael Wand, Thomas Hueber, Dean J Krusienski, Christian Herff, and Jonathan S Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[9] Bruce Denby, Thomas Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[10] Thomas Hueber, Guido Aversano, Gérard Chollet, Bruce Denby, Gérard Dreyfus, Yacine Oussar, Pierre Roussel-Ragot, and Maureen Stone, "Eigentongue feature extraction for an ultrasound-based silent speech interface.," in *Proc. ICASSP*, 2007, pp. 1245–1248.

[11] Thomas Hueber, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone, "Phone recognition from ultrasound and optical video sequences for a silent speech interface," in *Proc. Interspeech*, 2008.

[12] Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," *Proc. of ISSP*, pp. 365–369, 2008.

[13] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[14] Bruce Denby and Maureen Stone, "Speech synthesis from real time ultrasound images of the tongue," in *Proc. ICASSP*. IEEE, 2004.

[15] Tamás Gábor Csapó, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó, "DNN-based ultrasound-to-speech conversion for a silent speech interface," *Proc. Interspeech*, pp. 3672–3676, 2017.

[16] Tamás Grósz, Gábor Gosztolya, László Tóth, Tamás Gábor Csapó, and Alexandra Markó, "F0 estimation for DNN-based ultrasound silent speech interfaces," in *Proc. ICASSP*. IEEE, 2018.

[17] Maureen Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.

[18] Min Li, Chandra Kambhamettu, and Maureen Stone, "Automatic contour tracking in ultrasound images," *Clinical linguistics & phonetics*, vol. 19, no. 6-7, pp. 545–554, 2005.

[19] Lorenzo Spreafico, Michael Pucher, and Anna Matosova, "Ultrafit: A speaker-friendly headset for ultrasound recordings in speech science," *Proc. Interspeech*, pp. 1517–1520, September 2018.

[20] Aciel Eshky, Manuel Sam Ribeiro, Joanne Cleland, Korin Richmond, Zoe Roxburgh, James M Scobbie, and Alan A Wrench, "Ultrasuite: a repository of ultrasound and acoustic data from child speech therapy sessions," in *Proc. Interspeech*, September 2018.

[21] Thomas Hueber, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone, "Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips," in *Proc. Interspeech*, 2007.

[22] Licheng Liu, Yan Ji, Hongcui Wang, and Bruce Denby, "Comparison of DCT and autoencoder-based features for DNN-HMM multimodal silent speech recognition," in *10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[23] Eric Tatulli and Thomas Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *Proc. ICASSP*. IEEE, 2017, pp. 2971–2975.

[24] Yan Ji, Licheng Liu, Hongcui Wang, Zhilei Liu, Zhibin Niu, and Bruce Denby, "Updating the silent speech challenge benchmark with deep learning," *Speech Communication*, vol. 98, pp. 42–50, 2018.

[25] Kele Xu, Pierre Roussel, Tamás Gábor Csapó, and Bruce Denby, "Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. EL531–EL537, 2017.

[26] Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond, and Steve Renals, "Towards robust word alignment of child speech therapy sessions," in *UK Speech Conference*, June 2018.