COMPUTER VISION-BASED TRACKING AND FEATURE EXTRACTION FOR LINGUAL ULTRASOUND

by

KHALID AL-HAMMURI
B.Sc., Yarmouk University, 2012

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of
Master of Applied Science
in the Department of Electrical and Computer Engineering

Supervisory Committee

COMPUTER VISION-BASED TRACKING AND FEATURE EXTRACTION FOR LINGUAL ULTRASOUND

by

KHALID AL-HAMMURI
B.Sc., Yarmouk University, 2012

Supervisory Committee

Dr. Poman. So, Supervisor
Department of Electrical and Computer Engineering
University of Victoria

Dr. Alexandra. Branzan Albu, Co-Supervisor
Department of Electrical and Computer Engineering
University of Victoria

Dr. Sonya. Bird, Outside Member
Department of Linguistics
University of Victoria

Abstract

Lingual ultrasound is emerging as an important tool for providing visual feedback to second language learners. In this study, ultrasound videos were recorded in sagittal plane as it provides an image for the full tongue surface in one scan, unlike the transverse plane which provides an information for small portion of the tongue in a single scan. The data were collected from five Arabic speakers as they pronounced fourteen Arabic sounds in three different vowel contexts. The sounds were repeated three times to form 630 ultrasound videos. The thesis algorithm was characterized by four steps. First: denoising the ultrasound image by using the combined curvelet transform and shock filter. Second: automatic selection of the tongue contour area. Third: tongue contour approximation and missing data estimation. Fourth: tongue contour transformation from image space to full concatenated signal and features extraction. The automatic tongue tracking results were validated by measuring the mean sum of distances between automatic and manual tongue contour tracking to give an accuracy of 0.9558mm. The validation for the feature extraction showed that the average mean squared error between the extracted tongue signatures for different sound repetitions was 0.000858mm, which means that the algorithm could extract a unique signature for each sound and across different vowel contexts with a high degree of similarity. Unlike other related works, the algorithm showed an efficient and robust approach that could extract the tongue contour and the significant feature for the dynamic tongue movement on the full video frames, not just on the significant single and static video frame as used in the conventional method. The algorithm did not need any training data and had no limitation for the video size or the frame number. The algorithm did not fail during tongue extraction and did not need any manual re-initialization. Even when the ultrasound image recordings missed some tongue contour information, the thesis approach could estimate the missing data with a high degree of accuracy. The usefulness of the thesis approach as it can help the linguistic researchers to replace the manual tongue tracking by an automated tracking to save the time, then extracts the dynamics features for the full speech behaviour to give better understanding of the tongue movement during the speech to develop a language learning tool for the second language learners.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my sincere appreciation to my supervisors, Dr. Poman So, Dr. Alexandra Branzan Albu and Dr. Sonya Bird for their guidance, support and cooperation. Their suggestions and positive feedback for further improvement of my thesis helped me to enhance the quality of my work.

I would also like to thank Mr. Patrick Szpak from the speech research lab in the Linguistic department who helped me during the recording of the ultrasound videos.

Finally, I want to express my deepest gratitude to my parents for being a constant source of encouragement and support during the studying years away from home. I do not think I would have succeeded without their inspiration.

Dedication

To my family,

friends,

and the whole world.

Chapter 1

Introduction

This research focuses on using computer vision techniques to develop a methodology of analyzing lingual ultrasound videos during speech. The method is developed to be used by linguistic researchers to track the tongue contour and extract the significant articulatory features for different speech sounds. Understanding the details of dynamic tongue movement is especially beneficial for teaching and learning novel speech sounds in pedagogical and clinical contexts.

Computer vision is a field that mimics the function of the human visual system by using computer science algorithms in image processing. Using computer vision in the linguistic research areas has many advantages. Computer vision techniques can enhance the quality of the ultrasound images which improve the clarity of the tongue contour. Besides, the algorithm can automatically detect and track tongue contour to reduce the time of manual contour monitoring as in the typical method. Furthermore, using computer vision algorithms is useful for tracing the rapid tongue movements and deriving missing data by building an estimation model to improve real-time tracking. This is difficult to be achieved by humans, especially with cases of long video frame sequences and huge databases, considering various missing information due to the ultrasound artifacts. Section 1.1 provides a general overview of the visual feedback systems and the use of the ultrasound linguistic studies. Section 1.2 discusses the main challenges facing the lingual ultrasound. Section 1.3 highlights the principal contributions in this research thesis.

1.1 Visual Feedback and Lingual Ultrasound

The tongue has an essential role for all oropharyngeal behaviours such as chewing, swallowing, breathing and speech. During speech articulation, the tongue has a significant contribution to the vocal tract shape which forms our acoustic signal [1]. The shaping of

the tongue during speech is a result of the coordination of a complex array of muscles by nerve motor activity [2].

The methodologies for teaching pronunciation have developed throughout history. The cited work [21] reviews the methods of teaching second language learners pronunciations and describes the significant changes at four periods . The period starting at the 1940s to the end of the 1960s was focused on using the audiolingual method, or using the sound segments to focus on the pronunciation. In the 1960s to early 1980s, the ability of learner communication and finishing specific tasks were the main areas other than relying on the pronunciations. In the 1980s to the 1990s, the pronunciation teachings regenerated with considering the articulation of specific sounds, stress, rhythm, intonation, and the voice quality for articulatory settings. After the 1990s, computer-aided pronunciation approaches were introduced to improve the development of visual articulation feedback teachings.

Computer-assisted visual articulation feedback techniques can be subdivided into three categories. The simulation, the indirect and the direct techniques [22] are the main categories for computer assisted tools. The simulation technique provides visual information for L2 learners (second language learners) by using computer simulation software that mimics the articulation process. In-direct feedback refers to the methods that provide visual feedback based on the analysis of the acoustic data. The indirect feedback refers to the visual feedback that is not provided at the same time of data collection as the data need further processing after collecting it, examples for the indirect feedback are spectrographic, formant and vocal tract resonance displays. In a parallel way, ultrasound and intra-oral techniques are considered as a direct visual feedback, which provides a real-time visualization for the articulation process.

The recent technology developments provide linguistic researchers with useful tools to simulate the articulation behaviours. Virtual reality (VR) has recently taken a role in the linguistic research field and has been applied to many studies [51]. Statistical data from the magnetic resonance imaging (MRI) and electromagnetic articulation (EMA) were used to build a 3D model for the tongue behaviour to enhance the learning process by providing corrective feedback visualized on a monitor in front of the learner [52].

The spectrograms uses the frequency and intensity of the speech waveform to represent the movements of the oropharyngeal structures such as tongue, larynx, lips and jaw. Figure 1 depicts the waveforms and spectrograms for two sounds “a” and “i”.

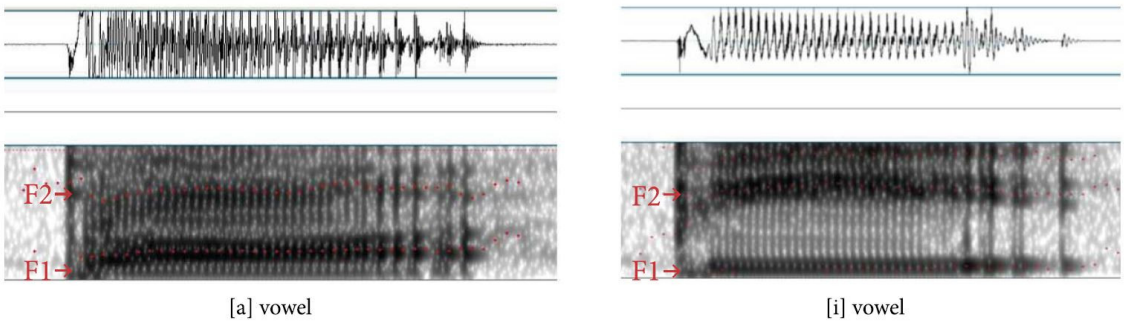


Figure 1. Spectrograms (On the bottom) and waveforms (on the top) for [a] and [i]. F1 and F2 are the first and second formant respectively [22].

Figure 2 shows the formant plot style for visualizing the articulatory speech. Formant uses a simplified visualization of the raw data by using the peak resonance information, which corresponds to tongue position for the utterance (speech) waveform.

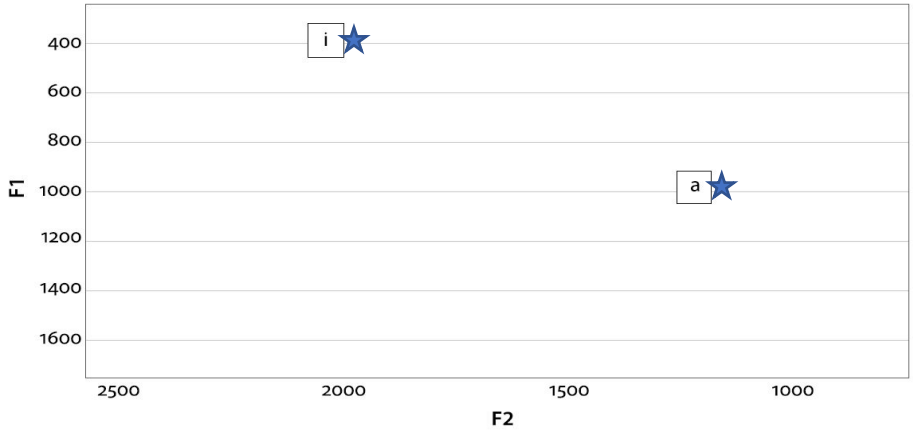


Figure 2. Formant display of [a] and [i] vowels. F1 and F2 are the Y and X positions of the tongue vertex (the highest point of the tongue) respectively [22].

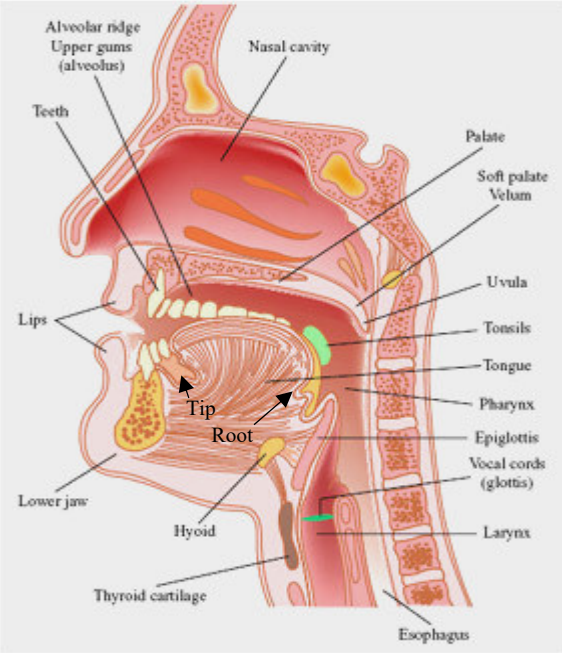


Figure 3. Vocal Tract Anatomy [24]

In the direct visual feedback techniques, information about the vocal tract (see Figure 3) is visualized at the same time as the speech to give real-time data visualization. Intra-oral techniques such as electropalatography (EPG), electromagnetic articulography (EMA), X-ray, MRI and ultrasound are the main direct approaches. EPG is made from an artificial palate with electrodes that track the location and timing of tongue contact with the palate. EPG is used for speech therapy, but there is no strong evidence of it being used in pronunciation teachings. In EMA, induction coils around the subject's head generate electromagnetic waves to stimulate the sensory coils in the mouth. The information from the stimulated coils is collected to provide measurements regarding the tongue movements by an external device that record and analyze the signal. MRI and X-ray can provide useful information about the vocal tract, but the main challenge is that those systems are not portable and they are just available in hospitals for clinical use. They are very large, expensive and it is difficult to use them in second-language learning process.

Among different imaging techniques, ultrasound imaging is currently considered one of the essential modalities that are used for articulation speech analysis. Lingual ultrasound imaging is deemed to be efficient, safe, portable and easy to use compared to other techniques. It has been used in linguistic phonetic research for a few decades and it has

also been increasingly applied as direct visual biofeedback in speech and language therapy [4]. The technique provides useful information about most of the tongue between the root and tip for quantifying the changes in tongue shape between different speech sounds [3].

In the case of the lingual ultrasound, the transducer is placed underneath the chin and the acoustic waves travel upward to be reflected from the tissue boundaries such as the upper tongue surface. The top side of the tongue surface is bound by either air or the palate bone. These have a substantial difference in their densities compared to the tongue which causes a strong echo (see Figure 4). The reflected echo from the tissue edges is displayed on the ultrasound image as a bright spot; the brightness increased as the difference between the tissue densities on the boundaries increased. In the tongue, there are also weaker echoes between muscle, fat and connective tissue interfaces. Ultrasound probe works as a transceiver, which both transmits and receives acoustic waves to measure the distance based on the time it takes for an echo to return. The recording session requires an average time from 30 to 60 minutes. A portable ultrasound system used during the recording session, typically has a head support system to stabilize the subject's head [1].

While the simulation process provides an estimation of the oropharyngeal behaviours, it is difficult to create a universal model to handle the variety of human tongue contour shapes. Although the intra-oral techniques EPG and EMA provide useful information about the tongue movement, they have downfalls. They are invasive techniques that require insertion inside parts of the human body and they are not efficient for the language learning process. However, it may be useful for research purposes such as using EMA to build a 3D model and extract statistical data for the tongue movement. Conversely, the ultrasound is an efficient method for the direct visualization of the tongue during the speech as it is portable, safe, and can provide accurate visual data for the tongue movement.

1.2 Main challenges for tongue ultrasound

Tongue ultrasound encounters several obstacles that can affect the quality of the final image and speech analysis. For example, high speckle noise is the appearance of small speckles (small dots, see Figure 5) on the ultrasound image when the ultrasound waves scatter from objects is smaller than the wavelength of the sound wave [6]. As a result,

ultrasound images have a low signal-to-noise ratio due to acoustic signal attenuation in the media of propagation and image artifacts (e.g. shadowing, mirroring, reflection and refraction) [5] — Furthermore, there is a need of a specialized person to obtain the image (See Chapter 2 for more details). However, language teachers and researchers used to record the lingual ultrasound, but there is no guarantee for the accuracy and quality of their recordings as handling the ultrasound instrument is complicated and extensive knowledge in this field is required. However, the ultrasound system is considered less complicated compared to other imaging modalities, but recording an excellent image requires adjusting many parameters and proper handling of the ultrasonic transducer. The tongue is deep inside in the oral cavity and to get image, the transducer should be placed beneath the chin. The transducer placement should be handled carefully as there are many obstacles in the path of ultrasound waves. Hyoid and mandible bones are the main obstacles in the oral cavity that create an acoustic shadow (black region) as they refract the acoustic waves away from the transducer. To avoid the acoustic shadow, we may push the transducer upward in the chin to make the transducer tip positioned in between the hyoid and mandible bones. Although this technique reduces the acoustic shadow, it may deform the posterior tongue segment, creating errors in its apparent shape, since the transducer may be pushed with an angle toward the tongue tip [1]. Figure 4 depicts the oral cavity and transducer placement in the midsagittal plane.

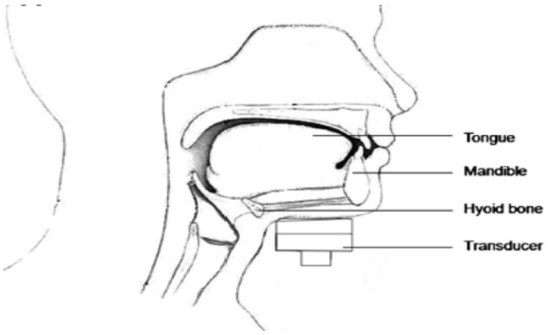


Figure 4. The oral cavity anatomy and the mid-sagittal placement of the ultrasound transducer [7].

Bone is considered a dense structure while the surrounding tissues have different densities that cause an impedance mismatch. This mismatch leads to a significant ultrasonic

loss due to the reflection and refraction of the acoustic waves. Besides, the bone is more absorbing than the soft tissues [8]. The consequence of the ultrasonic losses is missing some information from the apparent tongue on the ultrasound image especially that belongs to the tongue tip and the root. Figure 5 depicts the shadow effect of the mandible and hyoid bones marked by blue arrows.

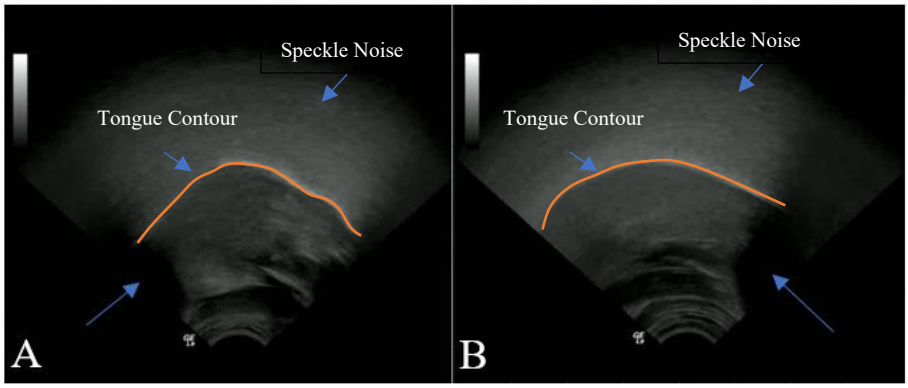


Figure 5. A. Ultrasound image for a female subject shows hyoid bone shadow effect on the tongue root noted by an arrow. B. Ultrasound image for a male subject shows mandible bone shadow effect on the tongue tip pointed out by an arrow. In orange the tongue boundary

The tongue surface appears as a bright line due to the high reflection at the tongue tissues air boundary. Tendons and blood vessels are also within the tongue structure which creates a bright profile. This bright profile may create an unreliable intensity profile along the tongue contour because they make the tongue thicker than normal. Monitoring tongue contour during speech is more challenging than observing it during silence. This is because tongue movements during the articulation deform the tongue to make it more concave during tongue muscle contraction while in the silence, the geometry is close to the smooth arc line (see tongue contour on Figure 5). During speech, the direction of the sound wave propagation with reference to the tongue surface angle becomes sharper or even parallel, which reflects the waves away from the transducer and causes a sporadic (discontinued) disappearance of the entire tongue structures as the ultrasound signal from some part of the tongue is missed.

The required time for manual tongue tracking is between one to two minutes for each frame while the automatic tracking is within few seconds for each frame. There is a need for automatic tongue contour tracking because manual tracking is 15 to 30 times slower than the automatic tracking. Tracking tongue contour requires automatically tracing the bright line. Although the apparent white line depicting tongue contour is the brightest one in a clear image, many other structures are visible in the image which causes a tremendous challenge for automatic detection. As depicted in Figure 6, the genioglossus muscle, geniohyoid, and mylohyoid muscle are dense structures that create a bright spot at the mouth floor which causes an obstacle for the automatic tongue tracking.

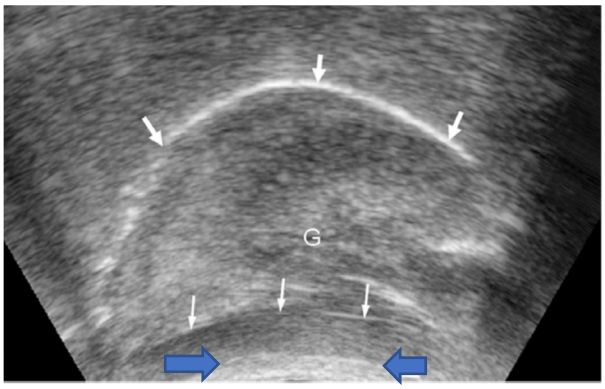


Figure 6. Midsagittal ultrasonography image showing the genioglossus muscle (G), geniohyoid (arrows), and mylohyoid muscles (arrowheads) at the mouth floor. The tongue surface noted by blue arrows [7].

To maintain a fixed position for the ultrasound transducer relative to the head, the ideal arrangement is to keep the head-transducer alignment fixed, while the jaw has free movement. In this research, a head and transducer support system has been used for the stabilization of the head and transducer but there is no guarantee that there will be no displacement on the head-transducer alignment. Any displacement can cause unreliable results for the tongue contour position. For this reason, the fixed position is essential to maintain the same reference for all ultrasound images at different recording sessions. Using the head support system can increase the accuracy of the tongue contour tracking data, however, it is claimed that a 0.5 cm shift on the tongue surface is caused due to transducer pressure in the upward direction [1]. Even during the jaw movements, the pressure effect

is almost the same as the transducer is fixed by a spring holder below the chin to keep the pressure during the movements. Figure 7 depicts the recording system used in research including the ultrasound and head-transducer support system.

Lingual ultrasound studies need to use video frame rates between 30Hz to 200Hz to detect the rapid tongue movements. Unfortunately, many systems are not able to record high frame rate videos due to the ultrasound system limitations which is typically 30Hz. This limits the ability to study some phonetic sounds such as the sounds that are related to the tongue tip, which can move very quickly. Having lower frame rates reduces the temporal resolution, which limits capturing the tongue signature in rapid tongue movement. The inadequate frame rate can be alleviated by the averaging method in [1], which adds additional frames between the frame sequences to increase the sampling rate. This is done by taking the average between two sequential frames [1]. The ultrasound system used in the thesis project has a 60Hz frame rate, which is acceptable for most of the articulation sounds. Additionally, in the thesis project, a new approach was proposed to handle the low frame rate issue by transferring image data from the individual video frame sequences to one 2D full concatenated data to represent the whole speech behaviour, not just focusing on the most prominent signature as in the typical method. The tongue contour from the sagittal plane is important as it can provide useful information about the deformation of the tongue surface in a 2D plane during the speech, which is sufficient for lingual analysis [20]. This will be discussed in more detail in Chapter 4. Besides, the ability to process the large video frame sequences was not applicable in many related works due to some technical limitations for each method (see Chapter 3 for more details) which can hinder the system capability for analyzing the long recording sessions. It can also cause difficulty in recording continuous speech sounds rather than just a single voice. More details about the related works limitations will be discussed in Chapter 3. In this thesis project, there were no limitations for analyzing any ultrasound videos regardless their size or length.

Figure 7. Ultrasound with the head-transducer support system.

Furthermore, analyzing the collected data has many challenges due to the varying traits of the participants such as age and gender as the morphological shape of the oropharyngeal structure varies with these factors [53,54]. Due to the diversity of human tongue contour shapes, the oropharyngeal shape was also diverse; for instance, jawbones for the females were relatively smaller than the males'. Tongue muscle shape and flexibility were also different for each person. Even for the same speech sound, some sounds had a source other than the tongue like the pharyngeal sounds, and the tongue movements did not make the main contribution for the final sound. This means that analyzing that kind of sound from the tongue ultrasound images was more complicated than analyzing the sounds that generated mostly from the tongue movement. All the previous challenges required additional processing and analysis which will be discussed in more detail throughout the following chapters.

1.3 Contribution and research methodology:

This research project proposed a new approach for automatic tracking, processing and analyzing of the tongue ultrasound images. The main contributions were data collection and algorithm design. Ultrasound data was collected with a collaboration of the University of Victoria linguistic department by using the speech research lab. The algorithm can be explained by breaking it down into four main blocks. First: image denoising algorithm by using a combination of the curvelet transform and shock filter; second: automatic selection of the tongue region of interest; third: missing data estimation and curve fitting; fourth: data transformation from the image space to one concatenated signal and features extraction.

The following chapters are structured as follows. Chapter 2 explains the theory of the ultrasound, Chapter 3 reviews the related works, Chapter 4 describes in detail the thesis proposed work, Chapter 5 discusses the experimental results, and Chapter 6 outlines the conclusion and suggests future works.

Chapter 2

Theory of Ultrasound Imaging

Sound wave is divided into three frequency ranges, namely the subsonic, sonic and ultrasonic. The subsonic wave is less than 20Hz and it is below the human audible range, whereas the sonic wave range is between 20Hz and 20 kHz and it is within the human audible range. The ultrasonic wave or ultrasound refers to the ultra-high acoustic frequency wave (larger than 20KHz) which is in the inaudible range of the human. In medical applications, the diagnostic ultrasound frequency ranges from 1MHz-20MHz. This Chapter discusses the theory of ultrasound imaging from the ultrasound generation to acquisition. Ultrasound imaging modality is portable, non-invasive, and cost-effective compared to other imaging modalities and it is being used for different medical applications such as capturing human anatomy, morphology, blood velocity and testing the function of the heart valves.

This chapter is structured as follows. First: generation of the ultrasound waves, second: ultrasound interaction with tissues, and third: ultrasound data acquisitions modes.

2.1 Generation of ultrasound waves

Ultrasonic waves are transmitted and received by piezo-electric crystals electro-mechanical energy conversion. When an AC voltage is applied to a piezoelectric crystal, it expands and contracts rapidly to transmit ultrasonic waves; this operation converts electrical energy to mechanical energy. While receiving an acoustic wave, energy conversion goes in the opposite direction in which an ultrasonic wave causes the piezoelectric crystal to generate an AC voltage. The amplitude of the acoustic wave is maximal when the thickness of the crystal is precisely half the wavelength. During the generation of the ultrasound, the waves interact with the medium or tissues that are close to the probe surface. To transmit the acoustic wave smoothly and to reduce the energy losses in the acoustic wave, the acoustic impedances (see section 2.1.1) of the piezo-electric crystal and the medium of the acoustic wave target (in our case the human tissues) should

be matched or the difference between the two impedances should be minimal. Unfortunately, acoustic impedance of human tissues and piezo-electric crystal are quite different. This impedance mismatch causes a significant loss of ultrasound energy as only a small portion of the source energy can propagate into the tissues to provide clinically desired information. Energy cannot propagate into the tissue and reflected back to the source. To solve this issue, a matching layer that has a specific acoustic impedance (the square root of the multiplication between crystal acoustic impedance and tissues acoustic impedance) is used to enhance energy transmission into the tissue and reduce the reflection. Furthermore, the acoustic gel is being used to create a bond between the transducer and the skin to eliminate the air between the probe and the skin. The air is considered as a strong reflector which create an obstacle for the wave propagation.

Figure 8 depicts the structure of the ultrasound probe components and the dual capability of the piezoelectric crystal in the generation and detection operations.

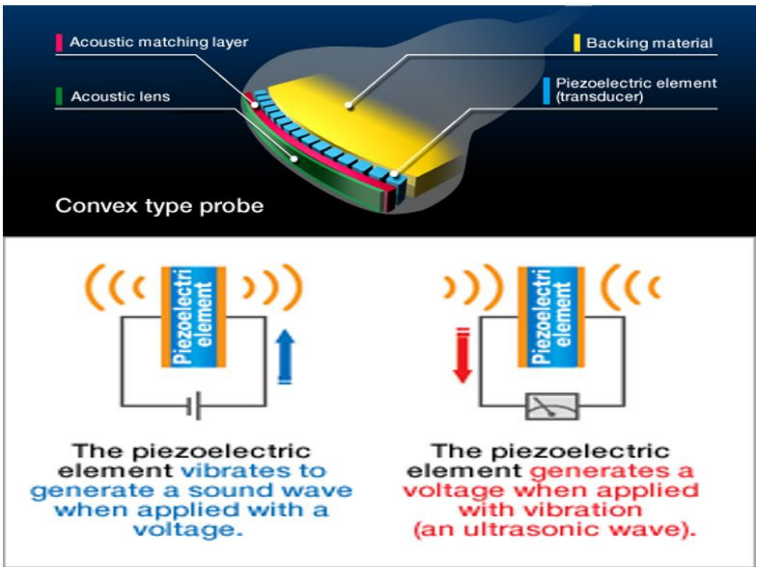


Figure 8. Top: ultrasonic probe structure, bottom: the basic principle of piezoelectric crystals [26].

2.1.1 Ultrasound wave propagation

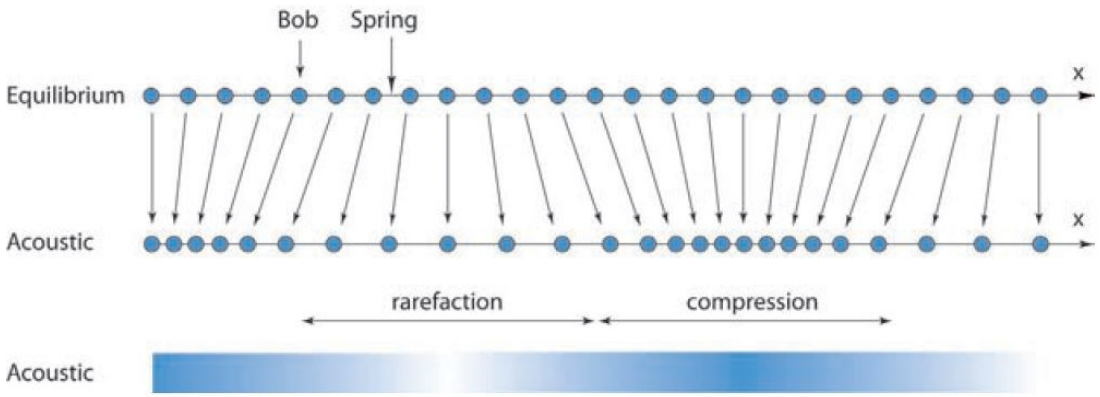


Figure 9. Schematically, a longitudinal wave represented as particles connected by massless springs that displaced from their equilibrium position [27].

Ultrasound waves carry kinetic and potential energy and propagate in the longitudinal direction, which can spread in all kinds of materials while the transverse propagation can only be in solid materials. The energy moves with the wave and any particles in the path of the propagated waves oscillate between compression and rarefaction (space) until the oscillation ends by damping (reaching equilibrium after losing the energy). Figure 9 depicts the longitudinal wave propagation.

The velocity of ultrasound waves depends on the density and elastic properties of the material. The substances that have higher densities can have higher acoustic speed and higher acoustic impedance. The ratio of the acoustic pressure to velocity of particle is characterized as the acoustic impedance Z.

$$\boldsymbol{Z = \rho c} \qquad (1)$$

Where c is the velocity of the acoustic wave and ρ is the density of the substance. Table 1 shows acoustic wave velocity and impedance for different substances.

Table 1. The values of acoustic impedance *Z*, and the acoustic wave velocity *c* of some substances [27]

Substance	*c* (m/s)	$Z = \rho c$ (10^6 kg/m2 s)
Air (25◦ C)	346	0.000410
Fat	1450	1.38
Water (25◦ C)	1493	1.48
Soft tissue	1540	1.63
Liver	1550	1.64
Blood (37◦ C)	1570	1.67
Bone	4000	3.8 to 7.4
Aluminum	6320	17.0

2.1.2 Pulse echo operation

In the pulse-echo operation, the ultrasound wave is transmitted as pulses with a certain delay between each pulse to allow for echo detection, which is the reflected pulses from the target. The pulse is generated in two to three cycles. The elapsed time between the transmission of the pulse and detection of the echo is related to the distance or depth of the target object (reflector). This is shown in equation (2).

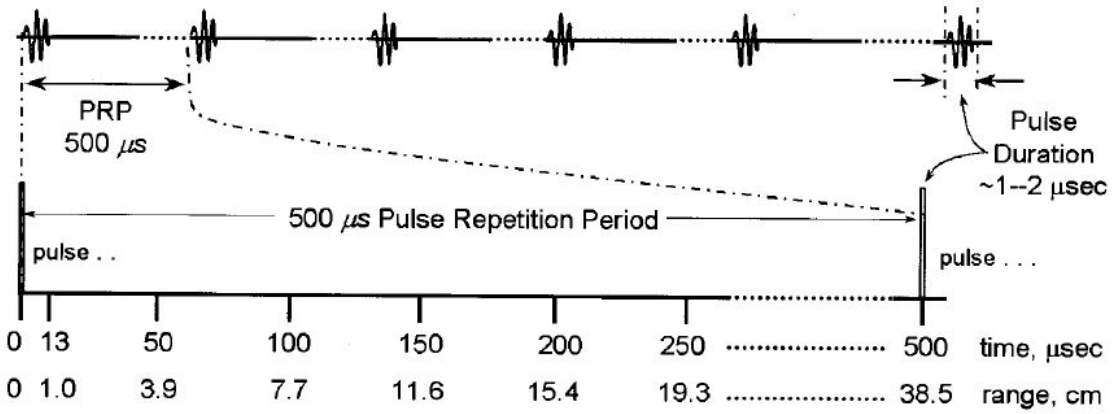
$$Elapsed\ time = \frac{Round\ trip\ distance}{Speed\ of\ sound} \tag{2}$$

The number of pulses per second is known as the pulse repetitions frequency (PRF). The time between pulses is the pulse repetition period (PRP) which is equal to the inverse of the PRF. The maximum PRF is limited by the elapsed time for the echo signal from the most distant structures to reach the transducer. PRP should be carefully identified because if the pulse transmission occurs before the detection of the echo from the previous transmitted pulse, part of the echo data from the most distant structures may be missed.

The maximal range of detection is one-half of the product of the speed of sound and PRP (the factor of 2 accounts for round-trip distance), as in Equation (3).

Maximal range= ½ × Speed of sound × PRP (3)

The penetration depth of the acoustic wave is inversely proportional to the ultrasound pulse repetition frequency. Hence a lower PRF signal has a deeper detection range. Figure 10 depicts an example of calculating the PRF from the PRP.



$$PRF = \frac{1}{PRP} = \frac{1}{500\,\mu s} = \frac{1}{500 \times 10^{-6}\,s} = \frac{2000}{s} = 2\,\text{kHz}$$

Figure 10. This figure shows the initial pulse occurring in a very short period with a pulse duration of 1 to 2 micro-sec, and the PRP is 500 micro-sec. PRF is 2 kHz. Range calculated at the speed of sound = 1,540 *m/sec* [29].

2.2 Ultrasound Interaction with tissues

Acoustic wave loses energy or changes its direction as it propagates through tissues. Reflection, refraction, scattering and attenuation are the four main phenomenon. While the reflected acoustic wave energy is detected by the ultrasound transducer, the transmitted acoustic energy attenuated by the tissue. Figure 11 depicts the four-phenomenon occurred along the path of wave propagation in the tissues.

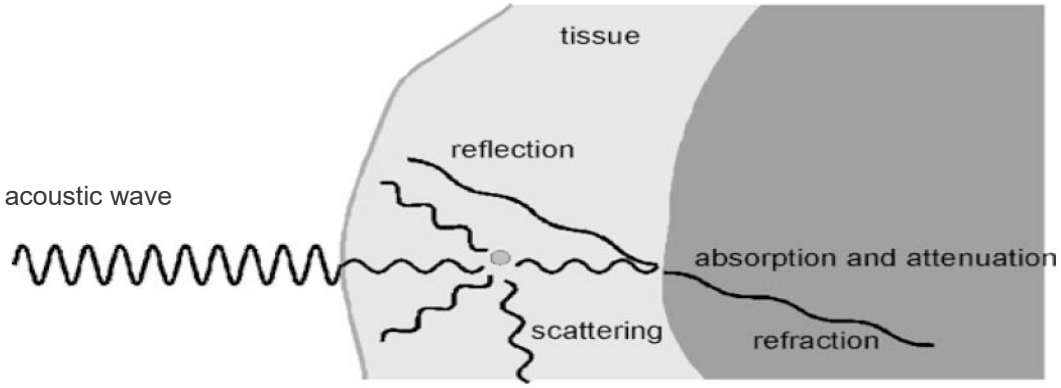


Figure 11. Interaction between the transmitted ultrasound wave and human tissues [28]

- **Reflection and Refraction:**

Acoustic wave propagating through two media with different acoustic impedances gets reflected at the interface separating the two media. The amount of reflection are given in equation (6). For a wave propagating in a medium with acoustic speed c_1 and density ρ_1 to a second medium with acoustic speed c_2 and density ρ_2, part of the wave is reflected and the other part is transmitted. If the direction of the incident wave is not parallel to the normal, the trasnmitted wave is refracted as described by Snell's law which describes the relation between the travelling wave angles with respect to the normal of the planar surface.

$$\frac{\sin \theta_i}{c_1} = \frac{\sin \theta_r}{c_1} = \frac{\sin \theta_t}{c_2} \qquad (4)$$

c_1 and c_2 represent the acoustic wave velocities in medium 1 and 2; while the $\theta_t, \theta_r, \theta_i$ are the angles of transmission, reflection and incident waves respectively. Figure 12 depicts the incident, reflected and refracted waves in the vicinity of a planar surface.

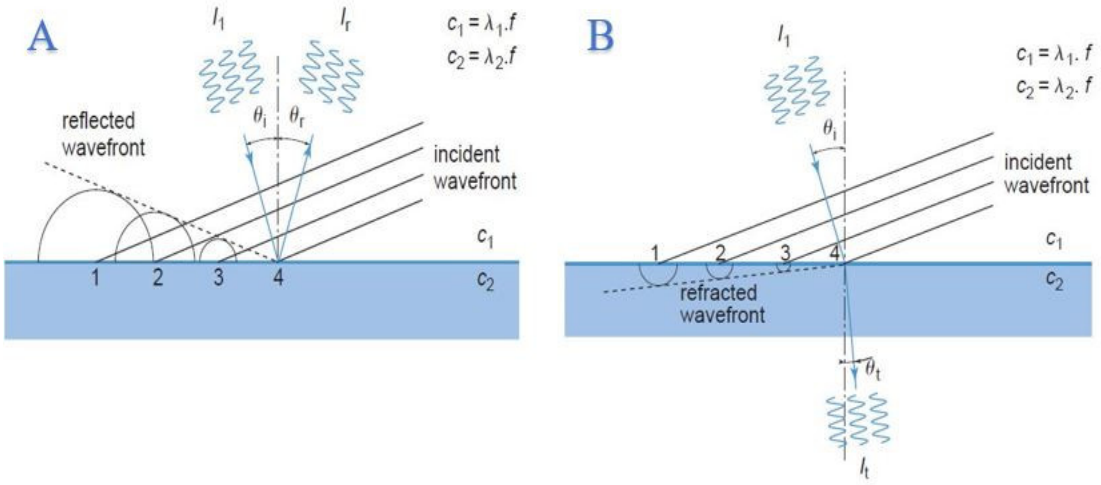


Figure 12. A, reflection of the acoustic wave from the planar surface in blue, interface separating the two media. B, refraction of the acoustic wave at the planar [27].

In addition to the change of directions, wave amplitudes also change. The amplitudes change are expressed in **T** and **R**, which are the transmission and the reflection coefficients respectively:

$$\boldsymbol{T = \frac{A_t}{A_i} = \frac{2Z_2 \cos\theta_i}{Z_2 \cos\theta_i + Z_1 \cos\theta_t}} \quad (5)$$

$$\boldsymbol{R = \frac{A_r}{A_i} = \frac{Z_2 \cos\theta_i - Z_1 \cos\theta_t}{Z_2 \cos\theta_i + Z_1 \cos\theta_t}} \quad (6)$$

Where A_t, A_r *and* A_i are the transmitted, reflected and incident wave amplitude respectively, Z_1 and Z_2 are the acoustic impedances of the two media.

Scattering:

The shape and size of the target object determine acoustic wave effect of wave scattering. There are three types. First, scattering from an object that is significantly larger than the acoustic wavelength, and is classified the specular scattering. Which the ultrasound wave is partially reflected from or transmitted through the object surface. Second, diffuse scattering occurs when the object is tremendously smaller than the acoustic wavelength. In this situation, the incident acoustic wave is scattered equally in all directions. Third, if the size of the object is similar to the incident acoustic wavelength, the diffractive scattering

occurs. In this case, the intensity of the scattered acoustic energy depends on the propagation of the scattered wave. Figure 13 depicts the three types of acoustic wave scattering.

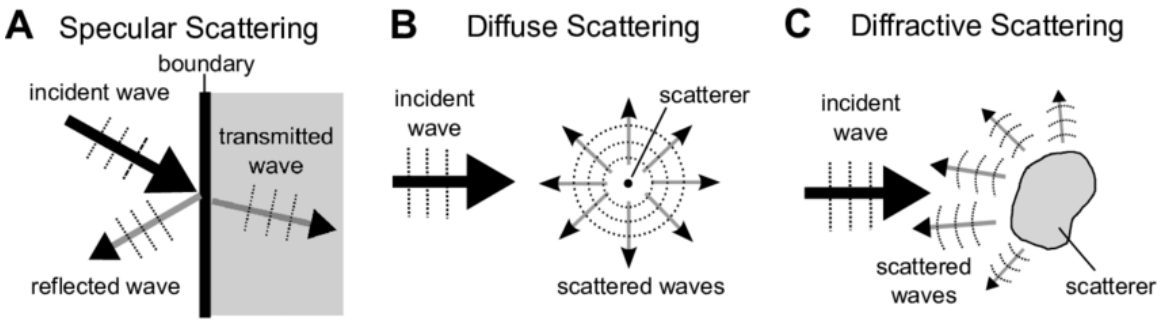


Figure 13. A, specular scattering. B, Diffuse scattering. C, Diffractive scattering [49].

- **Attenuation**:

Attenuation is characterized as the losses of the acoustic wave energy along the path of wave propagation. The amplitude of the propagating wave is exponentially reduced due to the viscosity of the tissues which convert the acoustic energy into heat. Attenuation is explained as a function of frequency and distance as in Equation (7).

$$\boldsymbol{H(f,z) = e^{(-\alpha z)} \equiv e^{(-\alpha_0 f z)}} \qquad (7)$$

Where z is the distance propagated by the acoustic wave in the tissue, α_0 is a material constant, f is the frequency of the acoustic wave and $\alpha = \alpha_0 f$ is the attenuation coefficient. Table 2 shows the values of α_0 for some typical biological substances.

Table 2. The values of the attenuation coefficient α_0 for some typical biological substances.

Substance	α_0 (dB/ (cm MHz))
Lung	41
Bone	20
Kidney	1.0
Liver	0.94
Brain	0.85
Fat	0.63
Blood	0.18
Water	0.0022

2.3 Ultrasound data acquisition modes

In this section, four different data acquisition modes for ultrasound imaging are discussed. They are A-mode, B-mode, M-mode and the doppler ultrasound-mode with continuous and pulsed wave.

2.3.1 A-mode

A-mode or amplitude-mode is a simple ultrasound mode that represents the reflected echoes from the tissue boundaries in the form of a digital signal proportional to the echo amplitude as a function of time. Digital signal amplitude varies as the attenuation of the acoustic wave varies based on the tissue densities and distance from the transducer. A-mode is used to localize the tissue interfaces along the path of the ultrasound beam, considering time and depth as almost linearly related as sound velocity is roughly constant in tissues. A-mode is rarely used nowadays, but it has been used in the past to determine the midline position of the brain for detecting the possible mass effect of brain tumours. A-mode is currently used in ophthalmology applications for precise distance measurements of the eye. Figure 14 depicts A-mode amplitude as a function of time.

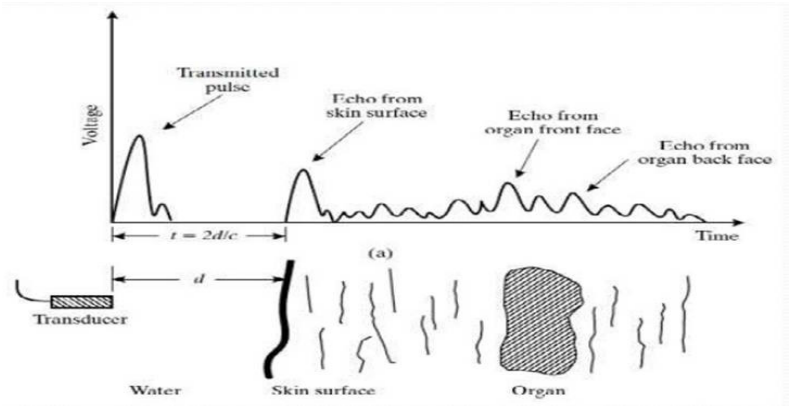


Figure 14. A-mode represented in voltage and time scale [30].

2.3.2 B-mode

In B-mode or brightness mode, the image is obtained by either translating or tilting the transducer between two A-mode acquisitions. Then the detected signal is modulated into dots with different brightnesses for each dot to resemble the density of the structures along the path of the acoustic wave. The brightness profile for each dot is proportionally related to the acoustic wave amplitude. B-mode acquisition is used in the application of the heart

or abdominal imaging. In B-mode, the ultrasound transducer can be either translated or tilted. Ultrasound acquisition by transducer translation mode is efficient for most of the applications, but it is not efficient for heart imaging and lingual ultrasound. Tilting the transducer with an angle is useful for cardiac imaging for capturing the heart image through the small space between the rips. In lingual ultrasound, tilting the transducer is useful for scanning the tongue between the mandible and hyoid bones. Because bone has a high attenuation coefficient, acoustic wave has a small penetration depth through bone. Figure 15 depicts the B-mode image acquisition in translating and tilting operations.

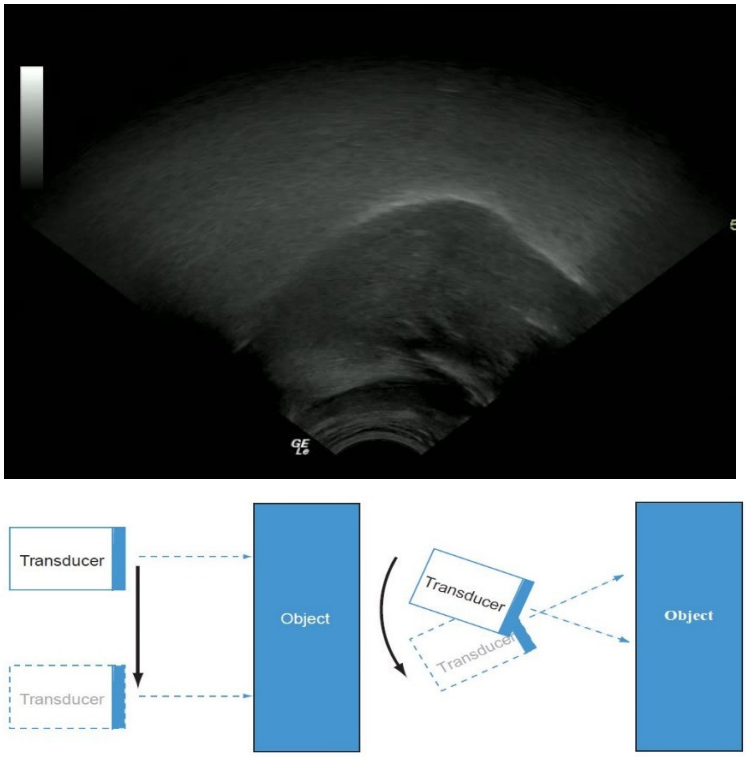


Figure 15. Top: B-mode image acquisition image. Bottom: translating or tilting the transducer in B-mode [27].

2.3.3 M-mode

M-mode, or motion mode, is a repeated B-mode measurement that uses a very high sampling frequency. The acoustic data from a single ultrasound beam is displayed as a function of time to monitor object movement which is represented by the object depth on the vertical axis and time on the horizontal axis. M-mode provides excellent temporal

resolution of motion detection. With the high temporal resolution, M-mode can be used to monitor the heart functions especially the movement of heart valves. Figure 16 depicts M-mode signals and compares it with A-mode and B-mode.

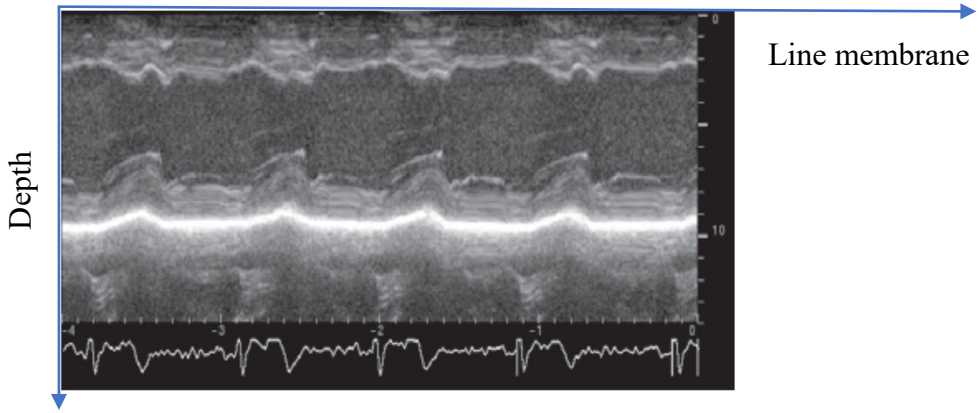


Figure 16. A display for M-mode data acquisition of the heart wall assessment. The black regions is the blood and the bright region is the heart membrane [27].

2.3.4 Doppler imaging

This section will explain two main types of Doppler imaging. First, there is continuous wave Doppler imaging (CW) which measures the frequency shift between the transmitted and received pulses. The frequency shift is used to measure the Doppler shift or Doppler frequency f_D as shown in Equations (8,9), where the f_R is the received frequency, f_T is the transmitted frequency, c is the speed of sound and V_a is tissues velocity.

$$\boldsymbol{f_D = f_R - f_T} \tag{8}$$

$$\boldsymbol{f_D = \frac{-2V_a}{c + V_a} f_T} \tag{9}$$

Doppler shift can be used to measure cardiac and blood velocity. Note that the Doppler frequency f_D is in the audible frequency range for human. The received signals are subdivided into segments. The Fourier transform for the signals give information about the frequency spectrum at subsequent time intervals. The spectral amplitude is encoded as a gray value in the image; the image is called the sonogram or spectrogram. The problem with CW Doppler is that it can just give velocity measurements without any depth

information. Figure 17 depicts the spectrogram of the CW Doppler image showing the velocity profile of the blood flow through a heart valve.

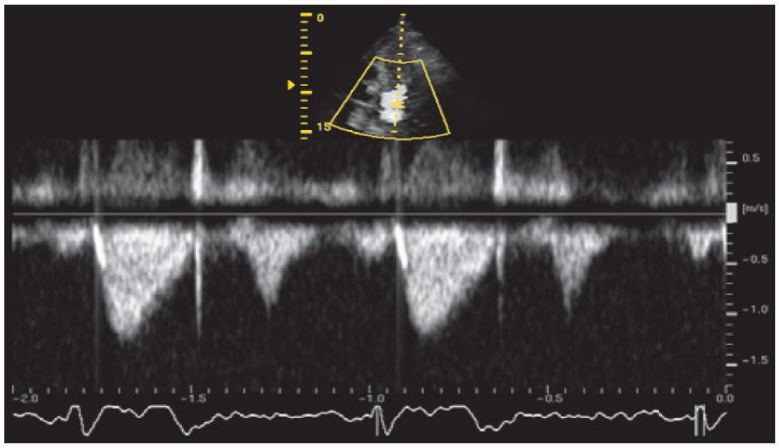


Figure 17. Velocity profile of the blood flow through a heart valve acquired by the CW Doppler spectrogram. The CW Doppler shows real-time information of the blood flow velocity profile [27].

The second type is the pulsed wave Doppler (PW). In the PW, the Doppler principle is not used (Doppler is the measure of frequency shift to estimate the velocity of the moving object). Instead, the received signal frequency is assumed to be the same as the frequency of the transmitted signal ($f_R = f_T$); the received signal is scaled and delayed from the transmitted one. If the scattering object moves away from the transducer with a constant axial velocity V_a, the distance increases between subsequent pulses with $V_a\, T_{PRF}$, where T_{PRF} is the pulse repetition period, $T_{\mathrm{PRF}} = \frac{1}{PRF}$. Equation (10) shows the relation between the Doppler frequency (f_D) and the velocity of the moving object V_a.

$$\boldsymbol{f_D = \frac{-2V_a}{c} f_T} \qquad (10)$$

Similar to the continuous wave Doppler image, Fourier transform is used to measure the frequency spectrum and the received signal is shown as a spectrogram. Figure 18 depicts a typical PW Doppler of blood flow in the aortic valve.

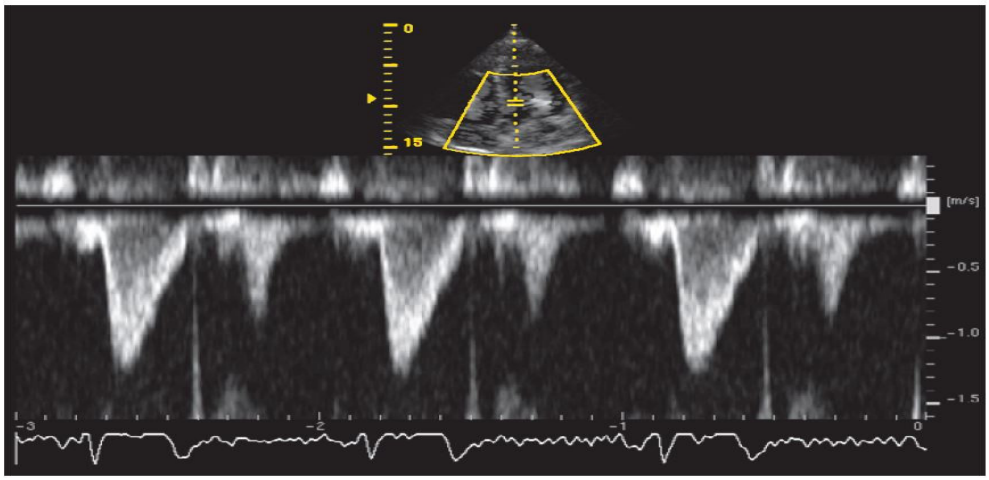


Figure 18. Typical PW Doppler spectrogram of blood flow through the aortic valve.

Chapter 3

Related Works

Automatic tongue monitoring is addressed in many research experiments by using different techniques as discussed in Chapter 1. However, the procedure for tongue tracking by using the lingual ultrasound is limited and can be characterized into two categories: techniques that require training, and techniques that do not. This chapter will review the related research experiments done on lingual ultrasounds and will be divided as follows. The first section reviews the non-training-based algorithms for tongue tracking. The second section reviews the training-based algorithms for tongue tracking.

3.1 Non-Training Based Algorithms for Tongue Tracking

Tracing the tongue motion in the ultrasound images requires using many image processing techniques like the image segmentation and pattern recognition algorithms. The base algorithm that is used for most of the tongue tracking methodologies is called the snake [43] which is an active contour that adapts to get closer and closer to the edges of the object, based on the selected energy threshold or gradient information. The cited work in [44] successfully used the snake to automatically segment and track the tongue contour. The first contour in the first video frame should be identified manually by an expert. The algorithm optimized the snake's external and internal energy functions for the first frame tongue contour to refine the detection of the tongue contour edges and the contour concavity respectively. The optimized contour from the current frame was used to initialize the snake in the next frame sequentially until the last frame. The snake algorithm has some drawbacks as it is highly sensitive for the speckle noise and even with this optimization, the algorithm may fail to detect the correct contour surface.

The cited work [13] proposed an improvement to the previous works [43,44] by using the available software Edgetrak. EdgeTrak is widely used by linguistics as it is free and easily accessible online. The improvements consider the gradient and intensity profile information in the local region around each snake element; while in the classical snake

[43], they just consider the gradient information. The enhancements optimized EdgeTrak to identify the tongue contour lower boundaries on the ultrasound image and take into consideration the tongue contour orientation to reject any undesirable edges.

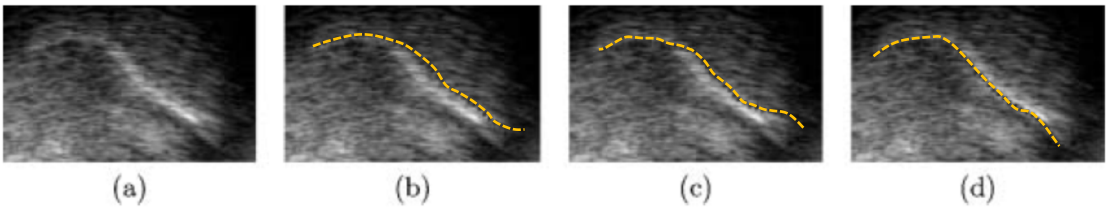


Figure 19. Extraction of tongue contour. (a) Ultrasound tongue image. (b) Snake initialization. (c) Edge extracted without optimization. (c) Edge retrieved after optimization.

EdgeTrak has many limitations. For example, the active contour and optimization processing time are computationally extensive. EdgeTrak also does not have denoising capability which decreases the tongue tracking accuracy because the snake's efficiency degrades with the noisy images. Furthermore, EdgeTrak could not process more than 80 frames at a time to limit the usage on just short frame sequences. Furthermore, EdgeTrak in some cases may identify the tongue contour outside the actual region of interest because the method does not use temporal smoothness with the snake's minimized internal energy in order to give more flexibility to detect tongue curvature. The detected tongue contour on the current frame is used to initialize the next frame and because of the limitations of the snake algorithm, EdgeTrak can fail in identifying rapid tongue movements. This is because the snake energy limits the adaptation of the snake from the current contour position to the new position of the tongue contour in the next frame. Figure 19 shows the efficiency of the optimization process to identify the lower tongue edge on the ultrasound image as in (d) and compare it with the results without using optimization.

Another research experiment handled the tongue tracking issue by introducing the TongueTrack software which is cited in [14,15]. TongueTrack formulated the tongue tracking as a global optimization problem that used the higher-order Markov random field (MRF) [17] to capture important information about the image which improved the detection of the tongue contour.

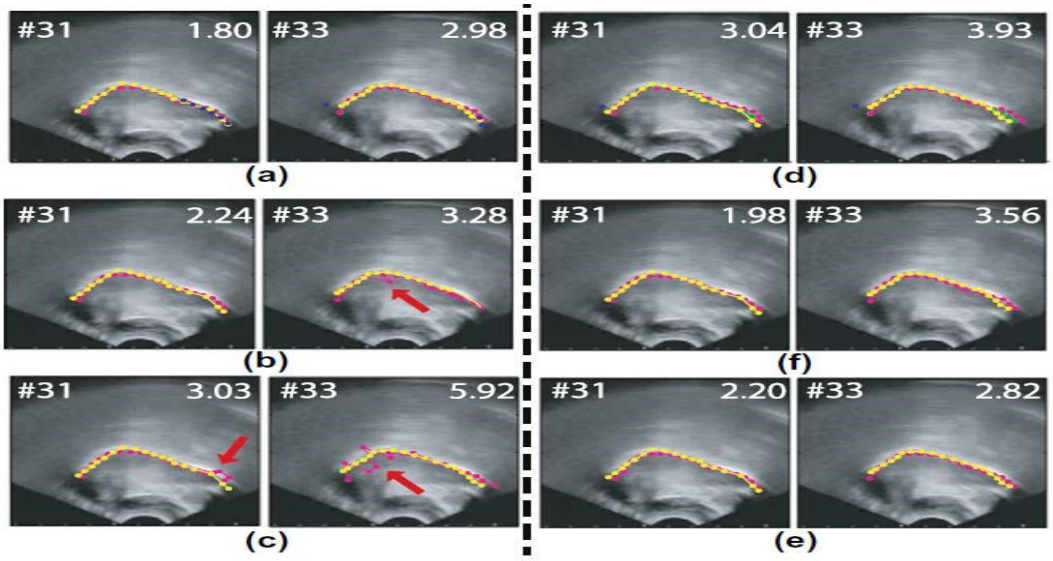


Figure 20. In pink, contours extracted by TongueTracks; in yellow, the manual reference. At the top left are mean errors and at the top right is the frame number. (a–f) depicts the effect of different input parameters contour and the arrows show the inaccuracies for some cases in the automatic method [14].

In the TongueTrack algorithm, three to five points should be manually identified on the tongue contour in the first frame to construct a template for the first contour. Then the points were fitted by a polynomial-curve fitting function to create a smooth and continuous contour. The algorithm generated an estimation model of the tongue contour deformation during the movements and stored them as a solution-space label set. After that, the current frame was compared to the estimated solution-space label set by an optimization process that minimized the MRF iteratively until reaching the user-defined threshold; which in the case of TongueTrack, is 2mm. TongueTrack was validated by comparing the results with the manually detected ground truth data of 63 ultrasound video sequences from two different groups of datasets. The reported accuracy of the mean sum of distances was 2-3mm based on the cited work [14]. Although TongueTrack provided some good results, there are some limitations to the software. The algorithm should optimize 9-parameters and to achieve a more accurate solution, the algorithm iterates 20 times by using different parameters to optimize the results. This makes it computationally expensive. In some cases, like the rapid tongue movement and noisy images, the algorithm may fail to detect the tongue contour and promptly ask for a manual reinitialization to correct the tongue contour

shape. Figure 20 depicts the TongueTrack segmentation and compares it with the manual reference and shows the effect of changing the parameters on the detected contour.

In a different manner of handling the tongue tracking, the tongue movement defined as mechanical system expressed in ordinary equations. To analyze the mechanical system a biomechanical model has been used in the cited work [20]. To use the biomechanical model, the algorithm was initiated by drawing a closed contour around the tongue surface and tongue internal points. Harries features detector was used to identify the most significant corners on the tongue contour edges. One hundred corner points were sorted in descending order based on the quality of each point. Then the displacements of these points in each subsequent frame were measured based on the optical flow algorithm which can derive the velocity and direction for each point. To reduce the number of features and the displacement error, the displacement of each corner feature position was estimated to be only in the neighbourhood of each corner point (about 15-30 pixels) in order to narrow the area of the estimated tongue movement in case of any external noise or missing information. The Covariance matrix was also computed to get rid of any unreliable features by computing eigenvalues of the matrix (the essential elements in the image). The authors of this work claimed results with the mean sum accuracy varied from 0.62mm to 0.97mm, based on changing different parameters. Although the reported results have a low error margin, they did not mention the size of the dataset used for validation. Furthermore, it is difficult to include all tongue deformation cases by this technique based on a biomechanical model that estimates tongue deformation during the speech. Many constraints and parameters were used to restrict the estimated motion which cannot be tuned easily. Although using Harries features to detect the most significant features is useful, there is no guarantee that the identified features will be visible in the next frame, especially in the fast tongue movement and low frame rate recordings. This limits the ability of this method to trace the feature or corner displacement from frame to frame. Figure 21 depicts the detected feature and approximated tongue contour by the biomechanical model method.

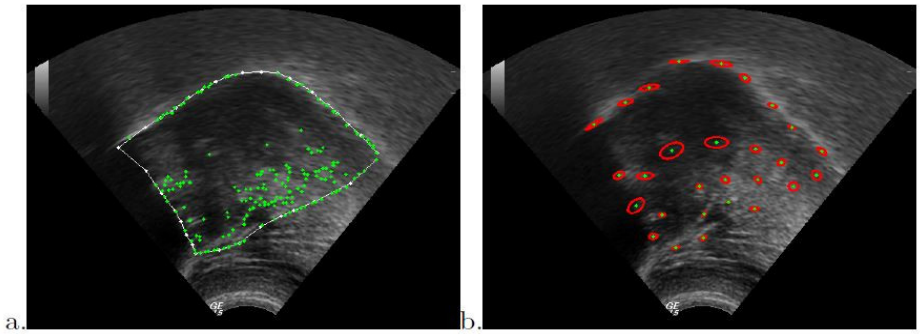


Figure 21. (a) The tongue region delineated on the first image of the sequence and the extracted features. (b) Covariance computed on the selected features with 20 pixels displacement on another image of the sequence [20].

3.2 Training Based Algorithms for Tongue Tracking

Because of the efficiency of artificial intelligence algorithms in many applications, there are many research projects in the lingual ultrasound tongue contour tracking that use machine learning and deep learning algorithms [11,16,19,48]. However, the training-based algorithms require collecting a huge database in order to be more accurate and valid for clinical use. Some promising results were reported in different kinds of literature to make the training-based algorithms applicable to be expanded in future research. The cited work [16] named as Autotrace suggested using the translational deep belief network (tDBN) [18] which is a machine learning algorithm that contains multilayers of the graphical model used to extract useful features from the image. The datasets were trained in two stages. The first stage extracted the features by using unsupervised learning (machine learning training method with known input and unknown output). The restricted Boltzmann machine was used for this purpose [16]. In the second stage, the algorithm learning was done by using supervised learning which is when the output is known for each known input data. In this case, a human-labelled tongue contour was used as a supervised learning output. The output results were validated by testing 8640 images collected from 7 subjects. The reported results showed that the mean sum distance (MSD) between the tDBN and hand-labelled contours is 2.5443±0.056 pixels (each pixel 0.295 mm). However, the results showed a small error margin but this approach is still dependent on the training data and any images outside of the testing and training datasets will not be accurate, especially for bad quality

images. Figure 22 depicts the tDBN output and the human-labelled tongue contours for different input ultrasound images by the Autotrace.

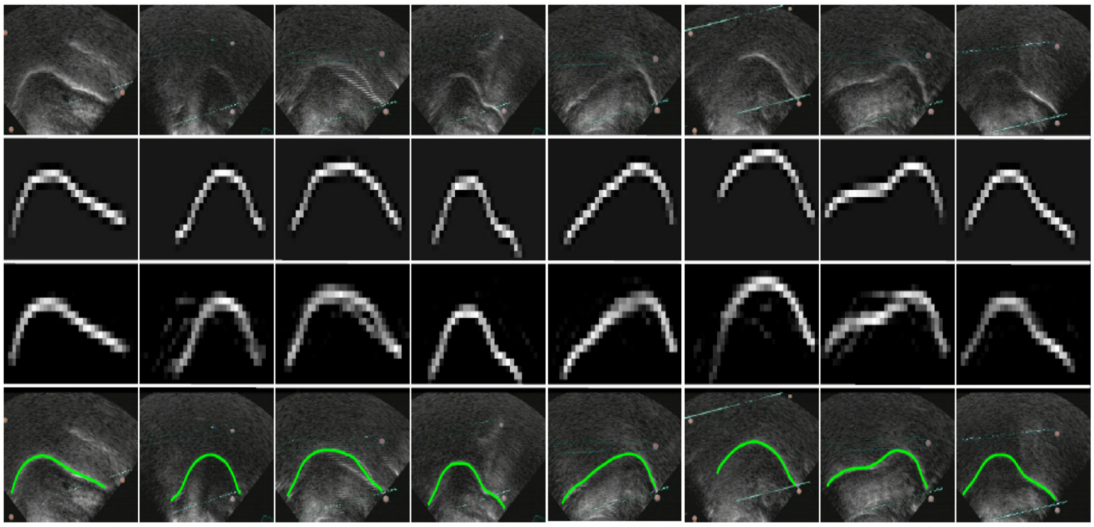


Figure 22. Top row: Ultrasound inputs. Second row: Manually delineated labels. Third row: tDBN outputs. Bottom row: Extracted contours [16].

Following the work that was done in the Autotrace [16], the cited work in [19] segmented the tongue contour by using the same deep learning method that was used in [16] with some modifications in the dataset training and labelling. Instead of using manual labelling, automatic labelling was used to reduce the labelling time. The algorithm proposes two phases for the data training. The first phase is learning the relationship between the ultrasound image and the extracted tongue contour. In this phase, the input is reshaped in a reduced image of 33x30 pixels to form a concatenated ultrasound image and binary tongue contour image. Secondly, the network is trained from the ultrasound image itself by using the ultrasound image to decode the hidden information in the image. Then the algorithm should be able to reconstruct the image and the tongue contour by comparing the second phase with the first one. Figure 23 depicts the two training phases used in the algorithm.

The method was validated by using 17000 example images (15000 used for training and 2000 for validation). Then another 50 images were selected randomly from the same recorded image session to test the results. The results showed that the average mean sum distance (MSD) was 1mm as a comparison between the extracted contour and hand-

labelled contours. While the MSD for the automatically extracted reference data and the detected contour was 0.8mm, they reported better results for the automated reference data than the hand reference. This can be arguable as we do not have enough information about the expert who labelled, therefore, we cannot determine how accurate the ground truth data is.

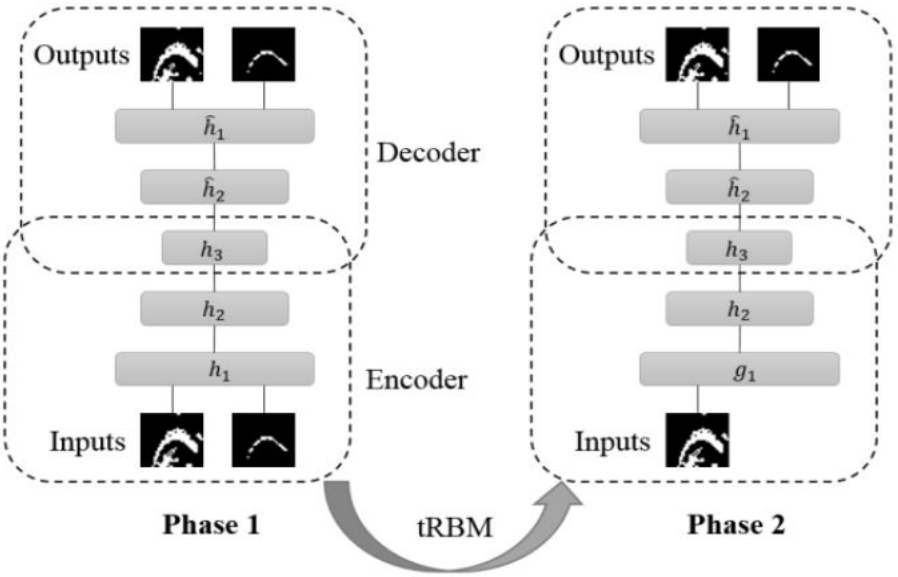


Figure 23. The two stages of learning. In the first one, the network learns the relationship between US images and the contour. The second one uses the relationship in the first phase to reconstruct the contour[19].

The cited work [47] derived an active appearance model from 700 X-ray images of the tongue to estimate the tongue contour position. Although they claim high accuracy for their estimation model, it did not seem to be a valid method for a universal tongue tracking application due to the dependency of this model on the X-ray images for the training. The work cited on [48] proposed what is called the active shape model. It was acquired from the lingual ultrasound segmented images to train the model based on the principal components analysis. Then it was used to estimate the tongue position only by ultrasound and without X-ray images. After using the active shape model to estimate the tongue contour position, the active shape model adjusted by using snake to fit the tongue contour edges.

To the best of my knowledge, the cited work [11] is the most recent work in tongue tracking and segmentation. It proposed a multi-hypothesis approach to extract tongue contours. The approach is based on using a motion model, particle filtering, and the active contour model (snake). The motion model was derived from the previously extracted tongue contours by manual labelling of different ultrasound images. To create the motion model, the extracted contours were normalized with respect to the position and length. Then the mean shape and principal component analysis were computed. From the previous step, information from different frames including the tongue position, scale and shape was used to compute the covariance matrix that contains the most important information of the tongue motion to use it in the model learning. To initialize the tongue tracker, at the first frame a manual set of points should be identified on the contour. Then the snake used these points to detect the full tongue contour edges. The segmented tongue contour from the first frame was copied for a certain number of copies called particles. In the next frame, the derived motion model was used to create multi-hypotheses for the tongue position, scale and coarse shape for each copied particle from the previous frame. The snake was used to adapt the estimated model contour position at each particle image to fit it properly on the actual tongue position. To select the best particle or the best-detected contour, each snake was optimized by using a band energy constraint that was suggested in [13] to ensure the contour is below the bright white band that represents the upper surface of the tongue [1].

The study was validated by using videos from healthy subjects and subjects with Steinert's disease, which is a form of myotonic dystrophy leading to slow speech, distorted vowels and consonants. The achieved mean sum of distances errors is 1.69 ±1.10 mm and they claimed that it was not highly dependent on the training data. However, the approach showed a novel methodology for the tongue tracking, but the accuracy was still dependent on the number of particles which increased the computational time in addition to the computational complexity of the snake algorithm. Figure 24 depicts the block diagram of the particle filter approach for tongue tracking and segmentation.

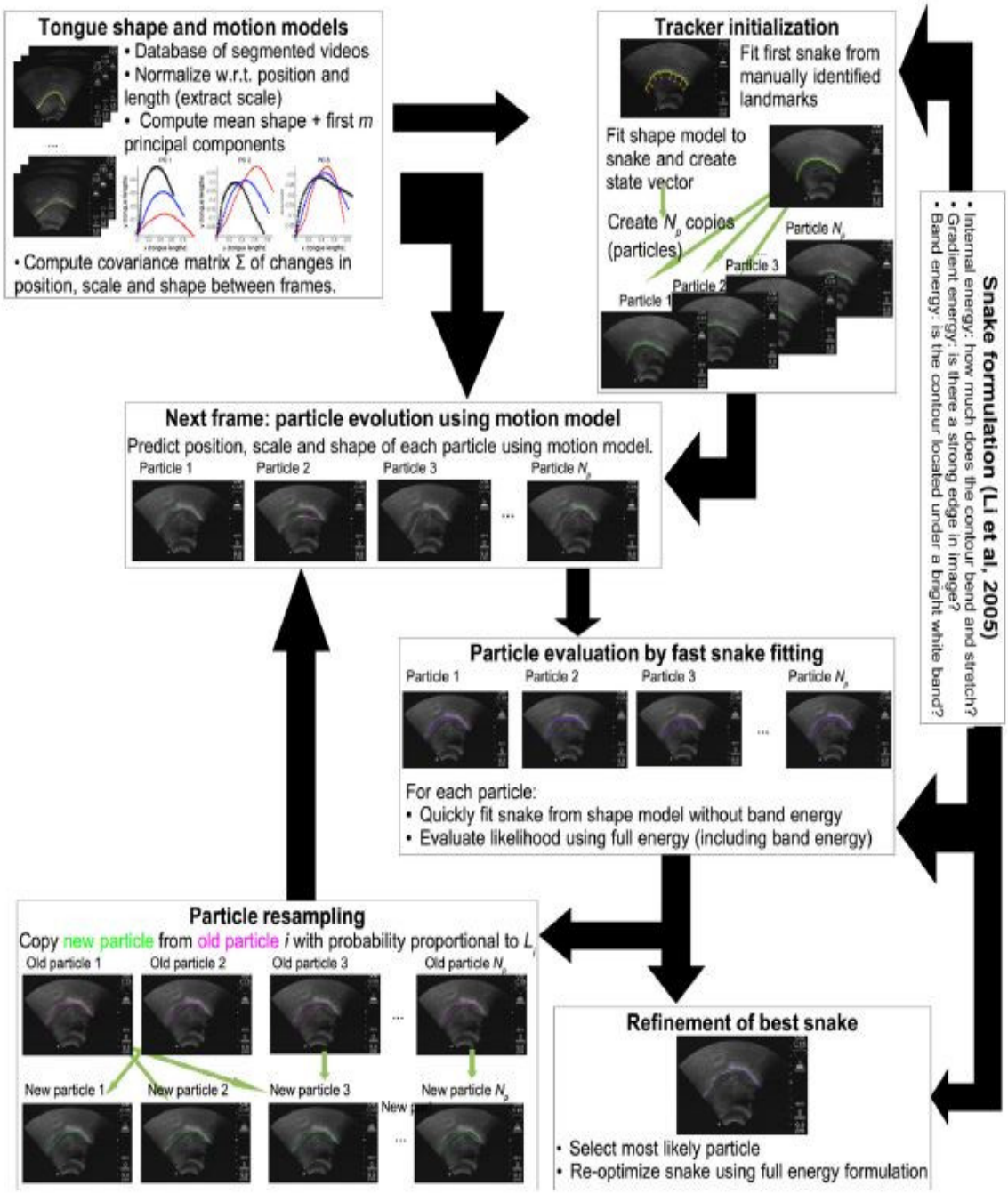


Figure 24. Particle filter tongue tracking and segmentation block diagram [11].

3.3 Summary

In this literature review, the tongue tracking approaches have been discussed in two main categories: the training and non-training-based tongue tracking algorithms. The non-training-based algorithms do not require any training database. Although the authors claimed some good results, there are many drawbacks for the associated works. For instance, EdgeTrak [13] and TongueTrack [14,15] need manual reinitialization during the processing as the algorithms fail during the rapid tongue movements. The processing frame length is limited also for both of them; Edgetrack can process up to 80 frames and TongueTrack is validated for 500 frames. The non-training based algorithms depend on modifying different parameters which make them expensive in the computation time. In addition, none of them implement a denoising technique which makes it difficult to analyze the ultrasound videos in the presence of high speckle noise and low signal to noise ratio. The biomechanical model proposed in [20] also suggested a new approach to handle the tongue tracking issue by comparing the tongue Harries features with a biomechanical model derived from 3D deformation model to estimate the tongue shape during the speech. They claim good results but the model will undoubtedly be limited to specific cases and will not estimate the variation of different subjects or even different speech for the same person. Also, Harries feature detector efficiency is low in the noisy images.

Although the training-based algorithms are more robust than the non-training based algorithms due to the capability of the machine and deep learning algorithms in features extractions, they are highly dependant on the training dataset. The accuracy of tracking tongue contour from any data other than the data from the same recording session or the same dataset is not guaranteed due to the difference in the ultrasound image quality and human variation. The work done in [11] showed the most valuable training methodology by using the particle filter and claimed to have less sensitivity towards the training data. However, the particle filter like other techniques required adjusting different parameters to make it computationally expensive and not valid for the real-time processing. Table 3 shows a summary for each method in the literature review.

Table 3. Related works review revised from [11].

Method Name	Data Training	Parameters	User Input
Automatic extraction and tracking [44]	NA	▪ Snake energy weights, dynamic programming search region size	Few points on the initial frame
Edgetrak [13]	NA	▪ Snake energy weights. ▪ Band energy penalty dynamic. ▪ Programming search region size.	Few points on the initial frame
TongueTrack [14,15]	NA	▪ Curvature weight. ▪ Spatial smoothness weight. ▪ Length preservation weight. ▪ Temporal smoothness weight. ▪ Profile-based data energy weight. ▪ Width of Gaussian kernel. Gradient-based data energy weight.	3-5 points on initial frame
Biomechanical Model [20]	NA	▪ Minimal distance between features. ▪ Optical flow algorithm parameters. ▪ Minimum edge length. ▪ Maximum edge length. ▪ Time step. ▪ Material constant.	Full tongue on the initial frame (Tongue surface and inner tongue)
Autotrace [16]	Yes, deep learning	▪ Deep neural network and learning hyperparameters. ▪ Training data subset selection parameters.	NA
Deep Neural Network [19]	Yes, deep learning	▪ Deep neural network and learning hyperparameters.	NA
Active appearance model [47]	Principal component analysis (PCA)	▪ Number of texture principal components, number of shape principal components	Few points on the initial frame
Active shape model [48]	Principal component analysis (PCA)	▪ Edgetrak's parameters + number of shape principal components	Few points on the initial frame
Particle Filter [11]	Principal component analysis (PCA)+ multivariate Gaussian	▪ Edgetrak parameters. ▪ Number of shape principal components. ▪ Minimum number of particles. Maximum number of particles.	Few points on the initial frame

Chapter 4

Thesis Work

This chapter discusses in detail a novel approach developed in this thesis for tongue tracking and feature extraction from the lingual ultrasound by using computer vision techniques. The automatic tongue tracking was done in four main steps: denoising image, identifying desired region of interest (ROI), detecting tongue contour, and estimating missing data and image transformation from discrete and static 2D frames to one full concatenated signal that contain information about the full speech recording not just single frame. Unlike the conventional method which analyzes the speech by identifying the most significant static tongue signature from one video frame as a sound feature, the thesis approach implements a new technique that converts the static images frames from the whole video into one concatenated signal to represent the dynamic tongue movements in one signal. This allows linguistics researchers to track the tongue behaviour during the full speech. The static signature is the tongue contour shape in one video frame that has the highest tongue vertex among the rest of the video frames (see section 4.5 for more details). Figure 25 shows the flowchart of the proposed algorithm.

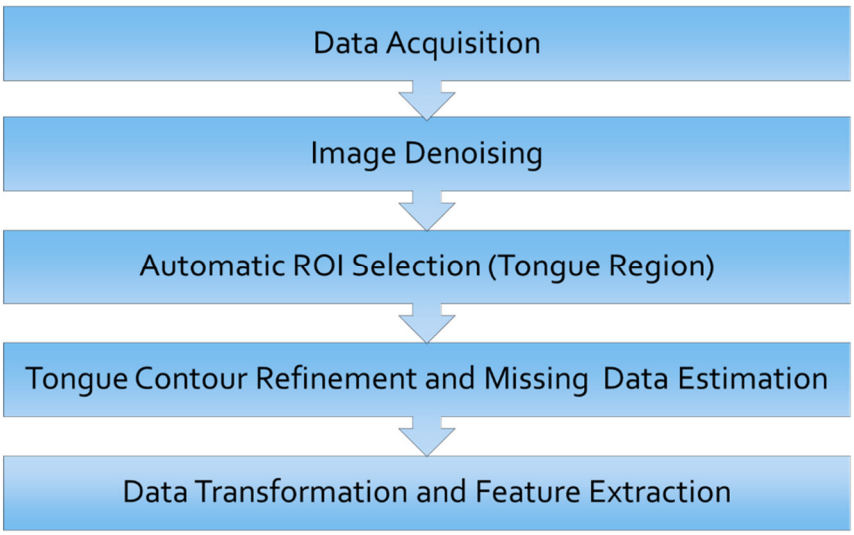


Figure 25. Thesis algorithm flowchart

4.1 Data Acquisition

The data were recorded with the collaboration of the University of Victoria speech research lab. There were five Arabic speakers consisting of three males and two females. The recorded datasets were collected from five subjects, each subject had to speak fourteen different Arabic vowel-consonants-vowel sequences (VCV sequences), where V was either /i a u/ and C was one of /k q t…./. Each VCV sequence was repeated three times to form 630 video sequences. The speaker should pause for about three seconds between each repetition; then, they should stop for a while before speaking the next VCV sequence to make sure that the tongue returned to the resting position. The ultrasound system used for the recording was the GE Logic e with the system frame rate at 60 Hz and the micro-convex ultrasound transducer was used with a bandwidth of (4-11) MHz. Table 4 lists the sets of the Arabic letters used during the lingual ultrasound recording sessions.

Table 4. The Arabic letters set used in the recording sessions (columns 3-4-5 use the international phonetic alphabet)

Letter set	Arabic letter	VCV_1	VCV_2	VCV_3	Repetition time
1	ك	aka _اكا	iki_اكي	uku_اوكو	3
2	ق	aqa_اقا	iqi_اقي	uqu_اوقو	3
3	س	asa_اسا	isi_اسي	usu_اوسو	3
4	ص	asˤa_اصا	isˤi_اصي	usˤu_اوصو	3
5	د	ada_ادا	idi_ادي	udu_اودو	3
6	ض	adˤa_اضا	idˤi_اضي	udˤu_اوضو	3
7	ت	ata_اتا	iti_اتي	utu_اوتو	3
8	ط	atˤa_اطا	itˤi_اطي	utˤu_اوطو	3
9	ث	aθa_اثا	iθi_اثي	uθu_اوثو	3
10	ل	ala_الا	ili_الي	ulu_اولو	3
11	لّ	alˤa_اللا	ilˤi_اللي	ulˤu_اوللو	3
12	هـ	aha_اها	ihi_اهي	uhu_اوهو	3
13	ح	aħa_احا	iħi_احي	uħu_اوحو	3
14	خ	aχa_اخا	iχi_اخي	uχu_اوخو	3

- ˤ for pharyngealization, i.e., tongue root retraction (into the pharynx) during the sound.
- ħ for the “hard h” sound, the pharyngeal fricatives
- χ for the uvular fricative, as in German “Bach”

The lingual ultrasound videos were recorded from Arabic speakers to automatically track the tongue movement during the speech of different Arabic sounds, which allowed for understanding the Arabic phonetics to improve a visual feedback tool for the Arabic

second-language learners. The algorithm was not limited to a certain language and it was designed to automatically track the tongue contour for any speaker and extract the dynamic tongue behaviour for each sound. The algorithm not restricted to any specific language as it analyses the ultrasound images by using a computer vision techniques without using any training data which make it not related to any specific language as detecting the tongue contour from ultrasound images was the same for all languages from computer vision perspective.

4.2 Image denoising and enhancement

In the lingual ultrasound applications, detection of the tongue contour was challenging with the existence of the noise. Unfortunately, ultrasound images are noisy by nature; high speckle noise and low signal to noise ratio degrade the quality of the image. To design an automated contour tracking algorithm, the noise should be suppressed as low as possible. To alleviate the noise problem, the thesis proposes an efficient methodology for noise filtering and image enhancement by using the combined curvelet transform [31-36] and shock filter [37-39]. Curvelet transform is a powerful tool for the image denoising which can produce a sharp image and reconstructs geometrical objects that have high directional structures or curvy edges (see section 4.2.1 for more details). The shock filter was used for the image segmentation and contrast enhancement. The combination of curvelet transform and shock filter were important in the lingual ultrasound tongue contour extraction as they combine the curvelet denoising capability with shock filter segmentation and contrast enhancement to extract the tongue contour. The curvelet transform can remove speckle noise as well as preserve ultrasound image textures and boundaries. Although the curvelet transforms succeeded in denoising the ultrasound images and preserving tongue contour edge details, the tongue contour was still connected with the surrounding objects or image background. To separate the tongue contour from the image background or any surrounding structures, the shock filter was used for tongue contour segmentation and enhancing the ultrasound image. The implementation of the shock filter was improved by using the Laplacian of Gaussian kernel [37-39] , rather than the Gaussian kernel as in the conventional method for further contrast enhancement to distinguish between the tongue

contour and the surrounding structures (see Figure 27 for curvelet and Shock filter denoising output).

4.2.1 Curvelet transform

The curvelet transform, is the most current version from the development of the wavelet [31] and ridgelet transforms [31]. To understand why using the curvelet transform is important in this research, the limitations of the wavelet and ridgelet transforms will be addressed as well as the advantages of the curvelet. Although the wavelet is used in many applications of the signal and image processing applications, it is unable to reconstruct or represent multidirectional structures; it is an isotropic function (isotropic, symmetrical object boundaries from all sides). In the image processing applications, many image structures are anisotropic (asymmetrical). Thus, the wavelet would fail to reconstruct the objects with a complex geometry and could not preserve all edge details. However, many complex wavelet algorithms were proposed to solve this problem, as in the multiresolution wavelet. However, they are not efficient for the real-time applications such as the lingual ultrasound because it is complicated and computationally expensive. Conversely, ridgelet transform was introduced as an anisotropic geometrical wavelet transform that can represent the multidirectional straight-line singularities (discontinuities). This feature means that the ridgelet can reconstruct geometries that have multidirectional straight lines, which is more efficient than the conventional wavelet. Unfortunately, the multi-directional straight-lines are rare in real-life applications because the geometries are more complex. After that, the same researchers improved the previous method by using the block-ridgelet, which divided the image into different partitions and then each partition was divided into different blocks. This method was the first generation from the curvelet transform and it was more efficient than the ridgelet transform. The applications of the curvelet transform first generation are limited because the geometrical representations of the ridgelet itself are not clear enough and the blocking effect becomes clearly visible on the image structure boundaries after the image reconstruction. The blocking effect is a common problem in image processing because it degrades the image quality at the blocks boundaries when it is divided into different blocks. The second generation of the curvelet transform was also proposed by the same authors, but at this time, the image reconstruction is based on the

frequency partitions. The curvelet transform is an efficient tool for reconstructing anisotropic geometries and can preserve curvy edges. It is widely used in different applications of image processing like denoising, segmentation and features extractions. The implementation of the curvelet transforms used in this project as a toolbox is called CurveLab [33]. Figure 26 depicts the flow graph of the discrete curvelet transform.

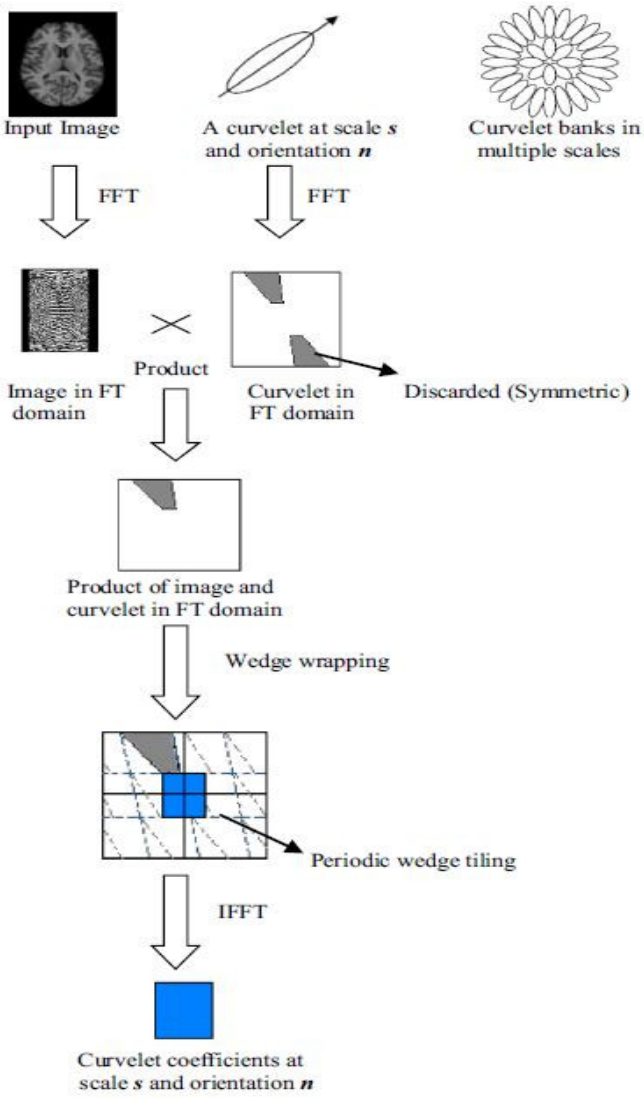


Figure 26. Flowgraph of discrete curvelet transform [50]

The CurveLab toolbox used the fast digital curvelet transform via the wrapping method [33,34]. The wrapping method assumes an input 2D image in a form of a Cartesian array.

$f[n1, n2]$, $0 \leq n1 < L_{1_{(l,j)}}, 0 \leq n2 < L_{2_{(l,j)}}$. $L_{1_{(l,j)}}, L_{2_{(l,j)}}$ are the length and width for the wrapping window respectively. The wrapping method generates a number of discreate curvelet coefficients which are indexed by orientation l and scale j. Equation (11) defines the formula for the discreate curvelet coefficients.

$$\boldsymbol{C^D(j,l,k) = \sum_{\substack{0 \leq n1 < L_{1_{(l,j)}} \\ 0 \leq n2 < L_{2_{(l,j)}}}} f[n1, n2]\ \varphi^D_{(j,l,k)}[n1, n2]} \qquad (11)$$

Where $\varphi^D_{(j,l,k)}[n1, n2]$ is the digital curvelet transform [34]. To capture the curved edges efficiently, the curvelet is implemented in the frequency domain and used pyramid structures with multi-scaling and different orientations at each scale. The curvelet waveform looks like a needle shape and it is very fine at the high scales [50].

The curvelet and the input image should be transformed into Fourier domain. After that, the product of image and curvelet was computed in the Fourier domain. Then the wedges were wrapped in a rectangular form because the frequency response of the curvelet is in a non-rectangular wedge. Finally, the inverse Fourier transform was computed to obtain the discrete curvelet coefficients.

The architecture of the fast digital curvelet transform via wrapping was arranged as follows [34]:

1. Apply the 2D FFT to get the Fourier samples $\hat{f}\ [n_1, n_2], -n/2 \leq n_1, n_2 < n/2.$
2. For each angle l and scale j , form the product $\widetilde{U}j_{,l}\ [n_1, n_2]\ \hat{f}[n_1, n_2]$.
3. Obtain the results from wrapping this product around the origin,

$\widehat{f}_{j,l}$, $[n_1, n_2]$ $W = \left(\left(\widetilde{U}_{j,l}\ \ \hat{f}\right)\right) \hat{f}[n_1, n_2]$. Where the range for n_1 and n_2 is now $0 \leq n_1 < L_{1,j}$ and $0 \leq n_2 < L_{2,j}$ (for θ in the range $(-\pi/4, \pi/4)$).

1. Apply the inverse 2D FFT to each $\widehat{f}_{j,l}$, to collect the discrete coefficients $C^D(j,l,k)$.

4.2.2 Shock filter

Although the curvelet denoising can optimally preserve image details, for further image enhancement and segmentation, the curvelet should be combined with the hybrid method [36]. This means that the curvelet should be used in a combination with another filter for better results. The advantage of combining two filters is a significant enhancement of the shock filter response when the image is blurred which can be achieved after the curvelet transform denoising filter. The curvelet blurring is helpful to reduce speckle noise and improve the shock filter response. The shock filter is a powerful tool for the image structure segmentation and contrast enhancements in order to segment and enhance the tongue contour from the surrounding objects.

The shock filter is a morphological process. It iteratively shocks image edges by dilation and erosion to create discontinuities or ruptures at the pixel's edges. The discontinuities are desired for the object segmentation and in the case of lingual ultrasound, the ruptures on the edges are useful to segment the tongue contour from the background or any connected edges that do not belong to the tongue, as they have weak connections with the tongue edges. The filter also improves the segmentation of multidirectional or flow-like patterns; it has been used for the applications of fingerprints and medical imaging [37]. The implementation of the shock filter can be explained by the Osher-Rudin equation [38,39].

$$u\_t = -sign(\Delta_u . \times | \nabla_u |) \qquad (12)$$

u is the image, Δ_u is the Laplacian of u, the Laplacian is the derivative function that is used to find areas with an abrupt change at edges. While $| \nabla_u |$ is the image gradient and can be calculated by computing the Euclidian norm $\sqrt{u_x^2 + u_y^2}$ of the gradient in x and y directions, the $sign$ function is the Laplacian operator [40]. The Laplacian operator ranks the pixel edges by assigning [-1,0,1] for each pixel to decide whether the pixel should be either dilated or eroded [39]. The Laplacian function is sensitive for the noise and the image should be smoothed before using the Laplacian derivative. The shock filter is enhanced by using the Laplacian of Gaussian kernel rather than the conventional Gaussian kernel to improve the contrast enhancement and to remove high-frequency noise components. The response showed an excellent enhancement for the image contrast; the intensity profile of

the bright regions was significantly different compared to the dark regions' intensity profile.

In the application of lingual ultrasound, the use of the curvelet-shock filter was efficient for the image denoising and segmentation to remove the speckle noise, enhance the image contrast and segment the tongue from the different objects in the mouth. As depicted in Figure 27, on the left is an example of the original tongue ultrasound image. In the middle is the speckle noise reduction after the curvelet transforms with preserving image details. On the right, the tongue contour is visible as a bright white arc after the shock filter segmentation.

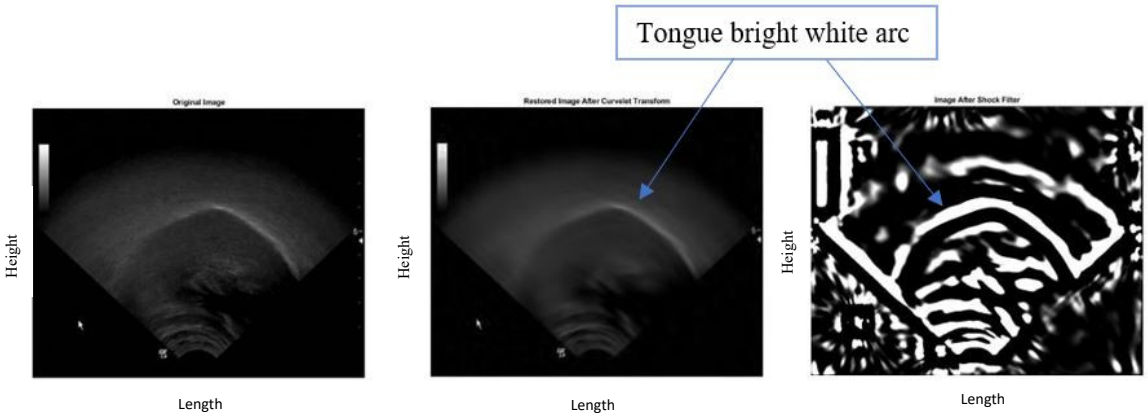


Figure 27. Left: Original image, middle: image after curvelet denoising, right: image after the shock filter.

4.3 Automatic selection of the region of interest

Within the ultrasound image, there are many structures in the oral cavity that can be visible in addition to the tongue contour. As mentioned in Chapter 1, the strong reflection from the palate, vessels, genioglossus muscle, geniohyoid, and mylohyoid muscles at the mouth floor are visible as bright structures, which is challenging for the automatic tongue tracking as the algorithm may be confused between the tongue and other structures. To overcome or alleviate the previously mentioned problem, any structure that interferes with tongue contour or has a strong intensity profile and is clearly visible on the ultrasound image,

should be discarded by selecting the desired tongue region. Figure 28 depicts the automatic region of the interest selection workflow.

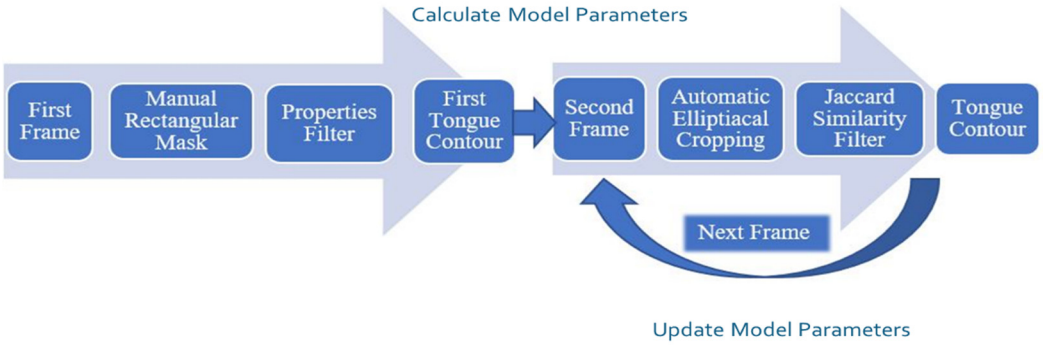


Figure 28. Automatic region of the interest selection workflow

To avoid any undesired structures, the proposed methodology was improved to handle this issue by designing a novel algorithm for the automatic region of the interest selection (tongue contour area). The algorithm had to be initialized only on the first frame by using a manual selection of a rectangular mask around the tongue contour to identify the tongue boundaries and to handle the variety of tongue shapes of different speakers. Due to the variety of human anatomy, it was difficult to create a universal mask for all subjects. In the first video frame, the tongue was in a relaxed position, the tongue contour was clearly visible and it was the largest segment in the image. It looks like an arc without any discontinuities except for some shadowing from the mandible and hyoid bones at the tongue tip and root. The main tongue body was evident as mentioned in Chapter 1. Besides, during the silence or tongue resting position, the undesired structures are not significantly visible and can be easily avoided. However, it is more difficult to distinguish some tongue segments from other structures during the speech, especially in the case of rapid tongue movements. After extracting the tongue contour on the first frame, the algorithm identified three extrema points (representing tongue tip, root and vertex) to create two elliptical masks for the consequent frames. The first mask was inside the tongue contour arc, starting from the tongue tip to end with the root below the tongue vertex. The second one was around the tongue contour. The elliptical masks were used as it is geometry that can resemble the natural concavity of the tongue contour shape. Also, many structures close to the tongue contour boundaries can be discarded easily by the elliptical masks with preserving tongue

contour textures. Figure 29 depicts selecting the square mask and identifying the three extrema points on the first contour and then estimating the dimensions of the two ellipses on the next frame.

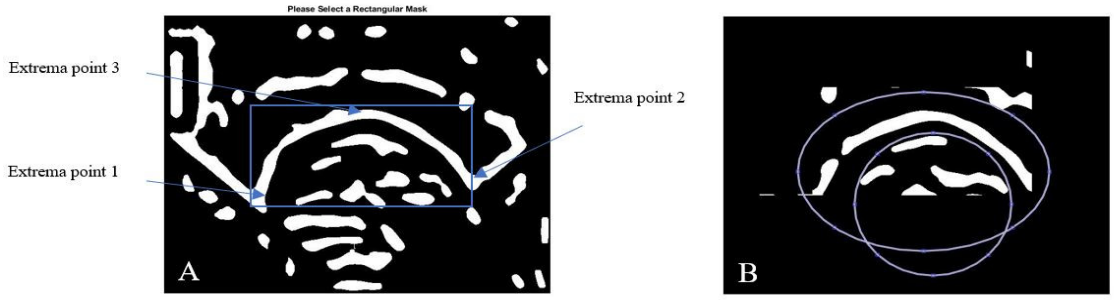


Figure 29. A: selecting a rectangle mask and identifying three extrema points on the first contour. B: estimating the position of two elliptical masks on the next frame.

The automatic region of interest selection used the three extrema points to estimate the dimensions of the two ellipses masks by calculating the four ellipse parameters denoted as xe, ye, d1 and d2, which are the x and y positions, transverse and vertical width respectively, as depicted in Figure 30.

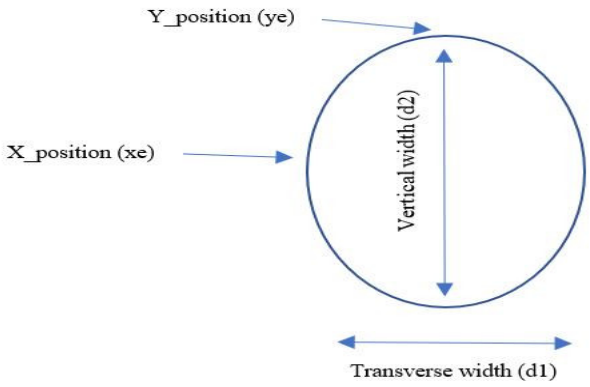


Figure 30. Elliptical cropping parameters

Equations (13-16) are used for calculating the first ellipse parameters d1, d2, xe and ye. The tongue root and tip positions were identified by x1 and y1, x2 and y2 which are the x and y positions for extrema 1 and extrema 2 points respectively. The points cx and cy are the X and Y coordinates for the tongue vertex or the highest point.

$$\mathbf{d1} = \frac{\mathbf{x2}}{\mathbf{x1}} - \left(\mathbf{0.45} * (\mathbf{x2} - \mathbf{x1})\right) \quad (13)$$

$$\mathbf{d2} = (\mathbf{y1} - \mathbf{cy}) + \mathbf{bias} \quad (14)$$

$$\mathbf{xe} = \mathbf{x1} + \frac{(|\mathbf{cx} - \mathbf{x1}|)}{\mathbf{2}} \quad (15)$$

$$\mathbf{ye} = \frac{\mathbf{cy} + \mathbf{y1}}{\mathbf{2}} - \frac{|\mathbf{y1} - \mathbf{cy}|}{\mathbf{5}} + \mathbf{bias} \quad (16)$$

In a parallel way, the second ellipse (identified around the tongue contour) parameters are measured by the equations (17-20):

$$\mathbf{d1} = \mathbf{x2} - \mathbf{x1} + \mathbf{bias} \quad (17)$$

$$\mathbf{d2} = (\mathbf{y1} - \mathbf{cy}) + \mathbf{bias} \quad (18)$$

$$\mathbf{xe} = \mathbf{x1} + \mathbf{bias} \quad (19)$$

$$\mathbf{ye} = \mathbf{cy} + \mathbf{bias} \quad (20)$$

The bias is a constant weight that is typically between 10 to 20 pixels. They were added to the equation to avoid the edges and to reduce the errors of the ellipse estimation model. Extrema points were updated sequentially at each frame, starting from the first frame as a reference point. Then the next frame extrema point was automatically estimated based on the previous one by using equations (13-20). The extrema points for the current frame were derived based on the previous frame because we had more clear information from the processed and segmented tongue contour of the previous one, rather than the non-processed tongue contour on the current frame. Many objects could be misclassified as a tongue segment if we had used the current and non-processed information to give inaccurate information about the tongue boundaries. This guarantee that the missing from the rapid tongue movement not affect the selection of the extrema points accuracy (rapid tongue movement may has a significant efficacy if it is faster than the ultrasound system acquisition frame rate).In addition, the tongue position displacement had not changed significantly between the two sequential frames which provided an excellent approximation for the current frame. To ensure that the estimation model was accurate, the algorithm used the first frame contour as a reference to compare it with the current tongue contour length in order to make proper adjustments of the elliptical parameters. The new estimation guaranteed that the next frame was not affected by the error or missing data from the current tongue contour. This means that if the contour length of the current frame

was too small compared to the first one, the next frame would not adjust its parameters based on the current frame and it would be adjusted based on the nearest good frame.

The automatic elliptical masks removed almost all undesired objects and retained the tongue contour. Although the automatic masks filtering could reject most of the unwanted objects, in some cases, objects that were close to the tongue contour were difficult and unsafe to be removed by the elliptical masks filtering. This was because if the ellipses masks expanded to remove the objects that close to the tongue boundary, the tongue may be affected by the filtering process and we may lose part of the tongue. Instead, the Jaccard similarity index was used to judge the segments that close to the tongue boundaries to see if they belonged to the tongue or not. The Jaccard similarity filtering was done by dilating the previous tongue contour and comparing it with the current segments. If the similarity index was above a certain threshold (the threshold was identified after studying different thresholds and selecting the best response based on the statistical data of different cases), the segment was retained as part of the tongue. However, if there were no similarities, the segment was rejected. The Jaccard similarity index was highly needed especially if we had missing parts from the tongue as in this case where the tongue was composed from different small segments. The Jaccard filter could retain the small and disconnected tongue parts efficiently, but this was difficult to achieve by using different filters like the properties filter. The Jaccard similarity index or coefficient is defined as the size of the overlapping area divided by the area of the union of the sample sets in order to measure the similarity between finite sample sets [41]. Figure 31 depicts the Jaccard similarity index principle.

Figure 31. Jaccard similarity index [41]

4.4 Tongue contour approximation and missing data estimation

4.4.1 Distance Transform tongue approximation

The strong reflectivity of the air above the tongue top surface reflected the ultrasonic waves and were visualized as a bright intensity profile on the 2D ultrasound image. This was the desired output for the tongue tracking. Unfortunately, the vessels, fat and tongue lower surface could also be visible within the tongue contour on the 2D image to create a thick contour. In some cases, the visualized tongue contour was larger than the exact tongue size due to some ultrasound artifacts as mentioned in Chapter 1, which gave wrong information about the contour boundaries. Because the intensity profile of the top surface and other structures were close to each other and the difference between their intensity levels were not easily distinguished either by the computer or the human eyes, the intensity profile should be transformed into a different level to approximate the correct tongue contour boundaries. Distance transform was used to transform the intensity profile and approximate the tongue top surface boundaries.

The distance transform ranked the intensity level for each pixel based on the distance from the nearest boundary. In other words, it calculated the distances between each pixel in the background and the nearest pixel in the image foreground on the binary image. The Euclidean norm was used to measure the distances, which is the straight-line distance between two points [42]. To improve the response of the distance transform, the transformation was squared to make the approximated contour thinner. Squaring is a helpful mathematical operation that can increase the value of the largest numbers and decreases the value of the smallest numbers (below 1) to discard pixels on the boundaries that have a weak profile and are probably not related to the actual tongue contour. Following to what was mentioned previously, the Canny edge detector is an image processing algorithm that can detect the object edges and suppress the noise used to extract the tongue contour edges.

Figure 32 depicts the distance transform of the segmented tongue contour on the binary image and shows the squaring effect on the transformation and then applying the Canny edges detection.

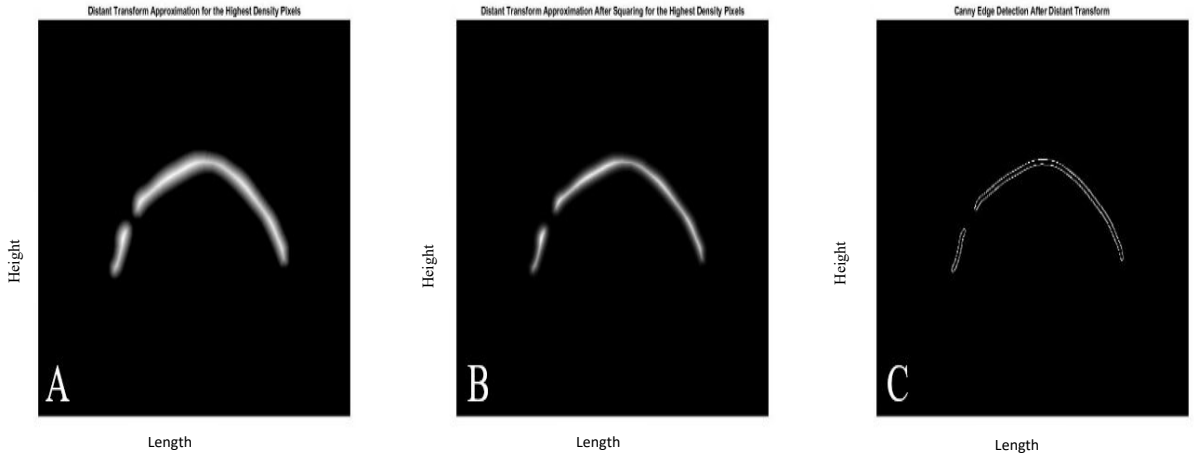


Figure 32. A. Distance transform image. B. Distance transform image after squaring. C. Canny edge detection for the image in B.

4.4.2 Missing tongue contour estimation and curve fitting

During the rapid tongue movement, the tongue contour may be detected as many small fragments due to the discontinuities which made it difficult to trace the contour. The main reasons for the missing tongue contour data were the angle of tongue relative to ultrasound transducer during the recording, unstable head position, ultrasound low frame rate, and shadowing caused by structures within the tongue, hyoid and mandible bones. To ensure high accuracy of the automatic tongue tracking, each contour should be as accurate as possible; the tracking is an adaptive process and depends on the correlation between sequential frames. Any missing frames can lead to wrong estimations of the next frame. Correcting the tongue shape required using an adaptive missing data estimation model to fill the gaps between the different disconnected tongue fragments and provide an accurate estimation of the tongue shape. The missing data adaptive estimation models divided into two cases is shown in Figure 33.

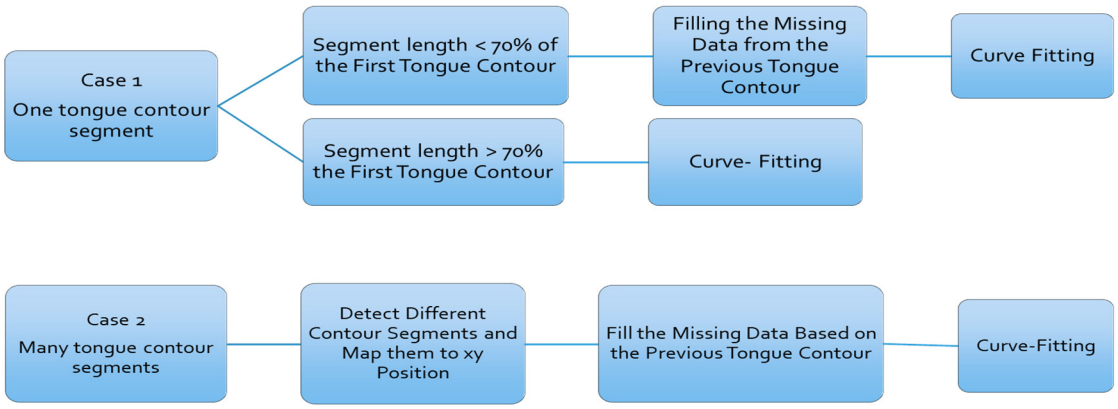


Figure 33. Missing contour data estimations and curve fitting workflow

Case-1: If the detected contour at the current frame was a single and short segment. The tongue contour segment length on the current frame was compared to the first tongue contour length which was used as a reference because it represented the tongue in the resting position. If the current contour length was less than 70% of the first one (the 70% length threshold chosen after comparing the length of the tongue contours that have missing data with the contours that don't have any missing to identify the shortage percentage), the missing data estimation algorithm estimated the missing tongue segments by mapping a new contour array for each tongue contour at each video frame based on the previous contour to fill the missing areas on the current frame from the previous contour. After filling the missing tongue contour parts, a polynomial curve-fitting with order-7 which is selected as it provides the most accurate fitting without missing any data or overfitting the curve (see Table 5 for polynomial curve fitting comparison). Polynomial curve-fitting was used for further smoothing and to guarantee that there were no missing parts if the information from the previous contour was not enough to create a perfect estimation for the tongue contour on the current frame.

Choosing the previous contour for the estimation was reliable because the tongue contour displacement within one frame lag was not significant compared to the previous frame position. The position for the two sequential frames were almost identical apart from

a small shift which was corrected by the curve fitting. Figure 34 depicts the Case 1 estimation result for a significant deficiency in tongue contour data and compares two curve fitting polynomial orders.

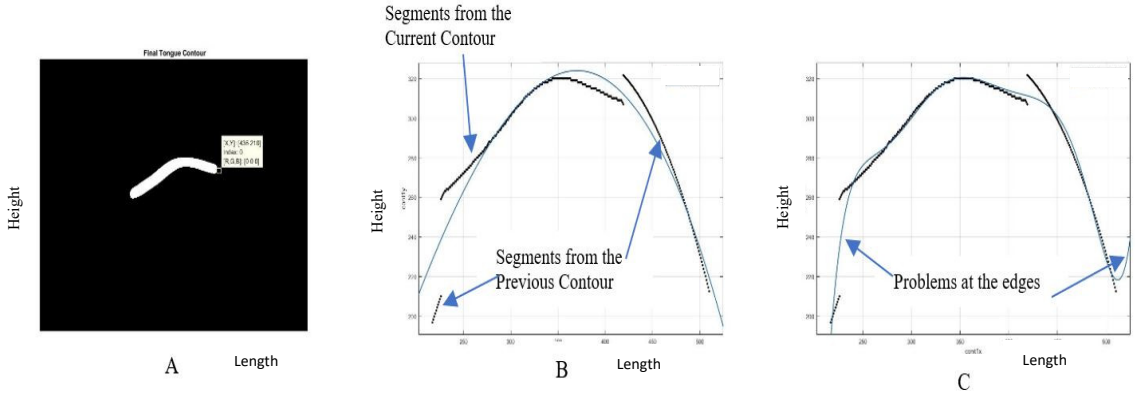


Figure 34. A) Case-1, Tongue contour on the current frame in white at the binary image. B) Case-1, segments from previous and current contours shown in solid black and curve-fitting poly-3 shown in solid blue line. C) Case-1, an example of using polynomial-9 curve fitting. Solid black and blue lines refer to tongue segments and curve fitting respectively. The problems at the edges in this poly-fitting order are handled by the proper selection of the rectangle mask on the first frame and using the poly fitting-order 7.

Case-2: If the current contour was more than one segment. In this case, the estimation model mapped the disconnected segments into a space of x-y position. Then the map of the current contour was compared to the previous contour map to fill any missing areas or gaps in between the current tongue contour segments from the previous tongue contour information. The curve-fitting algorithm with order-7 was also applied to this case for further smoothing. Figure 35 depicts the Case-2 estimation model with order-3 polynomial curve fitting. Then the estimation model was plotted over the original image.

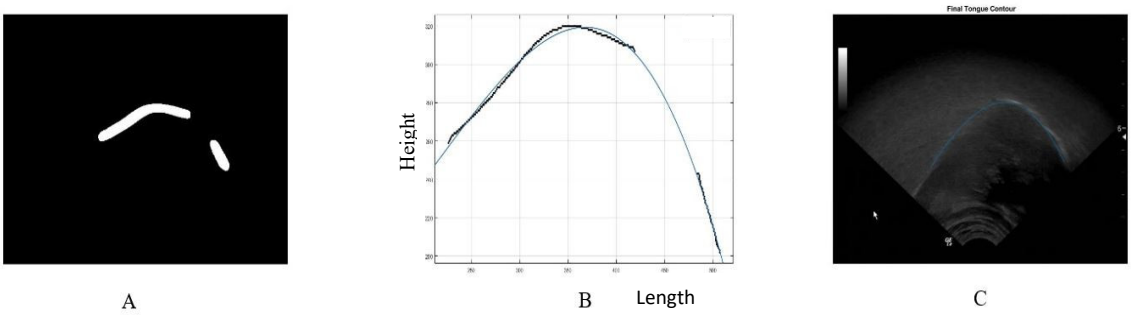


Figure 35. A) Example of Case-2 images, the binary image shows two contour segments in white. B) Order-3 polynomial curve fitting in solid blue line and the tongue segments are in solid black lines. C) In blue, fitted tongue contour overlay on the original image.

The order of the polynomial curve fitting (the number of the polynomial function coefficients) was very important to preserve the natural shape of the contour and to provide proper smoothing. The low rank of the poly-fitting function (ex. poly-fitting order-3) was good for fitting tongue contours that have a wide shape like the male subjects' tongue gestures and during the slow speed tongue movements. However, in the rapid tongue movements, the tongue curvature was concave and using low-order curve fitting would not represent the natural tongue shape, as the polynomial function was not flexible enough. With increasing the rank of the poly-fitting function, the natural tongue shape was closely represented by the poly-fitting rank increasing, so that even the complex tongue gestures could be fitted.

Table 5 shows the root mean square error for different poly-fitting orders. The error was significantly reduced after the order-7 to yield almost perfect matching to the tongue contour curvature. In this research project, the order-7 polynomial fitting was selected because it could provide a considerably low mean square error and smooth shape. Although, the order-9 provided the best root mean square error, the only drawback for this order was that there were some concavities on the tongue tip and root which could be reduced by the order-7. This problem may be alleviated by carefully selecting the rectangular mask around the tongue area on the first frame. The root mean square error (RMSE) for different poly-fitting coefficients was computed from a random 50 frame of five subjects. The RMSE was computed using a Matlab function cftool, which compared

the RMSE between the poly-fit function and the tongue contour segments before curve fitting.

Table 5. Root mean square error for different poly fitting coefficients computed from a random 50 frames of five subjects tongue contours.

Poly fitting coefficients	(RMSE) Root mean square error (mm)
2	4.5394
3	1.831
4	1.7072
5	1.1434
6	1.1458
7	0.5606
8	0.5422
9	0.4702

Figure 36 depicts the curve fitting of the tongue contour missing data estimation model. The polynomial-7 smoothing had the best result among the others. Figure 36 starts from the top left by polynomial-2 and sequentially ends with the polynomial-9 curve fitting at the bottom right.

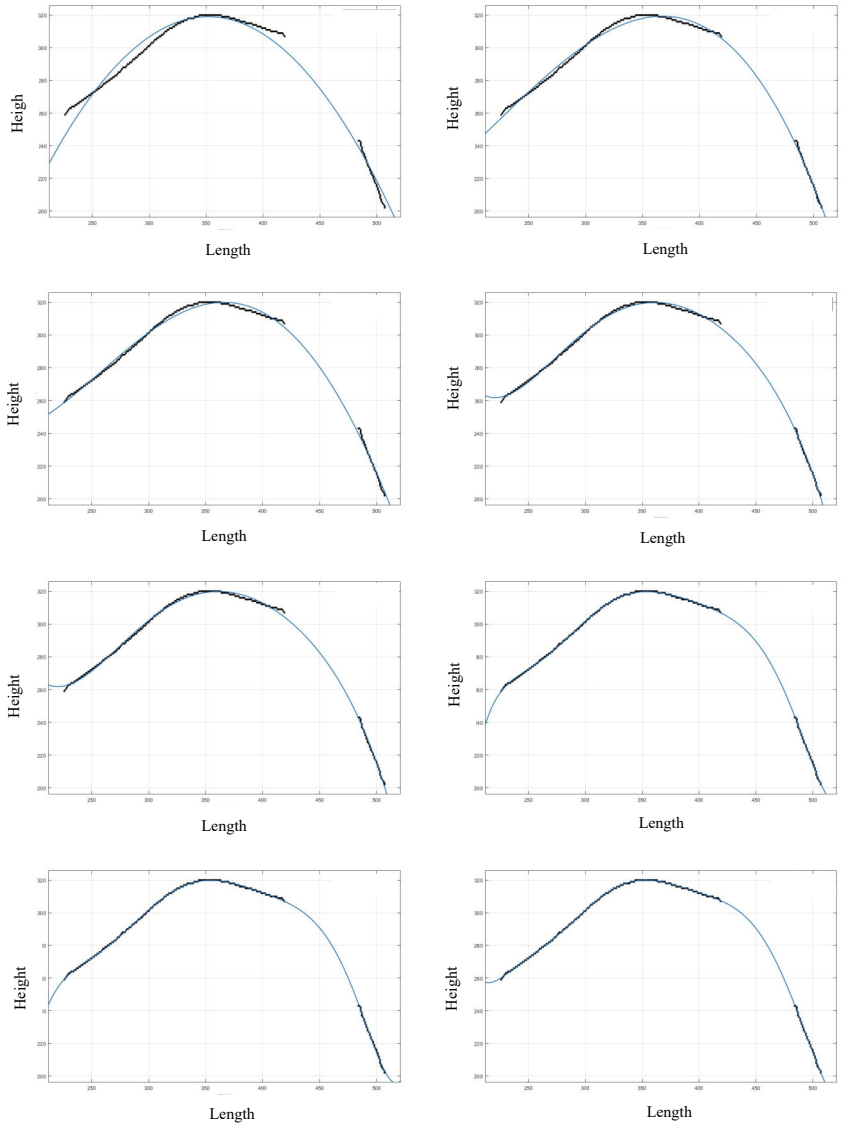


Figure 36. Polynomial curve fitting for the tongue contour missing data estimation models, starting from the top left by polynomial-2 and consequently ending with the polynomial-9 at the bottom right. In blue: curve-fitting. In black: tongue segment in current contour

4.5 Data transformation and feature extraction

Linguistic researchers are analyzing the tongue contour by identifying the most prominent tongue signature which provide them of information about the tongue contour deformation during the speech, or the tongue contour at one video frame that has the highest vertex (highest point) on the tongue body among the whole video frames for each sound [1]. Identifying the significant static gesture on one video frame was not an effective tool to provide useful information about the articulation speech sound and for this reason, many recent studies started using an augmented tool such as analyzing the audio or lip movements to give a better understanding of the speech behaviour.

The thesis proposes a dynamic and robust method to analyze the tongue movement for the whole video frames. Rather than using a single and static video frame for feature extraction, the segmented tongue contour on the 2D ultrasound images were transformed from the full video or frame by frame images to a dynamic 2D signal that represented the tongue shape and position in a sequential order of video frames (see Figure 38 for more details). To the best of my knowledge, there was no such proposed method that used the dynamic monitoring for the tongue contour tracing as visual feedback, except for some studies that used optical flow measurements to provide information about the tongue velocity and acceleration [55].

4.5.1 Data transformation

The segmented tongue contour from each frame was mapped to the X and Y positions and stored as a matrix, where X data represented the position of each point on the tongue contour and Y data represented the height of each point on the tongue contour. The width of the ultrasound image was 512 pixels and the width of the extracted contour was predicted to vary from frame to frame. It was typically less than 512 pixels as the areas from the left and right sides of the ultrasound image were black and the tongue contour data were concentrated around the centre of the image. The position information from the tongue contour points were arranged in a sequential index when they were saved on the matrix. The next frame data were arranged after the previous one until the end of the video sequence to form one line of tongue contour positions data for the full video. Because the

length of the tongue contour at different frames was not the same, the representation of the contour X positions data would not be accurate and the estimation of the actual position for each frame would not be achieved if the X positions data were stored in a sequential index. To handle the tongue contour positions issue, each extracted tongue contour measurement was padded with zeros (filling the empty spaces with zeros) from the left and right positions to preserve the positions information and maintain the same length for each contour. The padding with zeros preserved the accurate dimensions of the real tongue contours to represent the actual tongue behaviours. Figure 37 depicts the zero-padding process from the tongue contour on both sides.

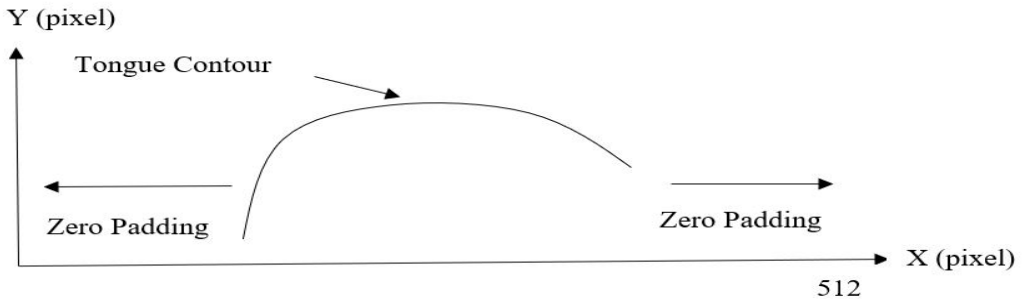


Figure 37. Zero padding illustration plot

Padding each tongue contour frame with zeros reshaped the mapping matrix to be formed from sequential tongue contours with the same length of each contour in one line to represent the full video sequences. Starting from the tongue contour on the first frame and connecting the end of the first contour tongue frame with the beginning of the tongue contour on the second frame and continue sequentially until the last video frame. After that, the tongue contour data was normalized to the unity to maintain the same scale for all tongue contour data and to make it more consistent for the comparison with different subjects. This was because each human subject had different oropharyngeal geometry. Even for the same sound, the scale of the segmented tongue contour could be different at different repetitions of the same sound because it is difficult to have exact sound loudness. The unity normalization rescaled the tongue contour to the reference point, which was the tongue root, to alleviate any variation of the transducer displacement or sound loudness. Figure 38 depicts an example of the dynamic full concatenated data of the tongue contour from all video frames.

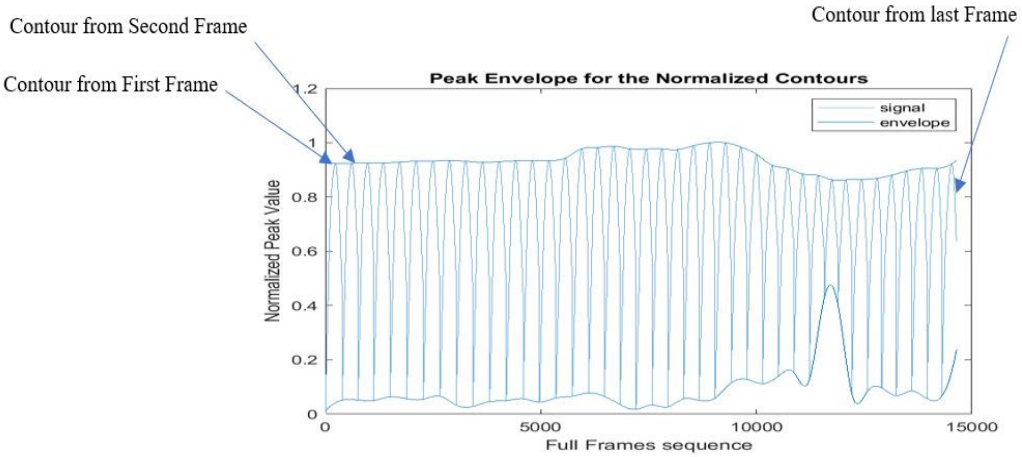


Figure 38. Example of the full concatenated signal of the tongue contour from all frames of sound "aka." The blue line at the top and bottom is the signal envelope.

4.5.2 Feature extraction

The thesis extracts a significant feature from a full ultrasound video concatenated signal to provide linguistic researchers with useful information about the dynamic tongue movement. This is in order to recognize the sound behaviour and provide the second language learners with a visual feedback system to assist the learning process, instead of just relying on the static tongue signature from the most essential frame (frame that has a significant vertex) as in the conventional method. The dynamic information allows the researchers to trace the static information for each frame sequentially in order to study and estimate the tongue behaviour for the full speech. The dynamic information is also useful to study the speech in a continuous and long sentence rather than just analyzing a single sound. Furthermore, many speech difficulties can be detected from identifying the inconsistency of the speech or the hesitation during the speech, which can be achieved by the dynamic feature extraction.

To analyze the continuous speech or a single sound, the thesis proposes a new technique for tongue feature extraction by using the signal peak envelope (signal peak tracing). The peak envelope reshaped the full concatenated signal to create a unique gesture for each sound and represented the full speech behaviour. As mentioned in Chapter 1, the tongue root and tip are not reliable for the dynamic tracing due to unreliability of the

apparent shape of the tongue root because of the transducer placement and the shadowing from the mandible and hyoid bones. The tongue body vertex information was less sensitive to the transducer placement and it was more reliable in the research compared to the information from the tongue tip and root as they are sensitive to the ultrasound transducer placement [1]. Tracing the signal peak envelope after padding each frame data with zeros was more efficient because it could preserve the peaks position for each frame while studying the vertex information with respect to the frame number. Not padding the signal with zeros provided inaccurate details about the vertex position. The signal envelopes information was normalized to the unity to make the data from different ultrasound recordings consistent with each other. Figure 39 depicts the envelope for two different sounds of the same speaker and shows the significant difference of the tongue behaviour for each case.

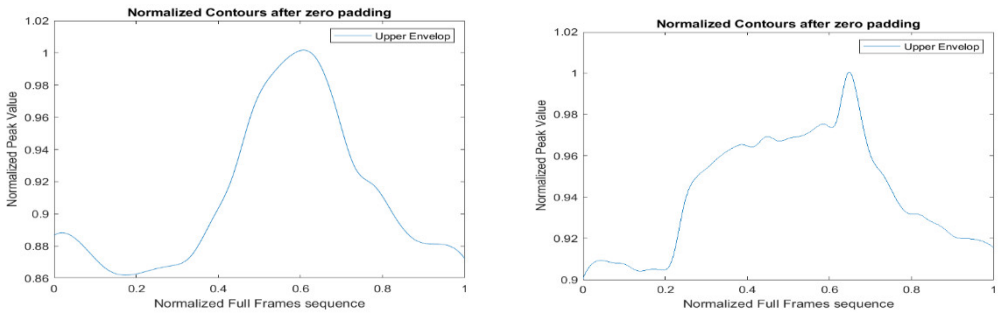


Figure 39. Envelope for the normalized full concatenated tongue contours data from all frames for two different sounds of the same speaker. The sound on the right is "uχu" The sound on the Left "uθu"

Feature extraction can alleviate any unreliability of the tongue root and tongue tip position information by providing extra smoothing for the whole tongue behaviour after considering the peak-envelope for the tongue body as a main feature. Furthermore, in the case of a significant missing of tongue contour, the tongue body could also be heavily affected to be almost unclear. The missing estimation is discussed in section 4.4.2. Missing tongue contour estimation and curve fitting may be not accurate enough. This made it difficult to build an accurate estimation model. In this case, the peak-envelope curve smoothing was useful to estimate an accurate and unique tongue signature.

Chapter 5

Evaluation and Results

5.1 Experiment Database

With reference to information mentioned in Chapter 4, the data were recorded from five Arabic speakers consisting of three males and two females. The recorded datasets were collected from five subjects, each subject had to speak fourteen different Arabic vowel-consonant-vowel sequences (VCV sequences), where V was either /i a u/ and C was one of /k q t…./. Each VCV sequence was repeated three times which yielded a total of 630 videos. Table 4 in Chapter 4 listed the different Arabic sounds that were used in this research.

5.2 Manual Tongue Contour Extraction

To ensure that the automatic proposed method is accurate, it should be compared to a ground truth reference that was extracted manually. Matlab software was used to implement the ground truth extraction contour tool to use it for the data validation. The manual tongue contour delineation was done by visually inspecting the actual tongue contour and then manually tracing the contour by identifying different points on the tongue surface. The tool then smoothed the manually identified points by using the polynomial-fitting function of order-7. Ground truth data were extracted from 10 randomly selected videos and about 13 frames from each one to form 130 ground truth data. Figure 40 depicts the process of manually selecting the ground truth data.

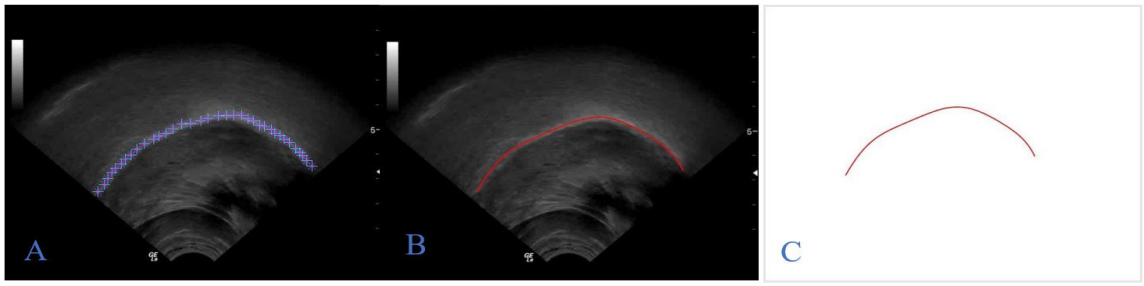


Figure 40. A) manual selection of tongue contour surface. B) contour smoothing. C) finally extracted contour.

5.3 Automatic Tongue Contour Extraction

The tongue contour was automatically extracted from the ultrasound images by using the proposed methodology in this thesis. Lingual ultrasound images were denoised and the tongue region of interest was selected. Then the missing data were estimated to represent the final tongue shape. The extracted contours were depicted over the original ultrasound image to show the results on different sequential frames. Even with cases that had a significant missing portion in the tongue contour data, which forms approximately 10% to 15% of the overall lingual ultrasound video frames for each recording session, the algorithm could detect the tongue, estimate the missing, and detect the tongue contour with a small error margin (0.955mm) compared to the manually extracted ground truth data. All undesired objects were removed to keep only the tongue segments. Then the tongue data were smoothed by using a polynomial curve fitting function to preserve the tongue's natural shape (The algorithm compared to ground truth data which is the accurate reference data that extracted manually by expert to compare the automatically extracted contours at each frame). Figure 41 depicts the ground truth and the thesis results are in blue and red, respectively at different frames. The extracted contour in red is almost identical to the ground truth contour in blue, which make it looks like the same contour.

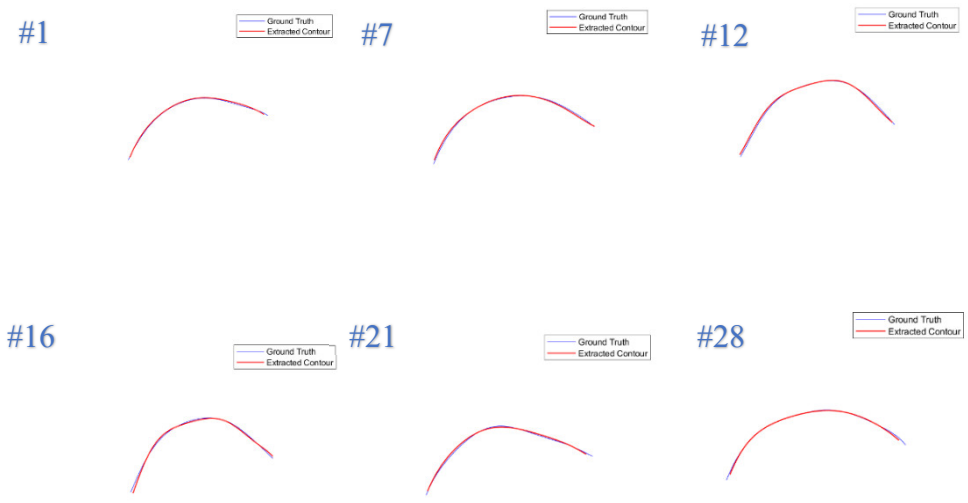


Figure 41. 29-year-old male subject automatic and manually extracted contours for different frames of the same recording session of sound "aka." Frame number is mentioned on the top left of each image in blue, the ground truth in solid blue, and the extracted contour in solid red.

Figure 42 depicts the ground truth and the thesis results are in blue and red, respectively on the original ultrasound image.

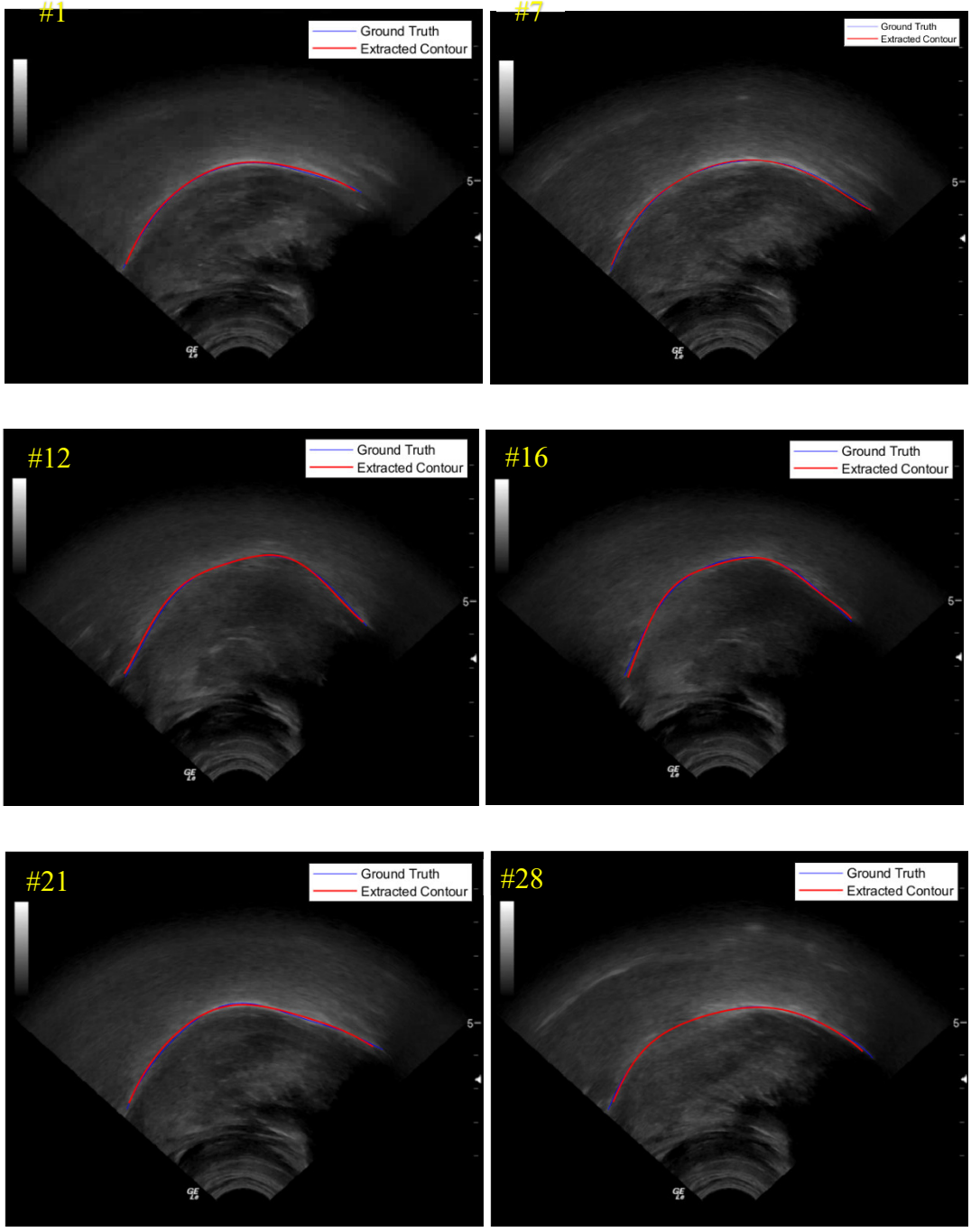


Figure 42. 29-year-old male subject automatic and manually extracted contours for different frames from the same sound "aka" depicted on the original ultrasound image. Frame numbers are mentioned on the top left of each image in yellow, the ground truth in solid blue, and the extracted contour in solid red.

5.3.1 Evaluation and Discussion for the Automatic Method

The tongue tracking results were evaluated by using the mean sum of distances (MSD) proposed by [13]. MSD was widely used in evaluating the accuracy of tongue tracking in different kinds of literature. The mean sum of distances compared automatically and manually extracted contours by measuring the distances in two steps. First, the distance was measured between each element at the automatically extracted contour with the closest element or point at the manually extracted contour. Second, the distance was measured between each point at the manually extracted contour and the closest point from the automatically extracted one. The summation of distances from these two steps was divided by the total number of elements of the manually and automatically extracted contours for normalization. Equation (21) shows the formula for the MSD.

$$\boldsymbol{MSD(U,V)} = \frac{\mathbf{1}}{\boldsymbol{m+n}}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{min_j}\ |\boldsymbol{v_j - u_i}| + \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{min_i}\ |\boldsymbol{u_i - v_j}|\right) \tag{21}$$

Where $\boldsymbol{n, m}$ is the contour length of the ground truth or manual and automatic extracted contours respectively, $\boldsymbol{v_j}$ is the manually extracted contour (ground truth), $\boldsymbol{u_i}$ is the automatically extracted contour, $\boldsymbol{min_i}$ and $\boldsymbol{min_j}$ denote the closest distances between each point on the contour and the nearest point on the other contour. MSD was measured in pixels and then converted to millimetres by assuming that each pixel is 0.295 mm [11]. There was an advantage of using MSD because the length of two contours was not identical and it was not efficient to use the conventional evaluation methods for the comparison like the mean sum of errors and the norm.

One hundred and thirty frames were randomly selected from the recorded dataset of different subjects to evaluate the proposed approach as listed in the Tables 6-10, see Appendix-A MSD Results. The average MSD for all testing samples, which resembled the mean error of the proposed algorithm, was 3.24 pixels and 0.955 millimetres. The error margin of the algorithm was proof that the proposed algorithms were efficient and the automatically extracted contours by the thesis algorithm were almost identical to the ground truth tongue contours for each one of them. The mean MSD for each case and the overall mean are discussed in Table 11, see Appendix-A MSD Results. However, the MSD

values were almost stable around the mean value for each case but at some frames, it was clear that the MSD increased significantly. This was because the accuracy decreased in the cases of the rapid tongue movement as the tongue contour was not clear and there was some missing. The worst reported result was MSD 6.18 and 1.82, in pixels and millimetres respectively. The best result was MSD 2.24 and 0.66, in pixels and millimetres respectively. The results showed that even in the cases that had significant discontinuities in the tongue contour, the algorithm was still able to estimate the missing and give results close to the actual contour with the least amount of error. Besides, the error margin of 1.82 mm was still acceptable (the acceptable range 2-3 mm as mentioned in the cited work [14]) and better than some reported results in different kinds of literature on the lingual ultrasound as will be discussed in the next section. Also, see Chapter 3 for more details about the accuracy of the related work methodologies.

5.3.2 Comparative analysis

This section will discuss and compare the MSD measurements of the proposed and related works. To be fair about their results, the proposed method will be compared with the mentioned results on their published works. Because each method used different recording datasets, which have different ultrasound video qualities and different video sizes, it was not efficient to rescale and resize the recorded videos in this thesis to be compatible with other methods, as they have a specific video size for the processing. Resizing the ultrasound videos not only degrades the quality of the ultrasound videos but also changes the actual tongue shape and position data to lead to a false analysis due to the unreliability of the tongue contour measurements. Furthermore, the methods that used the deep learning technique [11,16,19] required a huge dataset and this is outside the limit of the available data in this research.

The most recent work [11] used particle filter training for tongue tracking. The reported MSD was 1.69±1.1 mm. While the EdgeTrack [13] claimed some MSD results for different cases between 0.539mm-1.059mm, these results can be arguable as the EdgeTrack typically fails during the processing and it should be reinitialized manually. Also, there was

not enough information about the size of the validation data (The EdgeTrack wasn't validated by the thesis dataset as it is not compatible with recorded videos due to software limitations). Besides, in the case of a noisy ultrasound image, the EdgeTrack results were not accurate as it represented the tongue contour in a wrong position away from the actual tongue boundaries. Because of the previously mentioned argument, I included the results from the most recent work to review in this area [11]. It provided a good analysis of the many related works of tongue tracking and claimed that the MSD for the EdgeTrack was between 6.67±3.93mm. The TongueTrak [14] mentioned more fair results about their methodology and reported that the MSD was between 2mm -3mm. The review of [11] was close to what was mentioned in the original work. It measured the MSD for the TongueTrak as 3.48±1.5mm. In different literature, the MSD results claimed to be 0.75mm in the Autotrace [16]. This also can be arguable as the method was sensitive to the training data and any data from outside the training dataset or from different recording sessions. This gives different results. The review by [11] reported that the MSD for the Autotrace [16] was 2.61±1.22mm which could be more reasonable for this method. The biomechanical model approach [20] claimed that the mean distance varied from 0.62mm to 0.97mm with a median error of 0.7mm. The biomechanical model was derived from 700 X-ray images of the tongue to model the tongue behaviour, but there was a lot of unreliability in this method. The information about the validated dataset was not enough and the technique of tracking the tongue features by using Harris features was also not robust due to the high speckle noise and high signal to noise ratio in the ultrasound image. The cited work [19] also used the deep learning algorithm for tongue tracking. The validation resulted from 50 images that claimed to be about 1mm from the detected contour and the manual reference tongue contour. As mentioned before, any work related to deep learning is highly sensitive to the training dataset and the accuracy is not guaranteed for any other datasets. The criticism of the related work done based on the evaluation of their software's by using some of the online accessible and available dataset (were it applicable) and judging on their results based on their reported results and any reviews from other literatures by considering the research ethics for reporting an accurate result. To make the thesis algorithm applicable for future criticism the algorithm should be available and acceptable online.

The thesis method compared to the cited works still provided one of the best results with 0.955mm MSD. In a parallel way, the evaluation of each methodology was not limited to the MSD results that were claimed by each study There are many aspects to judge on each approach; Chapter 3 reviewed the capabilities and limitations for the related works in more detail and Chapter 4 explained the proposed work and demonstrated (see Appendix A for reported results) that the new approach handled the limitations for the other methodologies. The proposed approach was designed to denoise and automatically select the tongue area and even estimate missing data to make it a robust method for tracking the tongue contour on ultrasound videos. This was conducted regardless of the size or length of the videos and even with the existence of high speckle noise (the denoising efficiency evaluated by analysing how much accurate the automatic extraction will be after denoising the image , and by visually inspecting the images as it's impossible to have an accurate measure of the noise reduction because the amount of the noise in the ultrasound image can't be identified), the algorithm could derive good results without fail and without the need for manual reinitialization or training data. With reference to the information provided in this section, Table 13 in Appendix A shows the comparison results for the tongue detection accuracy between different kinds of literature and the thesis method.

5.4 Tongue Features Extraction Results

This section will present the results of extracting the unique feature from the dynamic tongue movement from the full video frames instead of just relying on the static feature from one frame as what is typically used in the lingual ultrasound research. The dynamic tongue movement unique feature was extracted after transforming the extracted contours from each frame to one concatenated signal. This preserved the tongue surface shape and the position for each contour element especially the tongue vertex, which is the highest point on the tongue body and the most important one. After transforming tongue contour data into one concatenated signal, the significant feature for each consonant (see Table 4 in Chapter 4) was extracted by tracing the peak-envelope of the signal to form a new signal that represented the unique tongue gesture for each certain sound. The advantage of this method was that it provided an analysis for the whole tongue movement during the speech

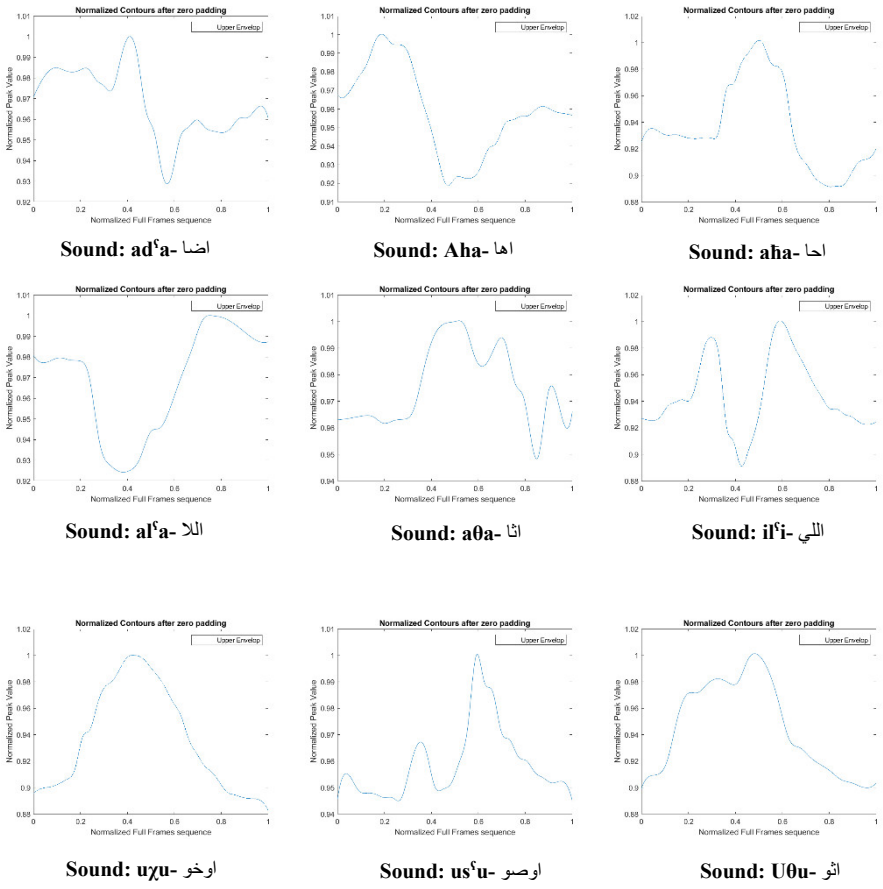


Figure 43. A sample of different sound signatures for the same subject. Below each signature is the sound in English and Arabic. The graphs show the normalized tongue contour. Y-axis is the normalized peak value. X-axis is the normalized full frame sequence.

and alleviated the problem of the detection uncertainty of the tongue tip and tongue root, which are unreliable due to many reasons as discussed in Chapter 1. A survey was done about the percentage of missing data and found that about 10-15% of the video frames contained missing data and sometimes the missing is huge enough to make the tongue almost unclear. Although the huge missing was estimated in the first line of estimation, which was mentioned previously in Chapter 4, there was a chance for some unreliability of the estimation, especially in the case of rapid tongue movement. The peak envelope feature provided full information of the tongue behaviour to make it less affected by any kind of bad frames or uncertainty of the tongue tip and root information. The results showed that the new approach could identify a unique signature for each subject at each certain sound with a high degree of similarity by evaluating different repetitions of the same sound

and compare the mean sum of errors between them. The dynamic feature extraction is a novel method that is proposed in this thesis and to the best of my knowledge, there is no method that has used the same technique. Figure 43 depicts different sound signatures for the same subject and shows that each sound has a unique signature that represents the tongue behaviour.

5.4.1 Evaluation and Discussion

To ensure the stability and reliability of the proposed methodology in extracting the unique tongue gesture, the sound was repeated three times to compare the sound repetitions with each other by measuring the mean squared errors (MSE). The mean squared error provided a quality estimation between the compared tongue gestures for different sound repetitions to see how similar they are to each other; the closer the MSE value was to zero, the better quality of the measure. Equation (22) shows the formula for calculating MSE.

$$\text{MSE}= \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i - Z_i})^{\mathbf{2}} \quad (22)$$

While n is the total number of tongue contour elements, $(\frac{1}{n}\Sigma)$ is the mean and $(Y_i\text{-}Z_i)^2$ is the squared difference between two tongue signatures. The comparison of the MSE results showed a high degree of similarity of the tongue signatures between different sound repetitions. As the results were close to zero, this means that the algorithm is valid and can detect the same information from each sound repetition without a significant amount of variations. The thesis methodology showed the robustness and reliability of identifying the dynamic sound signature and assigning it for each specific speech. Even for the same subject at a specific sound, the speech rate was varied. Due to the nature of human speech behaviour, there were some fluctuations on the tongue contour. This was caused by the speaker doing things such as pausing, increasing the voice loudness, or speaking in a rapid or slow speech rate. Additionally, discrepancies were caused by missing data and the ultrasound artifacts effect. However, the fluctuations were minor comparing to the main or significant feature which was still preserved and represented as a significant peak, representing the unique tongue gesture. The fluctuations were

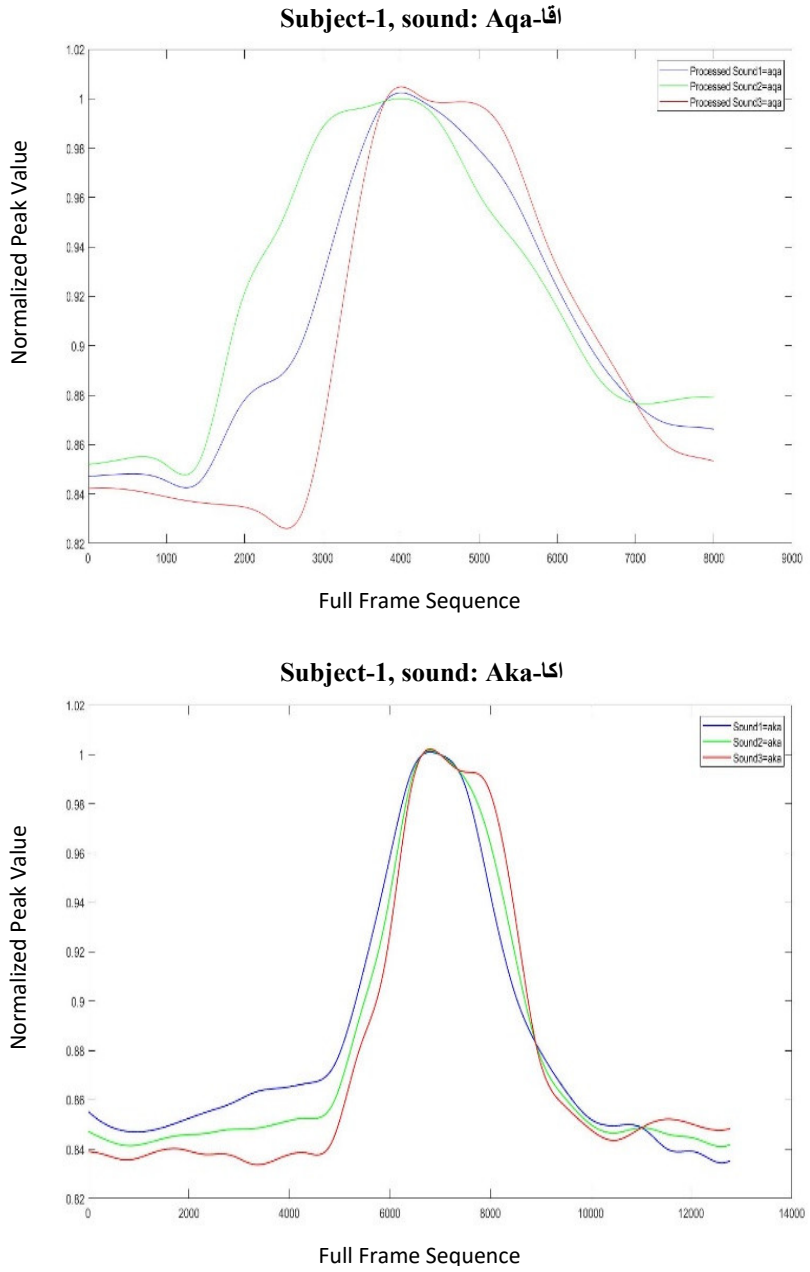


Figure 44. A sample of sound signatures for subject 1. Multiple repetitions of each sound are plotted on the same figure to see the similarity. Above each signature is the sound in English and Arabic

removed or alleviated by normalizing the unique signature and taking the average signature from many repetitions of the same sound. The average of the unique gesture was used to compare the different sounds with each other. Although the comparison showed a similarity between different subjects for the same sound on the main gesture shape, there was still a unique behaviour for each subject. The unique behaviour for each sound and

subject provided the proposed method with an ergonomic ability to identify the sound gesture for each subject by comparing the sound gesture with the average sound gesture database. This was also used to distinguish between different subjects based on their lingual behaviours from the whole database. This work needs further analysis and data collection to create a machine learning classifier. Table 12 shows the MSE results between each sound repetition for different subjects by selecting random cases from the datasets. The sound repetitions are depicted in Figures (44 –47) see Appendix B for Figures (45-47). The average gesture for each sound is depicted in Figures (48-52), see Appendix B for Figures (49-52)

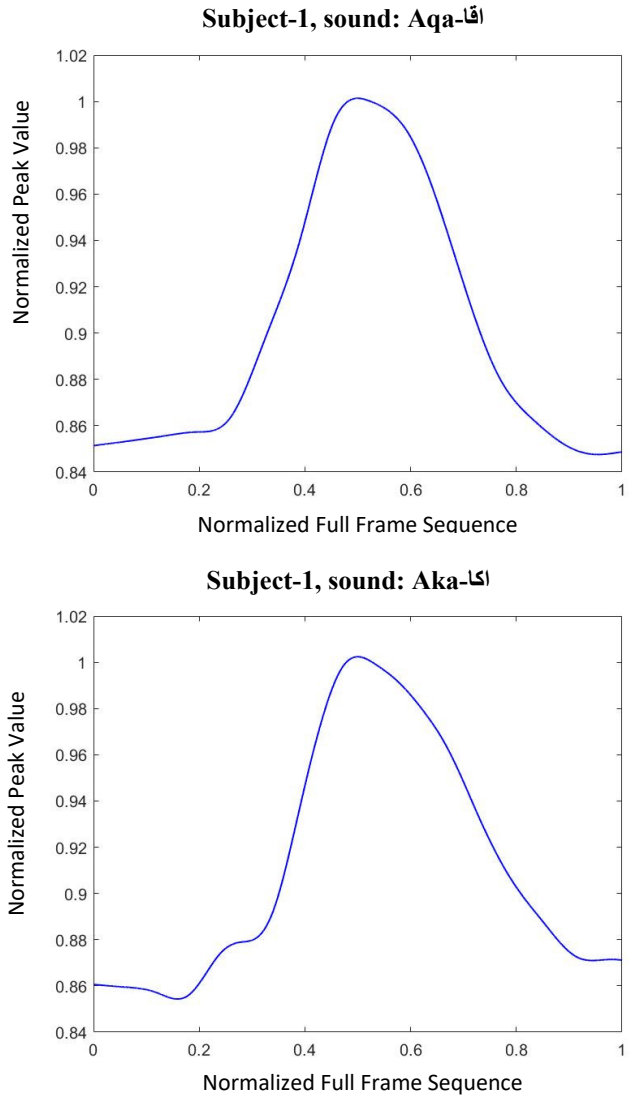


Figure 45. A sample of the average sound signature for subject 1. Above each signature is the subject and the sound in English and Arabic.

Table 6. Sample of MSE between each sound repetition for different subjects. The sound name mentioned with the repetition number for each subject.

Subject	Sound	MSE
Subject-1	aka2 .vs aka 3	4.6695e-04
Subject-1	Aqa1 .vs aqa 3	7.9922e-04
Subject-1	Aqa1 .vs aqa 2	0.0011
Subject-1	Aqa2 .vs aqa 3	0.0017
Subject-2	Aka1 .vs aka 3	5.1637e-04
Subject-2	Aka1 .vs aka 2	5.5823e-04
Subject-2	Aka2 .vs aka 3	2.5668e-06
Subject-2	Aqa1 .vs aqa 3	0.0010
Subject-2	Aqa1 .vs aqa 2	0.0013
Subject-2	Aqa2 .vs aqa 3	1.5195e-05
Subject-3	Aka1 .vs aka 3	2.7432e-04
Subject-3	Aka1 .vs aka 2	3.1158e-04
Subject-3	Aka2 .vs aka 3	8.5737e-05
Subject-3	Aqa1 .vs aqa 3	8.7904e-05
Subject-3	Aqa1 .vs aqa 2	1.0351e-04
Subject-3	Aqa2 .vs aqa 3	1.4993e-04
Subject-4	Aqa1 .vs aqa 3	0.0036
Subject-4	Aqa1 .vs aqa 2	0.0034
Subject-4	Aqa2 .vs aqa 3	3.3528e-04
Subject-4	Aka2 .vs aka 3	0.0012
Subject-5	Aqa1 .vs aqa 3	9.5798e-04
Subject-5	Aqa1 .vs aqa 2	0.0010
Subject-5	Aqa2 .vs aqa 3	3.2944e-04
Subject-5	Aka1 .vs aka 3	6.4298e-04
Subject-5	Aka1 .vs aka 2	7.7624e-04
Subject-5	Aka2 .vs aka 3	0.0016

Chapter 6

Conclusions

6.1 Summary

The thesis approach of using computer vision techniques for the lingual ultrasound tongue contour tracking and features extraction showed a robust, reliable and efficient design with a high degree of accuracy. The technique could process ultrasound videos without any limitations regarding the size or the length of the ultrasound video. The processing was done automatically for the lingual ultrasound video, there were no failures during the processing, and it did not need manual reinitialization during the processing. The thesis approach also did not need any training data to make it an efficient tool for the lingual ultrasound tongue tracking and feature extraction as it could process any ultrasound video regardless of the source of the recording. This was because it was not related to a certain dataset. The algorithm used an adaptive and regenerative model from the sequential frames of the ultrasound video to update the algorithm parameters for the tongue contour extraction and missing data expectation.

The main contributions of the thesis approach are structured as follows:

- Image denoising by using a combined curvelet-shock filter to remove high speckle noise and enhance the signal to noise ratio.
- Automatic selection of the tongue contour desired area by using the auto-regenerative model to exclude any undesired object around the tongue.
- Tongue contour approximation and missing data estimation by using adaptive model.
- Tongue contour data transformation from image space to full concatenated signal and feature extraction to identify a unique signature for each sound.

The results were validated by comparing the mean sum of distances (MSD) of the extracted contours from the thesis methodology with the ground truth data that were

delineated by hand. The average mean sum of distances was 0.955mm which means that the proposed automatic tongue contour extraction methodology was almost identical to the ground truth data. Furthermore, the results of the extracted sound gesture were validated by comparing the mean squared error (MSE) between different repetitions of the same sound. The average MSE showed a high level of similarity with 0.000858mm for all compared data.

6.2 Future Works

The ability of the thesis algorithm to automatically extract the tongue contour from lingual ultrasound videos with a high success rate and provide a unique signature for each sound after analyzing the dynamic tongue movement provides linguistic researchers with new and extraordinary information about the lingual ultrasound for further analysis. However, the algorithm is fast and didn't consume an extensive time during the processing, but the future work should focus on using GPU acceleration to make it more efficient in the real-time applications. As mentioned in Chapter 4 and Chapter 5, the average signature from each subject at specific speech provided unique information about the subject and the sound. The future works should focus on collecting an extensive database of various sounds from different speakers at different ages. This could then be used to build a machine learning classifier that is able to classify and analyze the dynamic tongue gesture and distinguish between different speakers and their sounds over long sound recordings. Furthermore, the extracted information should be used to provide a visual feedback to guide the second language learners during the speech learning process; a thermal image may be used. Besides, the future work should focus on improving the graphical interface of the algorithm to allow the linguistic researchers of using the algorithm functions and capabilities in an easy way.

Bibliography

[1]. Stone M. A guide to analysing tongue motion from ultrasound images. Clinical linguistics & phonetics. 2005 Jan 1;19(6-7):455-501.

[2]. Dawson KM, Tiede MK, Whalen DH. Methods for quantifying tongue shape and complexity using ultrasound imaging. Clinical linguistics & phonetics. 2016 May 3;30(3-5):328-44.

[3]. Zharkova N, Gibbon FE, Lee A. Using ultrasound tongue imaging to identify covert contrasts in children's speech. Clinical linguistics & phonetics. 2017 Jan 2;31(1):21-34.

[4]. Cleland J, Scobbie JM, Wrench AA. Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. Clinical linguistics & phonetics. 2015 Oct 3;29(8-10):575-97.

[5]. Tang L, Hamarneh G. Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on 2010 Jun 13 (pp. 154-161). IEEE.

[6]. Iskarous K. Detecting the edge of the tongue: A tutorial. Clinical linguistics & phonetics. 2005 Jan 1;19(6-7):555-65.

[7]. Hsiao MY, Wahyuni LK, Wang TG. Ultrasonography in assessing oropharyngeal dysphagia. Journal of Medical Ultrasound. 2013 Dec 1;21(4):181-8.

[8]. Pinton G, Aubry JF, Bossy E, Muller M, Pernot M, Tanter M. Attenuation, scattering, and absorption of ultrasound in the skull bone. Medical physics. 2012 Jan;39(1):299-307.

[9]. Csapó TG, Lulich SM. Error analysis of extracted tongue contours from 2D ultrasound images. InSixteenth Annual Conference of the International Speech Communication Association 2015.

[10]. Laporte C, Ménard L. Robust tongue tracking in ultrasound images: a multi-hypothesis approach. In Sixteenth Annual Conference of the International Speech Communication Association 2015.

[11]. Laporte C, Ménard L. Multi-hypothesis tracking of the tongue surface in ultrasound video recordings of normal and impaired speech. Medical image analysis. 2018 Feb 28; 44:98-114.

[12]. Orchard ME, Vachtsevanos GJ. A particle-filtering approach for on-line fault diagnosis and failure prognosis. Transactions of the Institute of Measurement and Control. 2009 Jun;31(3-4):221-46.

[13]. Li M, Kambhamettu C, Stone M. Automatic contour tracking in ultrasound images. Clinical linguistics & phonetics. 2005 Jan 1;19(6-7):545-54.

[14]. Tang L, Bressmann T, Hamarneh G. Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves. Medical image analysis. 2012 Dec 1;16(8):1503-20.

[15]. Tang, L., 2012. User guide for TongueTrack v1.2. http://tonguetrack.cs.sfu.ca/ TongueTrackUserGuide.pdf. Accessed: 2016-11-16.

[16]. Fasel I, Berry J. Deep belief networks for real-time extraction of tongue contours from ultrasound during speech. In2010 International Conference on Pattern Recognition 2010 Aug 23 (pp. 1493-1496). IEEE.

[17]. Lempitsky, V., Rother, C., Roth, S., Blake, A., 2010. Fusion moves for Markov random field optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1392–1405.

[18]. G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science,313(5786):504–507, 2006.

[19]. Jaumard-Hakoun, A., Xu, K., Roussel-Ragot, P., Dreyfus, G., Stone, M., Denby, B., 2015. Tongue contour extraction from ultrasound images based on deep neural net- work. International Congress of Phonetic Sciences.

[20]. Loosvelt, M., Villard, P.-F., Berger, M.-O., 2014. Using a biomechanical model for tongue tracking in ultrasound images. In: Proceedings of the International Sym- posium on Biomedical Simulation, pp. 67–75.

[21]. Gick B, Bernhardt B, Bacsfalvi P, Wilson I. Ultrasound imaging applications in second language acquisition. Phonology and second language acquisition. 2008 Mar 5; 36:315-28.

[22]. Bliss H, Abel J, Gick B. Computer-assisted visual articulation feedback in L2 pronunciation instruction. Journal of Second Language Pronunciation. 2018 May 31;4(1):129-53.

[23]. Dowd, A., Smith, J., & Wolfe, J. (1997). Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time. *Language and Speech*, 41(1), 1−20.

[24]. Pharynx and Larynx Anatomy Pleasant (throat Anatomy) Understanding the Basics of It with Diagrams (Dec,2018) [online]. Available: http://getreadyrossvalley.org/pharynx-and-larynx-anatomy/pharynx-and-larynx-anatomy-pleasant-%E3%80%90throat-anatomy%E3%80%91-understanding-the-basics-of-it-with-diagrams/.

[25]. Miller, A. and Finch, K. (2010). Corrected High frame rate Anchored Ultrasound with Software Alignment. Journal of Speech, Language and Hearing Research.

[26]. Basic principle of ultrasonic probes (Jan,2019) [online]. Available: http://www.ndk.com/en/sensor/ultrasonic/basic02.html

[27]. P. Suetens, Fundamentals of Medical Imaging, 2nd ed., Cambridge: Cambridge University Press, 2009.

[28]. Klaus D. Toennies. Advances in Computer Vision and Pattern Recognition, Springer-Verlag London Limited 2012.

[29]. Bushberg J, Seibert J, Leidholdt Jr E, Boone J. The essential physics of medical imaging. 2002. Eur J Nucl Med Mol Imaging. 2003;30:1713.

[30]. University of Manitoba , ultrasound imaging overview (Jan,2018) [online]. Available:http://umanitoba.ca/faculties/health_sciences/medicine/units/cacs/sam/8478.html

[31]. Ma J, Plonka G. The curvelet transform. IEEE signal processing magazine. 2010 Mar;27(2):118-33.

[32]. Starck JL, Candès EJ, Donoho DL. The curvelet transform for image denoising. IEEE Transactions on image processing. 2002 Jun;11(6):670-84.

[33]. Candes EJ, Demanet L, Donoho DL, Ying L. Curvelab toolbox, version 2.0. CIT. 2005.

[34]. Candes E, Demanet L, Donoho D, Ying L. Fast discrete curvelet transforms. Multiscale Modeling & Simulation. 2006 Sep 26;5(3):861-99.

[35]. Candès EJ, Donoho DL. New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences. 2004 Feb;57(2):219-66.

[36]. Devarapu KV, Murala S, Kumar V. Denoising of ultrasound images using curvelet transform. InComputer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on 2010 Feb 26 (Vol. 3, pp. 447-451). IEEE.

[37]. Vacavant, A. (2012) Fast Smoothed Shock Filtering. IEEE International Conference on Pattern Recognition (ICPR).

[38]. Osher, S.J. and Rudin, L.I. (1990) Feature-Oriented Image Enhancement Using Shock Filters. SIAM Journal on Numerical Analysis, 27, 919-940. http://dx.doi.org/10.1137/0727053.

[39] F. Guichard, J. Morel; "A Note on Two Classical Shock Filters and Their Asymptotics"; Michael Kerckhove (Ed.): Scale-Space and Morphology in Computer Vision, LNCS 2106, pp. 75-84; Springer, New York; 2001.

[40]. Kramer HP, Bruckner JB. Iterations of a non-linear transformation for enhancement of digital images. Pattern recognition. 1975 Jun 1;7(1-2):53-8.

[41]. Jaccard Index (Jan,2019) [Online]. Available: [https://en.wikipedia.org/wiki/Jaccard_index].

[42]. Solomon C, Breckon T. Fundamentals of Digital Image Processing: A practical approach with examples in Matlab. John Wiley & Sons; 2011 Jul 5.

[43]. Kass, M., Witkin, A. , Terzopoulos, D. , 1988. Snakes: active contour models. Int. J. Comput. Vis. 1 (4), 321–331.

[44]. Akgul, Y.S., Kambhamettu, C., Stone, M., 1999. Automatic extraction and tracking of the tongue contours. IEEE Trans. Med. Imaging 18 (10), 1035–1045.

[45]. Amini, A. A., Weymouth, T.E., Jain, R.C., 1990. Using dynamic programming for solving variational problems in vision. IEEE Trans. Pattern Anal. Mach. Intell. 12 (9), 855–867.

[46].Chalana V, Kim Y. A methodology for evaluation of boundary detection algorithms on medical images. IEEE Transactions on medical imaging. 1997 Oct;16(5):642-52.

[47]. Roussos, A. , Katsamanis, A. , Maragos, P. , 2009. Tongue tracking in ultrasound im- ages with active appearance models. In: Proceedings of the IEEE International Conference on Image Processing, pp. 1733–1736 .

[48]. Ghrenassia, S. , Laporte, C. , Ménard, L. , 2013. Statistical analysis of tongue shape in ultrasound video sequences: tongue tracking and population analysis. In: Ultra- fest VI, pp. 53–55 .

[49]. Mercado KP. Developing high-frequency quantitative ultrasound techniques to characterize three-dimensional engineered tissues. University of Rochester; 2015.

[50]. Shinde AA, Rahulkar AD, Patil CY. Fast discrete curvelet transform-based anisotropic feature extraction for biomedical image indexing and retrieval. International Journal of Multimedia Information Retrieval. 2017 Dec 1;6(4):281-8.

[51]. Pfeiffer T. Using virtual reality technology in linguistic research. In 2012 IEEE Virtual Reality Workshops (VRW) 2012 Mar 4 (pp. 83-84). IEEE.

[52]. Kedrova G, Anisimov N. MRI in linguistics and its applications: interdisciplinary approach. Stephanos. 2013(1):36-50.

[53]. Lai CS, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. 2001 Oct;413(6855):519.

[54]. Bowers JM, Perez-Pouchoulen M, Edwards NS, McCarthy MM. Foxp2 mediates sex differences in ultrasonic vocalization by rat pups and directs order of maternal retrieval. Journal of Neuroscience. 2013 Feb 20;33(8):3276-83.

[55]. Barbosa AV, Vatikiotis-Bateson E. Optical flow analysis for measuring tongue-motion. The Journal of the Acoustical Society of America. 2014 Oct;136(4):2105-.

Appendix A

MSD Results

Table 7 . Subject-1, MSD between the proposed method and the ground truth.

Subject	Video Sound	Frame #	MSD in pixels	MSD in mm
Subject 1	Aka	1	3.3652	0.992734
Subject 1	Aka	2	2.9091	0.858185
Subject 1	Aka	3	3.1875	0.940313
Subject 1	Aka	4	3.142	0.92689
Subject 1	Aka	5	3.3221	0.98002
Subject 1	Aka	7	3.6848	1.087016
Subject 1	Aka	10	3.5614	1.050613
Subject 1	Aka	12	3.3602	0.991259
Subject 1	Aka	15	3.6988	1.091146
Subject 1	Aka	16	3.5413	1.044684
Subject 1	Aka	20	3.4181	1.00834
Subject 1	Aka	25	3.1866	0.940047
Subject 1	Aka	27	3.1943	0.942319
Subject 1	Aka	29	3.6971	1.090645
Subject 1	Aqa	1	2.9293	0.864144
Subject 1	Aqa	2	2.7786	0.819687
Subject 1	Aqa	3	2.7173	0.801604
Subject 1	Aqa	7	2.7911	0.823375
Subject 1	Aqa	10	2.6066	0.768947
Subject 1	Aqa	12	3.2935	0.971583
Subject 1	Aqa	15	3.3932	1.000994
Subject 1	Aqa	18	5.215	1.538425
Subject 1	Aqa	20	3.062	0.90329
Subject 1	Aqa	22	3.2209	0.950166
Subject 1	Aqa	25	3.1169	0.919486
Subject 1	Aqa	28	2.9491	0.869985

Table 8. Subject-2, MSD between the proposed method and the ground truth.

Subject	Video Sound	Frame #	MSD in pixels	MSD in mm
Subject 2	Aka	1	2.9823	0.879779
Subject 2	Aka	2	3.7521	1.10687
Subject 2	Aka	5	2.8613	0.844084
Subject 2	Aka	7	2.6727	0.788447
Subject 2	Aka	10	3.3712	0.994504
Subject 2	Aka	12	3.7268	1.099406
Subject 2	Aka	15	3.7706	1.112327
Subject 2	Aka	18	6.1864	1.824988
Subject 2	Aka	22	2.7	0.7965
Subject 2	Aka	26	3.2605	0.961848
Subject 2	Aka	30	3.5635	1.051233
Subject 2	Aka	33	3.0491	0.899485
Subject 2	Aka	35	2.9528	0.871076
Subject 2	Aka	37	3.4982	1.031969
Subject 2	Aqa	1	3.1882	0.940519
Subject 2	Aqa	2	3.2593	0.961494
Subject 2	Aqa	5	2.853	0.841635
Subject 2	Aqa	8	2.8234	0.832903
Subject 2	Aqa	12	2.8855	0.851223
Subject 2	Aqa	15	3.2157	0.948632
Subject 2	Aqa	20	3.0155	0.889573
Subject 2	Aqa	30	3.6018	1.062531
Subject 2	Aqa	40	3.3460	0.98707
Subject 2	Aqa	50	3.3066	0.975447
Subject 2	Aqa	60	4.0165	1.184868

Table 9. Subject-3, MSD between the proposed method and the ground truth.

Subject	Video Sound	Frame #	MSD in pixels	MSD in mm
Subject 3	Ata	1	3.0466	0.898747
Subject 3	Ata	2	2.7115	0.799893
Subject 3	Ata	5	3.6123	1.065629
Subject 3	Ata	8	3.3208	0.979636
Subject 3	Ata	20	2.9399	0.867271
Subject 3	Ata	25	2.7092	0.799214
Subject 3	Ata	30	3.1107	0.917657
Subject 3	Ata	31	3.4805	1.026748
Subject 3	Ata	33	3.429	1.011555
Subject 3	Ata	37	3.0153	0.889514
Subject 3	Ata	40	4.4084	1.300478
Subject 3	Ata	45	4.1035	1.210533
Subject 3	Ata	49	4.3225	1.275138

Subject 3	Ara	1	2.9448	0.868716
Subject 3	Ara	2	2.6641	0.78591
Subject 3	Ara	4	3.8376	1.132092
Subject 3	Ara	6	3.5774	1.055333
Subject 3	Ara	8	3.6389	1.073476
Subject 3	Ara	10	3.9929	1.177906
Subject 3	Ara	12	3.7052	1.093034
Subject 3	Ara	18	3.8564	1.137638
Subject 3	Ara	22	3.2181	0.94934
Subject 3	Ara	30	3.1807	0.938307
Subject 3	Ara	40	2.9086	0.858037
Subject 3	Ara	46	3.2378	0.955151

Table 10. Subject-4, MSD between the proposed method and the ground truth.

Subject	Video Sound	Frame #	MSD in pixels	MSD in mm
Subject 4	Ulu	1	3.7702	1.112209
Subject 4	Ulu	2	3.9692	1.170914
Subject 4	Ulu	3	3.0605	0.902848
Subject 4	Ulu	5	3.2757	0.966332
Subject 4	Ulu	7	3.7031	1.092415
Subject 4	Ulu	10	2.6286	0.775437
Subject 4	Ulu	12	2.7982	0.825469
Subject 4	Ulu	15	2.4407	0.720007
Subject 4	Ulu	20	2.5518	0.752781
Subject 4	Ulu	22	3.3268	0.981406
Subject 4	Ulu	30	3.6196	1.067782
Subject 4	Ulu	35	2.7729	0.818006
Subject 4	Ulu	40	3.0384	0.896328
Subject 4	Ulu	41	2.7971	0.825145
Subject 4	Akha	2	2.7314	0.805763
Subject 4	Akha	3	2.5143	0.741719
Subject 4	Akha	4	2.5067	0.739477
Subject 4	Akha	5	2.9236	0.862462
Subject 4	Akha	6	2.7248	0.803816
Subject 4	Akha	7	3.1982	0.943469
Subject 4	Akha	8	3.209	0.946655
Subject 4	Akha	9	3.6697	1.082562
Subject 4	Akha	10	4.1183	1.214899
Subject 4	Akha	25	2.5851	0.762605
Subject 4	Akha	30	5.1662	1.524029
Subject 4	Akha	41	3.9904	1.177168

Table 11. Subject-5, MSD between the proposed method and the ground truth.

Subject	Video Sound	Frame #	MSD in pixels	MSD in mm
Subject 5	Ada	1	2.3883	0.704549
Subject 5	Ada	2	2.6432	0.779744
Subject 5	Ada	5	2.8965	0.854468
Subject 5	Ada	7	2.487	0.733665
Subject 5	Ada	10	2.9922	0.882699
Subject 5	Ada	12	2.6861	0.7924
Subject 5	Ada	15	3.0192	0.890664
Subject 5	Ada	18	3.8869	1.146636
Subject 5	Ada	23	2.9587	0.872817
Subject 5	Ada	26	3.6604	1.079818
Subject 5	Ada	33	3.5686	1.052737
Subject 5	Ada	39	2.6182	0.772369
Subject 5	Ada	42	2.2479	0.663131
Subject 5	Ada	43	2.3084	0.680978
Subject 5	Ata	1	2.2946	0.676907
Subject 5	Ata	2	2.9258	0.863111
Subject 5	Ata	3	3.3207	0.979607
Subject 5	Ata	5	3.3271	0.981495
Subject 5	Ata	8	2.8725	0.847388
Subject 5	Ata	10	4.0479	1.194131
Subject 5	Ata	13	2.8616	0.844172
Subject 5	Ata	15	3.2207	0.950107
Subject 5	Ata	18	2.9477	0.869572
Subject 5	Ata	20	2.5443	0.750569
Subject 5	Ata	24	3.5547	1.048637
Subject 5	Ata	29	3.0402	0.896859
Subject 5	Ata	33	3.058	0.90211
Subject 5	Ata	39	3.5129	1.036306

Table 12. The mean MSD of each case and the overall mean in pixels and millimetres.

Subject	Video Sound	Mean MSD in pixels	Mean MSD in mm
Subject-1	Aka	3.376	0.995
Subject-1	Aqa	3.172	0.935
Subject-2	Aka	3.453	1.018
Subject-2	Aqa	3.228	0.952
Subject-3	Ata	3.4	1.003
Subject-3	Ara	3.39	1.00
Subject-4	Ulu	3.125	0.921
Subject-4	Akha	3.278	0.967
Subject-5	Ada	2.882	0.85
Subject-5	Ata	3.1	0.914
	Overall Mean	**3.24**	**0.955**

Table 13. comparison of the tongue detection accuracy between different literatures and the thesis approach.

Method	Mean Accuracy (mm)
Particle filter [11]	1.69±1.1
EdgeTrack [13]	6.67±3.93
TongueTrak [14]	3.48±1.5
Autotrace [16]	2.61±1.22
Biomechanical model [20]	0.62-0.97
Thesis method	0.955

Appendix B

Results Figures

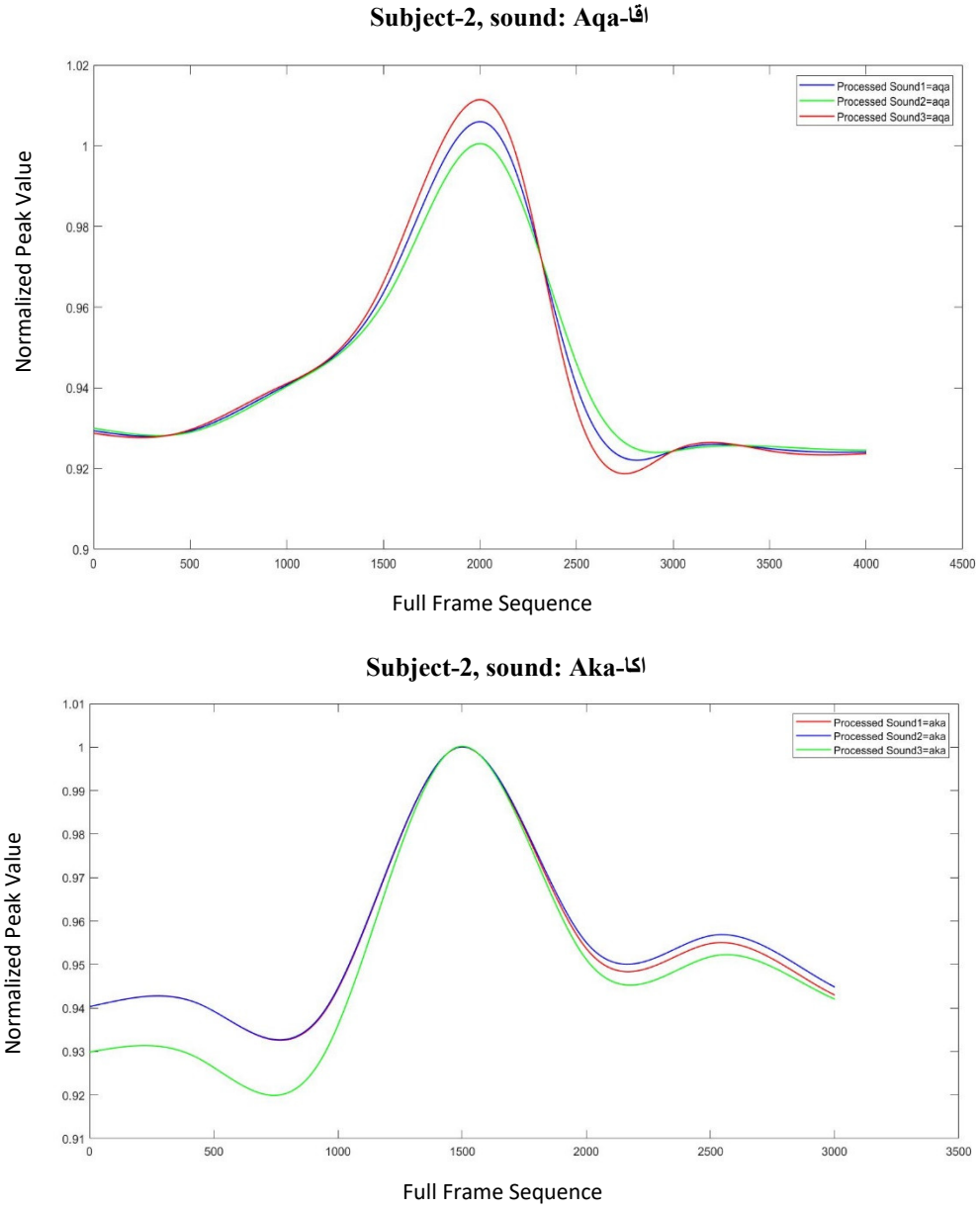


Figure 46. A sample of sound signatures for subject 2. Multiple repetitions of each sound are plotted on the same figure to see the similarity. Above each signature is the sound in English and Arabic.

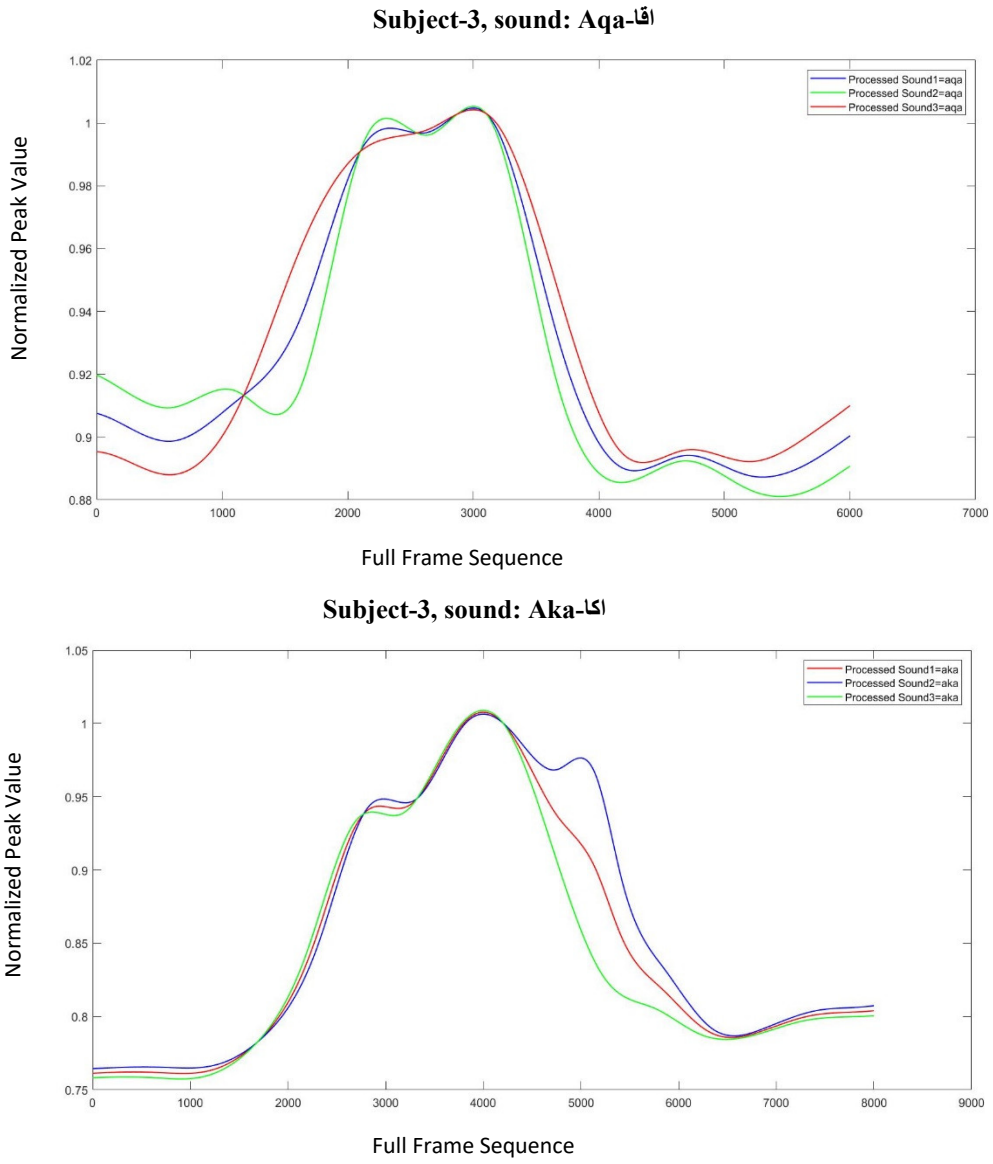


Figure 47. A sample of sound signatures for subject 3. Multiple repetitions of each sound are plotted on the same figure to see the similarity. Above each signature is the sound in English and Arabic.

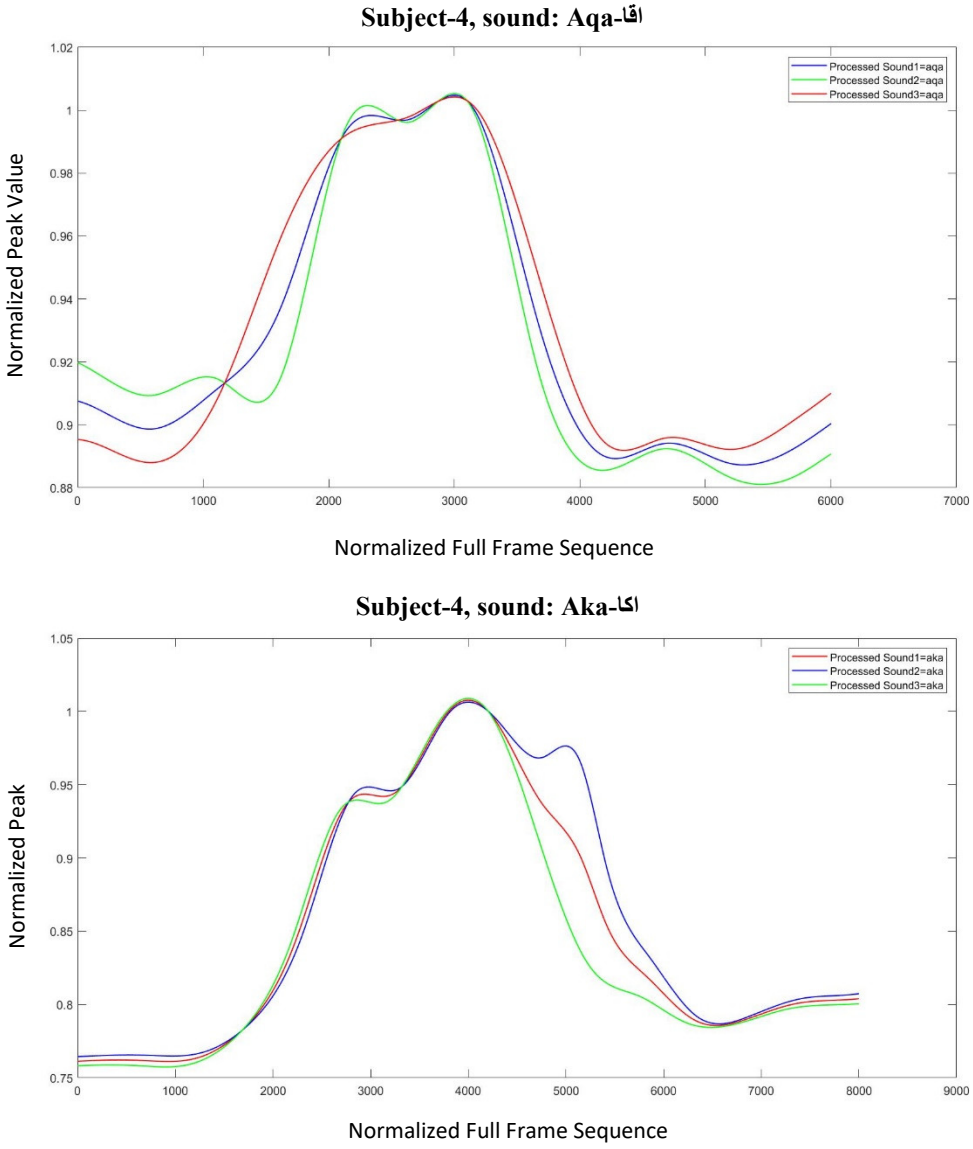


Figure 48. A sample of sound signatures for subject 4. Multiple repetitions of each sound are plotted on the same figure to see the similarity. Above each signature is the sound in English and Arabic.

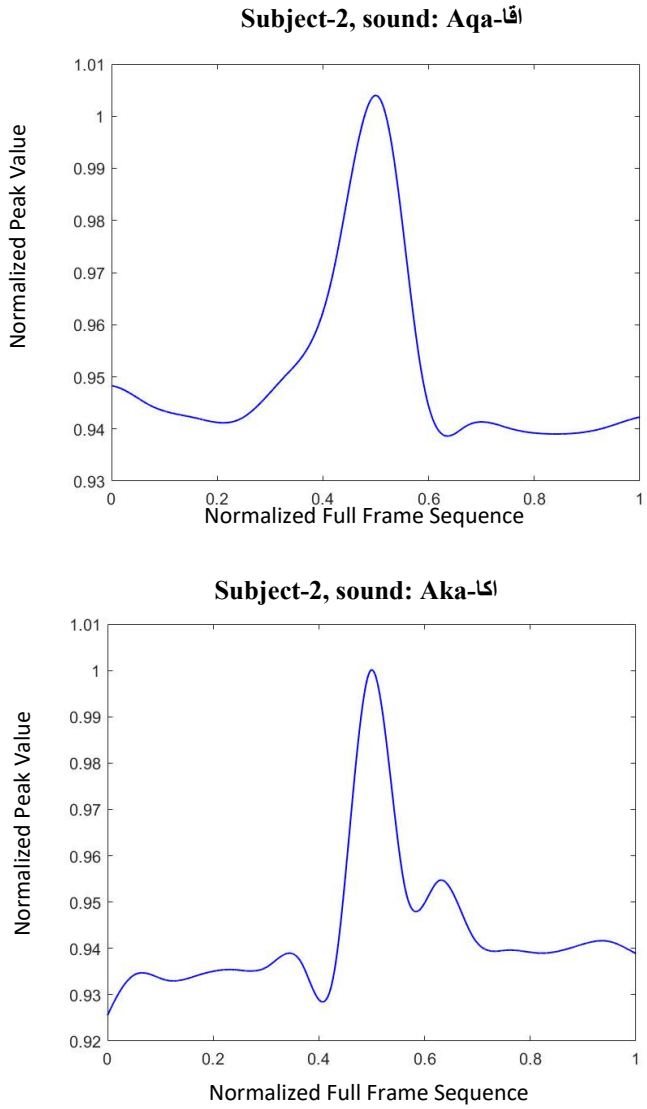


Figure 49. A sample of the average sound signature for subject 2. Above each signature is the subject and the sound in English and Arabic.

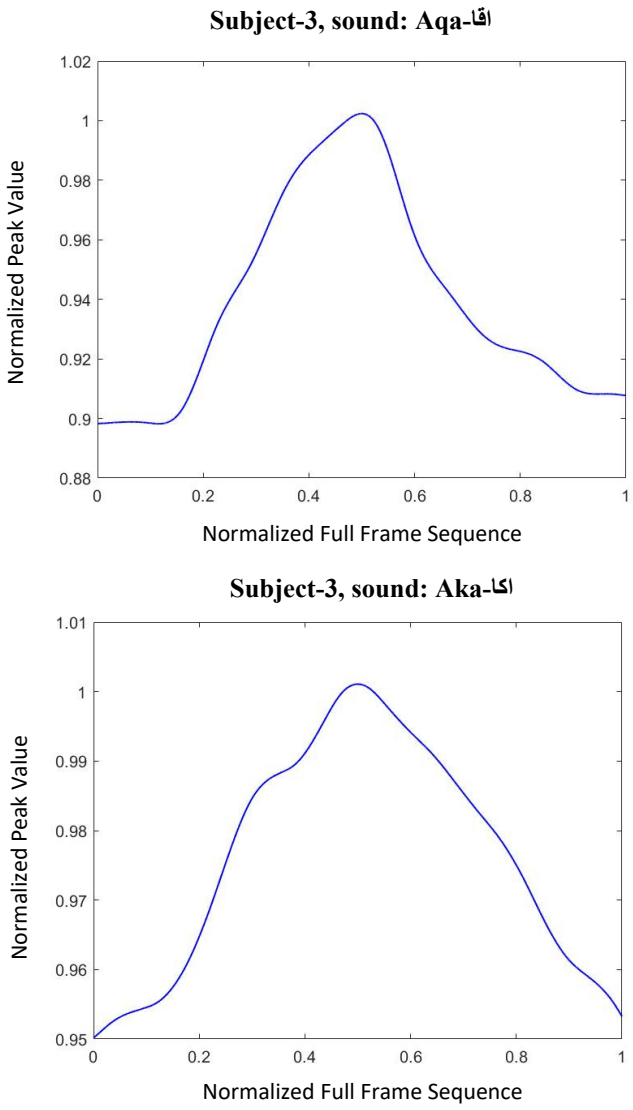


Figure 50. A sample of the average sound signature for subject 3. Above each signature is the subject and the sound in English and Arabic.

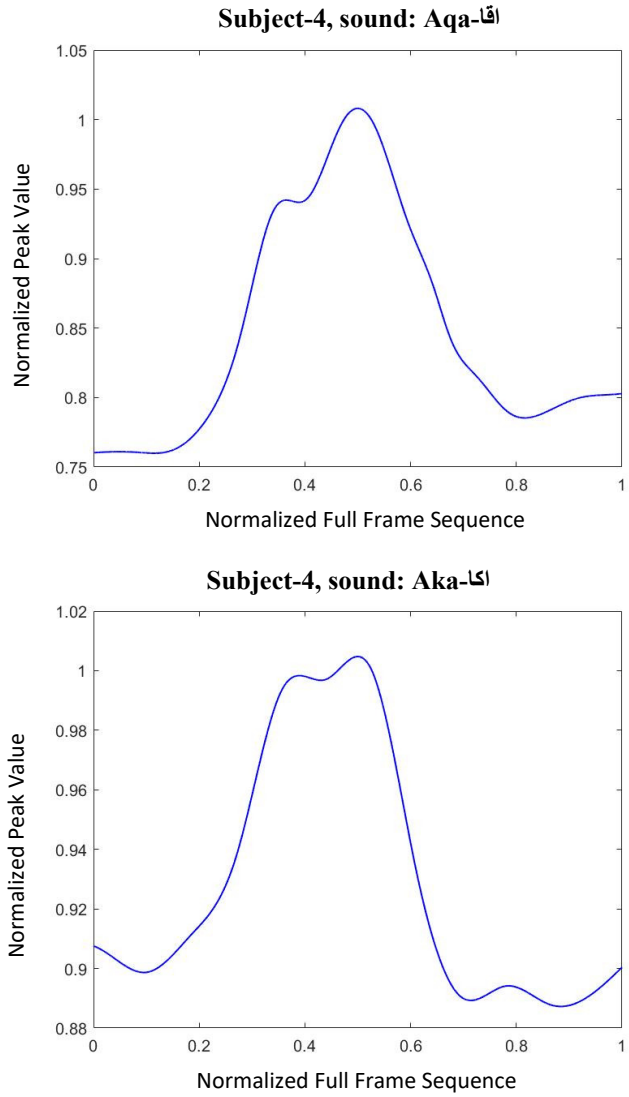


Figure 51. A sample of the average sound signature for subject 4. Above each signature is the subject and the sound in English and Arabic.

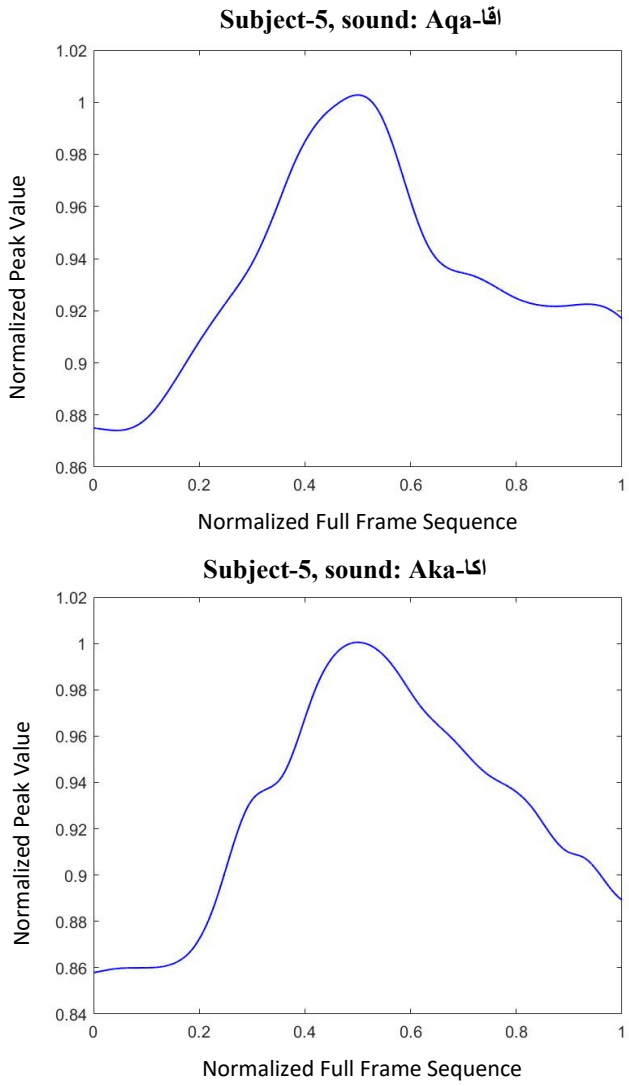


Figure 52. A sample of the average sound signature for subject 5. Above each signature is the subject and the sound in English and Arabic.