REAL-TIME ULTRASOUND-ENHANCED MULTIMODAL IMAGING OF TONGUE USING A 3D PRINTABLE STABILIZER SYSTEM: A DEEP LEARNING APPROACH

M. Hamed Mozaffari [*1] **and Won-Sook Lee** [†1]
[1]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

Résumé

Malgré une prise de conscience accrue de l'importance de l'articulation, il reste difficile pour les instructeurs de répondre aux besoins des apprenants en matière de prononciation. Il existe des outils pédagogiques pour l'enseignement et l'apprentissage de la prononciation bien qu'ils soient relativement rares. Récemment, une méthode multimodale améliorée par ultrasons (Enunciate) a été mise au point par des chercheurs de l'Université de Colombie-Britannique (UBC) pour visualiser les mouvements de la langue d'un apprenant, superposés à la face du locuteur. Des vidéos préenregistrées utilisant ce système ont été évaluées pour plusieurs cours de langues via un paradigme d'apprentissage mixte dans plusieurs classes de niveau universitaire. Bien que ce système multimodal ait été utilisé avec succès pour l'apprentissage de la prononciation, il nécessite encore des travaux manuels et des manipulations humaines, ainsi que la nature hors ligne du système, où les utilisateurs ne peuvent pas voir leurs mouvements de langue en temps réel. Dans cet article, nous avons développé un nouveau système d'entraînement à la prononciation multimodal complet, automatique et en temps réel, qui bénéficie de puissantes techniques d'intelligence artificielle, pour répondre aux difficultés des précédentes approches multimodales améliorées par ultrasons telles que le système UBC. Nous avons combiné les avantages de la technologie des ultrasons, de l'impression 3D et des algorithmes dde deep learning pour améliorer les performances des systèmes précédents. Plus précisément, notre système d'entraînement à la prononciation comprend plusieurs modules pour faciliter la personnalisation et le développement futurs par d'autres chercheurs. Il permet aux apprenants d'une langues d'observer automatiquement leurs mouvements de langue sur leur visage, en temps réel et avec une restriction minimale pendant la session de formation linguistique.

Mots clefs: Technologie ultrasons, formation à la prononciation, production de la parole et, visualisation de la langue, deep-learning, extraction automatique de contour, stabilisation de sonde.

Abstract

Despite renewed awareness of articulation importance, it remains a challenge for instructors to handle the pronunciation needs of language learners. There are relatively scarce pedagogical tools for pronunciation teaching and learning. Recently, an ultrasound-enhanced multi-modal method (Enunciate) has been developed by researchers at the University of British Columbia (UBC) for visualizing tongue movements of a language learner overlaid on the face-side of the speaker's head. Pre-recorded videos using that system was evaluated for several language courses via a blended learning paradigm at several university-level classes. Although that multi-modal system successfully utilized for pronunciation training, it still requires manual works and human manipulation as well as offline nature of the system, where users can not see their tongue movements in real-time. In this article, we developed a new comprehensive, automatic, real-time multi-modal pronunciation training system, benefits from powerful artificial intelligence techniques, to address the difficulties of the previous ultrasound enhanced multi-modal approaches such as the UBC system. We combined the advantages of ultrasound technology, three-dimensional printing, and deep learning algorithms to enhance the performance of previous systems. Specifically, our pronunciation training system comprises of several modules for easier future customization and development by other researchers. It empowers language learners to observe their tongue movements automatically, augmented on their face view in real-time with minimal restriction during the language training session.

Keywords: Ultrasound technology, pronunciation training, speech production, tongue visualization, deep-learning, automatic contour extraction, probe stabilization.

1 Introduction and Previous Works

Communication skill is one of the essential aspects of the second language (L2) acquisition so that it is often the first indication of a language learner's linguistic abilities [1, 2]. Pronunciation directly influences many social interaction skills of a speaker, such as communicative proficiency, performance, and self-confidence. Previous studies revealed that other aspects of L2 learning could be developed easier by accurate pronunciation [3, 4]. However, one of the most challenging skills to master in L2 training is to teach the correct pronunciation of tricky words [5] in traditional classroom settings.

∗. mmoza102@uOttawa.ca
†. wslee@uOttawa.ca

In practice, it is difficult for an L2 learner to utter difficult words or sounds precisely without any visual feedback of a native speaker and lack of awareness of how sounds are being articulated [5], especially in cases where the target sounds are not easily visible [1]. Visual feedback approaches have been developed over the past decades to facilitate L2 students to perceive moving speech articulators during speech or training sessions. These methods benefit from a range of instruments (called Electronic Visual Feedback (EVF)) [6], including ultrasound imaging, electromagnetic articulography (EMA), and electropalatography (EPG) [7]. Amongst those technologies, ultrasound imaging is non-invasive, safe, portable, versatile, user-friendly, widely available, and increasingly affordable. Besides, ultrasound modality offers high dimensional continuous real-time data with acceptable framerate.

Furthermore, ultrasound technology is capable of recording and illustrating the whole regions of the tongue (although the mandible sometimes obscures the tongue tip [1]) during both dynamic and static movements. Other imaging modalities such as MRI and X-ray (more specifically cinefluorography) are also capable of showing a mid-sagittal view of the tongue. However, these techniques are often prohibitively expensive, non-accessible, and invasive [8]. New technology-assisted language training methods [5,7,9–15], such as multimodal approaches using ultrasound imaging [1, 16, 17], have been successfully employed for language pronunciation teaching and learning, providing visual feedback of tongue gestures and poses. However, this technology is yet far from commercializing for use in every language education institutes.

In this study, we proposed a fully automatic pronunciation training system enables language learners to see their tongue on their face view in real-time. Our modular system, with a range of capabilities and facilities, provides a comprehensive toolbox for researchers applicable for different fields of linguistics applications from pedagogical to quantitative analysis. The authors observe several gaps in the current ultrasound multimodal approaches for L2 pronunciation training [1, 16–18]. In previous ultrasound-enhanced multimodal systems, manual synchronization between video and audio data, and image enhancement are essential parts [19]. In these systems, for quantitative analysis, one frame should be frozen manually. The whole super-imposing process of ultrasound and the side-face view is manual using editing software. All these manual works are time-consuming, subjective, and error-prone tasks, which require a knowledge of video and audio editing [1, 5, 17]. In our proposed system, manual modification of data is not necessary as well as all the super-imposing procedure are accomplished automatically and in real-time. For any further analytic study, users can work on video data in real-time without stopping the training session.

Besides manual works, users can watch only pre-recorded videos, usually created from the pronunciation of native speakers, and users can not compare their tongue gestures with native speakers simultaneously on the same system. Lack of information about the correct position, orientation, and scale of the ultrasound data forces researchers to stabilize the camera and head of users during video recording [1, 16, 17], without any flexibility for the user's head position, which makes language training non-conformable for the user. Therefore, these systems can not utilize in classrooms with many L2 students in real-time. For pedagogical applications, the position of an ultrasound probe is not crucial. However, for quantitative linguistic analysis, the ultrasound probe also should be fixed during video recording in the mid-sagittal plane. To alleviate these difficulties, we designed a simple stabilizer, including tracking markers called UltraChin. Using UltraChin and our deep learning tracking method, there is no need for fixing head movements. Therefore, users can observe their tongue gestures in real-time instead of watching pre-recorded videos.

1.1 Ultrasound Tongue Imaging

Nearly 50 years ago, one-dimensional ultrasound was first used effectively for illustration of one point at a time on the tongue's surface [20]. The two-dimensional ultrasound (B-mode settings for mid-sagittal or coronal view) has been employed in speech research since 40 years ago [21]. Nevertheless, due to the recent development of ultrasound imaging technology with greater image quality and affordability, it became an essential tool for imagining the articulators in speech research and pedagogical applications [1, 14].

Ultra-high frequency sound, both emitted and received by piezoelectric crystals of ultrasound transducer/probe, creates echo patterns that are decoded as an ultrasound image. Ultrasound signals penetrate and traverse linearly through materials with uniform density but reflect from dense substances such as bone. With the ultrasound transducer held under the jaw and with the crystal array lying in the mid-sagittal plane of the head, the ultrasound screen displays information about the superior surface of the tongue from the root to near the tip [22] (see Figure 1). Procedures and techniques of ultrasound image reconstruction and acquisition, specifically for tongue imaging, have been comprehensively described in [23].

Real-time tracking of tongue gestures and interpretation of ultrasound data by non-expert L2 learners is not always an easy task (see Figure 2 as an example of tongue contour in ultrasound device). Due to the noisy and low contrast images of ultrasound imaging, the tongue surface can be highlighted automatically in real-time for L2 learners, results in easier tracking of the tongue gestures (see the red curve in Figure 1).

One distinct linguistically valuable property of ultrasound imaging is the capability of simultaneous visualization of the front and back of the tongue [14]. Ultrasound has been utilized effectively in L2 pronunciation training [9]. For example, the efficacy of using ultrasound imaging on the pronunciation of North American /r/ and /l/ phonemes has been proven by depicting the complexity of the tongue's shape for L2 language learners [14, 24].

1.2 Ultrasound-enhanced Multimodal Approach

Recently, ultrasound-enhanced multimodal approaches have been applied for supporting L2 pronunciation students in

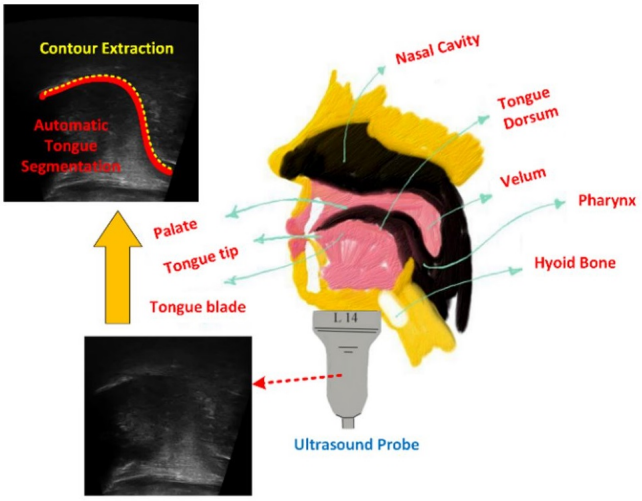

Figure 1: Tongue contour can be highlighted for better understanding of the tongue gestures in real-time video frames.

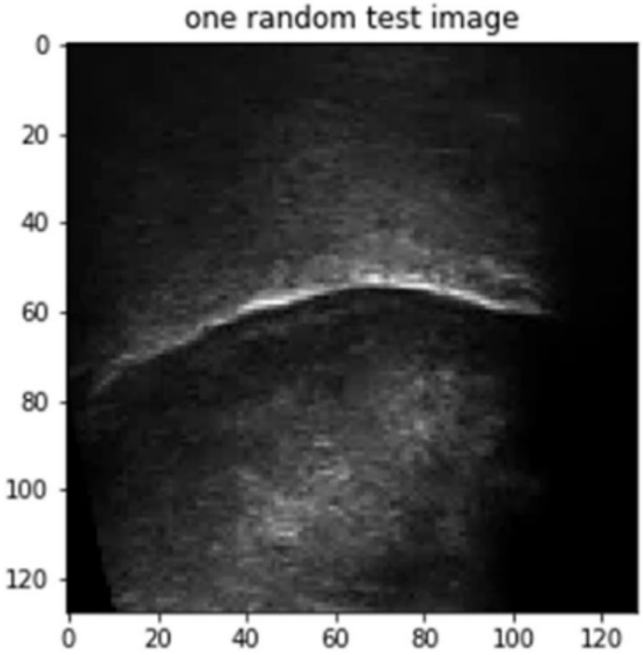

Figure 2: One sample of ultrasound video frame after cropping into a square sized format.

understanding and perceiving the location of their tongue from ultrasound video frames. Significant pioneer investigations of this method have been accomplished by researchers in the linguistics department at the University of British Columbia (UBC) (name of their system is eNunciate) [1, 5, 9, 14, 16, 17, 25]. The key technological innovation of this method is the use of mid-sagittal ultrasound video frames of the tongue, manually overlaid on the external profile views of a speaker's head. This approach allows L2 learners to observe pre-recorded videos comprising speech articulation of a native speaker superimposed on face profile [5,9,16]. In order to highlight the whole tongue region in recorded ultrasound video frames, the intensity of pixels related to the tongue region was changed manually to pink color [9, 16].

Benefits of a multimodal method for pronunciation language training have recently been investigated in a few studies [7, 18, 26]. However, manual work is yet extensive in many steps of these methods (pre-processing such as image enhancement, during the exam like overlaying ultrasound frames on RGB video frames, and post-processing including highlighting of the tongue region and audio/video synchronization). Besides, the overlaid videos come with some non-accuracy due to the lack of transformational specification (exact scale, orientation, and position information) of the ultrasound frame for superimposing on the face view.

In techniques such as eNunciate [16, 25] accurate transformation information cannot be generalized for real-time superimposing ultrasound data on face view. For this reason, the user's head should be restricted to one position during recording using stabilizers. Accurate synchronization between ultrasound data, video frames, and acoustic records is another challenge for those previous studies [19, 27]. A quantitative study of tongue movement only viable after freezing a target frame or during post-processing of recorded frames [5, 11]. Besides all these difficulties, language learners can only watch pre-recorded videos. In this study, we proposed a

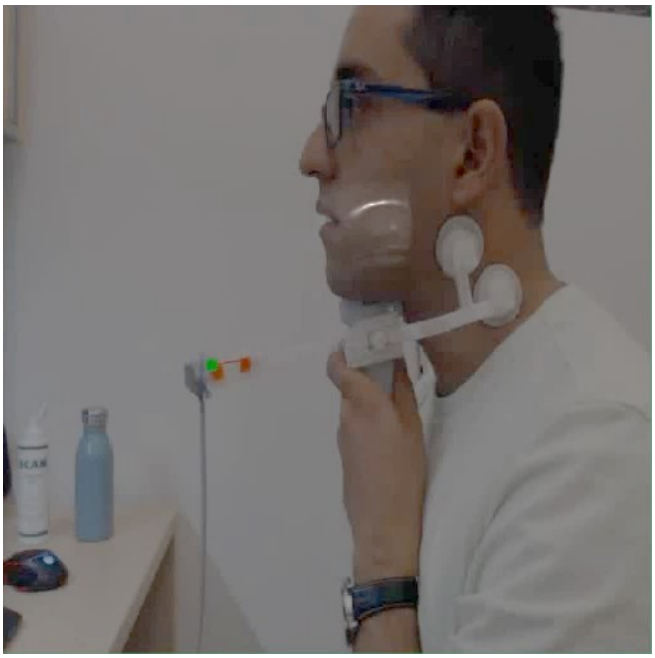

Figure 3: L2 learners can see their tongue gestures during a pronunciation training session using an ultrasound-enhanced multimodal approach. Real-time perceiving the tongue location in the mouth is significantly easier and more effective than looking at multimodal pre-recorded videos.

new system that enables users to observe their tongue movement in real-time on their face view without any considerable restrictions. Our ultrasound-enhanced multimodal system benefits from state-of-the-art deep learning techniques to calculate transformations required for real-time superimposing ultrasound data on face view. Figure 3 illustrates one frame captured using our ultrasound-enhanced multimodal technique.

1.3 Artificial Intelligence for Second Language

Artificial intelligence (AI) is a branch of computer science when machines execute tasks that typically require human intelligence [28]. Machine learning is a subset of AI, where machines learn skills by experiencing and acquiring knowledge without human involvement. Inspired by the functionality of the human brain, artificial neural networks are trained using a large amount of data to perform a task repeatedly. Deep learning algorithms are artificial neural networks with many (deep) layers similar to human brain structure [29]. Deep learning-based methods and applications in the image processing field such as object detection [30, 31] and image segmentation [32] have been a research hotspot in recent years. Deep learning methods are robust in automatic learning of a new task. In contrast, unlike traditional image processing methods, they are capable of dealing with many challenges such as object occlusion, transformation variant, and background artifacts [31–33].

The tongue surface is the gradient from white to the black area at the lower edge [23] in the form of a thick, white, and bright curve in ultrasound data. Although the tongue contour region can be viewed in ultrasound data, there are no hard structure references. For this reason, it is a challenging task for non-expert users to locate the tongue position and interpret its gestures without any exercise [16, 23]. Furthermore, due to the noisy and low-contrast images of ultrasound technology, it is an even more laborious task for users to follow the tongue surface movements, especially in real-time applications [16]. Instead of using a guideline on the screen (usually adopting the palate [11]), employing an automatic tracking technique, a language learner can perceive the real-time location of the tongue respect to landmarks of the face.

The rest of this article is structured as follows. Section 2 describes our proposed automatic and real-time ultrasound-enhanced multimodal pronunciation system. In this section, we explained each module of our system separately in detail to address each problem of previous systems mentioned before. Section 3 summarizes our experimental results. Finally, section 4 discusses and concludes our paper as well as potential future directions.

2 Methodology

2.1 UltraChin for Stabilization and Tracking

In previous ultrasound-enhanced multimodal studies, the head of language learners should be stabilized during video and ultrasound data collection as well as it is necessary for accurate superimposing video frames [1, 5]. A consequence of this restriction is a considerable reduction of the user's head flexibility. Moreover, the language learner should concentrate on the user interface for a long time with those limited movements where it might result in body fatigue and eye strain, ultimately reducing the effectiveness of the L2 training session (see Figure 4 for some samples of ultrasound probe stabilization methods). In a recent study by [18, 26], researchers could alleviate this difficulty by tracking of the face profile automatically. Tracking information made it easier for researchers to overlay video frames manually. However, tracking of the face is a subjective idea with low accuracy due to the variant of face angles and characteristics. It is noteworthy to mention that stabilizing the ultrasound probe is not necessary for pedagogical applications.

Figure 4: An illustration of some stabilization methods for head and probe. Designed helmets keep the probe under the chin, a) [34] b) [35]. One helmet which is fixed to the wall [36], and d) Optical tracking system for the head alignment [22].

In our system, users can keep the probe under the chin without any extra pieces of equipment (we named this state of our system as freehand). However, our goal is to provide facilities for all aspects of speech research, including pedagogical, both qualitative and quantitative analysis. For this reason, the freehand method can not guarantee that the probe orientation is always in the mid-sagittal plane during articulation. In order to make our system independent from the user's face profile and video frame superimposing process more accurate, we designed UltraChin, which is a universal 3D printable device compatible with any ultrasound probes. In general, UltraChin is used for two reasons in our proposed system : I) As a reference marker for probe tracking module, II) for keeping the probe under the chin aligned with the mid-sagittal plane of language learner's head.

In order to overlay the ultrasound video frame on a user's face automatically, for each frame, three transformation parameters related to the probe location should be calculated for each frame in real-time, and UltraChin markers provide locational information for this calculation. Other types of tracking markers as a reference have been employed in few studies for the head, palate, and tongue alignment correction purposes [37, 38]. UltraChin was created after several generations of designing (using SolidWorks software), 3D printing (using MakerBot Replicator 2), and testing on language learners (see figure 5 for several generations of UltraChin). In the last generation (see figure 6), we used natural materials in the process of 3D printing for skin sensibility prevention due to the contact of human skin with plastic, which was not considered in previous similar devices [34, 39]. Furthermore, UltraChin is expandable easily by adding extra parts where users can attach other types of sensors, such as electromagnetic tracking sensors. Unlike the previous helmets and stabilizer devices [34, 39, 40] for ultrasound tongue imaging, UltraChin is fully printable without the requirement of extra components such as rubber bands as well as publicly available [1]. One unique characteristic of UltraChin is that the ultrasound

1. https ://github.com/HamedMozaffari/UltraChinDesigns

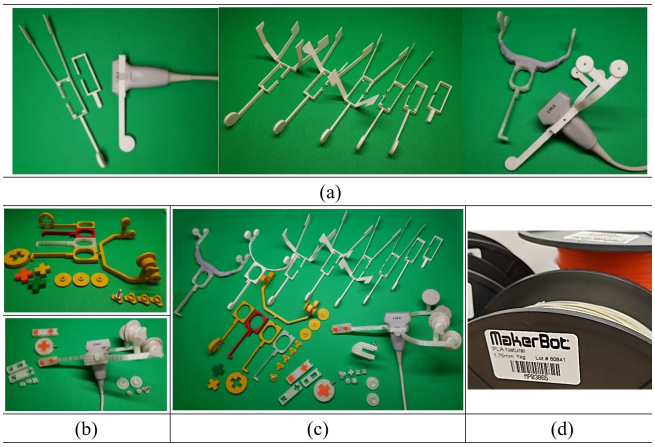

Figure 5: UltraChin : a) Integrated designs, b) Modular designs, c) Different versions and parts, d) Natural PLA materials for printing.

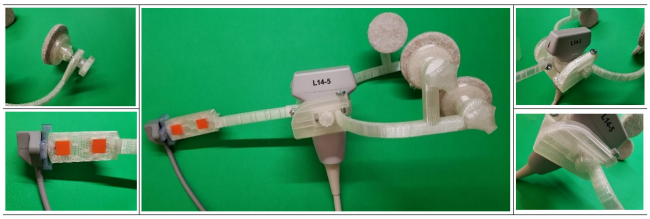

Figure 6: The last generation of UltraChin comprises of several modules. This version is universal and usable for different ultrasound probes as well as capable of being attached to other sensors.

probe is held by the user, which makes the process of data acquisition more comfortable and even more accurate after practicing the user to keep the probe precisely. At the same time, the optimized pressure of the probe on the chin is set by the user after a short training session, results in better image quality and less slippage of the probe and more comfortable sound articulation.

2.2 Automatic Tracking of Ultrasound Probe for Overlaying Videos

In our ultrasound-enhanced pronunciation training system, two video data are recording in real-time from the tongue using an ultrasound device and the face using a webcam camera. Scale and orientation of ultrasound video frames are almost identical because the ultrasound settings are fixed, and the probe is steady under the jaw. However, the user can move her/his head up, down, near, and farther from the webcam camera. For this reason, to project ultrasound frames on the face, we need to transform the ultrasound data on the dynamically changing face view of the user. Because users move in the same plane relative to the camera, having the position of two markers on UltraChin (three degrees of freedom DOFs) can be used to automatically calculate real-time location, scale, and orientation of ultrasound data on user's face in real-time.

Object localization (and detection) techniques determine where objects are located in a given image using bounding boxes encompass the targets. Various deep learning methods have been proposed for object localization in recent years [31, 41]. Similar to facial landmark detection [42], when several key features of the human face are detected as landmarks, we defined two key points on UltraChin extension leg as markers for the sake of probe tracking. Two landmarks (key points) are two upper-left corners of orange squares (see two embedded orange cubes in figure 6). The two markers are tracked automatically in real-time using our new deep convolutional neural network (named ProbeNet). In this method, positions of the two key points on UltraChin provide us transformational information in each frame, comprises of probe orientation, location, and a reference for scaling of the ultrasound data. We designed ProbeNet specifically for the probe tracking problem by inspiring from VGG16 network architecture [43].

Tracking of an ultrasound probe has already been accomplished using different kinds of devices such as electromagnetic, optical, mechanical sensors, and global positioning system (GPS) [44]. However, the primary motivation of those studies is to track the probe in three-dimensional space (usually with 6 degrees of freedom (DOF)). In this study, we considered a simplifying assumption : *ultrasound probe and language learner's face are aligned respect to the camera lens, both in two-dimensional planes* (see Figure 3 where ultrasound frame, segmented tongue dorsum with white color, and two orange markers are aligned respect to the camera lens). Under this assumption, tracking of the probe only requires the calculation of the location in a two-dimensional plane instead of three-dimensional space. For this reason, we selected two key points on the UltraChin, and the tracking problem was converted from three-dimensional space to two-dimensional space.

Figure 7 illustrates the detailed architecture of the ProbeNet for marker detection and tracking. Following advanced versions of VGG network architecture [43, 45], ProbeNet comprises of several standard convolutional layers followed by ReLU activation function and batch-normalization for more efficient network training. In the last block, we used a dense layer with four neurons, which provides 2D positions of the upper-left side of the two markers as well as a drop-out of 50 percent for better generalization over our annotated dataset. During a pronunciation training session, positional information of the two markers are predicted by ProbeNet. These data are used to find the best position, orientation, and scaling of the ultrasound frame overlaying on the language learner's face. Besides a specific calibration procedure is required to convert positional information into orientation and scaling data in the visualization module.

For calibration, we considered a ratio of the distance between the two markers and the width of the ultrasound probe head for scaling of the current ultrasound frame. Triangular geometric considerations between the two markers provide ultrasound frame translation and orientation information over the user's face. It is noteworthy to mention that all procedures of calibration and super-imposing of transformed ultrasound frames on face view are fully automatic and in real-time.

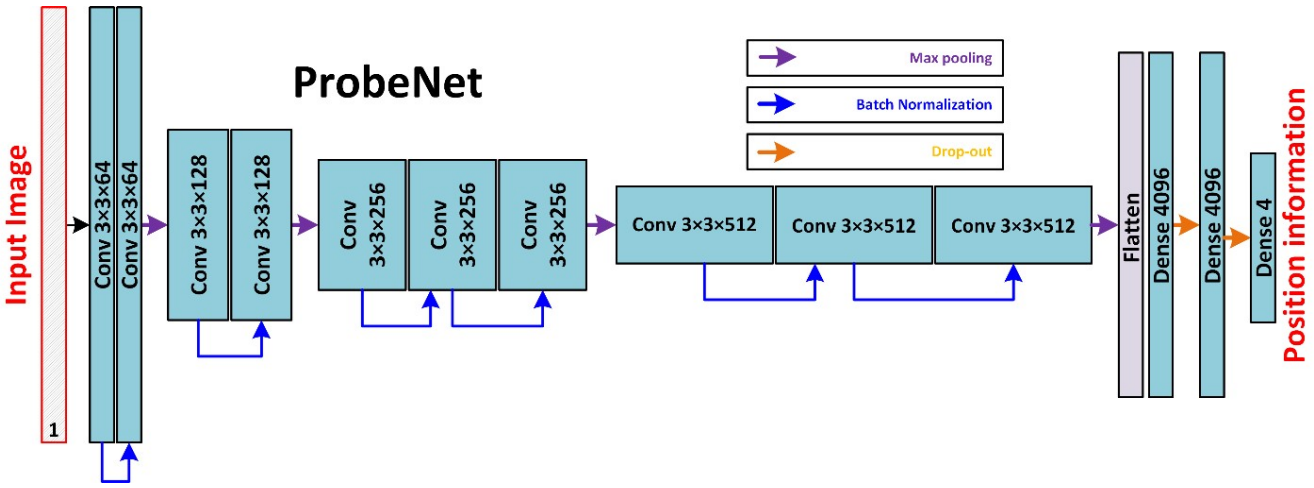

Figure 7: ProbeNet architecture for tracking landmarks on the UltraChin. Output layer provides two sets of numbers as vertical and horizontal positions of each landmark.

2.3 Automatic Tongue Contour Tracking

Typically, during tongue data acquisition, ultrasound probe beneath the user's chin images tongue surface in mid-sagittal or coronal view [46] in real-time. Mid-sagittal view of the tongue in ultrasound data is usually adapted instead of a coronal view for illustration of tongue region, as it displays relative backness, height, and the slope of various areas of the tongue. Tongue dorsum can be seen in this view as a thick, long, bright, and continuous region due to the tissue-air reflection of ultrasound signal by the air around the tongue (see Figure 2). Although the Mid-sagittal view of the tongue in ultrasound alone (e.g., see devices of Articulate Instruments Co.) helps the trend of L2 pronunciation learning, projects such as eNunciate [47] indicate that a multimodal ultrasound-enhanced system is more effective for interactive lingual articulation feedback [17].

The proposed manual coloring of the tongue region with pink color is not applicable for automatic and real-time applications in previous ultrasound-enhanced multimodal studies [17]. In this work, we utilized the most recent fully-automatic and real-time image segmentation method in this literature using deep learning techniques called BowNet [48, 49]. BowNet is used to track the surface of the tongue in video frames (as a continuous highlighted thick region). We used the tongue surface instead of the whole tongue region to facilitate the linguistics researcher for both qualitative and quantitative speech investigations. A detailed description of the BowNet architecture is beyond the scope of the present paper, and we described only several critical aspects of that model briefly for the sake of presentation. Curious readers can refer to studies by [49, 50].

Benefiting from different deep learning tools, including dilated convolutional layers and skip connections, the BowNet model could reach to higher accuracy with a robust performance in the problem of ultrasound tongue contour tracking and extraction in comparison to similar methods [51]. At the same time, there is no compromising for other aspects of the BowNet model like computational cost, the number of trainable parameters, or real-time performance [49]. Figure 8 presents the network structure and structural connections between different layers of the BowNet model [49]. As can be seen from the figure, there is a collaboration between two parallel encoding-decoding networks in BowNet structure. In one path, dilated convolution provides an efficient receptive field while on the other path, deconvolutional layers reconstruct features from the results of the encoder block. The concatenation of both paths provides more flexibility for the BowNet network to train on the search space with better exploration and exploitation ability. Although the BowNet model has few learnable parameters, it performs similar to bigger network models such as U-net [49, 52–54].

Automatic enhancement of ultrasound frames by highlighting the tongue dorsum region (using segmentation technique) enables language learners to focus on managing the challenges of L2 pronunciation learning instead of the interpretation of ultrasound data in real-time. Besides, extracted tongue contours provide teachers and language researchers valuable information for quantitatively comparison studies. It is noteworthy to mention that tongue contour extraction is done after tongue region segmentation using an image processing technique such as skeletonizing or just keeping the top pixels of the tongue region [18, 26, 55]. For a sample of an ideal segmented tongue surface region and extracted tongue contour, see red and yellow curves in Figure 1, respectively.

2.4 Automatic and Real-time Pronunciation Training System

We deployed our pronunciation training system using Python programming language and several standard public libraries as a modular system to enable other researchers to improve or customize each module for any future research. Figure 9 represents a schematic of all modules and their connections, implemented in our pronunciation training system. As can be seen in the figure, there are two streams of data recording in our system, an off-line module that is used for recording videos by native speakers for teaching and an online module

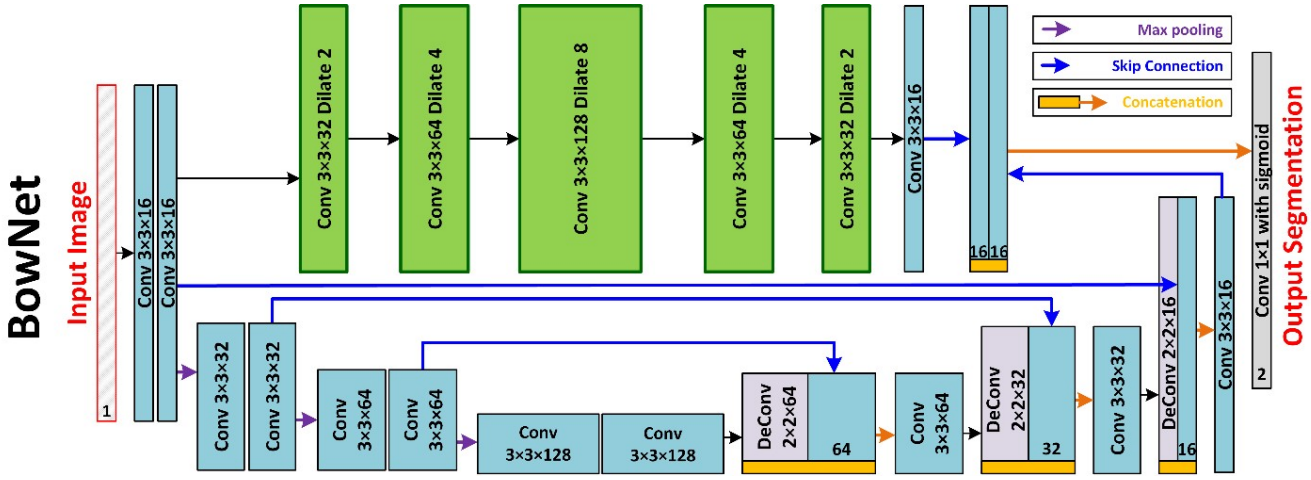

Figure 8: BowNet network architecture for tongue contour extraction automatically and in real-time [48]

for L2 learners, which is used for pronunciation training in real-time.

Ultrasound data acquisition and analysis generally involves capturing both acoustic and ultrasound video frames together so that the audio track can help with identifying target sounds on the ultrasound video stream. Synchronization of those two data can be done during recording or can be part of the post-processing steps, which is an integral and challenging part of having an accurate analysis. Our multimodal ultrasound-enhanced system encompasses different data and techniques during its performance, including tracking of UltraChin's markers, ultrasound data stream visualization, tongue surface segmentation, tongue contour extraction, audio recording and playback, calibration and superimposing video frames, learner's lips visualization (front side), and network connections between ultrasound and workstation. In order to have an approximately synced system, all these stages work together as the following procedure (see Figure 9 for more details) :

1. Face View Recording Module (FVRM) : The data stream from a high-definition webcam camera (Video and Audio) is captured and visualized in real-time. We used a Logitech Webcam with a framerate of 30 fps connected to our workstation (a personal computer with a CPU of 7 cores and 16 GB of memory equipped with a GPU of NVidia GTX1080). Video and audio are already synced in this stage.
2. Ultrasound Data Acquisition Module (UDAM) : Ultrasound stream video data is acquired and sent to the same workstation using Microsoft Windows remote desktop software (freely available on Windows desktops). We employed a linear ultrasound transducer L14-38 connected to an Ultrasonix Tablet with settings of the tongue (depth of 7 cm, a frame rate of 30 fps [23]). It is noteworthy to mention that our ultrasound probe is inadequate for most speech applications, and we only use that ultrasound probe for a demonstration of our system performance. It is noteworthy to explain that instead of working on the ultrasound stream in our Python codes, we used a Windows capturing library to grab ultrasound video from remote desktop software. This method enabled our system to be an ultrasound device-independent where it can work with different ultrasound devices. The price of the ultrasound streaming license for our machine was around $10K in 2018, asked from Ultrasonix company. Our system can capture data for free from all ultrasound machines with a network output port.
3. Ultrasound Probe Tracking Module (UPTM) : The current RGB video frame is fed to our probe tracking module. The pre-trained ProbeNet network model provides locations of two markers on the UltraChin (see the green dot (first marker) in Figure 9 and the connected red line, the middle image between two markers). In a predefined automatic calibration process, position, orientation, and probe head length are determined, and then they are sent to the visualization module.
4. Ultrasound Tongue Contour Tracking Module (UTCM) : Simultaneously with the ProbeNet model, the current ultrasound video frame is cropped, scaled, and fed to the BowNet model for the sake of tongue region segmentation. In this work, we illustrated segmented regions in white color without any post-processing enhancement.
5. Results of three modules UDAM, UTCM, and FVRM, which are three video frames, including cropped ultrasound frame, segmented tongue region, and RGB video frame, are superimposed using calculated transformation information (calibration data) from UPTM. A superimposed video stream is made by weighting the transparency of three video frame data. The result is sent to the visualization module for illustration and recording.
6. Visualization Module (VM) : In this module, a

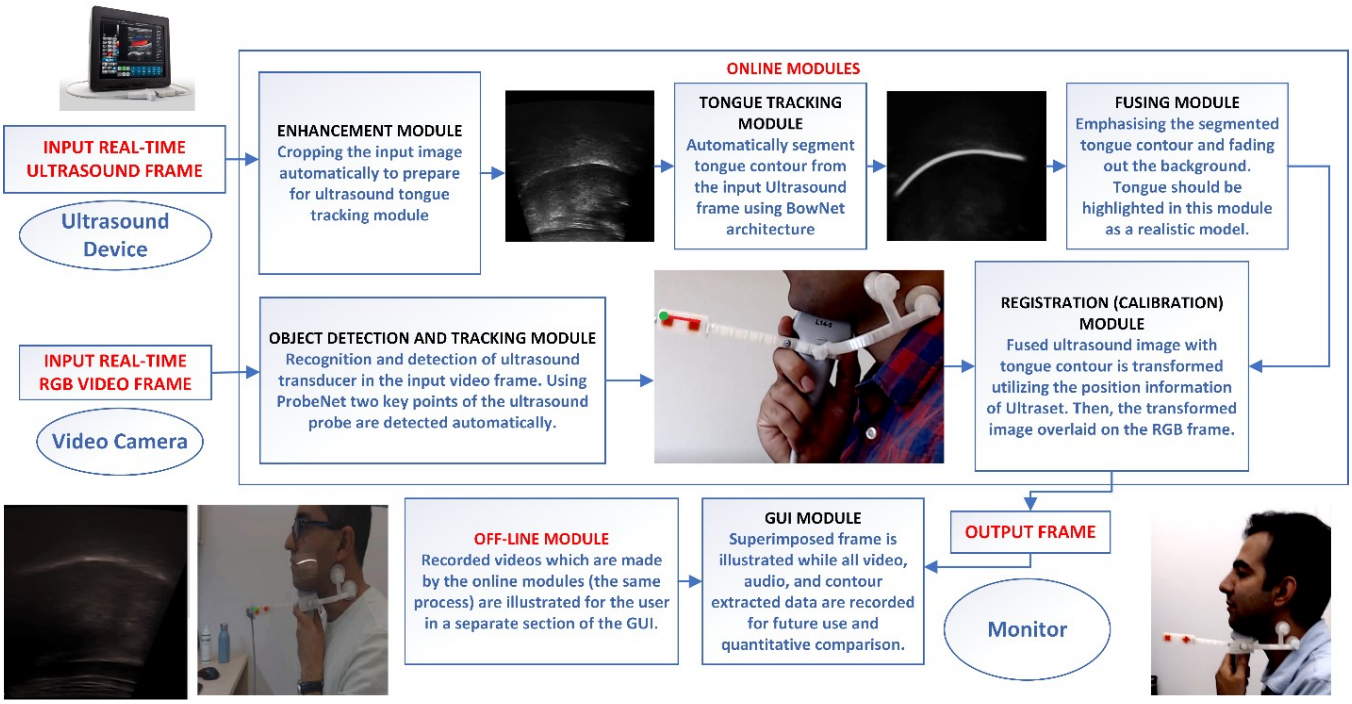

Figure 9: The detailed architecture of our multimodal, real-time, and automatic ultrasound-enhanced pronunciation training system comprises of two main online and offline modules.

simple designed graphical user interface (GUI) in Python language would illustrate several video streams including pre-recorded videos, superimposed video, individual frames from ultrasound and webcam (FVRM and UDAM with video and audio), and results of real-time quantitative analysis. This module is capable of showing data from different ultrasound machines simultaneously. There is also another camera for recording lip movements in front-view during a pronunciation training session. The development of this module is still in the early stages, and we show several windows side-by-side for the user and analytic data in the terminal window. Using two ultrasound devices, future GUI will provide an interactive panel between L2 learner and teacher as well as several comparison data between their tongue contours in real-time.

In our current system, a pronunciation learner or teacher can see several Windows in real-time separately on a display screen at the same time depends on the session target. For instance, as a speech investigation session, a researcher can see weighted ultrasound data superimposed on the face side in real-time, as well as separate non-overlaid ultrasound and RGB videos, accompany ultrasound tongue segmentation results. Having access to data from different modules of our system will assist researchers in comparing ultrasound tongue data qualitatively and quantitatively. Real-time data can be matched with recorded videos from native speaker pronunciation for evaluation of L2 learner's training progress. Moreover, real-time multimodal data could be compared with the data recorded in previous examination sessions as a follow-up study in the diagnosis of speech disorder, as a developed version of our current devices (refer to [56, 57]).

Due to the independence characteristic of our system, respect to the number of image processing streams as a multimodal system, in a different scenario, using two ultrasound devices, L2 language teachers and learners can see and compare their tongues in real-time. Moreover, our system is capable of illustrating the difference between their tongue contours automatically. Due to the lack of the second ultrasound, we can use recorded videos as the second reference video for our comparison studies and for capturing critical moments in the articulation. It is noteworthy to mention that we tested different ultrasound video data recorded by a curved ultrasound probe. The results of the UTCM module were even better in those videos (refer to our last UTCM module [58]). We also tested other webcams with different resolutions for the FVRM module. Results show that lower resolution webcams provide faster tracking but with lower accuracy.

3 Experiments and Results

In this study, we proposed a multimodal ultrasound-enhanced system with several modules, utilizing two different deep learning models. In this section, we explain the preliminary evaluation results of our system focus more on performance illustration of ProbNet, BowNet, and UltraChin. A comprehensive linguistic assessment is required for the evaluation of our system performance in terms of linguistics efficiency. In this section, we only report our pilot pedagogical evaluation from one user.

3.1 Ultrasound Probe Tracking Module

In order to train ProbeNet for tracking the markers on the UltraChin, we created a dataset comprises of 600 images of 3 different participants. Participants use our system for two minutes while a video is recorded from their UltraChin and face view. The recorded frames were annotated manually by placing two pre-defined key points on the upper-left side of orange markers on each frame. Dataset was divided into 80% training, 10% validation, and 10% testing sets. Finally, using our data augmentation toolbox (applying rotation, scaling, translation, and channel shift for images and the two key points), we created a dataset of 5000 images and their corresponding annotation information.

Adam optimization algorithm, with the first and second momentum of 0.9 and 0.999, respectively, was employed to optimize Mean Absolute Error (MAE) loss function [59] during training and validation. A variable learning rate with an exponential decay rate and an initial value of 0.001 was chosen for the training of the ProbeNet. We trained the ProbeNet model for ten epochs (each with 1000 iterations) with mini-batches of 10 images. Our experimental results revealed the strength and robustness of the ProbeNet in the landmark tracking task on the UltraChin device. We got an average MAE of 0.027 $\pm$ 0.0063 for ten times running of the ProbeNet on the test dataset.

3.2 Tongue Contour Extraction

Few previous studies have used deep learning methods for tongue contour extraction with acceptable results [51,52]. For the ultrasound contour tracking module (UTCM), we used one of the recent deep learning models in ultrasound tongue literature called BowNet [49]. Figure 8 represents the detailed architecture of the BowNet. For training settings of the BowNet, we followed the procedure in [49]. Similar to the ProbeNet, for the training of the BowNet model, we separated the dataset (identical to [49]) into 80% training, 10% validation, and 10% test sets.

The BowNet model was trained and validated using online augmentation, and then it was tested separately on the test dataset. Figure 10 presents a sample result of the BowNet model. Due to the more generalization ability of the BowNet network, it provides instances from different ultrasound machines with less false predictions. For more details about the performance evaluation of the BowNet, refer to the original study [49].

The BowNet model was trained to work on data recorded from two ultrasound datasets. The tongue contour tracking performance of our system might be dropped for new ultrasound data. A recent study in ultrasound tongue contour tracking literature [50] has investigated the usage of domain adaptation for several different ultrasound datasets, which can alleviate this difficulty significantly.

3.3 Accuracy Assessment of the UltraChin

Head and probe stabilization is not necessary if the system is only utilized as a pronunciation bio-feedback [60]. However, the accuracy of our system could be improved by adding 3D printable extensions to the UltraChin for head stabilization for a particular linguistic study. However, the main reason for using UltraChin is to track the two markers for the super-imposing of video frames. At the same time, UltraChin provides stabilization for ultrasound probe orientation. In order to evaluate our 3D printable design, we followed the method in [39]. We attached one magnetic tracking sensors, PATRIOT Polhemus Company (see Figure 11 and 12), on the UltraChin and participant's chin in two separate experiments. Six degrees of freedom (see Figure 12) were recorded after ten times repeating a similar experiment. For this experiment, the participant's head was fixed using the method in [36]. We asked the participant to repeat "ho-mo-Maggie" [39] and to open mouth to the maximum position for ten times. We calculated deviations of the UltraChin in terms of transnational (in millimeters) and rotational (in degree) slippages.

Table 1 shows the maximum error of the UltraChin after 10 times experiment. For a better understanding of the UltraChin performance, we checked two different settings where four screws of the device were loose (most comfort) or tight (dis-comfort).

Our experimental results showed that in the case of tightly firming four screws of UltraChin user's chin has a better long term translational and rotational unwanted slippage without losing a significant comfortability for the user's neck. Slippage errors might be even more due to the usage of cushions, skin deformations, and how the participant is keeping the probe under the chin. In compare to the system in [39], UltraChin has more long-term slippage in almost all directions. One reason is that UltraChin has fewer stabilizer arms than previous helmets. Nevertheless, UltraChin errors still are within acceptable deviation limits reported in [39].

3.4 A Preliminary Linguistic Evaluation of Our System

The positional information from real-time output instances of ProbeNet is used to calculate an estimation of the position, orientation, and scale of ultrasound frames on RGB video frames. In this way, real-time segmented tongue contours from ultrasound frames are predicted by the BowNet model and transferred on the face-side of language learners on RGB video frames. Therefore, the language learner can see real-time video frames created by superimposing raw RGB frames, transformed ultrasound images, and tongue segmented images. To illustrate the superimposed image, we considered different weights for the transparency of each image. Figure 13 shows several superimposed samples from our real-time multimodal ultrasound-enhanced system. In the figure, we considered transparency weights as 0.9 for RGB image, 0.4 for Ultrasound image, and 1 for the predicted map, respectively.

For the sake of representation, we also showed a guideline on the UltraChin to help language learners to keep the probe in a correct position in two-dimensional space (see red lines in Figure 13 between two orange markers). During the trai-

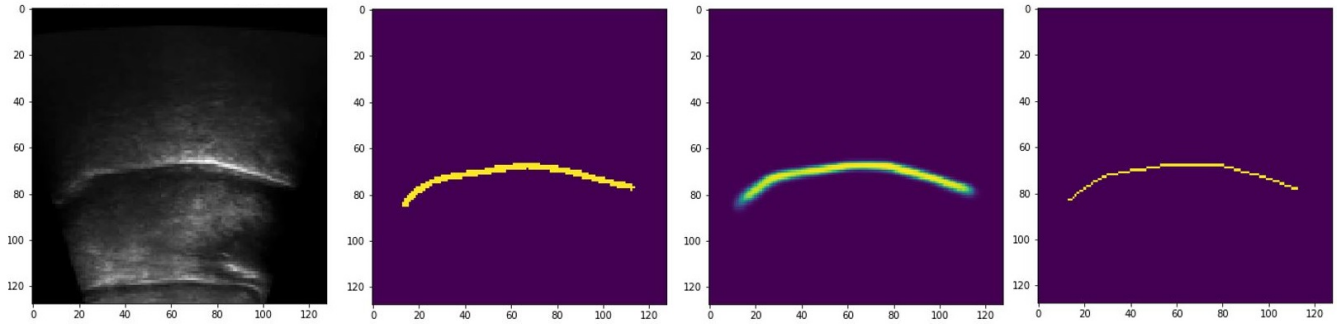

Figure 10: One test sample used for testing the BowNet network. From left to right columns are ultrasound image, ground truth image, predicted map, extracted contour from the predicted map.

Figure 11: First row : different views of magnetic tracking sensors attached on UltraChin. Second row : Different parts of UltraChin and magnetic tracking sensor can be seen in figure.

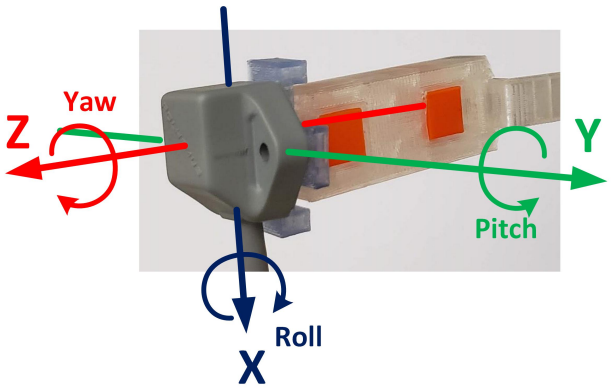

Figure 12: Tracking magnetic sensor was attached on UltraChin. Arrows show 6 degree of freedoms defined in our experiments.

ning session, users can check their face alignment with the camera using the red color guideline, which should always be between two markers. We simultaneously recorded acoustic data, superimposed video, individual video data from two RGB cameras (face-side and front-side view), real-time ultrasound frames, and tongue contour information for later experiments or follow-up study.

There are many methods and standards in the literature for testing L2 pronunciation acquisition methods [9, 11, 24]. It is possible to use the system for small numbers of L2 pronunciation training individuals [9] in a large classroom setting, either by providing individual ultrasound training to language instructors [61] or by presenting ultrasound videos as part of a blended learning approach [13], or even in a community-based settings [1]. For example, in [17], the previous ultrasound-enhanced system has been tested in several courses at UBC (named eNunciate). This kind of system is also tested for the training and revitalization of different indigenous languages [16]. The usability of ultrasound biofeedback in L2 pronunciation training has been comprehensively investigated in [11, 12].

Status of four screws	Max translational in milimeters			Max Rotational in degree		
	x	y	z	roll	yaw	pitch
Loose	$4.7 \pm 0.39mm$	$5.1 \pm 0.69mm$	$7.6 \pm 0.81mm$	$6.4 \pm 0.21^{\circ}$	$4.1 \pm 0.46^{\circ}$	$5.9 \pm 0.86^{\circ}$
Tight	$3.4 \pm 0.18mm$	$3.5 \pm 0.72mm$	$6.1 \pm 0.15mm$	$5.6 \pm 0.59^{\circ}$	$3.8 \pm 0.45^{\circ}$	$4.7 \pm 0.91^{\circ}$

Table 1: Maximum slippage of the UltraChin in 6DOF after 10 time testing on one participant. Values show the mean and standard deviation for each experiment. Screws of the UltraChin was loosely and tightly firm in two different experiments.

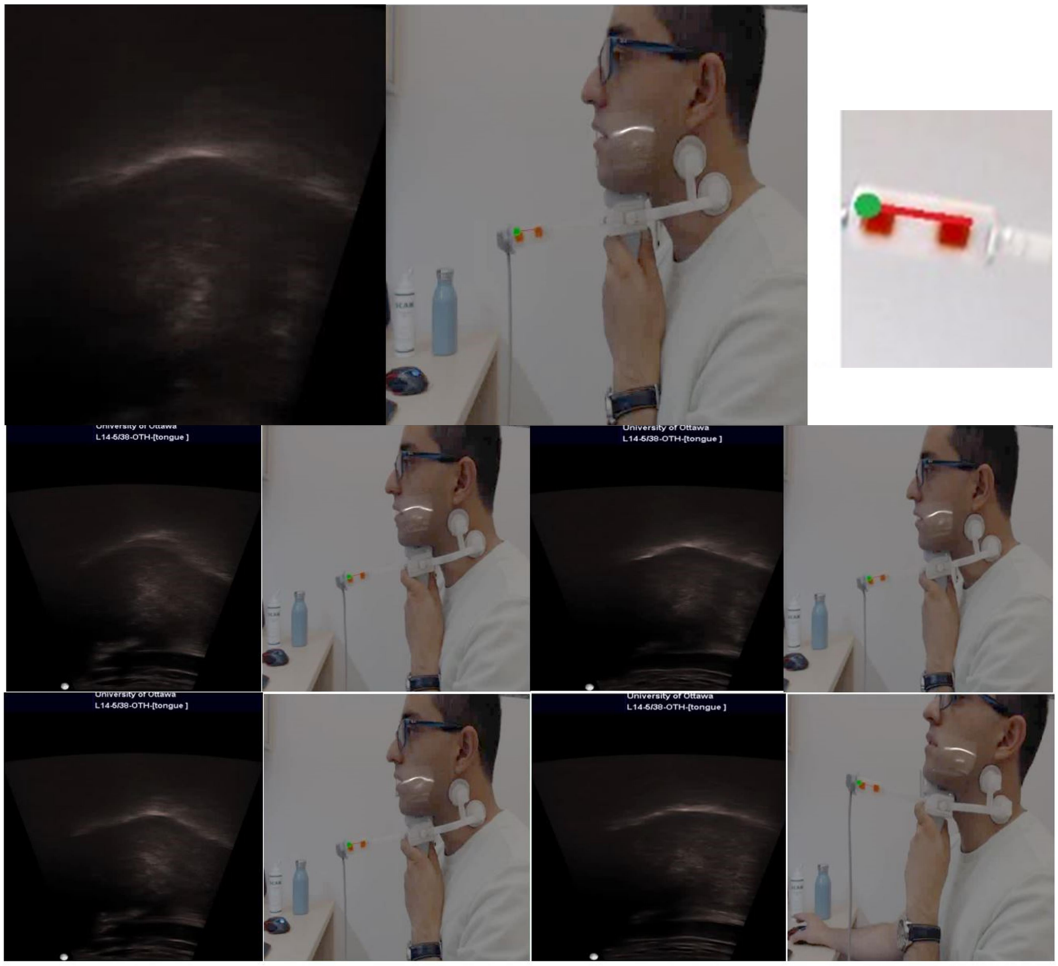

Figure 13: Sample frames of our real-time multimodal system. Thick white lines on the right-side images are extracted tongue contours from the corresponding left-side ultrasound images. Red line between two markers is illustrated as a guide.

Note that due to the limitation of this experiment, we can no conclude the effectiveness of our system on pedagogical aspects of L2 language, and linguistics experts should conduct a comprehensive collaborating user study. However, due to the lack of assessment facilities like a big classroom equipped with ultrasound machines, we followed the method in [9] for a participant experiment as a pilot evaluation study. In this technique, one approach is used repeatedly to measure the dependent variables from an individual. The dependent variables in this methodology consist of targets to be learned, such as vowels and consonants [9]. The main goal is to study articulator positions and segments, the accuracy of production, and speech intelligibility. Good candidates for ultrasound biofeedback are usually vowels, rhotic sounds, retroflex, velar and uvular consonants, and dynamic movements between tongue gestures [1].

Ultrasound has been utilized to teach individual challenging sounds, such as English /r/, in clinical settings [10, 24]. For this reason, we selected one individual participant to practice predefined sounds individually by comparing them with the pronunciation of the same statements by a native speaker. We utilized sample videos from the eNunciate project website UBC language department [47] as our truth pronunciation references. An Iranian L2 pronunciation learner volunteered to use our multimodal pronunciation system for ten sessions to improve pronunciation of /r/ sound. Each session contained 20 times repeating of /ri/, /ra/, and /ru/ and comparing with the video downloaded from [47]. Before the first session, we trained the participant for correct using our system and watching several training videos from the same website.

The benefits of our pronunciation system are not limited to only real-time and automatic characteristics. During pronunciation sessions, unlike other studies [15], there is no need for any manual synchronization. Furthermore, ultrasound frames, RGB video frames, audio data, overlaid images, and extracted contour information are recorded and visualized simultaneously. Our preliminary assessments showed that a language learner would fatigue slower than previous studies, in which the average time was 20 to 30 mins due to maintaining a relatively constant position [9]. In our system, non-physical restrictions such as using uniform backgrounds [18] in video recording have been addressed, and the system can be used in any room with different ambient features. Our system can provide researchers a real-time quantitative evaluation (such as mean sum of distances (MSD) in percentage) between tongue contours of language learner and teacher (requires the second ultrasound device or pre-recorded videos).
Testing the efficacy of our real-time automatic multimodal pronunciation system in detail remains in the early stages, and further research should be accomplished to create a fuller and more accurate assessment of our system with the collaboration of linguistics departments.

4 Discussion and Conclusion

In this study, we proposed and implemented a fully automatic and real-time modular multimodal ultrasound-enhanced pronunciation training system using several novel innovations. Unlike previous studies, instead of tracking the user's face or using tracking devices (see [44] for different tracking devices), the ultrasound probe position and orientation are estimated automatically using a 3D printable stabilizer (named UltraChin) and a deep learning model (named ProbeNet). ProbeNet was trained in advance on our dataset to track two markers on the UltraChin. This approach enables our pronunciation system to determine the optimum transformation quantities for multimodal superimposition as a user-independence system.
UltraChin makes the system universal for every ultrasound probe as well as invariant respect to the probe image occlusion. UltraChin errors due to the slippage of the device during a language pronunciation training session were within the standard range in the literature. At the same time, the pre-trained BowNet model [49], another deep learning model tracks, delineates, and highlights the tongue regions on ultrasound data. Different enhanced and transformed video frames from the different modules of our system are overlaid for illustration in the visualization module. Except for the preparation of training datasets, all modules in our system work automatically, in real-time, end-to-end, and without any human manipulation.
The application of our system can even be studied as visual biofeedback (VBF) for other applications like pronunciation training in different languages. Our system can be utilized for diagnosis and treatment planning of development speech disorders (SSDs), which is a common communication impairment in childhood who consistently exhibit difficulties in the production of specific speech sounds in their native language [62]. We believe that publishing our datasets, annotation package, deep learning architectures, and pronunciation training toolkit deployed on a publicly available Python programming language with an easy to use documentation will help other researchers in the different fields of linguistics.
Previous semi-automatic multimodal methods [26] revealed that language learners could understand the gestures of their tongue better in real-time using an ultrasound-enhanced multimodal visualization approach than previous recorded offline systems. Despite all the successful performance achievements of our proposed system, the development of our system is still in the early stages. An extensive pedagogical investigation of our pronunciation training system for teaching and learning should be accomplished to evaluate the efficiency and effectiveness of our system in different aspects of pronunciation training. Providing a comprehensive GUI for our system is also still under progress.

References

[1] Sonya Bird and Bryan Gick. Ultrasound biofeedback in pronunciation teaching and learning. In *Proc. ISAPh 2018 International Symposium on Applied Phonetics*, pages 5–11, 2018.

[2] Ron I Thomson and Tracey M Derwing. The effectiveness of l2 pronunciation instruction : A narrative review. *Applied Linguistics*, 36(3) :326–344, 2014.

[3] Khia Anne Johnson, Gloria Madeleine Mellesmoen, Roger Yu-Hsiang Lo, and Bryan Gick. Prior pronunciation knowledge bootstraps word learning. *Frontiers in Communication*, 3 :1, 2018.

[4] Tracey M Derwing and Murray J Munro. Second language accent and pronunciation teaching : A research-based approach. *TESOL quarterly*, 39(3) :379–397, 2005.

[5] Jennifer Abel, Blake Allen, Strang Burton, Misuzu Kazama, Bosung Kim, Masaki Noguchi, Asami Tsuda, Noriko Yamane, and Bryan Gick. Ultrasound-enhanced multimodal approaches to pronunciation teaching and learning. *Canadian Acoustics*, 43(3), 2015.

[6] Stephen Lambacher. A call tool for improving second language acquisition of english consonants by japanese learners. *Computer Assisted Language Learning*, 12(2) :137–156, 1999.

[7] Heather Bliss, Jennifer Abel, and Bryan Gick. Computer-assisted visual articulation feedback in l2 pronunciation instruction. *Journal of Second Language Pronunciation*, 4(1) :129–153, 2018.

[8] Lisa Davidson. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America*, 120(1) :407–415, 2006.

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

[10] Miwako Tateishi and Stephen Winters. Does ultrasound training lead to improved perception of a non-native sound contrast ? evidence from japanese learners of english. In *Proc. 2013 annual conference of the Canadian Linguistic Association*, pages 1–15, 2013.

[11] Barbara Bernhardt, Bryan Gick, Penelope Bacsfalvi, and Marcy Adler-Bock. Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics & Phonetics*, 19(6-7) :605–617, 2005.

[12] Tanja Kocjančič Antolík, Claire Pillot-Loiseau, and Takeki Kamiyama. The effectiveness of real-time ultrasound visual feedback on tongue movements in l2 pronunciation training. *Journal of Second Language Pronunciation*, 5(1) :72–97, 2019.

[13] Heather Bliss, K Johnson, Strang Burton, Noriko Yamane, and Bryan Gick. Using multimedia resources to integrate ultrasound visualization for pronunciation instruction into postsecondary language classes. *J. Linguist. Lang. Teach*, 8 :173–188, 2017.

[14] Ian Wilson and Bryan Gick. Ultrasound technology and second language acquisition research. In *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006)*, pages 148–152. Cascadilla Proceedings Project Somerville, MA, 2006.

[15] Thomas Hueber. Ultraspeech-player : intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training. In *INTERSPEECH*, pages 752–753, 2013.

[16] Heather Bliss, Strang Burton, and Bryan Gick. Ultrasound overlay videos and their application in indigenous language learning and revitalization. *Canadian Acoustics*, 44(3), 2016.

[17] Noriko Yamane, Jennifer Abel, Blake Allen, Strang Burton, Misuzu Kazama, Masaki Noguchi, Asami Tsuda, and Bryan Gick. Ultrasound-integrated pronunciation teaching and learning. *Ultrafest VII, Hong Kong*, 2015.

[18] M Hamed Mozaffari, Shenyong Guan, Shuangyue Wen, Nan Wang, and Won-Sook Lee. Guided learning of pronunciation by visualizing tongue articulation in ultrasound image sequences. In *2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pages 1–5. IEEE, 2018.

[19] Michael Aron, Nicolas Ferveur, Erwan Kerrien, Marie-Odile Berger, and Yves Laprie. Acquisition and synchronization of multimodal articulatory data. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[20] CA Kelsey, RJ Woodhouse, and FD Minifie. Ultrasonic observations of coarticulation in the pharynx. *The Journal of the Acoustical Society of America*, 46(4B) :1016–1018, 1969.

[21] Barbara C Sonies, Thomas H Shawker, Thomas E Hall, Lynn H Gerber, and Stephen B Leighton. Ultrasonic visualization of tongue motion during speech. *The Journal of the Acoustical Society of America*, 70(3) :683–686, 1981.

[22] Fiona Campbell, Bryan Gick, Ian Wilson, and Eric Vatikiotis-Bateson. Spatial and temporal properties of gestures in north american english/r. *Language and Speech*, 53(1) :49–69, 2010.

[23] Maureen Stone. A guide to analysing tongue motion from ultrasound images. *Clinical linguistics & phonetics*, 19(6-7) :455–501, 2005.

[24] Marcy Adler-Bock, Barbara May Bernhardt, Bryan Gick, and Penelope Bacsfalvi. The use of ultrasound in remediation of north american english/r/in 2 adolescents. *American Journal of Speech-Language Pathology*, 2007.

[25] Heather Bliss, Sonya Bird, Ashley Cooper, Strang Burton, and Bryan Gick. Seeing speech : Ultrasound-based multimedia resources for pronunciation learning in indigenous languages. *Language Documentation & Conservation*, 12 :315–338, 2018.

[26] M. Hamed Mozaffari., Shuangyue Wen., Nan Wang., and Won-Sook Lee. Real-time automatic tongue contour tracking in ultrasound video for guided pronunciation training. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 1 : GRAPP,*, pages 302–309. INSTICC, SciTePress, 2019.

[27] 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*, pages 365–369, 2008.

[28] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *Metabolism*, 69 :S36–S40, 2017.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553) :436, 2015.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[31] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning : A review. *IEEE transactions on neural networks and learning systems*, 2019.

[32] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4) :834–848, 2017.

[33] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding : A review. *Neurocomputing*, 187 :27–48, 2016.

[34] Lorenzo Spreafico, Michael Pucher, and Anna Matosova. Ultrafit : A speaker-friendly headset for ultrasound recordings in speech science. In *International Speech Communication Association*, 2018.

[35] James M Scobbie, Jane Stuart-Smith, and Eleanor Lawson. Looking variation and change in the mouth : developing the sociolinguistic potential of ultrasound tongue imaging. *developing the sociolinguistic potential of Ultrasound Tongue Imaging*, 2008.

[36] Lucie Ménard and Aude Noiray. The development of lingual gestures in speech : Experimental approach to language development. *Faits de langues*, 37 :189, 2011.

[37] Jeff Mielke, Adam Baker, Diana Archangeli, and Sumayya Racy. Palatron : a technique for aligning ultrasound images of the tongue and palate. 2005.

[38] Amanda L Miller and Kenneth B Finch. Corrected high–frame rate anchored ultrasound with software alignment. *Journal of Speech, Language, and Hearing Research*, 2011.

[39] James M Scobbie, Alan A Wrench, and Marietta van der Linden. Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In *Proceedings of the 8th International seminar on speech production*, 2008.

[40] Donald Derrick, Christopher Carignan, Wei-rong Chen, Muawiyath Shujau, and Catherine T Best. Three-dimensional printable ultrasound transducer stabilization system. *The Journal of the Acoustical Society of America*, 144(5) :EL392–EL398, 2018.

[41] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking : A survey. *Acm computing surveys (CSUR)*, 38(4) :13, 2006.

[42] Yue Wu and Qiang Ji. Facial landmark detection : A literature survey. *International Journal of Computer Vision*, 127(2) :115–142, 2019.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[44] Mohammad Hamed Mozaffari and Won-Sook Lee. Freehand 3-d ultrasound imaging : a systematic review. *Ultrasound in medicine & biology*, 43(10) :2099–2124, 2017.

[45] Li Deng, Dong Yu, et al. Deep learning : methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4) :197–387, 2014.

[46] Eric Slud, Maureen Stone, Paul J Smith, and Moise Goldstein Jr. Principal components representation of the two-dimensional coronal tongue surface. *Phonetica*, 59(2-3) :108–133, 2002.

[47] Ka-Wa Yuen, Wai-Kim Leung, Peng-fei Liu, Ka-Ho Wong, Xiao-jun Qian, Wai-Kit Lo, and Helen Meng. Enunciate : An internet-accessible computer-aided pronunciation training system and related user evaluations. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 85–90. IEEE, 2011.

[48] M Hamed Mozaffari and Won-Sook Lee. Bownet : Dilated convolution neural network for ultrasound tongue contour extraction. *arXiv preprint arXiv :1906.04232*, 2019.

[49] M Hamed Mozaffari, David Sankoff, and Won-Sook Lee. Ultrasound tongue contour extraction using bownet network : A deep learning approach. In *Proceedings of Meetings on Acoustics 178ASA*, volume 39, page 020001. Acoustical Society of America, 2019.

[50] M Hamed Mozaffari and Won-Sook Lee. Domain adaptation for ultrasound tongue contour extraction using transfer learning : A deep learning approach. *The Journal of the Acoustical Society of America*, 146(5) :EL431–EL437, 2019.

[51] Catherine Laporte and Lucie Ménard. Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. *Medical image analysis*, 44 :98–114, 2018.

[52] Jian Zhu, Will Styler, and Ian Calloway. A cnn-based tool for automatic tongue contour tracking in ultrasound images. *arXiv preprint arXiv :1907.10210*, 2019.

[53] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net : deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1) :67, 2019.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[55] TY Zhang and Ching Y Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3) :236–239, 1984.

[56] Fiona E Gibbon and Alice Lee. Articulation : Instruments for research and clinical practice. *Cleft lip and palate : Speech assessment, analysis, and intervention*, pages 221–238, 2011.

[57] Fiona E Gibbon and Sara E Wood. Visual feedback therapy with electropalatography. In *Interventions in speech sound disorders*. Paul H. Brookes Publishing Co., Inc., 2010.

[58] M Hamed Mozaffari, Md Ratul, Aminur Rab, and Won-Sook Lee. Irisnet : Deep learning for automatic and real-time tongue contour tracking in ultrasound video data using peripheral vision. *arXiv preprint arXiv :1911.03972*, 2019.

[59] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1) :79–82, 2005.

[60] Bryan Gick, Sonya Bird, and Ian Wilson. Techniques for field application of lingual ultrasound imaging. *Clinical Linguistics & Phonetics*, 19(6-7) :503–514, 2005.

[61] Masaki Noguchi, Noriko Yamane, Asami Tsuda, Misuzu Kazama, Bosung Kim, and Bryan Gick. Towards protocols for l2 pronunciation training using ultrasound imaging. In *Poster presentation at the 7th annual Pronunciation in Second Language Learning and Teaching (PSLLT) Conference. Dallas, TX*, 2015.

[62] Manuel Sam Ribeiro, Aciel Eshky, Korin Richmond, and Steve Renals. Ultrasound tongue imaging for diarization and alignment of child speech therapy sessions. *arXiv preprint arXiv :1907.00818*, 2019.