3D tongue motion visualization based on the B-mode ultrasound tongue images

Kele Xu

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Électronique

École Doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Kele Xu

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Visualisation tridimensionnelle de la langue basée sur des séquences d'image échographique en mode-B

soutenue le 13 décembre 2016

devant le jury composé de :

M. Bruce Denby	Directeur de thèse
M. Pierre Badin	Rapporteur
M. James Scobbie	Rapporteur
M. Mohamed Chetouani	Examinateur
Mme. Sylvie Le Hégarat	Examinateur

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Bruce Denby, whose expertise, patience and understanding, added considerably to my experience of study during the whole thesis. I appreciate his vast knowledge and skills in many areas, as the topic of this thesis is highly interdisciplinary. I appreciate all of his contributions of time and effort to make my thesis possible.

Very special thanks go out to Prof. Gérard Dreyfus and Pierre Roussel, without whose support and help in my daily life, I would not have completed my thesis. I would like to thank the members of our group: Aurore Jaumard-Hakoun, Clémence Leboullenger, for the friendships between us, as well as the good collaborations. I must also acknowledge Prof. Maureen Stone from University of Maryland Dental School and Prof. Yin Yang from University of New Mexico, who helped a lot for my thesis.

I would like to thank the China Scholarship Council (CSC), the ESPCI-ParisTech and Université Pierre et Marie Curie for their financial support for my thesis.

For this thesis, I would like to acknowledge the members of the jury, Prof. Pierre Badin, Prof. James Scobbie, Prof. Mohamed Chetouani, and Prof. Sylvie Le Hégarat, for their time, help and constructive suggestions.

Lastly, I would also like to thanks my parents, three sisters and my wife for the support through my life.

Abstract

A silent speech interface (SSI) is a system to enable speech communication with non-audible signal, that employs sensors to capture non-acoustic features for speech recognition and synthesis. Extracting robust articulatory features from such signals, however, remains a challenge. As the tongue is a major component of the vocal tract, and the most important articulator during speech production, a realistic simulation of tongue motion in 3D can provide a direct, effective visual representation of speech production. This representation could in turn be used to improve the performance of speech recognition of an SSI, or serve as a tool for speech production research and the study of articulation disorders.

In this thesis, we explore a novel 3D tongue visualization framework, which combines the 2D ultrasound imaging and 3D physics-based modeling technique. Firstly, different approaches are employed to follow the motion of the tongue in the ultrasound image sequences, which can be divided into two main types of methods: speckle tracking and contour tracking. The methods to track speckles include deformation registration, optical-flow, and local invariant features-based method. Moreover, an image-based tracking re-initialization method is proposed to improve the robustness of speckle tracking.

Compared to speckle tracking, the extraction of the contour of the tongue surface from ultrasound images exhibits superior performance and robustness. In this thesis, a novel contour-tracking algorithm is presented for ultrasound tongue image sequences, which can follow the motion of tongue contours over long durations with good robustness. To cope with missing segments caused by noise, or by the tongue midsagittal surface being parallel to the direction of ultrasound wave propagation, active contours with a contour-similarity constraint are introduced, which can be used to provide “prior” shape information. Experiments on synthetic data and on real 60 frame per second data from different subjects demonstrate that the proposed method gives good contour tracking for ultrasound image sequences even over durations of minutes, which can be useful in applications such as speech recognition where very long sequences must be analyzed in their entirety.

Using speckle tracking, the motion information can be extracted by following the speckles, which can be used to drive a generic 3D Finite Element Model (FEM) directly. Modal reduction and modal warping techniques are applied to model the deformation of the tongue physically and efficiently in 3D, which can handle with large deformation while retaining calculation efficiency. Nevertheless, the performance of speckle tracking was found to be somewhat unstable with comparison to contour tracking method in ultrasound tongue image sequences, which leads to unrealistic deformation of the 3D tongue model. Contour tracking can be more stable for the characterization of the motion of the tongue. However, obtaining the correspondence between contours of different frames is of great difficulty and registration between the 2D ultrasound image and 3D tongue model a major challenge. In this thesis, we show that these challenges can actually be converted into a "3D shape search" problem, based on which a more robust and realistic simulation is achieved. Compared to 2D images, such a 3D tongue motion visualization system can provide additional visual information and a quantitative description of the tongue's 3D motion. This work can be helpful in a variety of fields, such as speech production, articulation training, speech disorder study, etc.

Résumé

Une interface vocale silencieuse (SSI) est un système permettant une communication vocale à partir d'un signal non audible. Un tel système emploie des capteurs qui enregistrent des données non-acoustiques, pour la reconnaissance et la synthèse vocales. Cependant, l'extraction des caractéristiques articulatoires robustes à partir de ces signaux reste un défi. La langue est une composante majeure de l'appareil vocal, et l'articulateur le plus important dans la production de parole. Une simulation réaliste du mouvement de la langue en 3D peut fournir une représentation visuelle directe et efficace de la production de parole. Cette représentation pourrait à son tour être utilisée pour améliorer les performances de reconnaissance vocale d'un SSI, ou servir d'outil dans le cadre de recherches sur la production de parole et de l'étude des troubles de l'articulation.

Dans cette thèse, nous explorons un nouveau cadre de visualisation en trois dimensions de la langue, qui combine l'imagerie échographique 2D et une technique de modélisation tridimensionnelle fondée sur la physique. Tout d'abord, différentes approches sont utilisées pour suivre le mouvement de la langue dans les séquences d'images échographiques, qui peuvent être regroupées en deux principaux types de méthodes : le suivi de la granularité et le suivi de contour. Les méthodes de suivi du chatoiement (*speckle tracking*) comprennent le recalage de déformations (*deformation registration*), le flux optique, et la méthode de transformation de caractéristiques visuelles invariante à l'échelle (*Scale-invariant feature transform*, ou *SIFT*). En outre, une méthode de suivi réinitialisation basée sur l'image est proposée afin d'améliorer la robustesse du suivi du chatoiement.

En comparaison avec le suivi de chatoiement, l'extraction du contour de la surface de la langue à partir d'images échographiques présente des performances supérieures et une meilleure robustesse. Dans cette thèse, un nouvel algorithme de suivi de contour est présenté pour des séquences d'images échographiques de la langue. Cet algorithme permet de suivre le mouvement des contours de la langue sur de longues durées avec une bonne robustesse. Pour résoudre la difficulté causée par les segments manquants dus au bruit ou celle causée par la surface mi-sagittale de la langue qui est parallèle à la direction de propagation de l'onde ultrasonore, nous proposons d'utiliser des contours actifs avec une contrainte de similitude de

contour, qui fournissent des informations *a priori* sur la forme de la langue. Des expériences sur des données synthétiques et sur des images réelles acquises sur différents sujets à la cadence de 60 images par seconde montrent que la méthode proposée donne un bon contour de suivi pour ultrasons des séquences d'images, même sur des durées de quelques minutes. Cette technique peut par conséquent être utile dans des applications telles que la reconnaissance vocale où de très longues séquences doivent être analysées dans leur intégralité.

Pour le suivi du chatoiement, l'information de mouvement peut être extraite en suivant des tavelures, qui peuvent être utilisées pour piloter un modèle 3D générique utilisant la méthode des éléments finis (*Finite Element Model*, ou *FEM*) directement. Des techniques de réduction et de déformation modales sont appliquées pour modéliser efficacement de manière physique la déformation de la langue en 3D, permettant de traiter avec une grande déformation tout en conservant l'efficacité de calcul. Néanmoins, les performances du *speckle tracking* a été jugée plutôt instable par rapport à la méthode de suivi de contour dans des séquences d'images échographiques de la langue, ce qui conduit à une déformation irréaliste du modèle 3D de la langue. Le suivi de contour peut être plus stable pour caractériser le mouvement de la langue. Cependant, l'obtention de la correspondance entre les contours des différentes images est d'une grande difficulté et trouver la correspondance entre l'image de l'échographie 2D et modèle de langue 3D est un défi majeur. Dans cette thèse, nous montrons que ces défis peuvent effectivement être convertis en un problème de "recherche de forme 3D", sur la base duquel une simulation plus robuste et réaliste est atteinte. Par rapport à des images 2D, un tel système de visualisation de mouvement 3D de la langue peut fournir des informations visuelles supplémentaires et une description quantitative du mouvement 3D de la langue. Ce travail peut être utile dans une variété de domaines, tels que l'étude de la production de la parole, l'orthophonie, l'étude des troubles de la parole, etc.

Table of Contents

List of Figures

List of Tables

Abbreviation

SSI	Silent Speech Interface
MRI	Magnetic Resonance Imaging
EMA	Electromagnetic Articulography
SNR	Signal-to-Noise Ratio
FEM	Finite Element Model
FPS	Frame Per Second
SSIM	Structural Similarity Metric
CW-SSIM	Complex Wavelet Structural Similarity Metric
RF	Radio Frequency
SONAR	Sound Navigation and Ranging
PSF	Point Spread Function
STE	Speckle Tracking Echocardiography
SAD	Sum of Absolute Differences
MAD	Minimum Absolute Differences
SSD	Sum of Squared Differences
NCC	Normalized Correlation Coefficient
ML	Maximum Likelihood
PDF	Probability Density Function
MSE	Mean Square Error
PSNR	Peak Signal-to-Noise Ratio
SIFT	Scale-Invariant Feature Transform
SIFT Flow	Scale-Invariant Feature Transform flow

Chapter 1
Introduction

1.1 Silent speech interface concepts

Speech is the vocalized form for human-to-human communication, which is the most common and useful interface for human daily communication. Unfortunately, traditional natural speech interfaces present several problems:

- Speech is one-to-many modality, which can give rise to problems of users' interference and communication security;
- If there is a high level of background noise, the quality of speech communication degrades rapidly;
- The speech modality may be impossible when a speaker is incapacitated by illness or injury, either temporarily (laryngitis, flu, etc.) or permanently (cancer, laryngectomy, pulmonary insufficiency, accident, etc.);
- Speech communication may be impossible when the parties involved do not share a common language.

All of these difficulties arise from the propagation nature of the acoustic speech signal. The situation is similar with Automatic Speech Recognition (ASR) for machines, although this technique has suffered considerable evolution in last decades. When the audio signal is corrupted by environmental noise, the speech recognition performance degrades rapidly, which makes the communication unfeasible.

Were it feasible, however, to capture an exploitable speech signal at the production stage, before an audible speech is produced, or indeed suppress completely the audible speech signal by interdicting the use of the vocal chords, which could overcome the abovementioned difficulties. Such a system is referred to as a silent speech interface (SSI) [1], which enables

speech communication with non-audible signals (Figure 1.1 gives the framework of the ultrasound and optical imaging-based SSI [2]), that employs sensors to capture non-acoustic features for speech recognition. As the SSIs system can capture the signals before the speech production, they have the potential to be background noise insensitive, natural sounding.

Although the performance of SSI system is not stable yet, the potential applications of SSI seem evident in several different domains, just to name a few, telecommunication, medical fields, speech production (acoustic-articulatory inversion), et al. Over the past several years, such a SSI concept had gained more public acceptance, and more and more imaging techniques have been employed for the silent speech recognition problem. Indeed, the feasibility of SSIs for practical communication began to be shown.

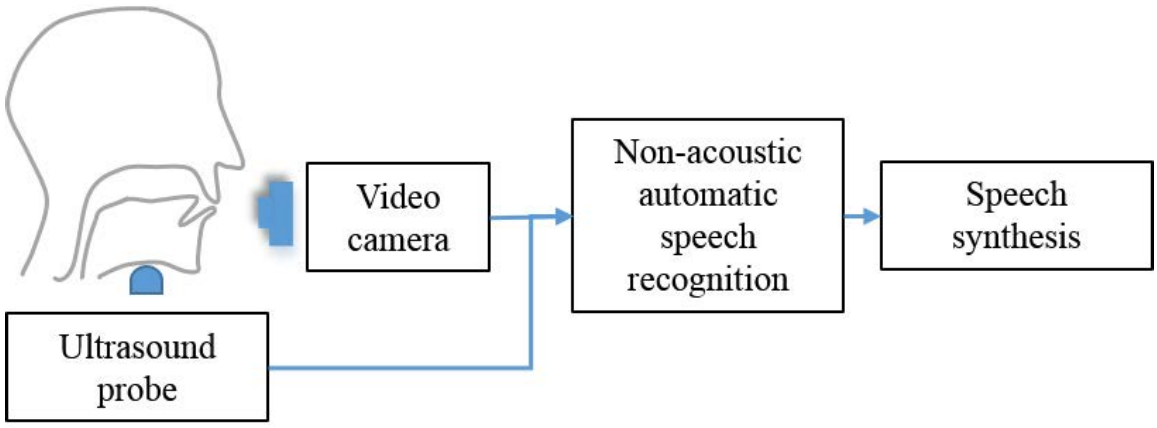
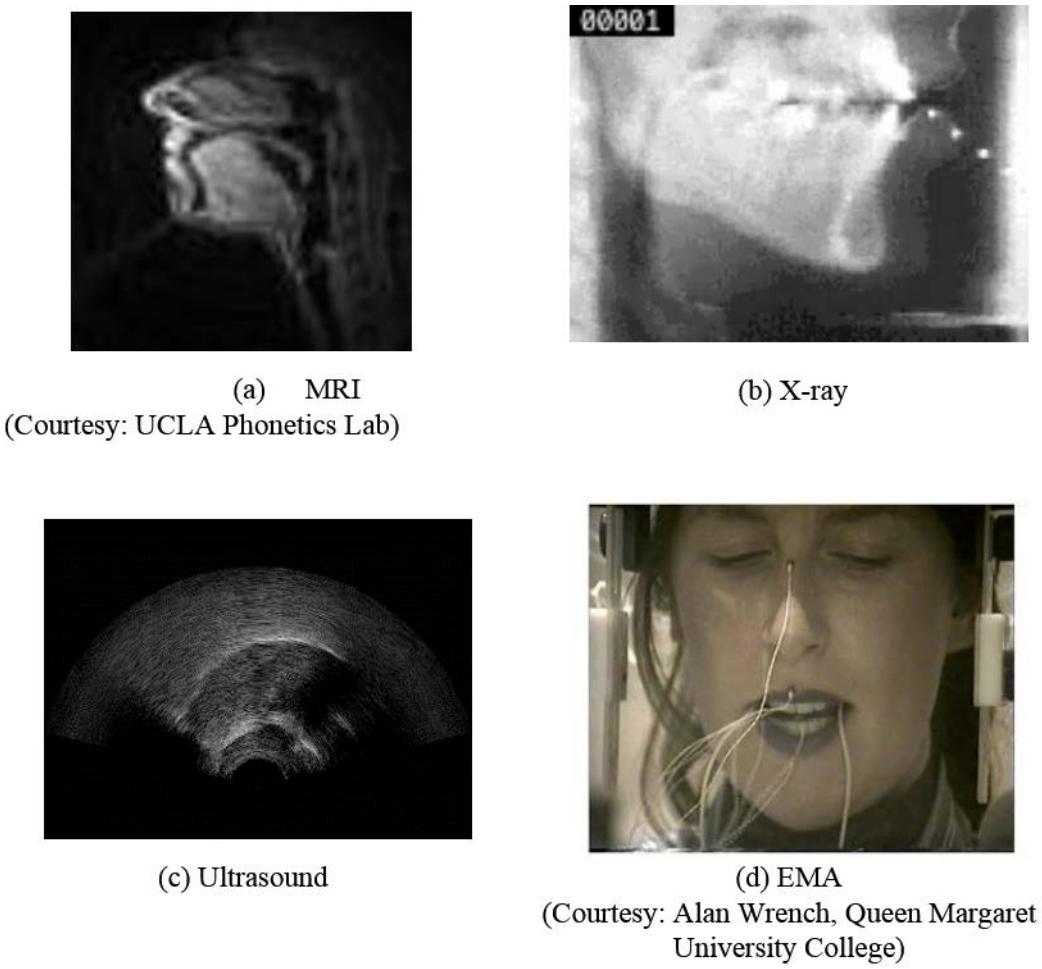


Figure 1-1 Ultrasound-based SSI (schematic).

1.2 Related work

To some sense, the SSIs are the speech recognition systems based on the analysis of the non-acoustical signal recorded during speech production. Thus, understanding and modeling the procedure that produces speech is essential to advance speech production science, which may also be helpful to improve the performance of SSI. Indeed, speech production has been studied over several decades using a variety of different types of sensors. During speech production, the tongue is the most important component of the vocal tract [3] for forming consonants and vowels. If we can recover the motion of the tongue motion quantitatively and

robustly, the SSI systems' recognition performance may be further improved. However, measuring tongue's motion directly is difficult since the tongue lies within the oral cavity and is inaccessible to most instruments. Various imaging techniques have been used to analyze the movement indirectly, including Magnetic Resonance Imaging (MRI) [4] (as shown in Figure 1-2 (a), accessed from [5]), X-ray [6] (as shown in Figure 1-2 (b)), ultrasound [3] (as shown in Figure 1-2 (c)) and electromagnetic mid-sagittal articulography (EMA) [7] (as shown in Figure 1-2 (d), accessed from [8]).



(a) MRI
(Courtesy: UCLA Phonetics Lab)

(b) X-ray

(c) Ultrasound

(d) EMA
(Courtesy: Alan Wrench, Queen Margaret University College)

Figure 1-2 Different imaging techniques used to visualize the vocal tract.

X-ray imaging has better temporal resolution, but exposes subjects to radiation and is a through-transmission technique, which projects the entire 3D head onto a single 2D image. MRI system captures tongue movement with good resolution (as can be seen in Figure 1-2

(b)), but requires summation of repetitions to get good spatiotemporal resolution. MRI images are recorded in supine position, which is atypical for speech. EMA data can provide directly motion information by measuring the motion trajectory of the tongue, however, its invasive property makes it difficult for natural speech production recording. Ultrasound is another widely used tool in the speech production research. Nevertheless, the Signal-to-Noise Ratio (SNR) of the ultrasound image is quite low, and the speckle noise degrades the images by concealing fine structures and reduces the signal to noise level. The strength of ultrasound imaging is that it images tongue motion at a fairly rapid frame rate (60Hz), which can capture subtle and swift movement during speech production. Furthermore, ultrasound imaging is noninvasive, less expensive than other imaging systems, and convenient for experimentation. In this thesis, due to its appealing properties, we employ the ultrasound to capture the motion information of tongue. To recover the continuous motion of the tongue, robust tracking approach is in high demand for ultrasound-based SSI system.

Tracking the tongue in an ultrasound image sequence is a challenging task due to the poor image quality and fast, irregular motion. Many literatures aimed to solve aforementioned problems. The classical approach to quantify the motion of the tongue is to extract the upper surface of the contour in the ultrasound image sequences. A non-exhaustive literature summary (Table 1-1) is conducted on the contour tracking approaches in ultrasound tongue images. As can be seen from the table, a variety of processing techniques can be used to track the contours of the tongue in the ultrasound images, for example, active contour models (also called as “Snake” model) [9], [10]; active appearance models (AAM) [11]; machine leaning-based tracking [12], [13], [14]; ultrasound image segmentation-based approaches. Most of the algorithms were applied to a static frame.

Table 1-1 Summary of contour extraction methods in ultrasound tongue images

Authors	Title	Methods
Yusuf Sinan Akgul, Chandra Kambhamettu, and Maureen Stone [9]. (1999)	Automatic extraction and tracking of the tongue contours.	Active contour model
Li Min, Chandra Kambhamettu, and Maureen Stone [10]. (2005)	Automatic contour tracking in ultrasound images.	Active contour model
Anastasios Roussos, Athanassios Katsamanis, and Petros Maragos [11]. (2009)	Tongue tracking in ultrasound images with active appearance models.	Active appearance model
Ian Fasel and Jeffrey Berry [12]. (2010)	Deep belief networks for real-time extraction of tongue contours form ultrasound during speech.	Machine-learning based
Lisa Tang, Tim Bressmann, and	Tongue contour tracking in dynamic	Ultrasound image

Ghassan Hamarneh [15]. (2012)	ultrasound via high-order MRFs and efficient fusion moves.	segmentation-based
Diandra Fabre, Thomas Hueber, Florent Bocquelet, and Pierre Badin [13]. (2015)	Tongue Tracking in Ultrasound Images using EigenTongue Decomposition and Artificial Neural Networks.	Machine-learning based method
Aurore Jaumard-Hakoun, Kele Xu, Gerard Dreyfus, Pierre Roussel, Maureen Stone and Bruce Denby [14]. (2015)	Tongue contour extraction from ultrasound images based on deep neural network.	Machine-learning based method

Since the revolution of the neural networks [16], machine-learning based contour tracking method has made great progress. However, the training depends on the large number of hand-labeled frames, which are not easy to obtain. Due to their ability to be guided by constraint forces, active contours may be particularly useful for contour tracking in images of the tongue. Indeed, tongue contour tracking using energy-minimization-based active contours, or "Snake", has been used extensively in previous research. In the contour tracking algorithm proposed in [9], the snake model was used on ultrasound tongue images for the first time, introducing gradient information in the definition of an external energy term. By including an intensity-related constraint, [10] proposed a new contour tracking system, named **EdgeTrak**, which works very well for sequences in which the entire contour always remains visible. If a part of the contour disappears in some images, however, due to poor acoustic coupling or a decrease in reflected energy, the obtained contour can become erroneous and require manual re-initialization to get back on track. This can become problematic for applications where long speech sequences are to be analyzed.

To help cope with low SNR in ultrasound images, some researchers have proposed to use other imaging modalities (e.g., X-rays) to obtain prior tongue shape information [11]. However, these modalities may use different frame rates, and registration between different modalities can also be difficult, making such an approach impractical (the use of X-ray is also nowadays banned). Thus, the contour tracking problem still poses a challenge in ultrasound tongue image sequences.

For a very long time, tongue motion analysis has been essentially limited to the midsagittal plane, but progress of the 3D imaging system (e.g. 3D MRI imaging system and 3D ultrasound imaging system) and the progress of the 3D computer simulation in anatomical and physiological field, have brought into this domain led to the point where dynamic 3D visualization of tongue motion has been unavoidable. 3D tongue modeling based on finite-

element-modeling (FEM) became more and more popular in the research field of speech production. Compared to the 2D image sequences, the effective 3D visualizing motion of the human tongue can provide extra visual information. A summary on the related literature of 3D vocal tract modeling are given in the table below.

Table 1-2 Related literature of 3D vocal tract modeling

Authors	Title
Maureen Stone [17]. (1990)	A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data.
Reiner Wilhelms-Tricarico [18]. (1995)	Physiological modeling of speech production: Methods for modeling soft-tissue articulators.
Olov Engwall [19]. (1999)	Vocal tract modeling in 3D.
Olov Engwall [20]. (2000)	A 3D tongue model based on MRI data.
Olov Engwall [21]. (2001)	Using linguopalatal contact patterns to tune a 3d tongue model.
Olov Engwall [22]. (2003)	Combining MRI, EMA and EPG measurements in a three-dimensional tongue model.
Ian Stavness, John E. Lloyd, Sidney Fels [23]. (2012)	Automatic prediction of tongue muscle activations using a finite element model.
John E. Lloyd, Ian Stavness, Sidney Fels [24]. (2012)	ArtiSynth: A fast interactive biomechanical modeling toolkit combining multi-body and finite element simulation.
Yin Yang, Xiaohu Guo, Jennell Vick, Luis G.Torres, Thomas Campell [25]. (2013)	Physical-based deformable tongue visualization.

In brief, previous efforts made 3D tongue modeling focus on three categories: static tongue modeling using geometry data-driven method ([17], [18], [19], [20], [21] and [22]), dynamic tongue modeling using muscle activation approach ([23], [24]), and motion-driven 3D tongue modeling [25]. In more detail, static tongue modeling aims to recover the 3D tongue shape using multi-slice using different imaging techniques, the slice come from both the midsagittal plane and the coronal plane. For dynamic tongue modeling, researchers aim to model the tongue's motion by simulating the stimulus of muscle, which can be used to drive the 3D tongue model in a dynamic manner. But, the tongue is complicated to model due to its large global and local deformation and intrinsic muscular activation. However, the treatment of muscle activation in the tongue still presents a number of challenges, despite many attempts to characterize the bio-mechanical properties of the tongue. Indeed, our understanding of human tongue bio-mechanical property is still very limited. Motion-driven deformable model is an alternative method for dynamic tongue modeling, and it has been

widely used for computer animation. Rather than muscle-driven 3D tongue modeling, motion-derived 3D modeling is used in our framework, as an alternate type of dynamic tongue modeling. This kind of inverse dynamics-based [25] approach can govern the deformation of the tongue model by using the motion information as position constraints, then forward dynamics by using the prescribed target motion trajectory.

As our goal is to recover the continuous motion of the tongue, the static modeling method may not be applicable in our case. Moreover, as the tongue's movement is swift, the 3D dynamic tongue model should generate different gestures in a short time-step. Thus, muscle-driven based approach may be also unsuitable for our work.

1.3 The thesis work

Based on the summary of previous work on contour tracking and 3D tongue modeling, we can see that 3D tongue modeling using ultrasound image sequences is a great challenge even through sustainable efforts have been made. However, if we can model the motion of the tongue quantitatively in 3D, extra information may be obtained to help improve the performance of the ultrasound-based silent speech interface.

Indeed, up until now, ultrasound-based silent speech interfaces have remained experimental due to the difficulties in rendering the sensor data independent of experimental conditions. Ultrasound images are noisy and difficult to interpret even under the best of conditions, and are highly dependent on exact sensor placement. Moreover, the ultrasound imaging quality varies between subjects, the woman's imaging quality is usually better than male subject is, the younger subject is always better than the elder [3]. Speaker-independent recognition (or multi-speakers' recognition) poses a greater challenge, and how to extract the robust and distinctive feature is of importance for the success of the SSI recognition system. On the other hand, devising a stable sensor acquisition platform is an extremely challenging task, as is indeed the inverse procedure of trying to correct for sensing problems using real-time post-processing.

This can be accomplished by making use of non-acoustic sensing of the articulator movement to create an acoustic vocal tract model (the component in the rectangle with the dotted line in Figure 1-3), and driving the model in software with an artificial vocalization

signal. The breakthrough required in order to make this type of instrumentation viable for silent speech/song production modalities will consist of devising imaging processing techniques capable of accurately and reliably measuring articulator movement in real time in a series of standard video and ultrasound images of the vocal tract. This will require, for example, bringing image processing algorithms on the difficult problems of noise, occlusion, acoustic contact ambient lighting conditions, etc., and also incorporating a priori anatomical knowledge about the articulators via 3D finite element physical models of these organs. Thus the thesis topic, the development of an active real-time model of the vocal tract, sits at the crossroads of the data acquisition, image processing, feature extraction, 3D modeling and human anatomy.

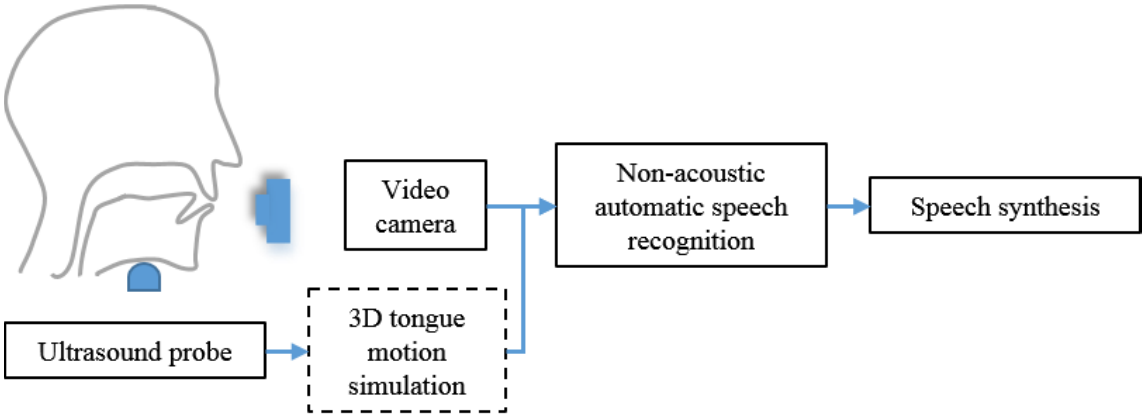


Figure 1-3 New potential ultrasound-based SSI framework.

In this thesis, we propose a novel 3D tongue visualization framework based on ultrasound image sequences of the tongue, which combines the 2D motion information extraction from ultrasound imaging system and dynamic 3D physics-based modeling strategy. And it can be useful for speech production research and the enhancement of silent speech recognition. The framework of the thesis work is given in Figure 1-4.

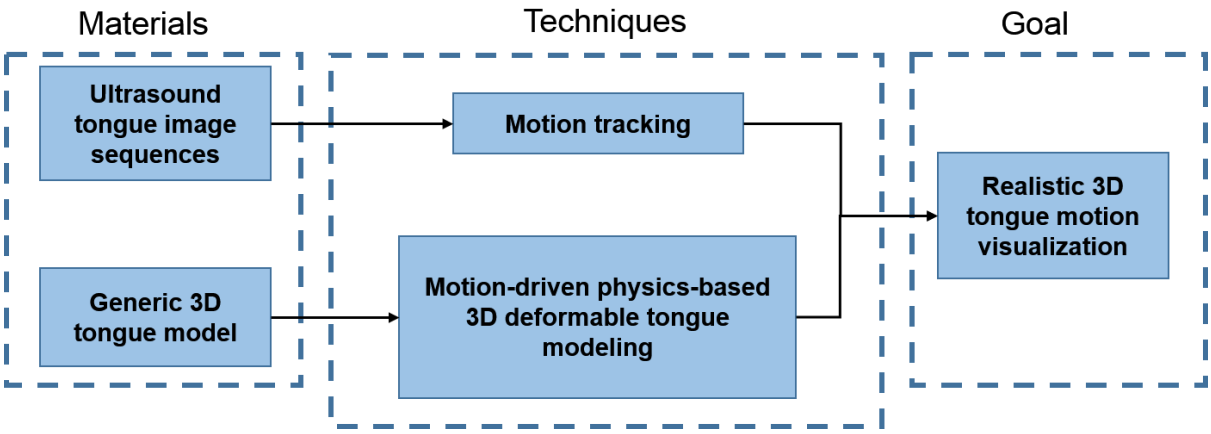


Figure 1-4 Framework of the whole thesis.

As can be seen from the Figure 1-4, the materials used throughout this thesis include ultrasound tongue image sequences and a generic 3D tongue Finite Element Model (FEM), which is extracted from Artisynth [24]. The original tongue model consists of 1,803 nodes and 8,606 tetrahedral elements. In this thesis, to create more smooth mesh and obtain a better visualization result, we subdivide the original generic tongue mesh, which consists of 12,967 nodes and 43,930 tetrahedral elements. (as shown in Figure 1-5).

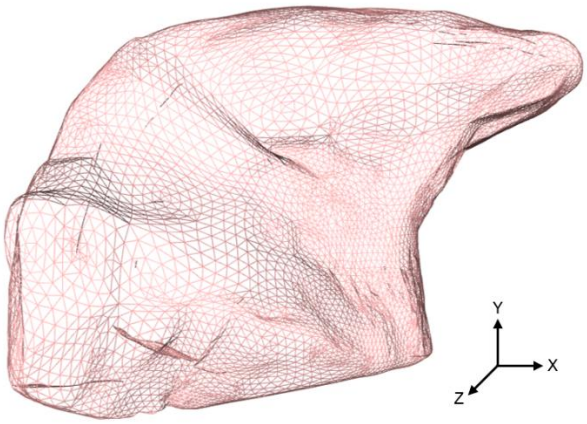


Figure 1-5 The tetrahedral mesh used for the simulation, and smoothed generic tongue mesh consists of 12,967 nodes and 43,930 elements.

In the early period of the thesis, we aims to track the continuous motion of the tongue using ultrasound with high robustness for long duration. The attempt is twofold: speckle tracking and contour tracking:

Speckle tracking: Speckle tracking methods aim to track the backscattered echoes produced by ultrasound scatters in the tissue of the tongue. The principle of speckle tracking is quite simple: two-dimensional speckle is defined as the spatial distribution of gray values in the ultrasound image, which is commonly referred as the speckle pattern. If the position within the segment of tongue tissue changes, it can assume that the position of its acoustic fingerprint (texture) will change accordingly. By tracking these patterns, we can follow the motions of tissue in real-time. In this thesis, we analyze the fundamental principles of speckle tracking. Then, different speckle tracking techniques have been implemented and tested on the ultrasound tongue sequences. Moreover, the complex wavelet structural similarity-based automatic speckle tracking re-initialization is also presented in this thesis. However, the performance of the speckle tracking was found to be somewhat unstable.

Contour tracking: Compared to speckle tracking, the extraction of the contour of the tongue surface from ultrasound images exhibits superior performance and robustness. The goal of contour tracking is to track the surface of tongue in the ultrasound image. However, accurate, robust tongue contour extraction, remains a challenging problem for ultrasound sequences of long duration, due to acoustic effects, speckle noise and poor signal-to-noise-ratio (SNR). Despite significant research efforts, manual refinement is usually needed, which is impractical for large-data tracking systems. Thus, in this thesis, a novel automatic contour tracking algorithm had been developed, wherein active contour model with contour group similarity constraint is used. Experiments on synthetic data and on real 60 frame per second (FPS) data from different subjects demonstrate that the proposed method gives good contour tracking for ultrasound image sequences even over durations of minutes, which can be useful in applications such as speech recognition where very long sequences must be analyzed in their entirety. Moreover, image-similarity-based contour tracking is also presented in this thesis, which achieves state-of-art performance with comparison to previous approaches.

In the later part of the thesis, we aim to employ the motion information extracted by using speckle tracking or contour tracking to drive the 3D Finite-Element Model (FEM) to simulate the motion of the tongue. The 3D tongue modeling technique is based on the deformable modeling technique. In more detail, the 3D model can be driven by imposing extra positional constraints at specified finite element nodes to enforce their displacements to some user-specified values.

Speckle tracking-based modeling: For speckle tracking, to drive the 3D tongue model, the modal displacement will be calculated by using the movement of the patterns directly, which will be transmitted to certain nodes of on the midsagittal tongue model surface in order to drive the 3D model at the acquisition rate of ultrasound image sequence. However, as mentioned above, the performance of speckle tracking is not robust yet, which make this kind of method unfeasible for practical application.

Contour tracking-based modeling: Using contour tracking to drive the 3D tongue model is the alternative method. However, due to the change of the contour tracking quality, the length of the contour extracted changes dramatically between adjacent frames, which makes it difficult to obtain the correspondence between contours of different frames. Thus, the registration between the 2D ultrasound image and 3D tongue model poses a great challenge. In the thesis work, we show that these challenges can actually be converted into a "3D shape search" problem, and the deformation simulated with the "3D shape search" framework is informative and qualitatively realistic compared with speckle tracking-based animation.

Overall, the main contributions of our work can be summarized as:

1. Based on the framework of [25], which make use of EMA data to drive the 3D tongue model, a physics-based deformable tongue visualization framework is presented using the finite-element-method (FEM) driven by ultrasound image sequences. Compared to the 2D image sequences, this technique can deliver much enriched visual information and can be used for several different speech tasks. Compared to EMA-based motion visualization, extra visualization can be obtained.
2. To overcome the problem of faint contour and against the signal-level noise, the similarity constraint is added to the active contour model. Moreover, complex wavelet structural similarity (CW-SSIM) measurement technique is used to reset contour tracking automatically, which can improve the robustness of the contour tracking algorithm.
3. By employing the image similarity-based constraint, a novel contour tracking method is proposed in the work of the thesis, which achieves state-of-art performance. On the other hand, a comparative study is conducted on the different contour tracking algorithms with automatic re-initialization.
4. Different motion tracking-based 3D tongue modeling methods are explored. By using the contour extracted from the ultrasound tongue image sequence, a novel shape similarity-

based 3D tongue dynamic modeling technique is proposed, based on which a more realistic motion simulation is implemented.

Moreover, although our original target is to design a platform to visualize the 3D tongue motion using the ultrasound data, this platform can also be extended to other imaging modals (e.g. MRI, EMA, X-Ray) to assist the studies on speech production.

1.4 Structure of the thesis

The organization of this thesis is given below, which is mainly based on our attempts to dynamic model the tongue motion in 3D using the ultrasound image sequences.

In detail, in this chapter, after an initial introduction to the rich variety of SSI applications, the ultrasound-based SSI problem is discussed specifically. Then a non-exhaustive literature review is conducted on both the contour tracking in ultrasound tongue image sequences and 3D tongue modeling technique. Then the thesis work is summarized, highlighting the contributions and improvements with comparison to previous solutions.

Chapter 2 describes the fundamentals of B-model ultrasound tongue imaging, including the basic principles of B-mode ultrasound imaging, as well as the ultrasound-tongue tissue interaction, which laid the foundation of speckle tracking and contour tracking in ultrasound tongue image sequences.

Chapter 3 presents the application of speckle tracking techniques in ultrasound tongue image sequences. Firstly, the fundamental principles of speckle tracking are given. Then, the practical implementations of speckle tracking techniques are discussed. Moreover, an image-similarity-based automatic speckle tracking re-initialization method is proposed in this chapter.

In chapter 4, the contour tracking approaches are investigated for ultrasound tongue image sequences. One of the methods is based on the modified active contour model, with the aim to cope with missing segments caused by noise, or by the tongue midsagittal surface being parallel to the direction of ultrasound wave propagation. Also, the image-similarity based automatic contour tracking re- initialization method is discussed in this chapter.

In chapter 5, the summary of the physical-based deformation modeling algorithm adopted in our framework is presented. Speckle tracking and contour tracking based 3D dynamic modeling of the tongue is given in this chapter. The experimental results and analyses are presented.

The last chapter concludes the thesis, outlining the lessons drawn from our study, its limitations, and proposing directions for future research.

Chapter 2
Principles of ultrasound tongue imaging

2.1 Introduction

Compared to other medical imaging modalities (such as MRI, X-ray or EMA), the ultrasound imaging system is portable, non-invasive and relatively inexpensive. Over the last decades, the ultrasound imaging has drawn widespread acceptance and witnessed continuous development in the diagnosis and more complex arrangements designed to have been be used to enhance the imaging performance [26]. The primitive display modes for ultrasound, such as A-mode, and static B-mode have given way to real-time, high-resolution and ultra-fast imaging [27]. Moreover, some new ultrasound systems can visualize the internal structures in three dimensions. In essence, the basic principles of ultrasound imaging are still used in the modern medical ultrasound systems. In this chapter, we begin with the basic principles of ultrasound imaging, including the physics basis and principles of ultrasound imaging, which laid the foundations for ultrasound tongue imaging.

Ultrasound imaging has been used to image the human tongue for over three decades by phoneticians, predominately in a clinical context [17]. In the early stage, the ultrasound equipment was found in the hospitals, which made them difficult to be accessible for the speech production researchers. However, with the improvements in technology and reduced cost, the ultrasound is increasingly used as an imaging tool for the researchers in clinical linguists and phonetics fields [3], [28], [29]. Continuous efforts have been made to understand the motor control during speech production, and the ultrasound-based tongue motion analysis has witnessed great progress in last decades.

The classical method for tongue motion analysis aims to extract the contours of the upper tongue surface using ultrasound imaging. Another potential feasible approach in deriving correspondence is speckle-tracking method. All the motion-tracking approaches are based

upon the characterization between the ultrasound waves and the tongue tissue. Thus, in the second part of this chapter, the interaction of the tongue muscle and the ultrasonic waves is analyzed, laying the foundation for speckle tracking and contour tracking, which are the main topics for the next two chapters.

In more detail, the organization of this chapter is given as follows: firstly, in Section 2.2, the basic principles of the ultrasound B-mode imaging are presented, which include the basic physics of the medical ultrasound imaging, a brief introduction of the ultrasound pulse and a general preview of different ultrasound scan types. Section 2.3 discusses the interaction property between the tongue tissues and the ultrasound waves, and the data-acquisition platform is presented. While, in the last part of Section 2.3, the image distortions, which that may occur during the recording, are analyzed. In the end of this chapter, a conclusion is drawn.

2.2 Basic principles of medical ultrasound imaging

2.2.1 Basic physics of medical ultrasound imaging

In brief, ultrasound image imaging is based upon the echo-location principle. However, the connection between the medical ultrasound sound and the echo-location principle was not made until the mature of the underwater acoustics, which made use of SONAR (Sound Navigation and Ranging) to measure the depth of water at sea. The inventions of SONAR and medical ultrasound imaging can be traced to the sinking of Titanic [30] when scientists tried to detect icebergs underwater using echo ranging. Nevertheless, at that time, there were no practical ways to implement the ideas until the discovery of piezoelectricity. In 1916 and 1917, by making use of technologies of piezoelectricity, Paul Langevin and Constantin Chilowsky invented a high-power echo-ranging system to detect the submarines [31].

The recognition that ultrasound could benefit medical diagnoses can be traced to World War I. Afterward, ultrasound is progressively applied to the therapy and the surgery. In the 1970s, medical ultrasound witnessed a rapid expanding with the advent of 2D real-time systems. Color flow systems occurred in 1990s. Presently, active research field includes contrast agents, molecular imaging, tissue characterization, integration with other modalities, such as photo-acoustic imaging [32].

Despite sustainable efforts been made, the basic principles of medical ultrasound imaging are almost the same today as they were several decades ago [33]. The basic medical ultrasound imaging system is shown in Figure 2-1.

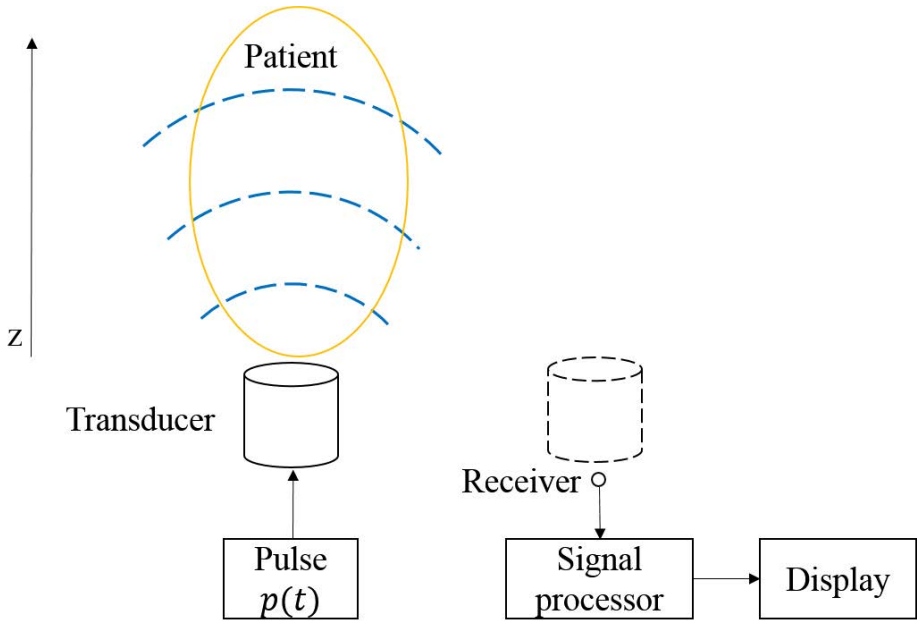


Figure 2-1 Framework of medical ultrasound imaging system.

The pulse is generated from the ultrasound transducers (or probes), which contain multi-piezoelectric crystals. When an electric field is applied to an array of piezoelectric crystal located on the surface of a transducer, mechanical vibration will happen in response. This phenomenon is called the piezoelectric effect, which was originally described by the brothers Pierre Curie and Jacques Curie [34]. The mechanical vibration will result in an ultrasound-high frequency wave, which will propagate through a medium by compression and rarefaction.

The pulse propagates into the body where it reflects off mechanical in the homogeneities regions. In more detail, like audible sound, the pulse will echo back from the interface between transmission mediums of different densities. Strong echoes mean large differences of densities, while weak echoes can occur at the tissues-to-tissues interfaces as they have similar density. Because the distance is the product between the velocity and the time, a

reflector at distance z from the transducer will cause a pulse echo at time $t = 2z/c$, where c is the sound velocity in the body.

2.2.2 Ultrasound pulse

The ultrasound waves can be described by three terms: frequency, wavelength and amplitude. The frequency and wavelength are inversely related. Modern medical ultrasound devices use sound waves in a range of 1-20 MHz. The wavelength of medical ultrasound varies from 1.5 mm at 1 MHz to 0.15 mm at 10 MHz, which enables good depth resolution. The propagation velocity is dependent on the compressibility of the medium. The average velocity of sound in soft tissues is 1540 meters/second. Frequency selection for the transducer is of importance to obtain optimal resolution for ultrasound images. High-frequency waves can image with higher axial resolution, but are more attenuated than the low frequency waves. Low-frequency provides images of lower resolution but can penetrate deeper in the structures.

A pulse excites the transducer with a short pulse in a particular direction, and the ultrasound waves are generated in pulses, which consists of 2-3 sound cycles of the same frequency. The pulse is often modeled as an amplitude modulated sinusoid.

$$p(t) = a(t)e^{i\omega_o t} \tag{2.1}$$

where $\omega_0 = 2\pi f_0$ is the carrier frequency, and $a(t)$ is the envelope.

The emitted ultrasound pulse can be viewed as the impulse function of the ultrasound medical imaging system. Suppose the echo pulse is used to represent the output of the ultrasound system during interrogation of an ideal point target. The echo pulse can also be regarded as the ultrasound system's point spread function (PSF), and received echo pulse can be considered as the impulse response of the biological medium. However, as the pulse travel along a straight path and it is quite short, the pulse is often referred as the ultrasound beam.

2.2.3 Ultrasound scan types

There are several different types of ultrasound scanning types, which include A-mode, B-mode, M-mode and some other formats. The display formats are often selected based on the

practical applications. In this part, we will give the preview of A-mode and B-mode scanning types.

- **A-mode imaging**

The A-mode scan type is the simplest mode, which is the basic element for other ultrasound scans. A-mode transducer scans a line though the tissue, and displays the amplitude versus depth for the echo signal, which is used to measure the distance accurately. Echoes will be generated from the tissue interfaces or tissue-air interfaces.

Suppose at depths $d_1,\cdots,d_n$ there are interfaces with reflectivities $R(d_1)$

$$R(d)=\sum_{n=1}^{N}R(d_n)\delta(d-d_n) \tag{2.2}$$

The signal received by the transducer can be modeled as:

$$v(t)=K\sum_{n=1}^{N}R(d_n)\,p(t-2d_n/c) \tag{2.3}$$

where K is the constant gain factor relating to the impedances of the transducer, electronic pre-amplification. A natural estimate of the reflectivity is

$$\hat{R}(d)=\left|v(2d/c)\right| \tag{2.4}$$

where $\left|v(t)\right|$ is the envelope of the received signal. The estimated reflectivity can be plotted by using the amplitude versus time, which can be regarded as a function of depth d. Hence, this type of ultrasound imaging is called as A-mode imaging [35].

The aforementioned model is a highly simplified model. However, in practical application, when ultrasound wave travels through tissues with different acoustic impedance (the definition is given below), reflection will occur. The acoustic impedances are the measures of the response of the medium to a given acoustic pressure, which is determined by the density ρ and stiffness of the medium. In the tissue, the acoustic impedance is defined as:

$$z=\rho c \tag{2.5}$$

For air, as the density and stiffness are low, thus, z is very small. While, for bone, z is much higher, as it has quite high density and stiffness.

- **B-mode scan**

The B-mode ultrasound image is produced by a transducer array, for example, a large number of small transducer elements are arranged in a straight line. B-mode scan can be viewed as a concatenation of A-mode scans. Normally, a complete B-model image is made up more than 100 lines.

The reflected ultrasonic waves are detected by the same transducer and converted into an electrical signal (Radio Frequency signal). Pulses that returned later are displayed in the image as far away from the transducer. The received echo pulse can be viewed as the impulse response of the medium. The raw RF data for ultrasound waves is not well suited for the interpretation by users. The exemplar ultrasound-processing pipeline for RF signal to B-mode conversion is given in Figure 2-2. Firstly, the envelope of the RF signal is detected and coded in a way given as follows: high-amplitude reflections are the bright pixels in an image and low-amplitude reflection is represented by dark. After frequency compounding, the envelope of RF signal is detected to represent the original signal. Afterwards, a nonlinear intensity map is applied to decrease the range of the data. In the end, several different filters specific are used to create a B-model image. In one sense, the B-mode ultrasound imaging was constructed from the echoes to form a cross-sectional image representing tissues and organ boundaries within the body [35]. The brightness of each pixel is related to the amplitude of the reflected pulse. Hence, this type of ultrasound imaging is called as B-mode imaging (Brightness-mode imaging).

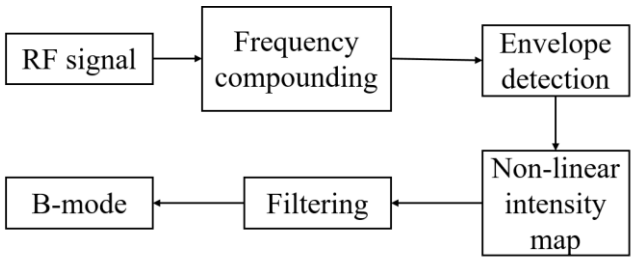


Figure 2-2 Exemplar ultrasound processing pipeline for RF signal to B-mode image conversion.

The temporal resolution of B-mode ultrasound image ranges from 30 frame-per-second (FPS) to 80-90 FPS. Although, recently, there are some progress on ultrafast ultrasound imaging [27], the ultrasound machines typically collect up 30-60 FPS. To get a deeper scan, the ultrasound machine settings will employ slower imaging rates as longer time is needed to wait for the returning pulses.

2.3 Ultrasound-tongue tissue interaction

Since last several decades, B-mode ultrasound imaging has been employed to visualize the motion of the tongue with considerable success [29]. When the ultrasound waves travel through a medium, some effects may occur, which include attenuation, reflection, refraction and scattering. The interaction between the sound waves and the medium is determined by the acoustic properties of the medium and the ultrasound imaging system. In this section, firstly, we analyze the interaction between ultrasound and the tongue tissue. Then we present our data recording equipment used throughout this thesis.

2.3.1 Ultrasound tissue interaction

The quality of an ultrasound tongue image is of utmost importance in determining its usefulness. The overall quality of the ultrasound tongue image is the end product of a combination of many factors originating not only from the imaging system but also from the stability of the recording system and the performance of the operator. All of the components within the ultrasound tongue imaging system, including the transducer, image processing, display and recording devices, impact on the ultimate quality of the ultrasound tongue image. It is necessary to analyze the interaction between the ultrasound and tongue tissue, which can be helpful to make better use of ultrasound imaging.

When we are employing ultrasound to measure the motion of the tongue, the transducer typically is placed under the chin (as can be seen from the Figure 2-3). The ultrasound waves will transmit through the tongue body until it is reflected back.

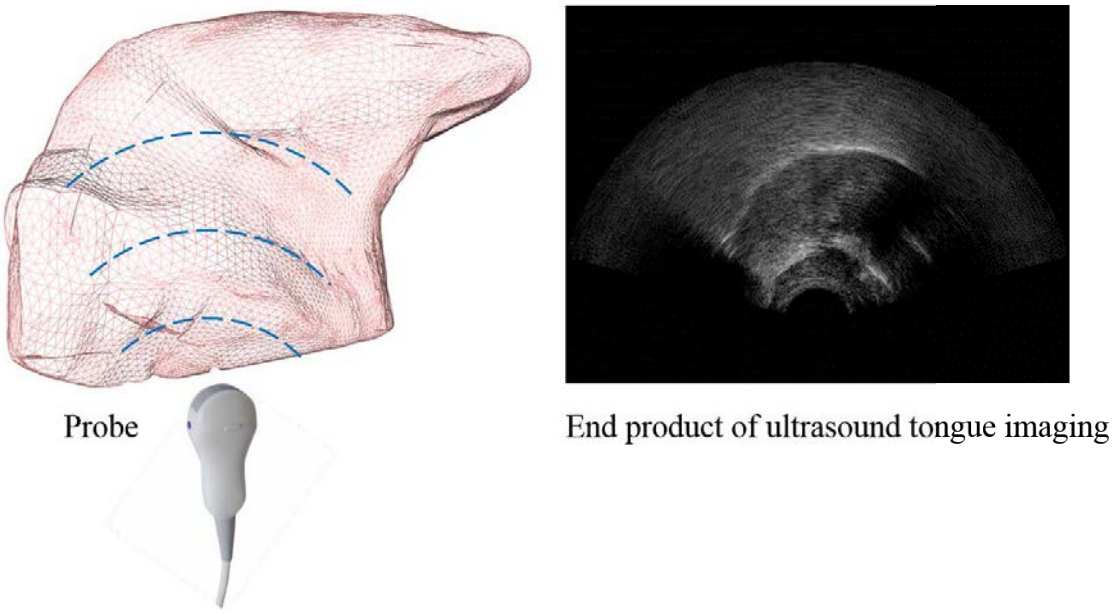


Figure 2-3 Ultrasound-tongue imaging.

When ultrasonic waves travel through the tongue tissues, waves are partly transmitted to deeper structures, partly reflected back to the transducer as echoes, partly scattered, and partly transformed to heat.

Reflection: Reflection of the ultrasound waves occurs when the sound reaches tissue boundaries (or tissue-air-interface) with different acoustic impedance. In the ultrasound tongue image, due to the difference of acoustic impedances between the air and the tissue, the brightest line in the B-mode ultrasound tongue image is the interface between the tongue surface and the air (as shown in Figure 2-3).

Scattering: In ultrasound tongue imaging, scattering refers to the interaction of ultrasound wave with microstructures of the tongue, which are much smaller than the wavelength. When the ultrasound wavelength is greater than the structure in the tongue, it will create a uniform amplitude scattering in all directions, which gives rise to the speckle phenomenon in the ultrasound tongue images (as shown in Figure 2-4). In brief, the “speckle” is the specific spatial distribution pattern of gray values in a local region of ultrasound tongue image. The “speckle” is also referred to as a speckle pattern. As the speckle is originated from the RF signal, the “speckle” can also be used to define the amplitude distribution of the RF signal. To distinguish between them, in the rest of this thesis, the gray-value distribution is referred as “speckle”.

Attenuation: Different tongue tissues have different acoustical properties, which enhance or decrease the propagation of the ultrasound waves. As the tongue contains considerable amounts of fat, refraction may occur and the returning echo is significantly attenuated.

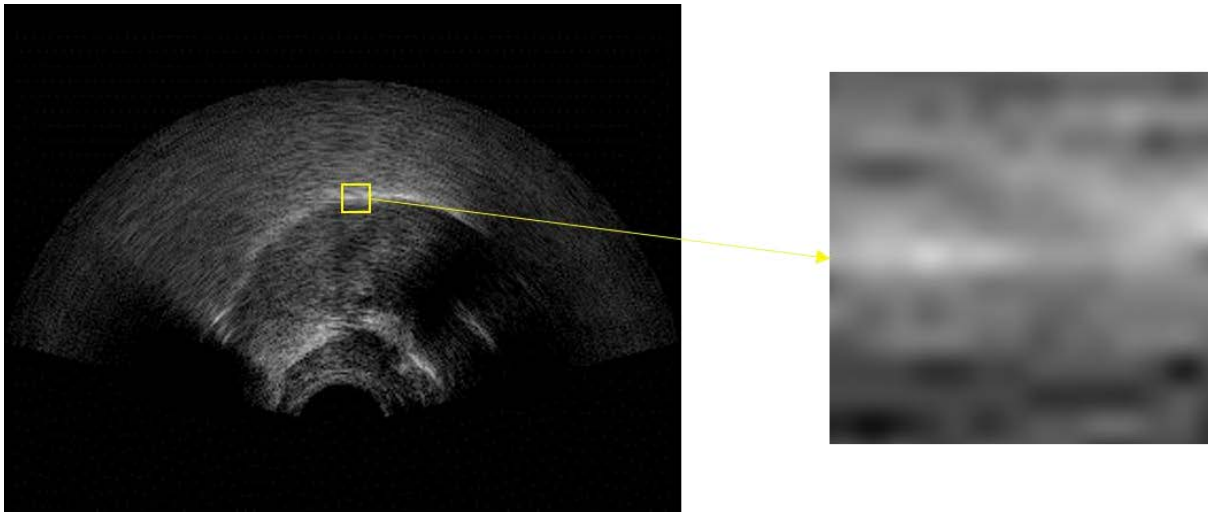


Figure 2-4 A particular segment of tongue results in a specific spatial distribution of gray values, the speckle pattern, in the ultrasound tongue images, which can be used as the acoustic marker of the tongue tissue.

The imaging quality varies between different subjects: such as the image quality of female subject is better than the male subject, while the younger subjects are better than the older subjects [28]. On the other hand, it is worthwhile to note that: the tongue tip is not visible in most cases, which makes the characterization of the tongue's motion more difficult.

2.3.2 Ultrasound tongue image distortions

Due to the sensitivity of the ultrasound imaging quality, there are several image distortions in ultrasound tongue images. In more detail, the distortions include speckle noise contamination, double edges, contour discontinuities, contour invisibility and inconsistent transducer placement and so on. Here, the focus is mainly on the influence of speckle and the contour invisibility.

Speckle noise: Due to the inherent contamination with the speckle noise, the analysis of ultrasound tongue image poses a great challenge. Although there are more attempts to employ this kind of information to follow the motion, speckle noise processing is still an open issue in ultrasound image processing. Indeed, the speckle noise degrades the ultrasound tongue image, concealing the fine structures [36], which leads to the difficulties in the motion tracking of the tongue.

Hidden from the view (or faint contour): When the tongue tissue goes perpendicular to the propagation direction of ultrasound waves, the quality of image is good. While, when the tongue tissue goes parallel to the propagation direction of ultrasound waves, the quality of the image is poorly and missing occurs. Take the /i/ as an example, due to the natural motion of the tongue, part of the tongue is invisible. This kind of invisibility increased the difficulties of ultrasound tongue interpretation.

2.3.3 Ultrasound tongue data acquisition

The ultrasound data acquisition devices used in this thesis belong to our SSI system's data acquisition part, which is comprised of a helmet to hold the ultrasound probe to capture the movement of the tongue at the frame rate of 60 Hz, a VGA, CMOS industrial camera for the lips, a microphone to record acoustic speech signals, an electroglottograph (EGG) to measure and record vocal fold contact movement during speech production. All the different modes of data are recorded synchronously [37].

The lightweight, adjustable helmet (Figure 2-4) is fitted with a micro-convex, 1 inch diameter, 128 element ultrasound probe for tongue imaging [37]. An adjustable platform is used to hold the ultrasound transducer in contact with the skin beneath the chin. The ultrasound machine chosen is the Terason T3000, a system which is lightweight and portable yet retaining high image quality, and allowing data to be directly exported to a PC via Firewire. Our entire SSI system can be placed in a small carrying case, thus enabling everyday applications.

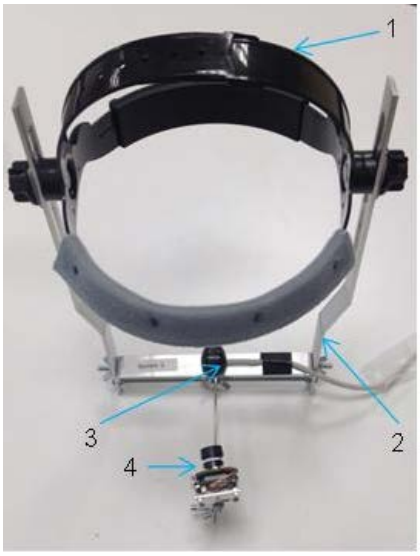


Figure 2-5 Multi-sensor Hyper-Helmet: (1) Adjustable headband. (2) Probe height adjustment strut. (3) Adjustable US probe platform. (4) Lip camera with proximity and orientation adjustment.

2.4 Conclusion

In this chapter, we briefly went through the basic principles of medical ultrasound imaging. Moreover, the ultrasound pulse is discussed, and the ultrasound scan types are presented. In the second part of this chapter, the interaction between ultrasound and tongue is discussed, which include reflection, scattering and attenuation in the ultrasound tongue image sequences. Also, the definition of speckle is given in this part, which is a quite important topic for the next chapter. Based on the interaction between the ultrasound and the tongue, we discussed some distortions in the ultrasound images, such as: contour invisibility, which is a major problem for different previous proposed contour tracking methods. In the end of this chapter, we presented the ultrasound data-recording platform used throughout this thesis.

Chapter 3
Speckle tracking in ultrasound tongue images

3.1 Introduction

In the speech production field, researchers have made sustainable efforts to improve the interpretation of the configurations of the articulators. As the tongue is one of the major components of the vocal tract, the quantification of the tongue's motion is important to understand the change of articulators over time during the natural speech. As the main objective of this thesis is to model the 3D tongue motion in a dynamic manner, robust correspondences between continuous frames need to be established in the ultrasound tongue image sequences, which will be used to drive the 3D physics-based deformable model dynamically. However, compared to other imaging methods such as x-ray, electromagnetic articulography (EMA), the ultrasound imaging technique can provide limited information on the correspondences. Indeed, the nature of the tongue motion made it even harder to obtain correspondences due to several reasons: complex tongue shape, hidden from the view, high-level speckle noise contamination and out-of-plane motion.

To address the aforementioned difficulties, different approaches can be used in the ultrasound image sequences to obtain point correspondences, one of which is speckle tracking. Speckle tracking has been an active field in ultrasound imaging for the last several years [38]. Specifically, with the advance of ultrasound image quality and the calculation power of computer, the speckle tracking approach matures into practical research to follow the motion. In the cardiology field, speckle-tracking echocardiography (STE) [39] is used to characterize the motion of tissues in the heart, and the utilities of STE are increasing recognized. In the clinical setting, more and more evidence was shown that the assessment of the cardiac deformation by speckle tracking techniques can provide incremental information.

However, compared to the progress in the cardiac motion estimation, only few attempts have been made to explore the potential application of speckle tracking on the ultrasound tongue images. To the best knowledge of us, only [40] explored speckle tracking as a method to obtain the point correspondences in ultrasound tongue images. In more detail, [40] proposed a deformation registration-based method to obtain point correspondences in order to estimate the displacement in the ultrasound tongue images. This method enables us to calculate the motion and hence obtain the displacements from first frame to any frame in the image sequences. Moreover, the speckle tracking method can be used to estimate the displacement along the curve, which may be helpful for us to obtain "virtual flesh point markers" on the tongue surface [40]. This kind of information is of great interest for the linguist and clinical phonetics. Moreover, by following the motion of the "virtual flesh point markers", 3D dynamic tongue modeling may be feasible.

This chapter aims to explore the speckle tracking method to obtain point correspondences further. In more detail, the organization of this chapter is given as follows: in Section 3.2, based on the imaging principles of ultrasound imaging, we make a summary of the fundamental principles of speckle tracking firstly. Afterwards, we present the applications of the different speckle tracking implementation techniques which include deformation registration, optical flow and local invariant feature-based method. A comparative study is conducted on the different speckle tracking approaches. Moreover, as demonstrated in the comparative study, speckle tracking is not always stable and tracking failure occurs during the tracking processing. To address the tracking error accumulation, an image similarity-based automatic speckle tracking re-initialization method is proposed in the Section 3.3, which aims to recover the tracking from the failure automatically. A conclusion is drawn in Section 3.4.

3.2 Speckle tracking in ultrasound tongue images

Two important categories of speckle tracking approaches have been used widely during the last decades, one is based on deformation registration, and the other is optical flow. Recently, the local invariant feature has been applied into tracking problem successfully. In this section, we first go through the fundamentals of speckle tracking, followed by the

implementation detail of different speckle tracking methods in the ultrasound tongue images. Then, a comparative study is conducted on tracking performance using different techniques.

3.2.1 Fundamentals of speckle tracking

The fundamental principle of 2D speckle tracking is quite simple: if the imaging system is of a sufficiently high frame rate (as aforementioned in previous chapter, in our data-recording platform, the Terason T3000 here is clocked at 60 FPS), the speckle patterns are preserved between subsequent image frames. Tracking characteristic speckle patterns can thus provide markers for tagging the soft-tissue motion. If the position of a segment of the tongue tissue changes, one may assume that, the position of its acoustic fingerprint (texture) will change accordingly. By tracking these patterns, we can follow the motions of tongue tissue in real-time.

Thus, before exploring the speckle tracking method further, we would like to discuss whether the assumption is reasonable or not, which are given in the followings. As presented in Chapter 2, ultrasound imaging is based upon the pulse-echo. The ultrasound reflections occur at tissue-air interface and the interface between different types of the tongue tissues. Moreover, as mentioned in Chapter 2, the scattering will also occur when the size of the local structures of the tongue (e.g., blood-muscle) is much smaller than the wavelength of ultrasound during the propagation. As tongue tissues may contain many scattering sites, the signal detected by the transducer is the interference of the individual reflections occurring at the individual scatters. Each scatter will reflect the incident wave but the amplitudes of the waves are relative low. As the distance between the transducer and the scatters is different, the reflected ultrasound waves will also be received at a different time. Thus, constructive interference and destructive interference occur occasionally. If the constructive interference occurs, a high-amplitude RF signal will be detected, while if destructive interference occurs, a low-amplitude RF signal will be detected [41].

The received RF signals will be converted to B-mode ultrasound tongue image. The speckles (spatial distribution of the gray values) are associated with the reflections from individual scatters within the tongue, which carry information about the unresolvable scattering structures. In other words, the characteristics of the speckle are determined by the positions of the exact scatter. Thus, the motion information and deformation of the tongue can be extracted by following the corresponding gray values distributions. If the segment of

the tongue is showing a scatter signal moving away from the transducer, the reflected waves will arrive later due to the increase of traveling time. However, an identical signal with time-delayed will be detected, and the same distribution of the gray values will occur at a position further away from the transducer. Thus, tracking the motion of the "speckle" does indeed follow the motion of the underlying tissue [41].

Based on the previous reasoning, it can be seen that speckle can only be preserved if the interference between the individual scatters in the tongue remain identical, which means the relative position between the transducer and the scatterings' sites should be the same. However, in practical applications, out-plane motion and deformations of the tongue will change the relative position of the particular tongue tissue and the transducer, which will induce speckle-decorrelation. To alleviate the speckle decorrelation, high imaging framerate is needed to limit the amount of the out-of-plane and deformation between adjacent frames [41].

3.2.2 Deformation registration

Non-rigid deformation registration methods are often applied to estimate the correspondences between landmarks in adjacent frames. With this kind of method, the correspondence between the tissue points is obtained by minimizing a similarity measure. The local similarity measure is usually based on the L^1 or L^2 norms comparison on the intensities of image blocks, such norms are appropriate for the image sequences, which are characterized by Gaussian statistics. In [42], Kontogeorgatis et al. proposed to use the minimum absolute difference (MAD) to estimate the motion of the local blocks in the ultrasound images. In [43], Yeung et al. suggested using the sum of absolute differences (SAD) criteria, and the single scale motion estimation was expanded to multi-scales. While, [44] made a comparative study on different similarity indices and claimed that sum of squared differences provided superior performance in low-level noise contamination. In [45], the author suggested to use the normalized correlation coefficients (NCC) for block matching.

However, the ultrasound image is contaminated by the multiplicative speckle noise, and the traditional approaches may be unsuitable in this case, as these methods assumed that the imaging system is contaminated by the Gaussian noise. Commonly, researchers proposed to

replace the assumptions of a Gaussian distribution by ultrasound-specific noise models, and a Rayleigh distribution is assumed.

Based on this assumption, Strintzis and Kokkinidis [46] used maximum likelihood (ML) to estimate the motion in the ultrasound images. In the method proposed in [46], it was assumed that only one image contains speckle noise while the reference image is not contaminated by the noise. In [47], Cohen and Dinstein proposed a novel similarity measure using the maximum likelihood method based on the assumption that two consecutive ultrasound images are both corrupted by multiplicative Rayleigh noise and the probability density functions (PDF) of the noise are independent of each other.

Suppose $x_{i,j}$ and $y_{i,j}$ is the intensity of every pixel in the two consecutive frames of ultrasound image, where i and j denote the position of the pixel in the image X and Y. Assume that the displacement between the two pixels matched is $v_{i,j}$.

$$v_{i,j}^{ML} = \arg\max p\left(x_{i,j} \middle| y_{i,j}, v_{i,j}\right) \tag{3.1}$$

where the conditional probability density function depends on the noise model. In ultrasound image, the probability density function of the noise model is multiplicative Rayleigh. If the noiseless pixel is denoted by $s_{i,j}$, the following model for the observed pixel in X and Y stands:

$$x_{i,j} = \phi_{i,j}^{1} s_{i,j} \tag{3.2}$$

$$y_{i,j} = \phi_{i,j}^{2} s_{i,j} \tag{3.3}$$

where $\phi_{i,j}^{1}$ and $\phi_{i,j}^{2}$ are two independent noise element with a Rayleigh density function given by:

$$p_{\phi}^{1}\left(\phi^{1}\right) = \frac{\phi^{1}}{\lambda_1} \exp\left\{-\frac{\left(\phi^{1}\right)^{2}}{2\lambda_1^{2}}\right\}, \quad \lambda_1 > 0 \tag{3.4}$$

$$p_{\phi}^{2}\left(\phi^{2}\right) = \frac{\phi^{2}}{\lambda_2} \exp\left\{-\frac{\left(\phi^{2}\right)^{2}}{2\lambda_2^{2}}\right\}, \quad \lambda_2 > 0 \tag{3.5}$$

Based on the Eq. (3.2), (3.3), (3.4) and (3.5), we can re-define the noise probability density function as follow:

$$p_\eta = 2\frac{\lambda_1^2}{\lambda_2^2}\frac{\eta}{\left(\eta^2+\left(\frac{\lambda_1}{\lambda_2}\right)^2\right)^2}, \quad \eta>0 \tag{3.6}$$

where $\eta_{i,j} = \frac{\lambda_{i,j}^1}{\lambda_{i,j}^2}$. Then we can obtain the following equation:

$$y_{i,j} = \eta_{i,j} x_{i,j} \tag{3.7}$$

Taking the natural logarithm of both sides of (3.7), we can obtain the following model for the pixel in the ultrasound images:

$$\tilde{y}_{i,j} = \tilde{x}_{i,j} + \tilde{\eta}_{i,j} \tag{3.8}$$

From Eq. (3.8), the original Rayleigh noise is transformed to an additive noise, whose probability density function is given as follows:

$$p(\tilde{\eta}) = 2\frac{\lambda_1^2}{\lambda_2}\frac{\exp(2\tilde{\eta})}{\left(\exp(2\tilde{\eta})+\left(\frac{\lambda_1}{\lambda_2}\right)^2\right)^2} \tag{3.9}$$

We suppose the independent noise of the successive frame follow the same distribution ($\lambda_1 = \lambda_2$). And the maximization of this equation (Eq. (3.10)) is equivalent to the Eq. (3.1). The objective function to measure the similarity of pixels is:

$$E(v_{i.j}) = \left\{\tilde{y}_{i,j} - \tilde{x}_{i,j} - \ln\left(\exp\left(\tilde{y}_{i,j} - \tilde{x}_{i,j}\right)\right) + 1\right\} \tag{3.10}$$

Follow the notation in [47], the motion estimator given in Eq. (3.10) is donated by CD2.

In Figure 3-1, we use the deformation registration schema to estimate the deformation between consecutive frames and sample results are given here. The goal of this method is to determine the deformation map between continuous frames based on the local similarity measurement. The cubic B-spline function is used to model the deformation field. In [40], the author proposed to use sum of square difference (SSD) to measure the difference between the

deformed previous frame and the next frame. We conducted our experiments by using both SSD and CD2.

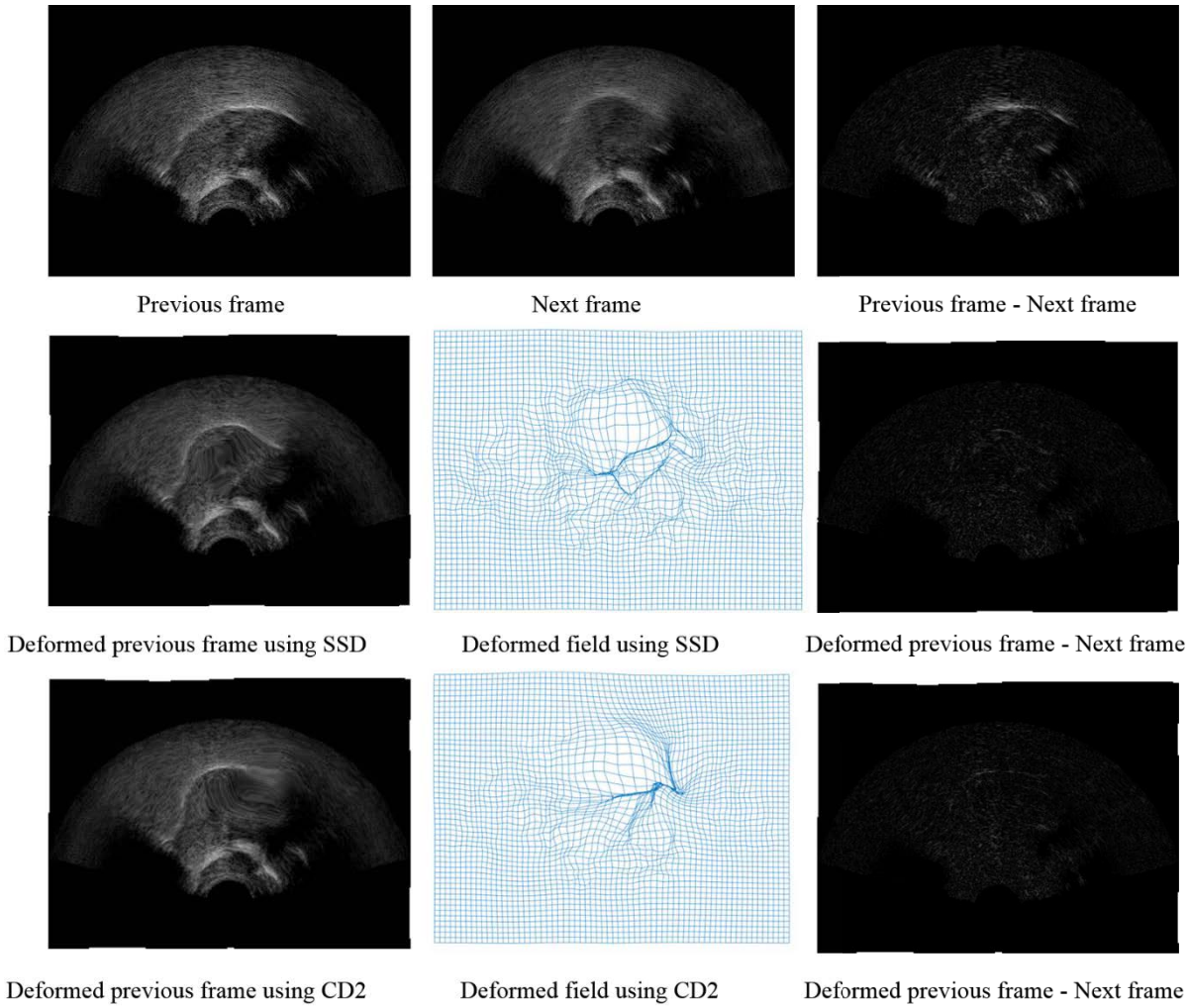


Figure 3-1 Deformation registration of ultrasound tongue images.

Based on visual observation, both SSD and CD2-based deformation registration work well even under large deformation, as no structures is visible in the difference images. Note, to highlight the difference, the time interval between the previous frame and the next frame is expanded while the experiment is conducted. By default, the time-interval should be identical to the imaging framerate of ultrasound machines (1/60s), while, the time interval is set to 1 second here.

In the practical application of speckle tracking in ultrasound tongue images, the virtual markers on the tongue surface are of great interest for the researchers. Figure 3-2 is given to illustrate the utility of the deformation registration in tracking “tissue points”. The original tissue points are manually selected in the previous frame. Before the deformation registration, the location of the tissue points is shown in the next frame. After the registration, the tracked markers were shown in the deformed previous frame. As can be seen from the figure, all the tracked markers lie near the tongue surface, which demonstrates the feasibility of the deformation registration-based speckle tracking.

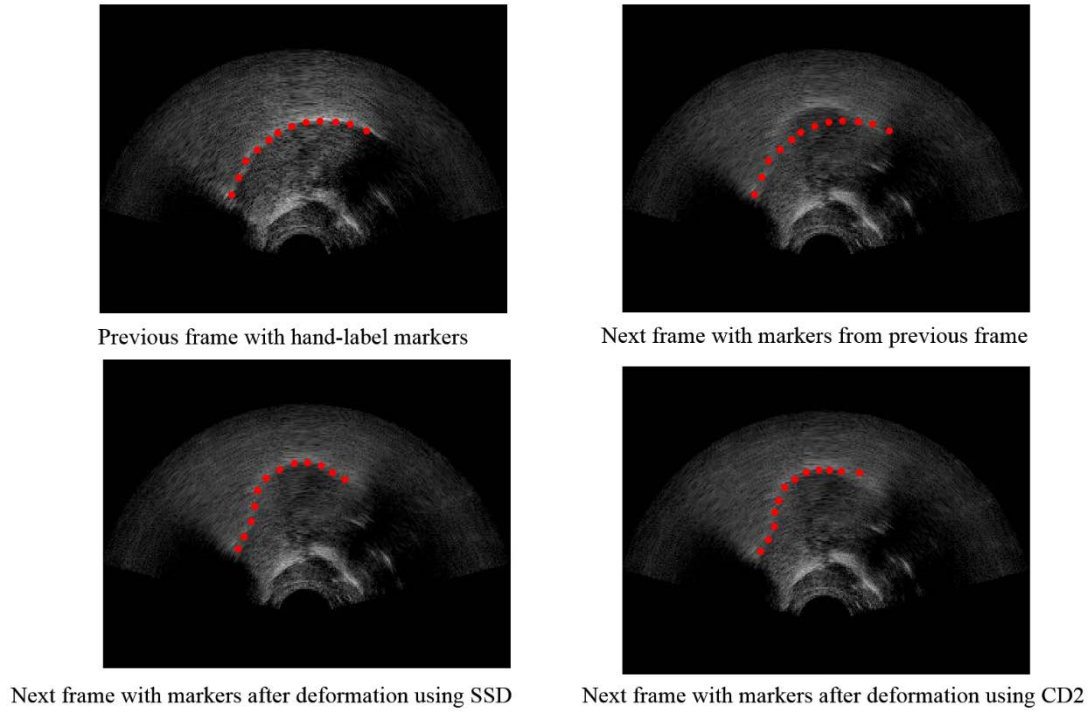


Figure 3-2 Deformation registration-based virtual land markers tracking in ultrasound tongue images.

In [40], the author proposed a qualitative analysis of deformation registration in the prosodic domain, In fact, a more comprehensive evaluation is needed to quantitatively assess the performance of deformation registration method, which will be given below after the general introduction of other speckle tracking methods.

3.2.3 Optical flow

Optical flow is another widely used motion tracking method in the video. In order to compute the optical flow between two adjacent ultrasound tongue images, the following optical flow constraint equation need to be solved:

$$I_x u + I_y v + I_t = 0 \tag{3.11}$$

where I_x, I_y, and I_t are the spatiotemporal image brightness derivatives. u is optical flow in the horizontal direction, while v is the vertical optical flow. Several different methods have been proposed to solve Eq. (3.11), such as Horn-Schunck method [48] and Lucas-Kanade [49] method. Here, we take the Lucas-Kanade method: to solve the optical flow constraint equation for u and v. The Lucas-Kanade optical flow method sub-divides the original image into smaller sections and assumes that velocity in each section is a constant. Then, this method performs a weighted least-square fit of the constraint equation to a constant model for $\begin{bmatrix} u & v \end{bmatrix}^T$ in each sub-region Ω. By minimizing the following equation, the fitting can be achieved.

$$\sum_{x \in \Omega} W^2 \left[I_x u + I_y v + I_t \right]^2 \tag{3.12}$$

where W is a window function. Then, we can formulate this minimization problem as given follows:

$$\begin{bmatrix} \sum W^2 I_x^2 & \sum W^2 I_x I_y \\ \sum W^2 I_y I_x & \sum W^2 I_y^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum W^2 I_x I_t \\ \sum W^2 I_y I_t \end{bmatrix} \tag{3.13}$$

In the Lucas-Kanade method, a difference filter is used to calculate I_t. Then, for each pixel, we can solve the optical flow constraint equation for u and v by solving 2-by-2 linear equations using the following approach:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \sum W^2 I_x^2 & \sum W^2 I_x I_y \\ \sum W^2 I_y I_x & \sum W^2 I_y^2 \end{bmatrix} \tag{3.14}$$

Then we can make use of the eigenvalues of A for optical flow estimation. Suppose the eigenvalues of A can be represented by λ_1 and λ_2. If both λ_1 and λ_2 are bigger than a set threshold, the equation can be solved directly by using Cramer's rule as A is non-singular. If

A is singular (one of λ_1 and λ_2 is smaller than the threshold), the gradient flow should be normalized to calculate u and v . If both eigenvalues are smaller than the threshold, the u and v will be set as 0. An example of optical-flow-based speckle tracking experiment is given in Figure 3-3.

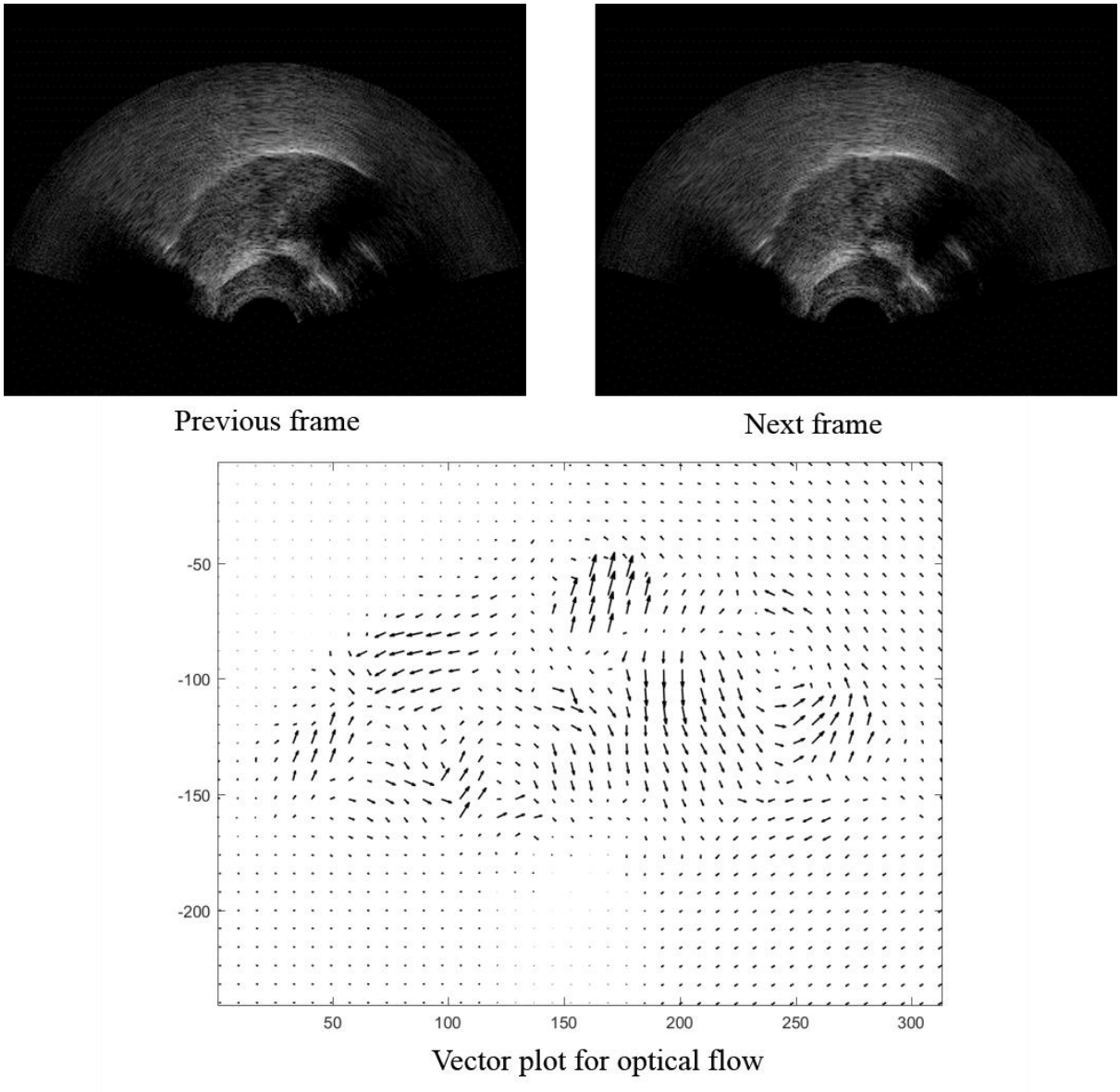


Figure 3-3 Application of optical-flow in ultrasound tongue images.

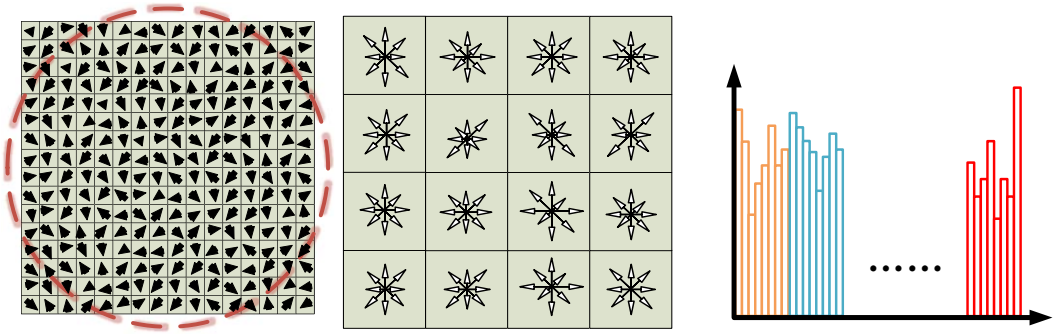
As can be seen on the figure, optical flow can capture some small motion from the ultrasound tongue images. However, the correctness of the motion vector calculated still

needs to be further explored, as there is no ground truth for evaluation. Similar to deformation registration, optical-flow can also be used to follow the tissue points on the surface of the tongue, whose performance can be evaluated. The detail evaluation and comparison will be given in the section 3.2.5.

3.2.4 Local invariant feature

Since last decade, local invariant feature has drawn many attentions due to its superior performance on several computer vision tasks, such as object tracking. Despite the numerous publications of studies on the image feature descriptors, the descriptors representing the distribution of small-scale features within the interest point neighborhood [50], which was introduced by Lowe, have shown robust performance to obtain points correspondences.

In this part, we take the Scale-Invariant Feature Transform (SIFT) algorithm [50] as an example, a state-of-the-art method of local invariant feature description, which can be divided into three main steps: feature detection; feature description and feature match. Here, feature description corresponds to speckle pattern description, while the feature match can be regarded as a speckle pattern similarity measure. More specifically, SIFT keypoints are detected as local extrema in scale-space, and keypoints are assigned to one or more orientations based on local image gradient directions. Each keypoint is described as a vector using neighborhood oriented gradient histogram information. A 16x16 neighborhood around the keypoint is taken. The block is divided into 16 sub-blocks of 4x4 size. For each sub-block, 8 bin orientation histogram is created. So a total of 128 bin values are available. It is represented as a vector to form keypoint descriptor. In addition to this, several measures are taken to achieve robustness against illumination changes, rotation. In the original step of feature matching, the Euclidean distances between vectors are calculated to obtain correspondences of the keypoint.



(a) Oriented Gradient (b) Sub-blocks (c) Descriptor feature vector

Figure 3-4 Describe the keypoints by using the feature vector.

Despite the excellent performance of SIFT algorithms, only sparse correspondence can be obtained after the matching between the descriptors in the ultrasound image due to the poor quality (as can be seen from Figure 3-5). As mentioned earlier, the goal of speckle tracking is to analyze the motion of the tongue, especially the tissue points lying on the surface of the tongue. Thus, this kind of approach may not be feasible for our purpose.

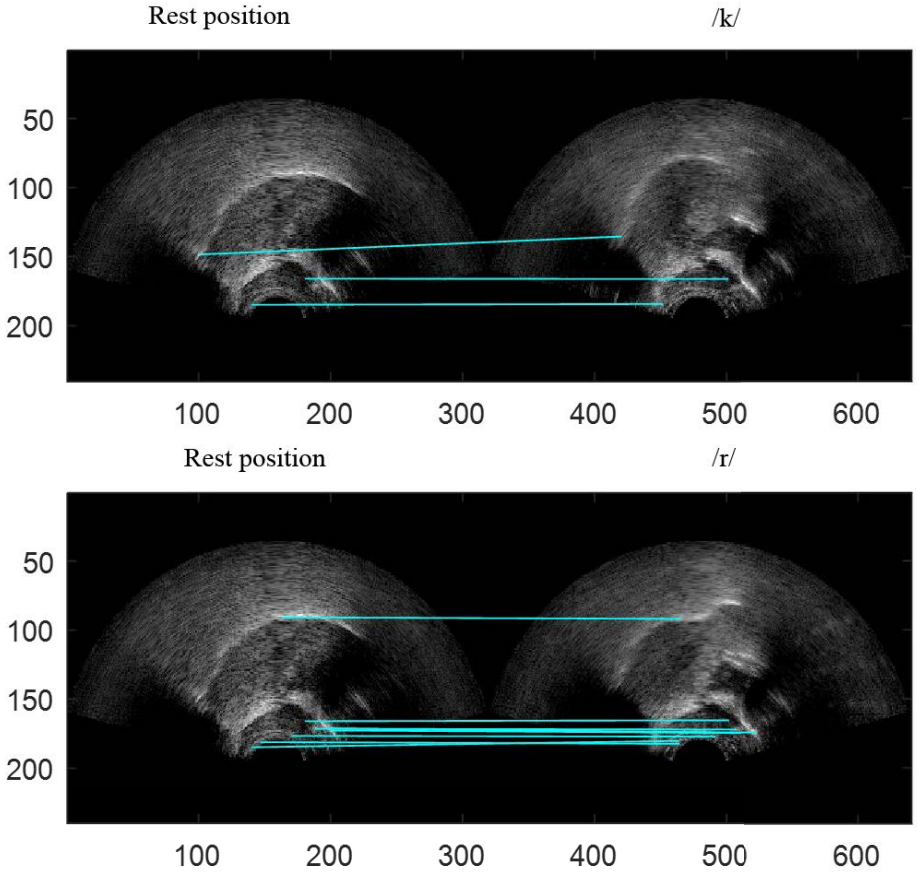


Figure 3-5 Examples of local invariant feature to obtain point correpondences.

Indeed, dense correspondence between two ultrasound tongue images is of a great need to characterize the deformation of the tongue. Analogous to optical flow, where an image is aligned to its temporally adjacent frame, [51] proposed SIFT flow to obtain dense point correspondences using SIFT descriptors. Here, we use SIFT flow, to build the dense correspondence between descriptors to estimate the motion of particular pattern.

The initial objective of SIFT flow is to align the query image in the retrieved set which consists of a large collection of a variety of scenes. If the dataset is large enough to cover all the possible scenes, the nearest neighbors would be visually similar to the query image. Based on the hypothetical scenario, a SIFT descriptor is extracted at each pixel in the image and encoded the pattern information, and optical flow approach is used to build the correspondences of the descriptors in the two frames. The use of SIFT descriptor can allow robust matching.

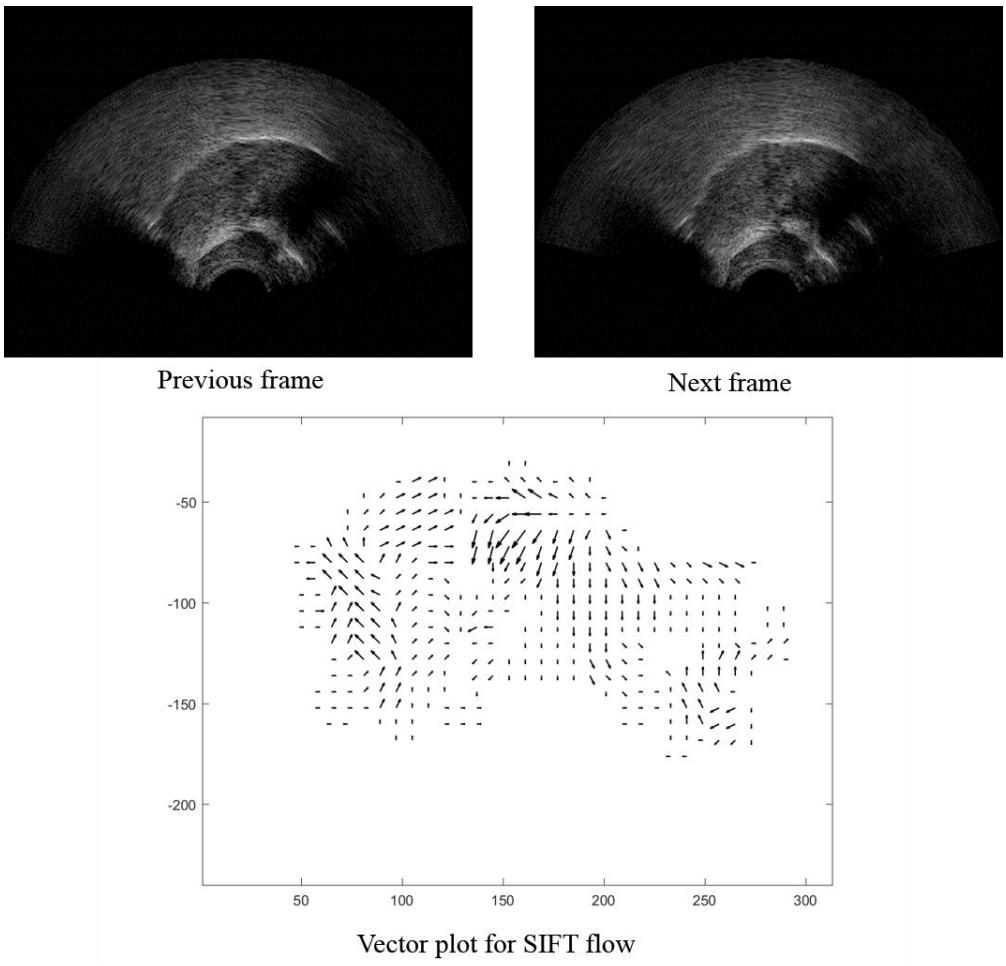


Figure 3-6 Application of SIFT flow on the ultrasound tongue images.

Based on the SIFT flow, some sample results are given in Figure 3-6. As can be seen from the figure, similar to the optical flow, the SIFT can capture some small motions in the ultrasound tongue image sequences. On the other hand, it can be observed from the figure that: the SIFT flow is more robust to the noise. Compared with the standard optical flow, SIFT works better (note the motion vectors in the black region in the ultrasound tongue images). The quantitative evaluation of SIFT flow-based speckle tracking method is given in next section.

3.2.5 Comparison between different speckle tracking methods in ultrasound tongue images

In the ultrasound tongue images, there is no ground truth to evaluate the performance of different speckle tracking methods, a specific evaluation method is to be designed. As the researchers are more interested in the upper surface of the tongue, an experiment has been carried out to compare the tracked speckles on the contour to hand-labeled curves. The idea behind this kind of evaluation is that: different speckle tracking method can be used to follow several "tissue point" on the tongue contour, and the virtual contour can be compared to with the hand-labeled curves. If the distance between the tracked contour and the labeled contour is small, the speckle tracking method is better in following the surface deformation.

For validation, we follow the widely used: mean sum of distances (MSD), defined as:

$$\mathrm{MSD}(\mathbf{U},\mathbf{V}) = \frac{1}{2n}\left(\sum_{i=1}^{n}\min\left|\mathrm{v}_i - \mathrm{u}_j\right| + \sum_{i=1}^{n}\min\left|\mathrm{u}_i - \mathrm{v}_j\right|\right) \tag{3.15}$$

where **V** is the contour extracted automatically and **U** is the result of hand-labelling. Note once again the points on the compared contours are not physical tissue points but simply representative positions on the contour; MSD simply compares the similarity of two contours *i* and *j* by finding, for each point on contour *i*, the closest point to it on contour *j*, which, in general, will *not* correspond to a reference point on contour *j*. The comparison is conducted on a consecutive sequence, of 150 frames. The data is the recording from the female subject, and all the frames are hand-labeled. The MSD error for different approaches on the continuous sequences is given in the figure below.

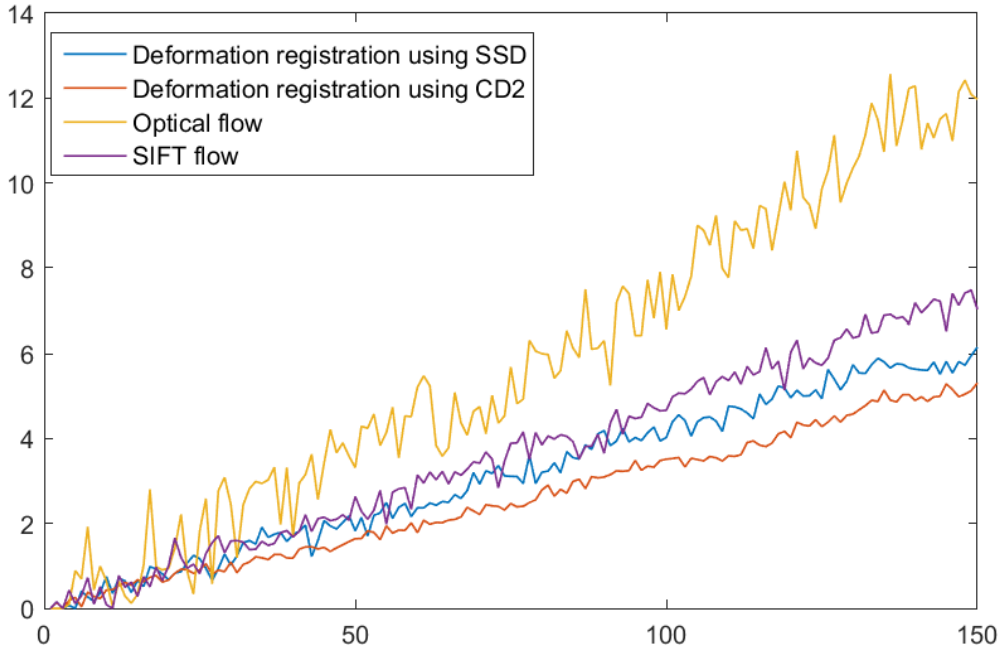


Figure 3-7 The MSD error of different speckle tracking methods.

As can be seen from the figure, except for the optical-flow method, the speckle tracking methods provide similar MSD error. Deformation registration-based speckle tracking method gives better performance with comparison to SIFT flow, and CD2-based local similarity measurement gives a slightly better result obtained using the SSD. Based on the experiments, we can conclude that both deformation registration and the SIFT-flow-based speckle tracking methods can provide extra point correspondence information on the tongue surface with relative high robustness. On the other hand, optical flow or its variants may be unsuitable for the ultrasound tongue image due to the speckle decorrelation. Moreover, the deformation of the tongue is large and the motion of the tongue is quite fast, which poses an even greater challenge to follow the motion of the tongue by using the optical flow as standard optical flow cannot handle with large deformation. On the other hand, it is worthwhile to notice that the tracking error was accumulated during the tracking processing. If hand-refinement is done for the speckle tracking method, the performance can be further improved.

3.3 Similarity-based automatic speckle tracking re-initialization

As demonstrated in previous sections, the tracking error was accumulated on the whole sequence. Thus, hand-refinement is often needed to reset the tracking [10]. The goal of this section is to explore this issue: is there a way to re-initialize the speckle tracking automatically, thus improving the robustness of the speckle tracking method. Unlike the general tracking problem in the natural video sequence (such as pedestrian detection), this similarity-based automatic reset method may be feasible in the ultrasound tongue image sequences as the tongue's motion trajectory is limited. For example, the tongue goes back to the rest position frequently during the speech production. The key idea of automatic re-initialization is the following (as shown in Algorithm 1): before speckle tracking is carried out, the image similarity coefficient (defined below) is calculated, between current frame and the hand-labeled frames (note that the number of frames can be more than 1). If this coefficient exceeds a set threshold, the points of the contour are reset to those which were input for the hand-labeled frames. This provides a method to prevent accumulation of errors over long sequences, which can lead to erroneous tracking, and amounts to a sort of "automatic re-initialization" of the tracking points based on initial a priori information.

Algorithm 3-1: Tracking re-initialization method for the speckle tracking in the ultrasound tongue image sequences.

```
1: load  labeled frames L (L1, L2, ...)
2: for  frame-number F in [1, T] do
3:     load Fn (n = 1, T)
4:         if similarity index (Fn, L) > threshold
5:             Reset the contour to the hand-label contour
6:         else
7:             do Speckle Tracking
8:         end if
9: end for
```

In this section, we first introduce the image-similarity measurement method, and a comparative study is made on ultrasound tongue images using different image similarity indexes. Then, by making using of the more suited similarity index, the automatic speckle tracking re-initialization method is applied. To demonstrate the feasibility and robustness of the proposed method, the re-initialization method is incorporated into aforementioned speckle

tracking methods, and a quantitative comparison is made with and without automatic re-initialization. The results demonstrate that, using the proposed method can improve the performance by reducing the MSD error.

3.3.1 Ultrasound image similarity measurement

Before digging into the automatic speckle tracking re-initialization, it is desirable to explore a robust and accurate similarity index to measure the similarity between the current frame and the hand-label reference frames.

The simplest and most widely used similarity measure is the mean squared error (MSE), which calculates the mean of squared intensity differences between the input and reference image pixels. Let $x_i, i = 1,2,\cdots,N$ and $y_i, i = 1,2,\cdots,N$ be the intensity values of gray images $\mathbf{X}$ and $\mathbf{Y}$ respectively. We can calculate MSE as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 \tag{3.16}$$

Based on the MSE, the signal-to-noise ratio is another index, which can be used to measure the similarity.

$$PSNR = 10\log_{10}\left(peakval^2 / MSE\right) \tag{3.17}$$

where *peakval* is either specified by the user or taken from the range of the image datatype (the default value is set as 255). However, ultrasound images are influenced by speckle noise, which makes these methods unfeasible.

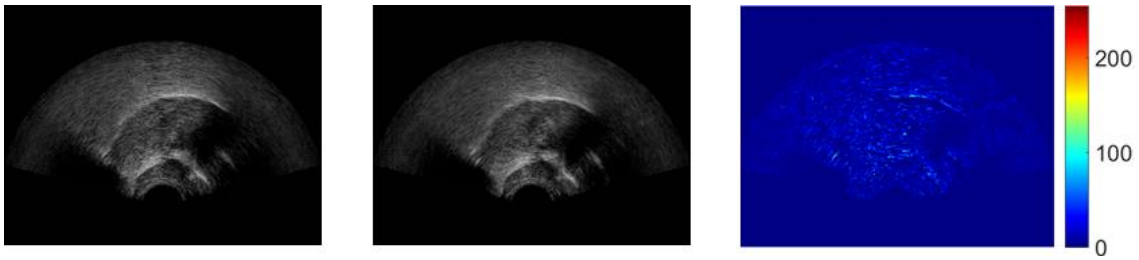


Figure 3-8 Two frames used to calculate the similarity index, the size of the frame is 320 × 240. (c) is the difference between Frame 1 and Frame 46. (a) Frame number: 1; (b) Frame number: 46; (c) Frame 46 - Frame 1 (colormap).

As the milestone of image similarity measures and image quality measures, the structural similarity [52], makes significant progress compared to the previous methods.

The SSIM index measures three kinds of visual impact of changes in luminance, contrast and structure between two images:

$$l(x,\mathrm{y}) = \left(2\mu_x\mu_y + C_1\right) \Big/ \left(\mu_x^2 + \mu_y^2 + C_1\right) \tag{3.18}$$

$$c(x,y) = \left(2\sigma_x\sigma_y + C_2\right) \Big/ \left(\sigma_x^2 + \sigma_y^2 + C_2\right) \tag{3.19}$$

$$s(x,y) = \left(\sigma_{xy} + C_3\right) \Big/ \left(\sigma_x\sigma_y + C_3\right) \tag{3.20}$$

where μ_x, μ_y, σ_x, σ_y, and σ_{xy} are the local means, standard deviations and cross-covariance for adjacent two frames x and y, and C_1, C_2 and C_3 are the constants. At each coordinate, the SSIM index is calculated within a local window. In our experiment, 11×11 circular-symmetric Gaussian weighting function, with standard deviation of 1.5 pixels, is normalized to sum to unity ($\sum_{i=1}^{N} w_i = 1$). The statistics, μ_x, μ_y, σ_x, σ_y, and σ_{xy} are then redefined as:

$$\mu_x = (1/N)\sum_{i=1}^{N} w_i x_i \tag{3.21}$$

$$\mu_y = (1/N)\sum_{i=1}^{N} w_i y_i \tag{3.22}$$

$$\sigma_x^2 = \left(1/(N-1)\right)\sum_{i=1}^{N} w_i \left(x_i - \mu_x\right)^2 \tag{3.23}$$

$$\sigma_y^2 = \left(1/(N-1)\right)\sum_{i=1}^{N} w_i \left(y_i - \mu_y\right)^2 \tag{3.24}$$

$$\sigma_{xy} = \left(1/(N-1)\right)\sum_{i=1}^{N} w_i \left(x_i - \mu_x\right)\left(y_i - \mu_y\right) \tag{3.25}$$

where x_i is the intensity for the pixel i in x while and y_i is the intensity value of pixel i in y. The overall index of similarity is a multiplicative combination of the three terms.

$$SSIM(x,y)=\left[l(x,y)\right]^{\alpha}\left[c(x,y)\right]^{\beta}\left[s(x,y)\right]^{\gamma} \tag{3.26}$$

If we set $\alpha=\beta=\gamma$ and $C_3=C_2/2$, the index can be simplified as follows:

$$SSIM(x,y)=\left(\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)\right)\Big/\left(\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)\right) \tag{3.27}$$

However, the SSIM index is very sensitive to the noise and the image distortions. During our recordings, as the tongue may not have the exact same appearance as in the previous position, there may be small distortions between the images of the same utterances, hence the SSIM's performance may be not stable for our case (as demonstrated in the later experiments). Complex wavelet structural similarity (CW-SSIM; [53]) is an extension of the SSIM method to the complex wavelet domain, which is a novel image similarity measurement and robust to small distortions.

To implement the CW-SSIM index for the comparison, the images are decomposed using a complex version of a multi-scale, multi-orientation steerable pyramid decomposition [54]. In more detail, to compute the CW-SSIM similarity between two ultrasound tongue images, suppose we can represent the complex wavelet coefficients of the two frames (x and y) by using $W_x=\left\{w_{x,l}\,|\,l=1,...,L\right\}$ and $W_y=\left\{w_{y,l}\,|\,l=1,...,L\right\}$, which are extracted at the same spatial location in the same wavelet sub-bands of the two images being compared, L is the level of decomposition (In our experiment, L is set as 5). Then we calculate the complex transform of them. The CW-SSIM similarity index between x and y is:

$$S(x,y)=\frac{2\left|\sum_{l=1}^{L}w_{x,l}w_{y,l}^{*}\right|+K}{\sum_{l=1}^{L}\left|w_{x,l}\right|^2+\sum_{l=1}^{L}\left|w_{y,l}\right|^2+K} \tag{3.28}$$

where w^* is the complex conjugate of w and K is a small positive stabilizing constant. Both SSIM index value and CW-SSIM index value range from 0 to 1, and 1 means the contents in two images compared are the same.

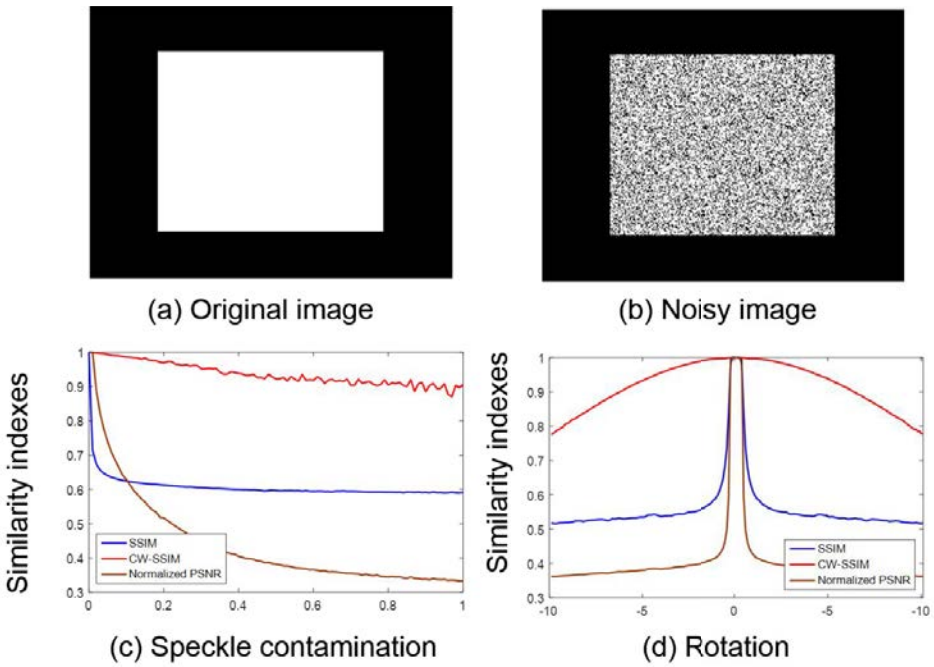


Figure 3-9 A quantitative comparison experiment was conducted on different similarity indexes using the synthetic image (the unit for the rotation is degree).

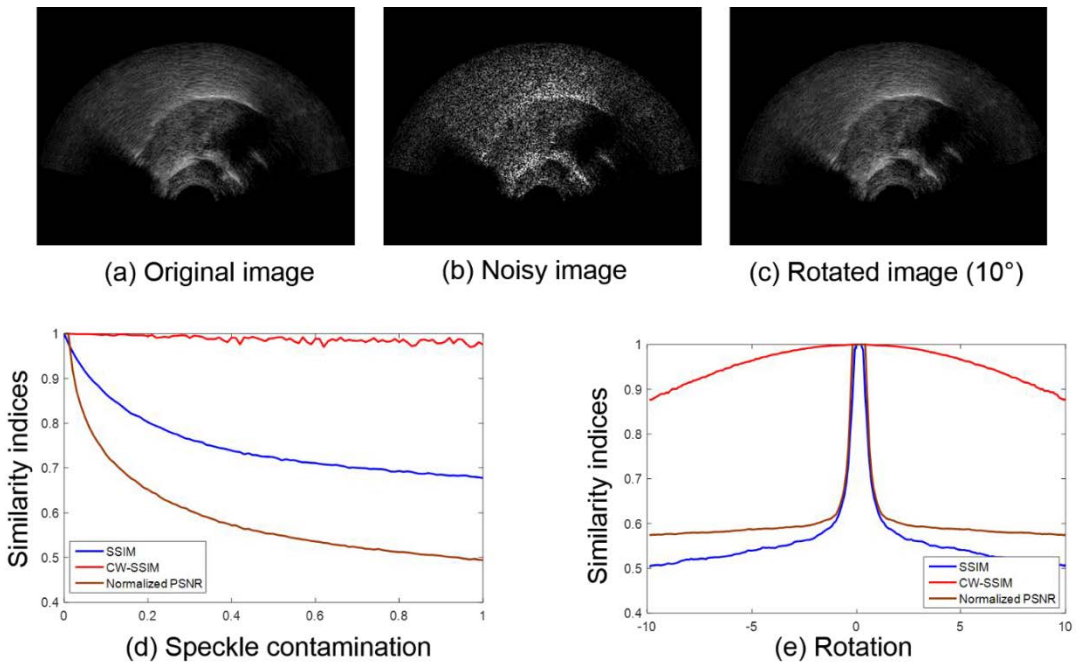


Figure 3-10 Comparison between different similarity indices on different situations using real ultrasound tongue images.

A quantitative comparison experiment was conducted on different similarity indexes (as shown in Figure 3-9 by using synthetic data. The original image in (a) is a white rectangle surrounded by a thick black bounding box, while the sample frame with added speckle noise is given in (b), and the change of the similarity index in different situations are given in (c) and (d). As can be seen from the Figure 3-9, compared with SSIM and PSNR, CW-SSIM has superior performance with the occurrence of the speckle noise and rotation about the center of the image. Note that we normalized MSE values and PSNR values here (by dividing the biggest value of the similarity index); as PSNR can be obtained from MSE, only normalized PSNR is used in our figures. Similar to the experiment given in Figure 3-9, another experiment is conducted by adding the noise and rotation to real-ultrasound tongue images, the results are given in Figure 3-10, which also demonstrates that CW-SSIM gives better performance.

The experiment given in Figure 3-11, is conducted on the sensitivity of the dissimilarity measurement. And the curves of the similarity index is given in the figure. In this thesis, the CW-SSIM will be used to explore the potential applications of the similarity measurement in automatic speckle tracking re-initialization.

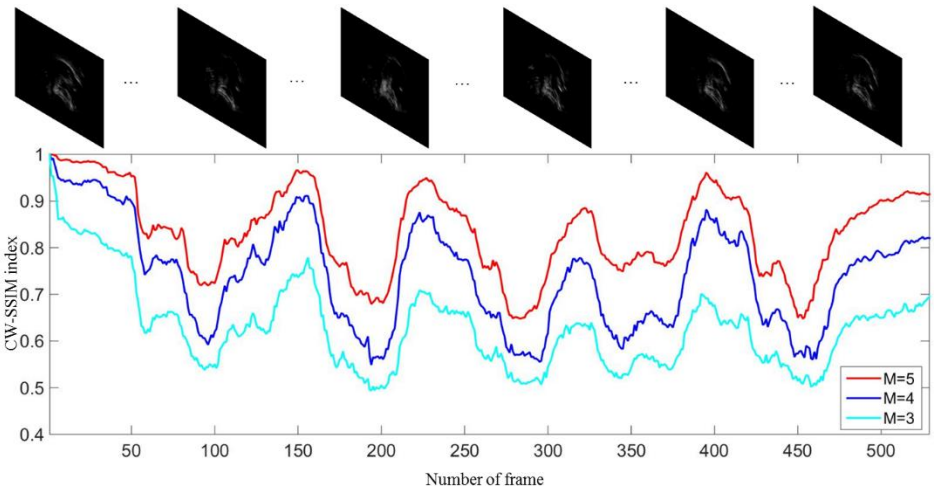


Figure 3-11 CW-SSIM index of the entire image in an ultrasound image sequence of five utterances of phoneme /k/. Three different levels of decomposition M are shown.

3.3.2 Ultrasound image-based speckle tracking re-initialization

Following the evaluation procedure, the MSD will be used to measure the performance of different speckle tracking methods with proposed re-initialization. To make the comparison more effective, the parameter and data were kept the same as used in the previous section. The results are given in Figure 3-12. As can be seen from the figure, the re-initialization method can dramatically reduce the accumulated error, thus improve the performance of the speckle tracking, which demonstrate the feasibility of proposed methods.

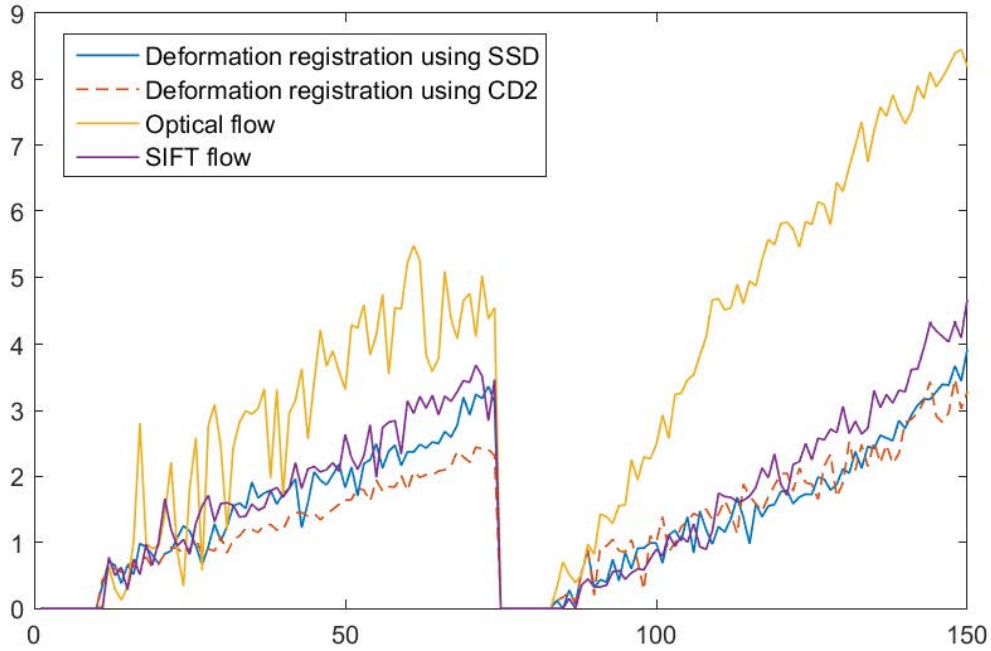


Figure 3-12 MSD error of speckle tracking methods with automatic re-initialization.

3.4 Conclusion

In this chapter, we first discussed the fundamental principles of the speckle tracking method, which presented the physical origin of the speckle. Then, different speckle tracking techniques have been tested on the ultrasound tongue image sequences, which include deformation registration, optical flow and local invariant feature. MSD is used to evaluate the performance, and the tracking error was accumulated during the tracking. To cope with this problem, by making use of the tongue's repetitive motion, the global image similarity is

employed to reset the speckle tracking automatically, with the ultimate goal to increase the robustness of the speckle tracking in ultrasound tongue images. This re-initialization method may also be useful for the contour tracking in the ultrasound tongue images, which will be discussed in Chapter 4. On the other hand, the displacements of the speckles will be used to drive the 3D tongue model in a dynamic manner, which will be given Chapter 5.

Chapter 4
Contour tracking in ultrasound tongue images

4.1 Introduction

Compared to the speckle tracking method, the contour tracking approach may be more robust to follow the motion of the tongue, as extra coherent motion information can be employed to guide the tracking procedure. Indeed, as aforementioned, a variety of processing techniques can be used to track the contour of the tongue, for example, active contour models [9], [10]; active appearance models [11]; machine learning-based tracking [12] [13]; and ultrasound image segmentation-based approaches [15]. More recently, some researchers have also proposed using physical properties of the tongue, contained in a realistic 3D model of the tongue, to help guide the contour extraction process [55]. Due to the physical motion of tongue, temporal prior shape information can be employed to help contour tracking processing.

Despite sustainable efforts, robust contour tracking still poses a great challenge in the ultrasound tongue images. In this chapter, a new contour-tracking algorithm is presented for ultrasound tongue image sequences, which can follow the motion of tongue contours over long durations with good robustness. To cope with missing segments caused by noise, or by the tongue midsagittal surface being parallel to the direction of ultrasound wave propagation, active contours with a contour-similarity constraint are introduced, which can be used to provide "prior" shape information. The idea is to extract prior shape information from the ultrasound image sequence itself, included as an extra force to guide the evolution of the hypothetical tongue contour to better handle images with missing/vague contours. As the same with speckle tracking re-initialization, to alleviate the problem of accumulated tracking error over long sequences, which can necessitate manual re-initialization, we also follow the automatic procedure, based on the CW-SSIM index [53]. The resulting algorithm has been tested on both synthetic data and real ultrasound image sequences from multiple subjects.

Results obtained demonstrate that the proposed method can improve robustness of active contours against missing segments, and has the ability to re-initialize the contour tracking automatically without manual intervention. Such an automatic tracking approach can act as a complement to hand-scanning and more traditional (but more labor-intensive) contour extraction tools, such as **EdgeTrak**, and can be a valuable improvement for research in areas where longer sequences must be analyzed, such as speech production, speech recognition in silent speech interfaces [1].

In more detail, in section 4.2, we first present the modified active contour model. Then in section 4.3, a comparative study is conducted on different contour tracking algorithms with the automatic re-initialization method incorporated. Furthermore, in section 4.4, we extend the automatic re-initialization method to the extreme by using hundreds of labeled frames to "extract the contour." In the end of this chapter, a conclusion is drawn based on our attempts on the contour tracking test on ultrasound tongue image sequences.

4.2 Active contour model with Contour group-similarity constraint

4.2.1 Active contour model with contour group-similarity constraint

An active contour model is a spline function obtained by minimizing an energy function that is intended to fit the spline to edges present in the image while retaining a reasonably regular shape. Suppose we have an active model of the tongue contour that can be represented by n discrete points $\mathbf{V} = [\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_n]$, with the total energy for snakes defined as:

$$E_{\text{total}} = E_{\text{int}} + E_{\text{ext}} \tag{4.1}$$

where E_{int} is the total internal energy (sum over the n contour points of the local internal energies defined in Eq. 4.2) and E_{ext} is the total external energy (sum over the n contour points of the local external energies defined in Eq. 4.2). In this thesis, we follow the definition of local internal energy and local external energy used by [9]:

$$\begin{cases} E_{\text{int}}\left(\mathrm{v}_j\right)=\alpha\left(1-\overrightarrow{\mathrm{v}_{j\text{-}1}\mathrm{v}_j}\cdot\overrightarrow{\mathrm{v}_j\mathrm{v}_{j+1}}\Big/\left|\overrightarrow{\mathrm{v}_{j\text{-}1}\mathrm{v}_j}\right|\left|\overrightarrow{\mathrm{v}_j\mathrm{v}_{j+1}}\right|\right)+\beta\left|\left\|\mathrm{v}_j-\mathrm{v}_{j\text{-}1}\right\|-d\right| \\ E_{\text{ext}}\left(\mathrm{v}_j\right)=1-\left|\nabla\mathbf{I}\left(\mathrm{v}_j\right)\right|/\mathrm{K} \end{cases} \tag{4.2}$$

where $j=2,\cdots,n-1$, and α, β are the weighting parameters for the internal energy; d is the average distance between two consecutive points on the contour; v_j is the j^{th} point of the active contour; $\nabla\mathbf{I}$ is the gradient of the image intensity; and K is a normalization constant.

Although **EdgeTrak** used dynamic programming in the optimization process, in practical applications, the performance of the active contour model is prone to corruption by missing boundaries due to movement of the tongue or to speckle noise. To help cope with this problem, a contour sequence similarity constraint is proposed in this work. In an ultrasound image sequence with a high frame rate (in our experiment, 60 fps), a contour extracted in a previous frame should normally be very similar to that obtained in the current frame, i.e., the deviation of contours extracted from adjacent frames should not exceed a certain threshold. Even when the local deformation of the tongue is large, a previous contour can act as a predictor to help regularize the movement of the active contour, since the true motion of the tongue must be physically reasonable.

Before the contour-similarity constraint can be added to the active contour model, an appropriate similarity measure must be defined. A classical method of measuring the similarity between two contours is to calculate distances between corresponding points. However, this is not suitable in our case as there is no straightforward way to identify actual corresponding physical tissue points in contour tracking in ultrasound tongue image sequences. On the other hand, existing techniques to compare contours in the absence of strict point correspondence, such as MSD (defined in previous chapter), would be difficult to integrate into an energy-based active contour approach as there is no efficient optimization method to minimalize the energy function due to the fact that the MSD is non-convex. Here, we explore the use of the rank of a matrix formed from a set of contours to measure the similarity of the contours of a contour sequence, as in [56]. Let a set of m consecutive contours be represented by vectors $\mathbf{c}_j$, for $j=1,2,\cdots,m$ (the size of the vector is $2n\times 1$). Each $\mathbf{c}_j=\left[\mathbf{c}_j^x,\mathbf{c}_j^y\right]^T$ is obtained by concatenating vectors $\mathbf{c}_j^x$ and $\mathbf{c}_j^y$, where $\mathbf{c}_j^x$ is the orthogonal projection of vector $\mathbf{c}_j$ on the x axis, and $\mathbf{c}_j^y$ is the orthogonal projection on the y axis. For an image sequence of a sufficiently high frame rate, it can be assumed that $\mathbf{c}_j$ is generated from $\mathbf{c}_{\text{j-1}}$ via an affine transformation. The vectors form a matrix $\mathbf{C}=\left[\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_j,\cdots,\mathbf{c}_m\right]$ (of size $2n$

$\times m$, where n is the number of discrete points and m is the number of consecutive contours), which is a low-rank matrix for any m. It was proved in [56] that $\text{rank}(\mathbf{C}) \le 6$.

To enforce similarity between the contours extracted from adjacent frames, a constraint term is added to the energy defined in Eq. (4.1), resulting in a new energy E_{sim} for each contour:

$$E_{\text{sim}} = E_{\text{int}} + E_{\text{ext}} + \gamma \, rank(\mathbf{C}) \tag{4.3}$$

where γ is a weighting parameter. Now, the rank of a matrix is a discrete quantity, which makes the optimization of E_{sim} difficult. One way to solve this problem is to use the nuclear norm regularized linear least squares technique to rewrite the constraint term as [56]:

$$E_{\text{sim}} = E_{\text{int}} + E_{\text{ext}} + \gamma \|\mathbf{C}\|_* \tag{4.4}$$

where $\|\mathbf{C}\|_*$ is the nuclear norm of $\mathbf{C}$ (the sum of singular values of the matrix $\mathbf{C}$). As the nuclear norm of a matrix is a good approximation to the rank of the matrix, it is common to substitute the rank minimization problem by the minimization of the nuclear norm, as has been widely done in low-rank modeling such as in [57]. The accelerated proximal gradient algorithm [56] is used for the optimization in this paper, as follows: the active contours are evolved at each iteration using image-based forces, after which the contour group similarity regularization is imposed by means of singular value thresholding. More detail can be found in [58] and [56]. Note that for the m-1 frames at the beginning of the sequence, no contour similarity constraint is added.

Compared to extracting the contour from a single frame, the new force included with the active contour model acts as prior information that influences the movement of the contour. In the proposed algorithm, the internal energy is used to keep the continuity of the contour; the external energy is used to attract the active contour to the real contour in the ultrasound tongue image, while the similarity-constraint acts as an additional force to limit the degrees of freedom of the movements of the active contour. When several adjacent contours are clear in the ultrasound tongue image sequences, the external energy will be the dominant term; whereas if dramatic deformations of the tongue occur in some frames, the similarity-constraint force dominates. In our experiments, the weight γ was chosen by hold-out validation: the value of γ that provided the minimum value for the Mean Sum of Distances – MSD, defined in Eq. (3.5) – for a subset of the data, and was subsequently used for the whole data set. As can be seen in Figure 4-1 (a) for synthetic contour data (with random speckle noise added), and Figure 4-1 (b) on some real ultrasound test data for frame when the tongue

goes parallel to the ultrasound beam; the similarity constraint makes the active contour more robust and physical.

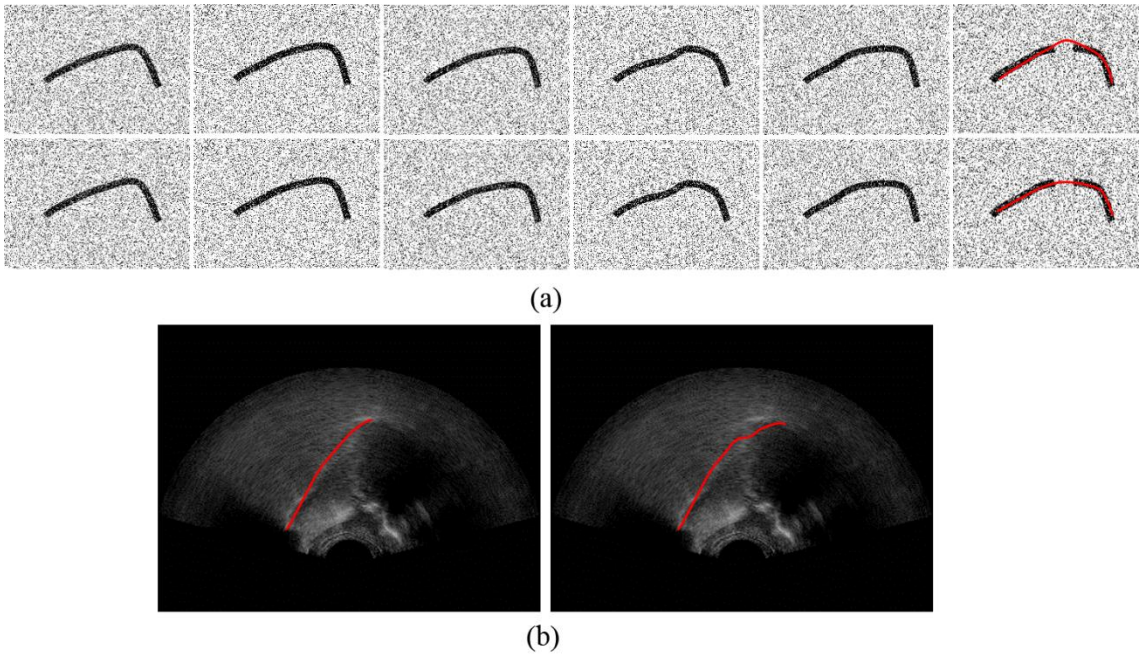


Figure 4-1 Evaluation of contour tracking on synthetic data. (a) Validation on the synthetic data, each row represents the image sequences and the red line represents the contour extracted from the image. The red line in the top row shows the contour extracted without similarity constraint while the one in the bottom row is the one with similarity constraint. In our experiment, m = 6 and n = 24 (see discussion in text). (b) Validation on the ultrasound tongue data. The red line in the left column shows the contour extracted without similarity constraint while the one in the right column is the one with similarity constraint.

4.2.2 Automatic re-initialization during contour tracking

As with other techniques to extract tongue contours in ultrasound image sequences [15] and [10], the input of our algorithm consists of several points (in our work, at least 12) placed on the surface of the tongue in the first frame of the sequence (NB: As mentioned earlier, there is no fixed relation between the points chosen and physical tissue points). Using this input along with the technique of image similarity measurement, a novel automatic tracking re-initialization is presented in this section. In this way, we may hope to dispense with the need to manually re-initialize the contour finding process due to accumulated tracking error over long sequences.

In this paper, to avoid erroneous re-initialization and make a compromise between accuracy and complexity, the CW-SSIM index threshold is set to 0.8, and $M = 4$. In the ultrasound image sequences, if the index exceeds this threshold, the locations of the discrete points will be re-initialized to the positions that were set manually in the first frame, thus improving the contour tracking automatically, without manual intervention. The example given in Figure 4 (from "Female 2", as defined in section 4) shows the automated re-initialization process. Before applying the active contour model on Frame 93, the CW-SSIM index is calculated between 93 and the first frame (with manually chosen contour points). As the index (0.81) exceeds the threshold we set (0.80), the contour is reinitialized as the first frame (in our tests, always the rest position). Without the contour tracking, the execution time of CW-SSIM is about 0.19 second for each pair of frames in our tests.

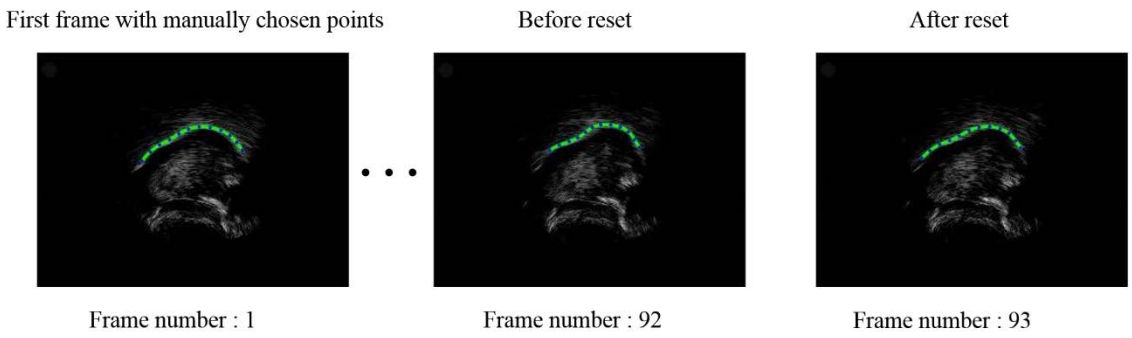


Figure 4-2 Example of automatic re-initialization in the contour tracking. As the CW-SSIM index between the first frame and the Frame 93 exceeds the threshold, the contour is re-initialized to the original position in the first frame. (The CW-SSIM index between Frame1 and frame 92 is 0.79).

4.2.3 Experiments and results

Four datasets involving 5 speakers were used in our tests:

- An unpublished dataset from speaker "Male 1";
- An unpublished dataset of isolated utterances from subject "Female 1";

- Portions of a POLYVAR corpus recorded at our laboratory in 2011 on three speakers, “Female 2”, “Male 2”, and “Male 3;
- Portions of a TIMIT corpus recorded in our laboratory in 2010, on speaker “Male 1” .

Our algorithm is implemented using MATLAB 2014b on a Windows 8 desktop with Intel 4-Core 3.7 GHz CPU, 16 GB RAM, and ATI Radeon HD 7800 with 6 GB DDR3 VRAM, Dual AMD Filepro 512 GB PCIe-based flash storage. Hand labelling for comparison to automatic tracking results was performed by a single labeller and took about 2 months.

First we examine the execution time of the proposed algorithm. The size of $\mathbf{C}$ is $2n \times m$, where n is the number of discrete points ($n = 12$) and m is the number of adjacent contours, here set to 10. The singular value decomposition calculation is fast, and the energy minimization using the decomposition in fact faster than the minimization processing without the contour similarity constraint; the computationally heaviest part of our work is the calculation of the similarity between the current frame and the first frame (the tongue in rest position) using CW-SSIM. The time performances for the different subjects are very similar, ranging from 220 milliseconds to 239 milliseconds per frame; thus, on average, a sustained rate of roughly 5 Hz can be maintained with the algorithm in its present form. On longer sequences, of course – and even more so with hand-scanning – hand re-seeding of contours will slow overall processing time down very dramatically, as compared to the proposed method. In this sense, the approach presented can be viewed as a complementary, more automatic approach, that can be of significant value in applications where long sequences must be analyzed in their entirety, as mentioned earlier.

Figure 4-3 shows an example result, analogous to the test of Figure 4-1 on synthetic data, on twenty contiguous frames from Female 1, in a segment of the data where the contour is rather faint, due to the orientation of the tongue. Red lines show the results obtained with the contour similarity constraint; yellow, those without.

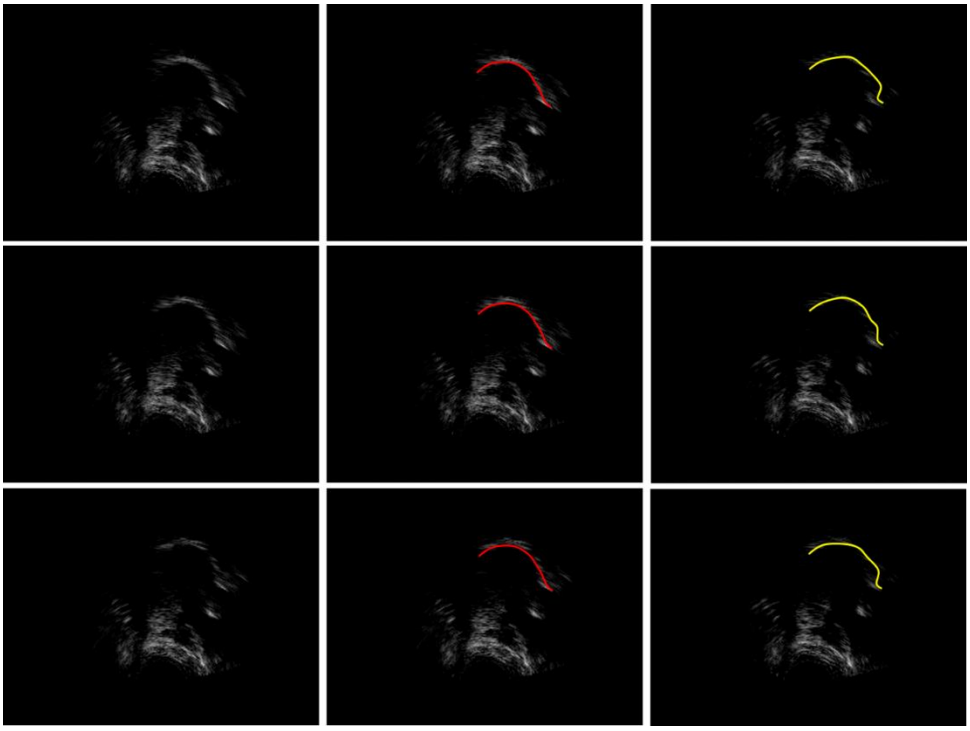


Figure 4-3 The comparison between the contour-extracted with contour similarity constraint (red line in second column) and without contour similarity constraint (yellow line in third column). As the number of frames is small, image-similarity-based re-initialization was not necessary here.

The tracking of another example sequence (Female 2) is shown in Figure 4-3, which lasts more than 3 minutes (over 17000 frames). The Figure shows the contours from a selection of frames in the sequence. No manual re-initialization was made during the tracking, aside from the initial seeding done on frame 1. On visual inspection of the tracked contours, the algorithm works quite well.

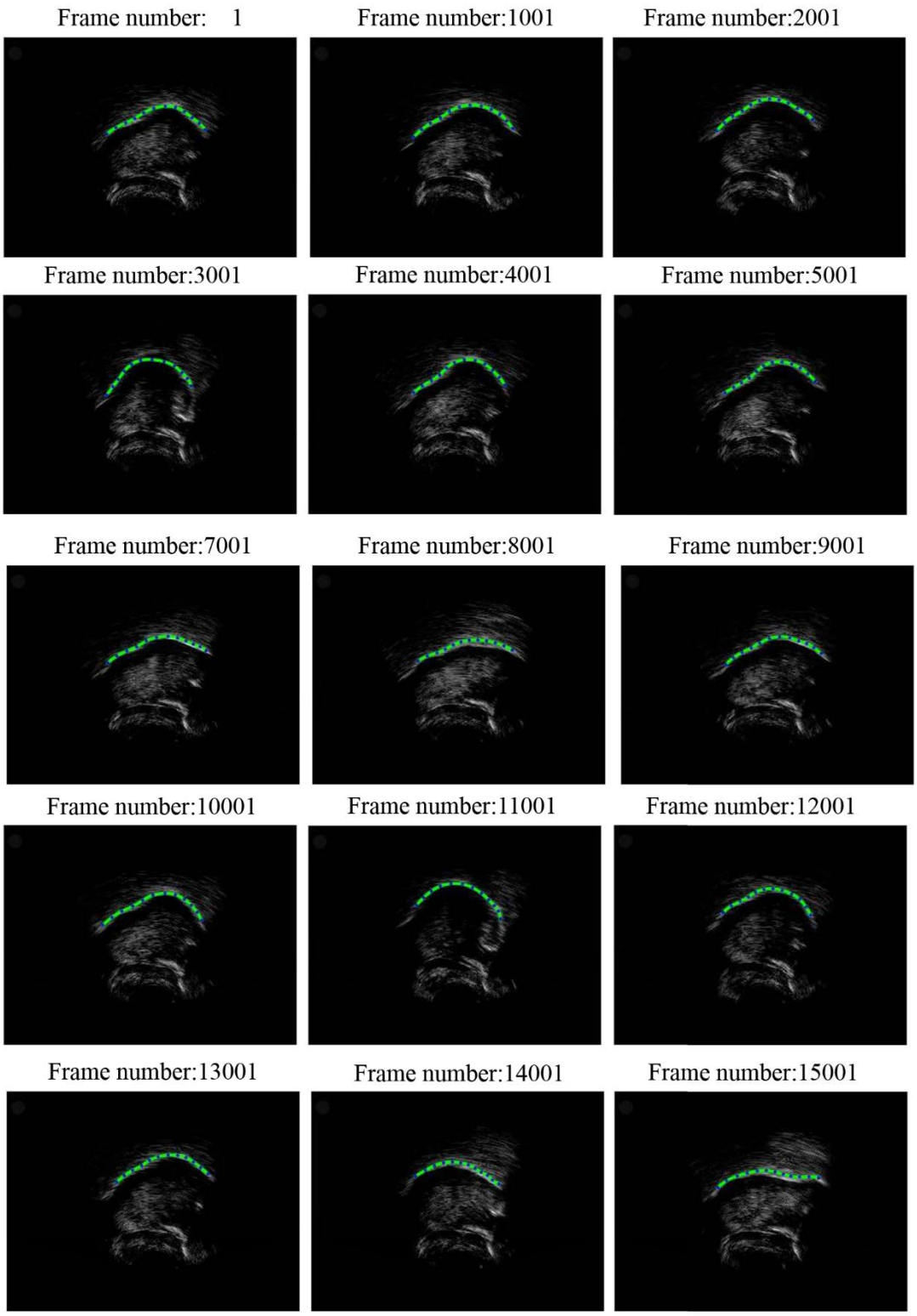


Figure 4-4 The results (Female 2) of contour tracking in the sequences of long duration (green lines are the contour tracked in each frame, while the blue points represent the points to represent the curve). To keep the original result, no contour extrapolation is made.

Occasionally, tracking errors do occur, as shown in Figure 4-5, due to the presence of a high level of noise or other anomaly in a particular region of the image over an extended period, during which the tongue does not return to the rest position, and the contour therefore not automatically reset.

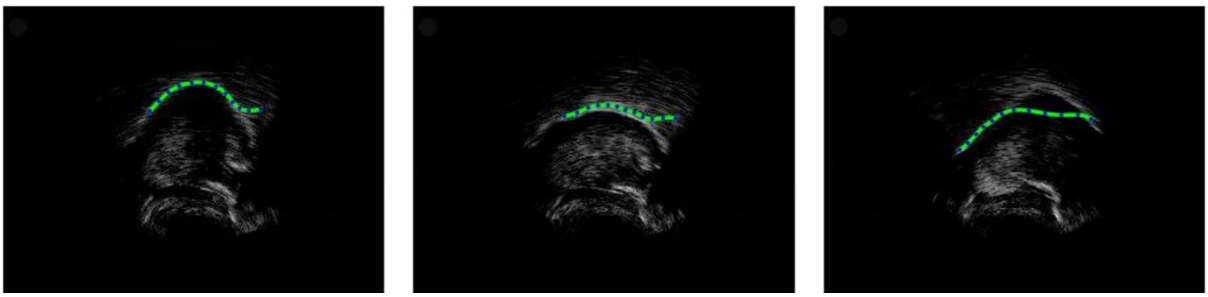


Figure 4-5 Examples for poor tracking (Female 2).

As our sequences consist of large numbers of image frames due to the high capture rate (60 fps), it is very laborious to extract contours manually for all frames. Therefore, 4000 frames were chosen randomly for manual contour extraction from the data recording of Female 2, 1000 frames for Male 2, and 2000 frames for Male 3. Compared to the manually extracted contours, the MSD errors between the contours extracted automatically by different methods are given in Table 4-1.

Table 4-1 Errors by using different methods for different subjects (For Female 2,400 contours were extracted manually, for Male 2, 1000 contours were extracted manually, while 2000 contours were extracted manually for Male 3.) The standard deviation is also given in this table.

Methods	MSD mean errors and standard deviation (pixels, 1 pixel = 0.295 mm)		
	Female 2	Male 2	Male 3
Similarity constraint + CW-SSIM	**3.36 ± 0.86**	**3.65 ± 1.02**	**2.96 ± 0.95**
Similarity constraint + SSIM	4.09 ± 2.01	4.29 ± 2.93	5.84 ± 3.40
No similarity constraint + CW-SSIM	18.96 ± 1.08	16.46 ± 1.29	18.64 ± 1.07
No similarity constraint + SSIM	22.43 ± 2.68	19.43 ± 3.47	21.27 ± 4.55
Similarity constraint	4.52 ± 2.53	5.52 ± 3.41	6.45 ± 2.87

It can be observed that the proposed method, with contour similarity constraint and automatic re-initialization, has the best performance (smallest mean and smallest standard deviation of MSD error). Some other examples of tracking results for Female 1 are given in Figure 4-5, showing a variety of tracking qualities. Some results for Male 1, Male 2 and Male 3 are given in Figure 4-6, Figure 4-7 and Figure 4-8. No manual re-initialization was performed on any dataset during the entire tracking procedure. The performance demonstrates the algorithm versatility on the different subjects. We note that since Male 3 had undergone a laryngectomy, no hyoid bone or shadow of the hyoid bone can be observed in the image sequences of this speaker.

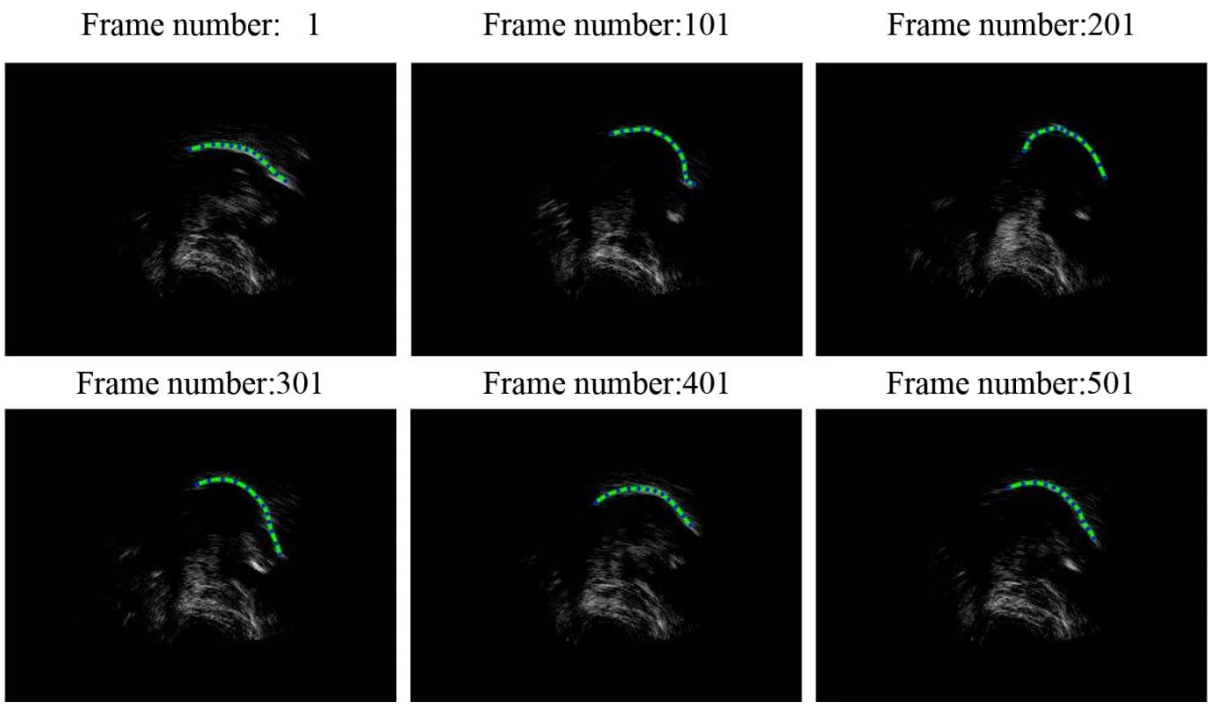


Figure 4-6 Some examples of results for Female 1

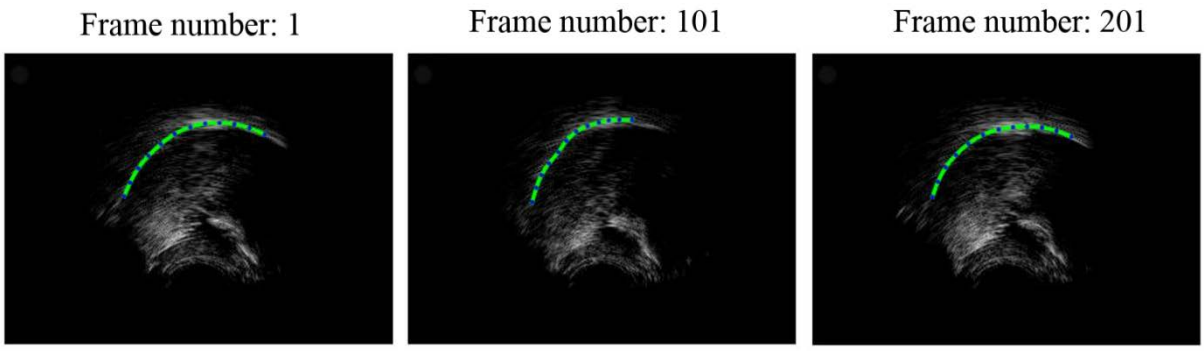


Figure 4-7 Some examples of results for Male 1.

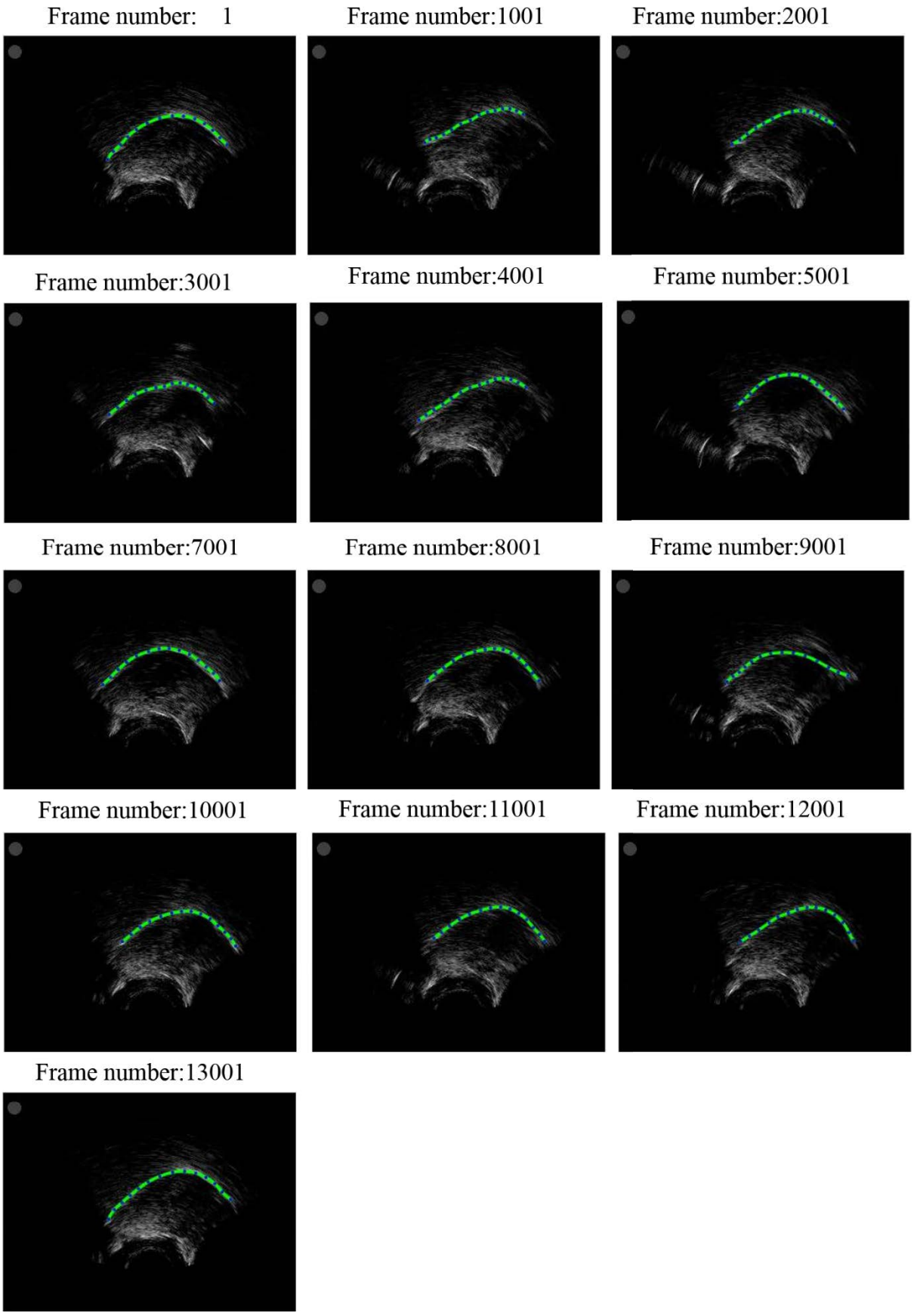


Figure 4-8 Some examples of results for Male 2.

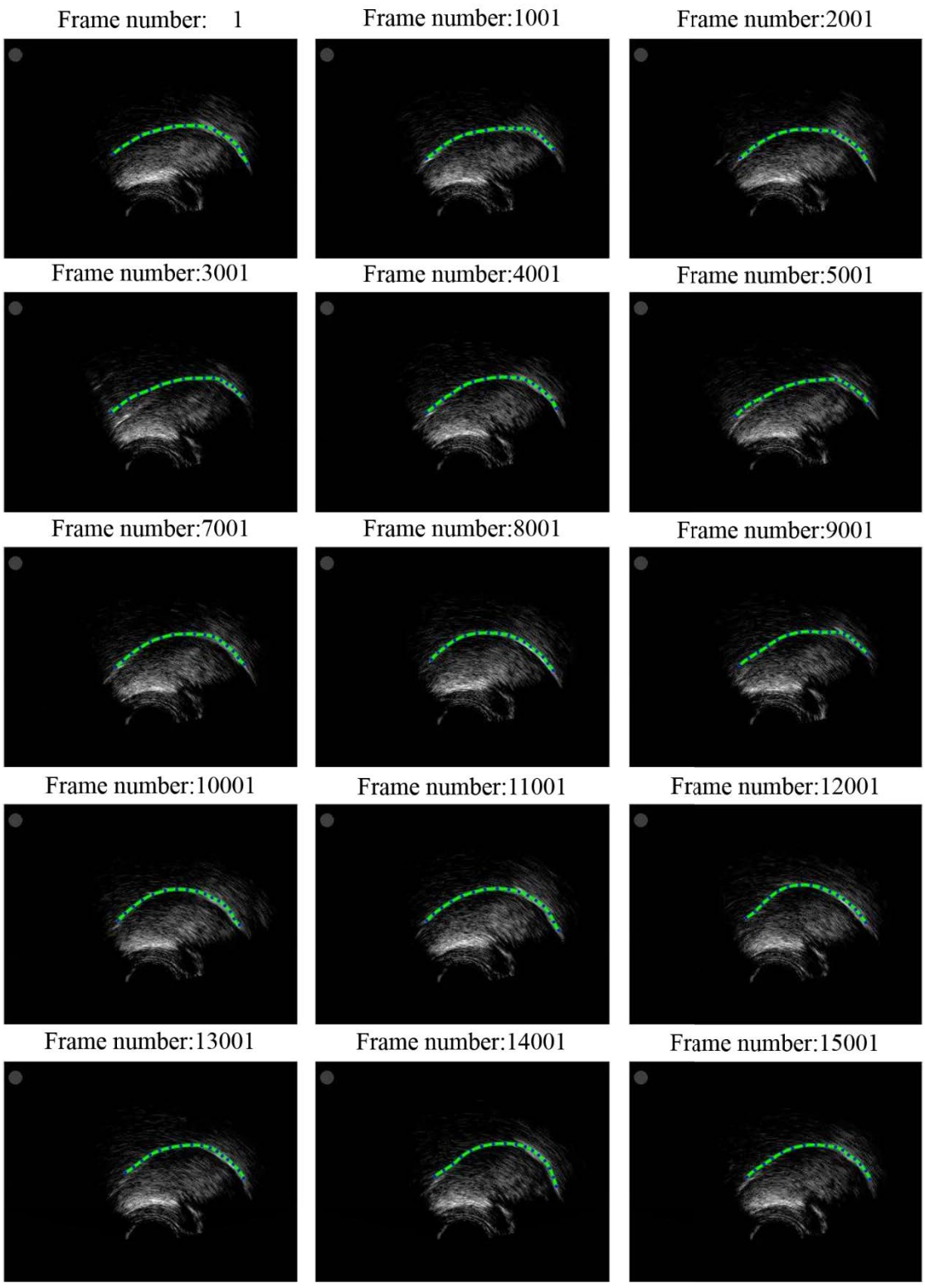


Figure 4-9 Some examples of results for Male 3.

4.3 A comparative study on the different contour tracking algorithms

4.3.1 Comparison of contour tracking methods with re-initialization

In this section, we will conduct a comparative study on the different contour tracking algorithms in the ultrasound tongue image with CW-SSIM-based automatic re-initialization. To make a comparative study on different contour tracking algorithms, the re-initialization method is incorporated into **EdgeTrak** and **TongueTrack** and a quantitative comparison is made with and without automatic re-initialization.

In more detail, as mentioned in the introduction section, unlike the general tracking problem in the video sequence, this automatic reset method may be feasible in the ultrasound tongue sequences as the trajectory of the tongue motion is repetitive (for example, the tongue goes back to the rest position frequently during speech production). If we label some contours manually in some frames, then these frames can be the seeds to trigger the automatic re-initialization using similarity measurement. Theoretically, if the number of the hand-labeled frames is large enough, which can cover all the potential trajectories of the tongue, the tracking problem can be converted to the use of a robust similarity measurement to find the correspondences between label frames and the current frame pending processing.

In more detail, the proposed comparison experiments are conducted as follows: Before contour tracking is carried out, the similarity coefficient is calculated, between the current frame and 5 hand-labelled reference frames, selected manually for each subject, in order to cover, in an ad-hoc way, the space of possible different tongue configurations for each speaker. If the similarity coefficient exceeds a set threshold (in our experiment, 0.85), the positions of the contour are re-initialized to those which were input for the reference frame.

Following the previous work in [59], we conducted a comparative study on the error analysis. The data recorded two female and two male subjects of normal speaking abilities. The utterance is also the same for each subject, with saying "*I owe you a yoyo*". The total number of images is 1145. Hand-labeled contours will be used as the ground-truth for evaluation. Our goal is not to compare the error induced by the tracers, but to evaluate whether the automatic re-initialization can improve the robustness of the contour tracking. In

this experiment, mean sum of distances (MSD) are used to evaluate the error. The experiments are conducted using different tracking algorithms with and without automatic re-initialization method and the results are given in Table 4-2.

Table 4-2 A comparison between with and without automatic re-initialization method (1 pixel = 0.295 mm)

MSD error (pixels)	**EdgeTrak**	**TongueTrack**	Method proposed in Section 4.2
Without automatic re-initialization	7.06 ± 2.77	5.59 ±3.04	4.68 ±2.81
With automatic re-initialization	3.46 ±1.04	3.60 ±0.96	4.05 ±0.97

As can be seen from the table, the automatic re-initialization method can improve the tracking performance by only using 5 hand-labeled frames. We can expect that: with more hand-label frames, the MSD error can be reduced further, which will be demonstrated in next section. The results demonstrate that, using the proposed method initialization can improve the performance by reducing the average mean sum of distances (MSD). Furthermore, we extend the automatic re-initialization method to the extreme by using hundreds of labeled frames to "extract the contour".

4.3.2 Similarity-based contour extraction

In this section, we expand the automatic re-initialization idea to the extreme for contour tracking: using hundreds of hand-labeled frames and CW-SSIM index to extract contours in ultrasound tongue image sequences.

Before the extraction processing is conducted on the ultrasound tongue data, an image dictionary needed to be constructed, which consists of 2000 manual labeled frames for each subject (three subjects' data are used in our experiment). The data used are the recordings of multi utterances from different subjects. Figure 4-10 gives an example of the label frames for different subjects in the database.

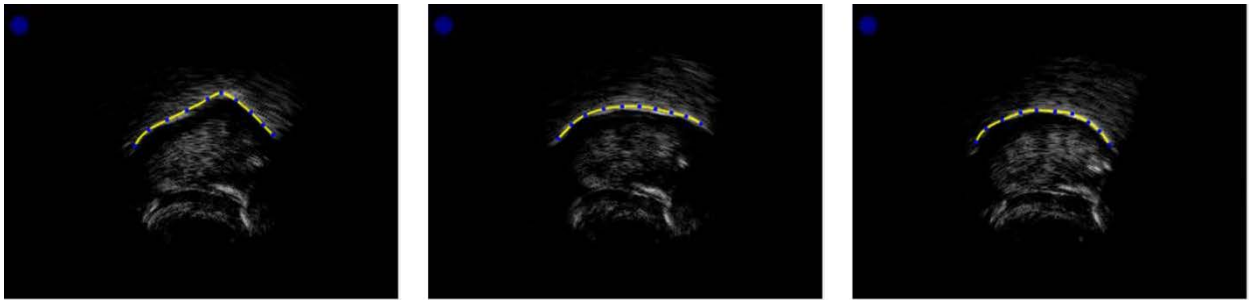
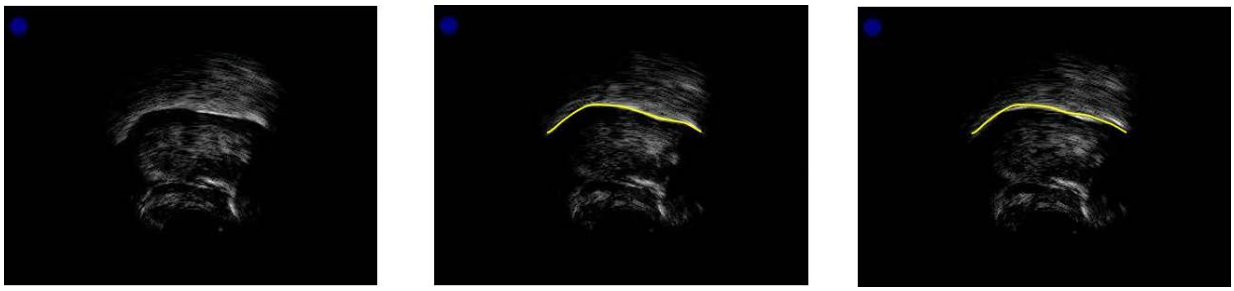


Figure 4-10 Sample frames from the image dictionary. The yellow curves represent the contours labeled manually.



(a) (b) (c)

Figure 4-11 The frame in (a) is the frame pending processing, while the two frames in (b) and (c) give the most similar hand-labeled frames selected from the database by using CW-SSIM index. The yellow lines are the contour label manually. The similarity index between (a) and (b) is 0.9569, while the similarity index between (a) and (c) is 0.9636.

In fact, the number of frames in the database, which need to be labeled, is directly linked to the degree of freedom of the tongue. Figure. 4-11 gives an example of most similar frames in the database of hand-label frames.

As our sequences consist of large numbers of image frames due to the high capture rates (60 frames per second), it's very difficult to label all frames for the validation on the whole dataset. Therefore, 500 extra frames were chosen randomly for hand labeling to evaluate the accuracy of the algorithm. Compared to the manually extracted contours, the MSD errors between the contours extracted automatically using different methods is given in Figure. 4-12. As can be seen from the figure, with 1000 hand-label frames as the database, the proposed similarity-based method achieves 2-3 pixels in MSD error. With 2000 hand-label frames, the

error drop to 1.5 – 2 pixels for different subjects. Moreover, we also noticed from the figure, the MSD error of contour extraction has faster error convergency for the subject who had under gone laryngectomy than the normal subjects, which may demonstrate that the degree of freedom of the tongue of this patient subject may be smaller than the normal subject.

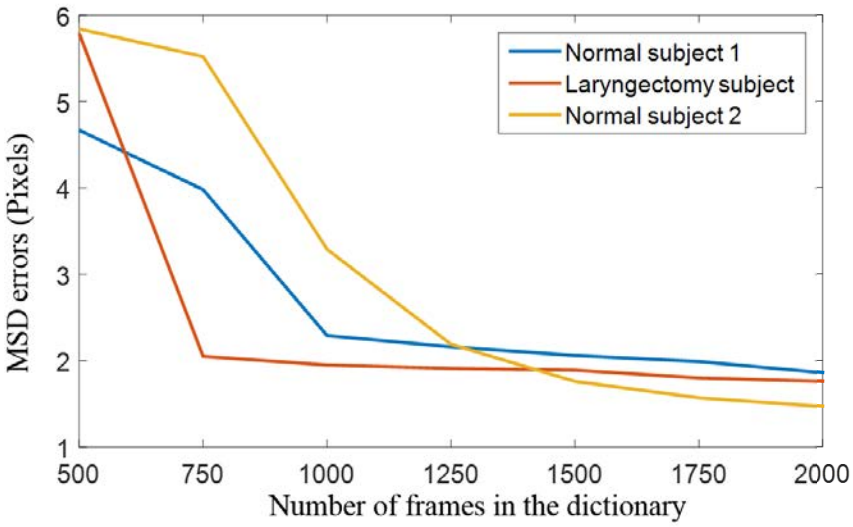


Figure 4-12 Errors by using different methods MSD Errors across (pixels, 1 pixel = 0.295 mm)

4.4 Conclusion

In this chapter, different automatic contour-tracking algorithms are tested for ultrasound tongue image sequences. Firstly, based on the active contour model, a novel active contour method with temporal regularization is presented to address the problem of missing or faint contours with a contour similarity constraint. Moreover, the image similarity-based automatic re-initialization technique is also incorporated into the contour tracking algorithms in this chapter, which can increase the robustness of the tracking. This method serves as a complementary approach to hand-scanning and existing semi-automatic scanners (e.g., **EdgeTrak**), and can be an important tool for applications where analysis of long sequences is important, such as speech production, speech recognition and the like.

Moreover, to make a comparative study between other different contour tracking algorithms, a comparative study is conducted on with and without automatic re-initialization method. The results demonstrate that, using the proposed method can improve the

performance by reducing the average mean sum of distances (MSD) from 5-6 to 4 pixels. Furthermore, we extended our automatic re-initialization method by using more frames, and select the labeled contour in the more similar frame to represent the contour in the current frame, which achieves superior performance compared with **EgdeTrak** and **TongueTrack**. Still, the scalability of the method is limited due to the necessity of extensive hand scanning.

Chapter 5
Physics-based 3D tongue motion modeling

5.1 Introduction

In speech production research, a realistic 3D tongue motion visualization is of importance, and an accurately quantified description of the 3D tongue motion may be helpful for the SSI systems. Furthermore, 3D dynamic tongue modeling can serve as a tool to study articulation training. However, despite considerable efforts, "seeing speech", as the process is often defined, remains a challenge.

Recently, advances in physics-based 3D modeling technique have advanced the technique to a point where ultrasound-based 3D tongue modeling may be feasible. Based on the motion tracking presented in the previous two chapters, in this chapter, we explore a generic tongue visualization framework, which combines the 2D ultrasound imaging and a physics-based 3D modeling technique. Although our primary goal is to design a platform for ultrasound data, the system can also serve as an interface for other imaging modalities (e.g., Magnetic Resonance Imaging (MRI)) to assist studies of speech production.

The organization of the chapter is given as follows: In section 5.2, we describe the principle of physics-based 3D modeling, based on which the overview of the interface is also presented in this section. The framework of the speckle tracking-based 3D tongue motion visualization is given in section 5.3, while in section 5.4, the contour tracking-based 3D tongue motion modeling method is presented, and section 5.5 provides conclusion.

5.2 Physics-based 3D tongue modeling

5.2.1 Theoretical foundations of motion-driven based 3D tongue modeling

Currently, existing state-of-the-art platforms (such as ArtiSynth [24]) focus on modeling driven by muscle activations. Nevertheless, despite many attempts to characterize the biomechanical properties of the tongue [60], our understanding of tongue muscle activations is still incomplete. Furthermore, most existing 3D tongue visualization frameworks are unable to simulate real-time tongue motion. Modal analysis based on a linear strain tensor is suitable for real-time simulation. However, it has been suggested that the deformation of the tongue could be large, and the hypothesis of small deformations may not be appropriate. Green's nonlinear strain tensor can model large deformation; however, time stepping of the resulting nonlinear system can be computationally expensive. Modal warping [61] technique is a good solution to this problem, which explores to handle rotational parts of deformation in the framework of model analysis, thus uniting the benefits of both modal reduction and stiffness warping. This technique is used in our framework to animate the tongue deformation.

The displacements of the nodes obtained from the motion tracking approaches are applied as a linear constraint, which can be integrated into the governing equation of the dynamic deformable tongue model using the Lagrange multiplier method [62].

This can be expressed as:

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f} \tag{5.1}$$

The mass, damping, and stiffness matrices $\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$ (of size $3n$ x $3n$, where n is the number of nodes) are determined by the material's intrinsic physical properties; $\mathbf{u}$ is the vector of the displacements of the nodes from their original positions on the mesh; $\mathbf{f}$ is the vector of external forces. Eq. (5.1) is a coupled system of ordinary differential equations, which typically cannot be solved in real-time. To address this problem, we adopted linear modal analysis to accelerate the computational efficiency by solving the generalized eigen problem. Suppose $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ (a diagonal matrix of eigenvalues) are the solution matrices for $\mathbf{K\Phi} = \mathbf{M\Phi\Lambda}$, where $\mathbf{\Phi}^{\mathrm{T}}\mathbf{M\Phi} = \mathbf{I}$, $\mathbf{\Phi}^{\mathrm{T}}\mathbf{K\Phi} = \mathbf{\Lambda}$. The modal displacement can be expressed as a linear combination of the columns of $\mathbf{\Phi}$.

$$\mathbf{u} = \mathbf{\Phi q} \tag{5.2}$$

Substitution of Eq. (5.2) into Eq. (5.1) followed by a left-multiplication by $\mathbf{\Phi}^{\mathbf{T}}$, results in:

$$\ddot{\mathbf{q}} + \mathbf{\Phi}^{\mathbf{T}}\mathbf{C}\mathbf{\Phi}\dot{\mathbf{q}} + \mathbf{\Lambda q} = \mathbf{\Phi}^{\mathbf{T}}\mathbf{f} \tag{5.3}$$

We can use only several dominant columns in $\mathbf{\Phi}$ (e.g. the ones associated with the smallest eigenvalues), thus the computation load of Eq. (5.3) is considerably reduced. Under the commonly adopted Rayleigh damping condition: $\mathbf{\Phi}^{\mathbf{T}}\mathbf{C}\mathbf{\Phi} = \xi\mathbf{I} + \zeta\mathbf{\Lambda}$ (where ξ and ζ are scalar weighting factors), the calculation can be carried out in nearly real-time.

As the linear modal analysis cannot deal with large magnitude deformations, the modal warping technique [61] is used to compute the nonlinear deformation term, so that the new representation of the rotational part becomes:

$$\mathbf{w} = \mathbf{W\Phi q} \tag{5.4}$$

where $\mathbf{w}$ is a vector representing the angular velocities of the nodes and $\mathbf{W}$ is the *curl* of the linear displacement. And Eq. (5.3) can be simplified as a simple linear system of:

$$\mathbf{A}\ddot{\mathbf{q}} = \mathbf{b} \tag{5.5}$$

which is to be solved repeatedly at each time step. The Newmark-average acceleration method is used in our framework to solve the equation and we can get:

$$\mathbf{A} = \mathbf{M} + (h/2)\mathbf{C} + (h^2/4)\mathbf{K} \tag{5.6}$$

$$\mathbf{b} = \mathbf{f} - \mathbf{C}\dot{\mathbf{q}}^* - \mathbf{K}\mathbf{q}^* \tag{5.7}$$

where $\dot{\mathbf{q}}^* = \dot{\mathbf{q}}^- + (h/2)\ddot{\mathbf{q}}^-$ are two predictors. The superscript "-" represents the values in previous frame. The modal displacements and velocities can be calculated by using:

$$\begin{cases} \mathbf{q} = \mathbf{q}^* + (h^2/4)\ddot{\mathbf{q}} \\ \dot{\mathbf{q}} = \dot{\mathbf{q}}^* + (h/2)\ddot{\mathbf{q}} \end{cases} \tag{5.8}$$

where h is the time interval which is set as 1/60 second, according to the actual data sampling rate of the ultrasound imaging devices. By using modal reduction and model warping, the 3D deformation of the tongue can be realistic and retain calculation efficiency.

5.2.2 Interface overview

An interface of our whole tongue modeling visualization framework has been developed. Figure. 5-1 provides an overview of the interface devolved to implement the whole framework. The main 3D view is given in the left part of the whole interface. Three different orthogonal perspectives (top, front and lateral) are embedded into the interface, which lie next to the main 3D view. Next the three auxiliary views, the ultrasound image sequences as well as the displacement, velocity, and acceleration rate of the speckle tracked, changes of the tongue's volume are also visualized in the right portion of the interface. All of the views are fully coupled.

The described dynamic 3D tongue motion visualization platform is implemented in Microsoft Visual C++ 2010 in a Windows 7 environment, using a PC with Intel Core i7, 8G DDR3L and an NVIDIA GHTX862M.

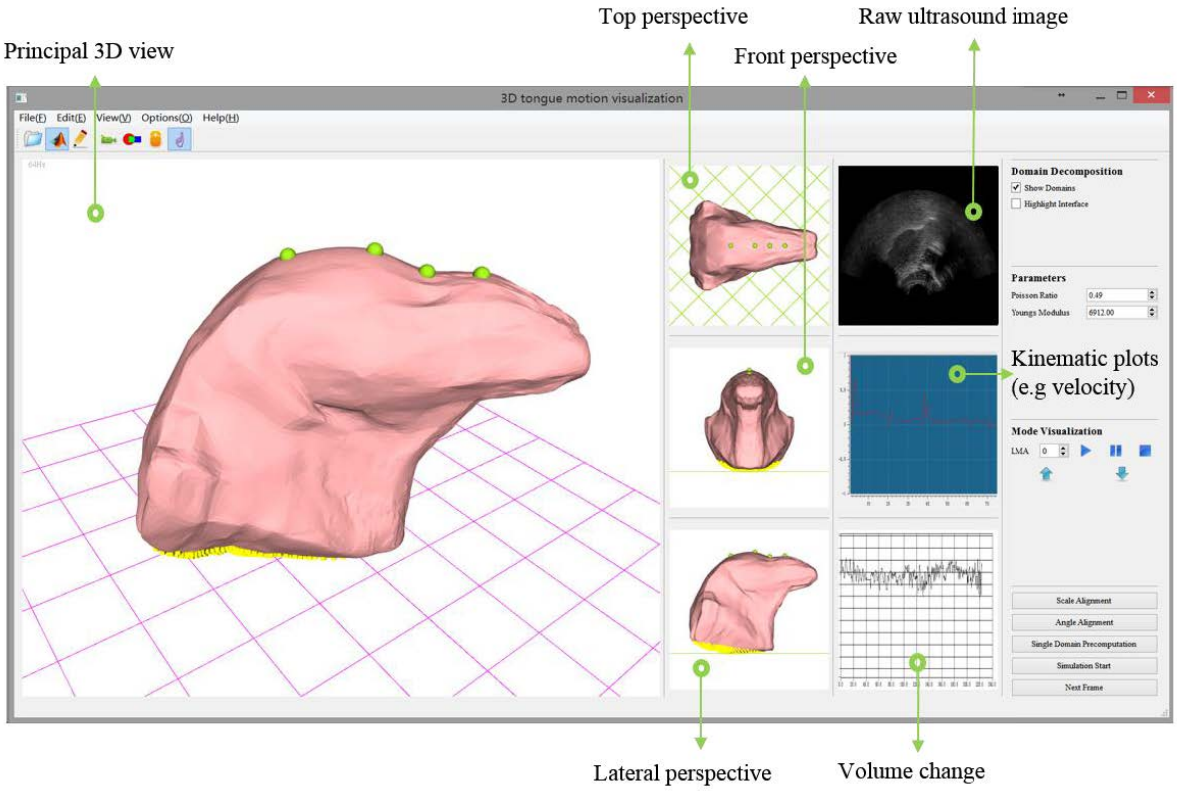


Figure 5-1 A snapshot of the user interface of the platform being developed.

5.3 Speckle tracking-based tongue motion simulation

To drive the 3D tongue model, the displacements of the constraint nodes (as can be seen in Figure 5-2) needs to be obtained by using aforementioned different motion tracking methods. In this section, we will explore this kind of approach.

5.3.1 Speckle tracking-based tongue motion visualization

Firstly, we summarize the framework to employ speckle-tracking method to drive the 3D tongue model, which can be roughly divided into three modules: pre-alignment, speckle tracking and motion-driven 3D tongue model modeling.

As the first step, the alignment is made between the generic tongue model and the ultrasound data in twofold: scale-alignment and angle alignment. The visible reference positions are selected in the vocal tract ultrasound sagittal scan (as shown in Figure 5-3), which include the "shadow" of hyoid bone and tendon. Based on the distance between the two reference positions, we can obtain the approximate scale ratio of between 3D tongue model and the 2D ultrasound image along the direction of X axis and Y axis (Suppose we can represent the scale ratio as : s_y / s_x). Also, we can make use of the contour extracted manually in the coronal scan to calculate the scale ratio in the direction of Y axis and Z axis and represent it by using s_y / s_z . Then a scale matrix can be obtained as follows:

$$S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & s_y / s_x & 0 \\ 0 & 0 & s_z / s_x \end{bmatrix} \tag{5.9}$$

By multiplying the scale matrix with the coordinates of the nodes, a novel 3D tongue mesh can be obtained, which can roughly be aligned with the size of the tongue of the subject. Here, the alignment is of importance as the size of the tongue varies dramatically in different subjects. Without this step, the modeling framework will fail during our experiments.

For the angle-alignment, suppose θ_1 donates the angle between the X axis and the vector which links the hyoid bone and tendon in the 3D tongue model, while θ_2 represents the same angle in the ultrasound image. We can use a rotation matrix R , which rotate the model

clockwise by an angle $\theta_1 - \theta_2$, to minimize the angle difference between the tongue model and the real ultrasound data. Multiplying the generic tongue model by the scale matrix and rotation matrix, a new tongue's geometry model can be obtained, which has been aligned approximately with the ultrasound image.

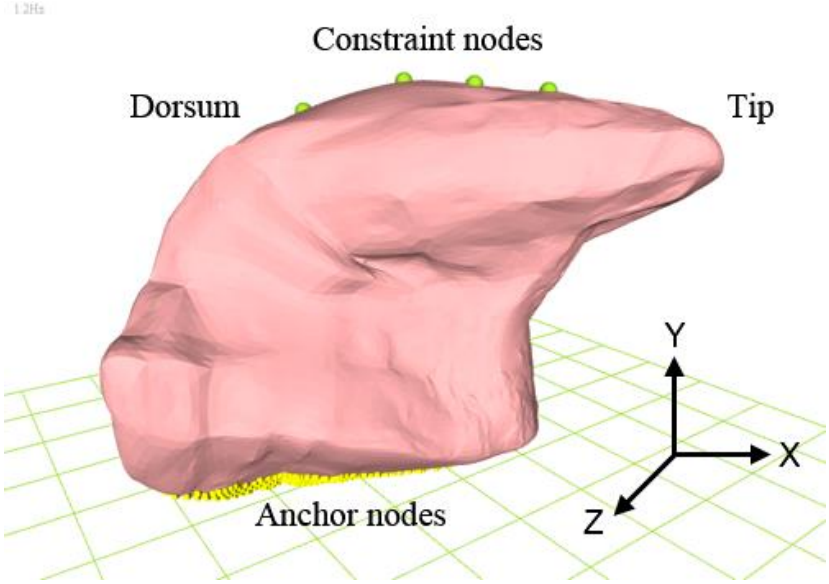


Figure 5-2 Generic tongue model with anchor (yellow) and mid-sagittal constraint nodes (green), for driving the model, are shown in the rest configuration. Anchor nodes' displacements are zero during the motion of the tongue model.

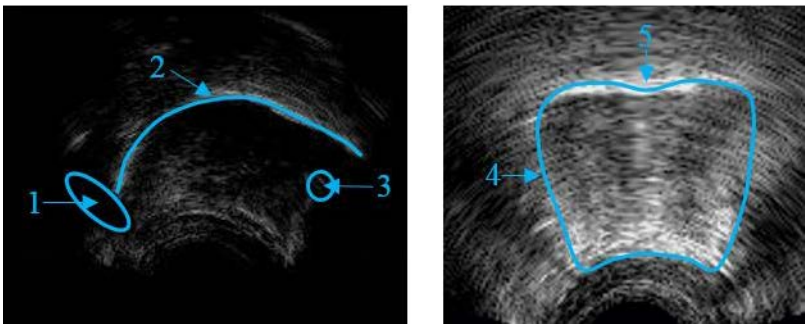


Figure 5-3 Vocal tract ultrasound scan: (1) the "shadow" of Hyoid bone; (2) upper tongue surface; (3) tendon; (4) tongue surface; (5) central groove.

After the alignment, for the second step, the speckle tracking method is applied to the ultrasound tongue images to obtain point correspondences. Theoretically speaking, the

displacement of the “tissue point”, extracted by using speckle tracking, can be subsequently transmitted to selected nodes on the midsagittal tongue model surface in order to drive the 3D model at the acquisition rate of the ultrasound image sequence. As speckle tracking can follow all the “tissue points” in the image sequences, we manually select several control points on the 3D tongue model, and all the points are associated with the speckle patterns tracked.

5.3.2 Experimental results

Although, as demonstrated in Chapter 3, optical flow-based speckle tracking method provide lower performance with comparison to other speckle tracking methods. In our framework, our initial test used optical flow to obtain correspondences between the patterns in 2D image sequences, from which the tongue’s motion is derived. The reason is that, optical-flow is computationally efficient, and this kind of method can track the “tissues” in real-time. Although somewhat unstable, the technique provides a simple method for initial tests of the 3D visualization platform, and some sample results are given in Figure 5-4.

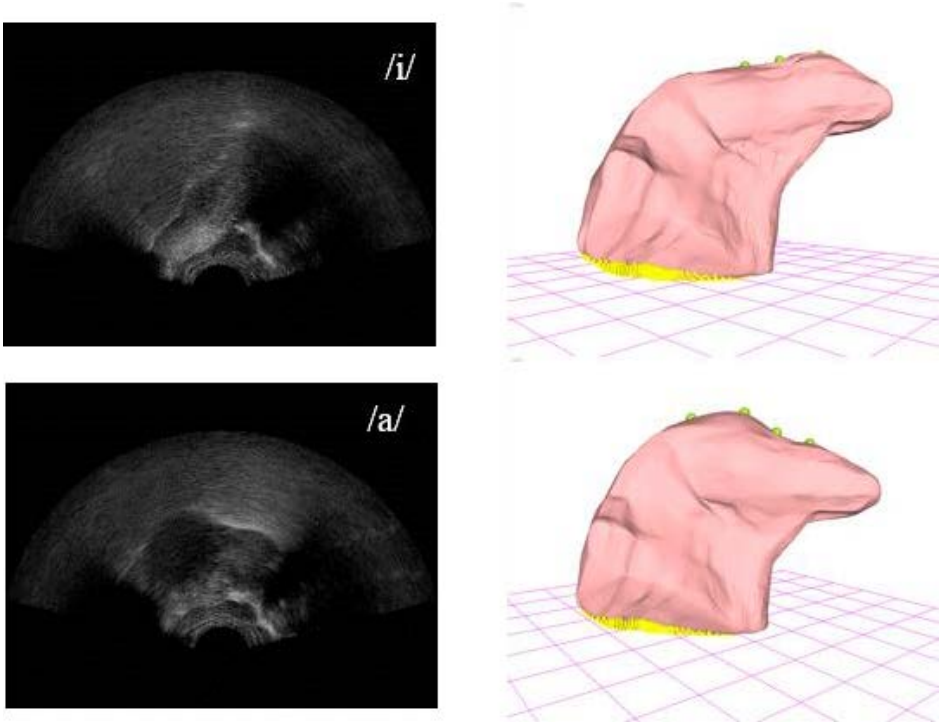


Figure 5-4 Some examples of the visualization results.

As the tongue should be volume-preserved, to test the change of tongue volume during deformations, an experiment was performed, with results as shown in Figure 5-5. It can be seen that the change in the tongue volume remains small, less than 2% in most cases.

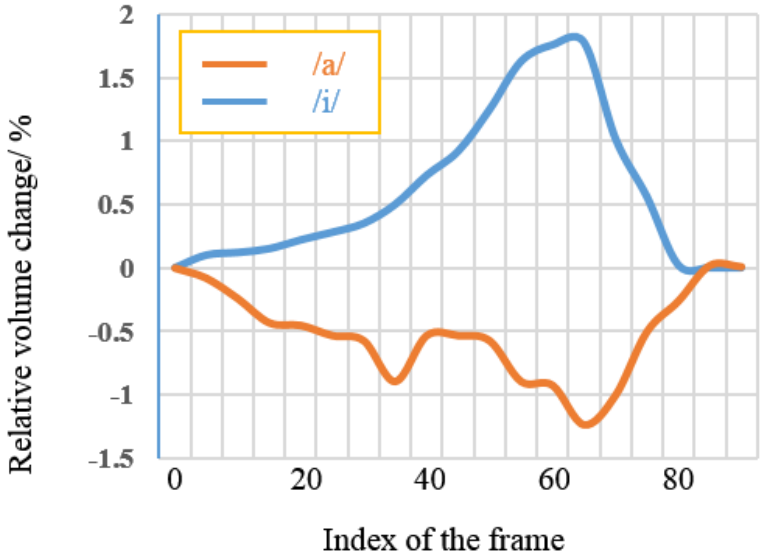


Figure 5-5 Volume change of the tongue model.

Using 150 modal bases (the number of the columns in matrix $\mathbf{\Phi}$), the simulation throughput is 43.2 fps. Thus although tongue deformation during speech can be rapid, the developed framework appears to be able to meet this demand, following the motion and generating tongue shapes in real-time with relatively good accuracy. However, as discussed in previous chapter, the tracking error was accumulated during the tracking even the image-similarity-based re-initialization method was used. Thus, several unrealistic deformations are generated. To solve this issue, more robust tracking method is needed. Compared to speckle tracking, contour tracking can characterize the motion of the tongue with higher accuracy. Thus, in next section, we will use contour tracking as the motion tracking method.

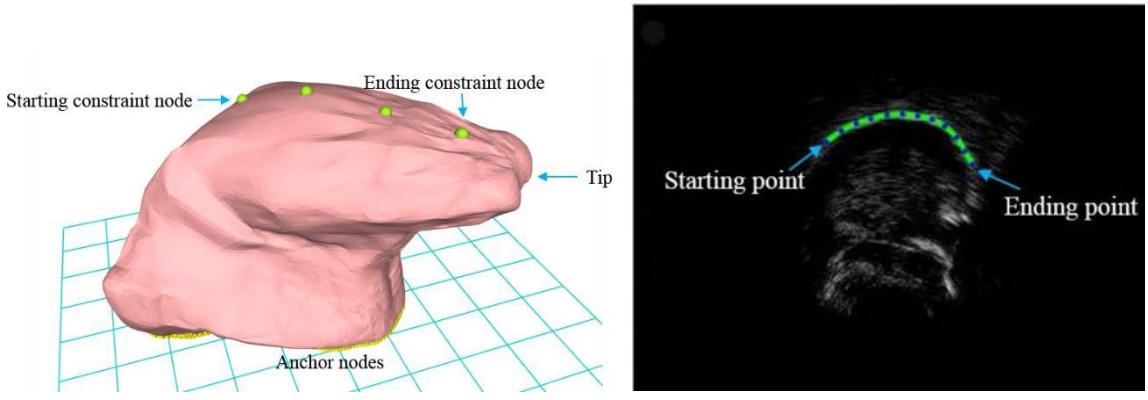
5.4 Contour-guided 3D tongue motion visualization

Here we use the extracted contours to drive the motion of the tongue, as the speckle tracking is found sometimes unstable, which will induce unrealistic deformation during the motion visualization.

5.4.1 Contour-based 3D tongue motion visualization

The 3D model can be driven by imposing extra positional constraints at specified finite element nodes to enforce their displacements to some user-specified values. To drive the 3D tongue model, the modal displacement needs to be calculated by making use of the contour extracted from the ultrasound image sequences. However, obtaining the correspondence between the tissue points of the contours of different frames is of great difficulty, and registration between the 2D ultrasound image and 3D tongue model is another challenge. Rather than using speckle tracking, in this section, we show that these challenges can actually be converted into a "3D shape search" problem. The detailed method is given as follows:

Step 1: Initialization. Four constraint nodes are selected manually (as shown in Figure 5-6(a)). In this section, we suppose the first and last nodes are associated to the starting points and ending points of the contour extracted from the 2D image (as shown in Figure 5-6 (b)).



(a). 3D tongue model used in our framework. (b). Ultrasound tongue image with contour extracted.

Figure 5-6 Elements used for the 3D visualization. (a) The 3D model used in our framework, the green circles denote the constraint nodes, whose displacements are associated with the modal displacement. The yellow nodes are anchor nodes whose displacements are zero

during the deformation of the tongue model. (b) Target curve extracted from the image, the green lines are the surface of the tongue.

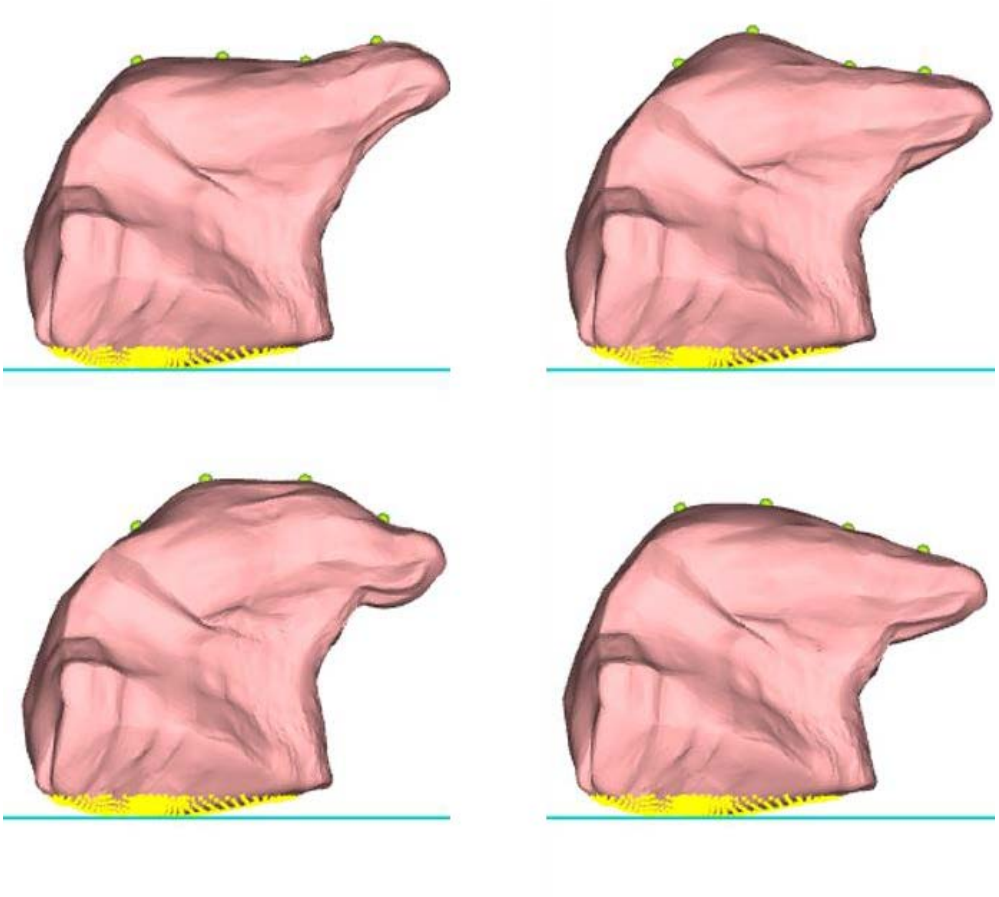


Figure 5-7 Sample frames in the 3D tongue shape dataset.

Step 2: Database Construction. Each constraint node on the 3D tongue model has 2 degrees of freedom. At each time step, a constraint point will be assigned a random displacement along the X-axis and Y-axis in the midsagittal plane. Because the movement of the tongue is smooth, we set up an upper threshold to the magnitude of the displacement so as to eliminate any discontinuous deformation. The 3D tongue model will then generate different tongue shapes, which are used to construct a 3D tongue shape database (some samples from the dataset are given in Figure 5-7.). As the displacement is random, some unphysical 3D tongue shapes will be generated, which will be discarded manually. For every 3D tongue shape in the database, a contour can be extracted from the model by using the nodes lying on the surface between the starting node and ending node in the mid-sagittal

plane. As the movement of the tongue can be viewed as symmetric, the 3D contours from the database can be projected into the mid-sagittal 2D plane, and compared to the target curve extracted from the 2D ultrasound image. In our experiment, the number of 3D sample tongue shapes in the database is 1000 presently.

Step 3: Contour Extraction. The method proposed in Section 4-2 is used to extract the contour in the ultrasound tongue image (as shown in Figure 5-6 (b)).

Step 4: Similarity Measurement. A measurement is made of the similarity between the contour extracted from the ultrasound image and the 2D contours projected from the 3D tongue shapes in the database. The definition of the similarity error is the mean sum of distances (MSD), which is defined in previous sections. The smaller the MSD error is, the better the similarity. Here, we review the detailed definition of MSD:

$$\mathrm{MSD}(V_1, V_2) = \frac{1}{2n}\left(\sum_{i=1}^{n} \min \left\| v_i^1 - v_j^2 \right\| + \sum_{i=1}^{n} \min \left\| v_i^2 - v_j^1 \right\|\right) \tag{5.10}$$

where V_1 is the contour extracted from the image and V_2 is the contour extracted from the 3D tongue shape in the database, v_i^1, v_i^2 are the elements of the contour V_1 and V_2 respectively. Here n is the number of the elements of the contours (In our experiment, $n = 12$). Four constraint points generate V_2, while 12 points are selected to represent V_1. Consequently, to make the MSD measurement feasible, V_2 is re sampled equidistantly to keep the number of elements in the two contours the same.

During simulations, very small distances between constraint points were found to generate pathological curves. To retain smoothness in the tongue model, a penalty term was therefore added to the MSD error, defined as follows:

$$P = \sum_{i=2}^{m}\left(\frac{1}{\left\| v_i^2 - v_{i-1}^2 \right\|}\right) \tag{5.11}$$

where m is the number of constraint nodes before re-sampling (here m is set as 4) and v_i^2 is the i^{th} constraint node. The overall objective function is now given as:

$$l = \alpha\left(1/\mathrm{MSD}(V_1, V_2)\right) + \beta \times (1/P) \tag{5.12}$$

where α and β are the weighting parameters (in our experiment, $\alpha = 0.8$ and $\beta = 0.2$).

At each time-step, this contour-based 3D deformation method is implemented to measure the similarity of the contour extracted from 2D image and the contours projected from the 3D tongue shape. The most similar 3D tongue shape (the biggest l) will be selected to represent the target curve shape associated with the ultrasound frame.

The key reason for selecting the contour similarity measurement to create an association between the 2D ultrasound image and 3D tongue model is that, compared to ultrasound image similarity measurements or other similarity measurements using a 3D tongue model, measuring the similarity between 2D curves is of high efficiency. At the same time, although motion feature extraction from ultrasound tongue image still has difficulties, the contour extraction method is fairly robust in comparison with tissue points tracking method (or speckle tracking).

5.4.2 Experimental results

The most time consuming step in our framework is the construction of the 3D tongue shape database, which was completed offline. The average processing time to build the association between current ultrasound frame and the 3D tongue model is about 1.2 seconds on our platform.

Here we select only four constraint nodes to drive the motion of the tongue on the 3D model's surface. In fact, the displacements of the constraint nodes must in reality be coupled since the tongue is a muscle-activated organ. However, the couple-relation is difficult to model. The compromise here is to use only four nodes to drive the model, with each node regarded as being independent of the others. Nevertheless, the deformation simulated with the proposed framework is informative and qualitatively realistic. Figure 5-8 presents some results of the visualization platform on different vocalizations.

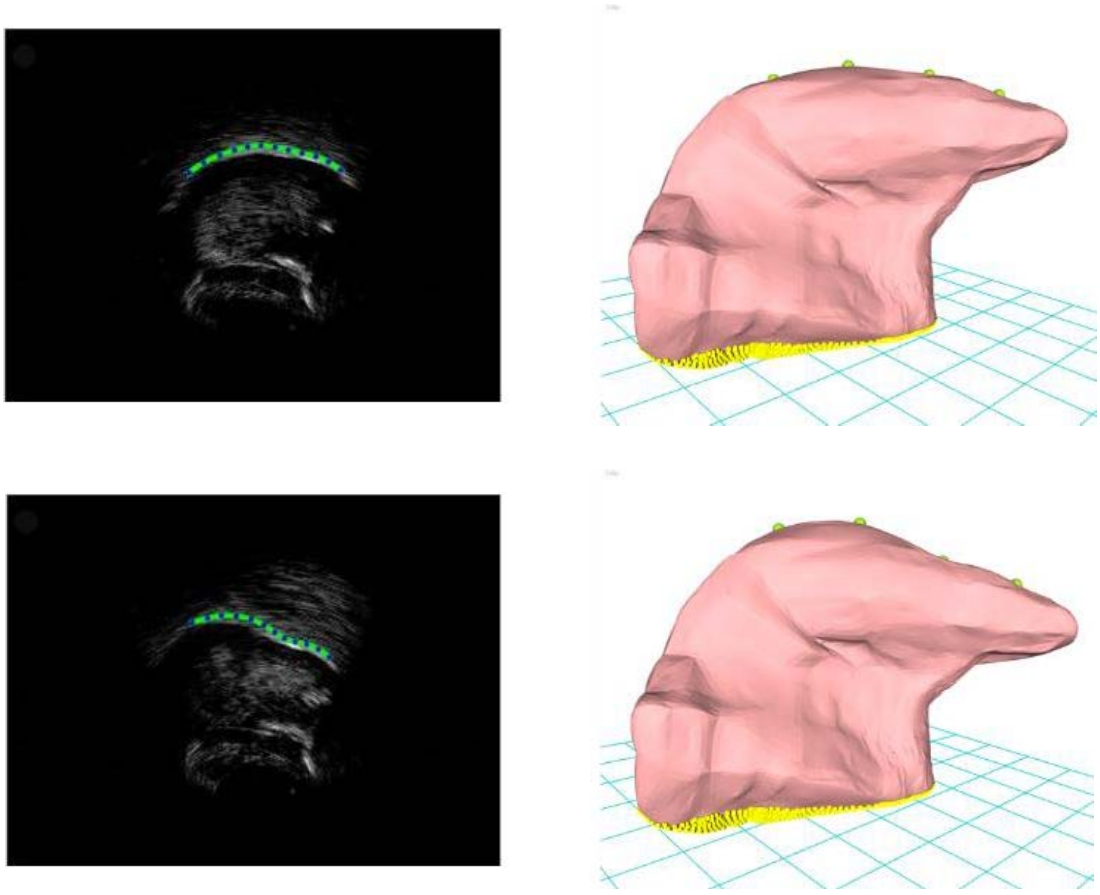


Figure 5-8 Sample frames of 3D tongue modeling. The ultrasound images are given in the left column. The meaning of the color line and points is the same as Fig. 1. The 3D tongue shapes are given in the right column, which are selected from the 3D tongue database based on the method proposed in section 4.

As there is no effective quantitative evaluation method for the 3D tongue motion visualization presently, to further demonstrate the feasibility of the proposed method, the midsagittal plane of the 3D tongue model can be extracted from the model after the deformation. If the midsagittal contour of the 3D model can be fit to the ultrasound image, the effectiveness of the method will be validated. Figure 5-9 gives some sample results, which demonstrate performance by visual observation.

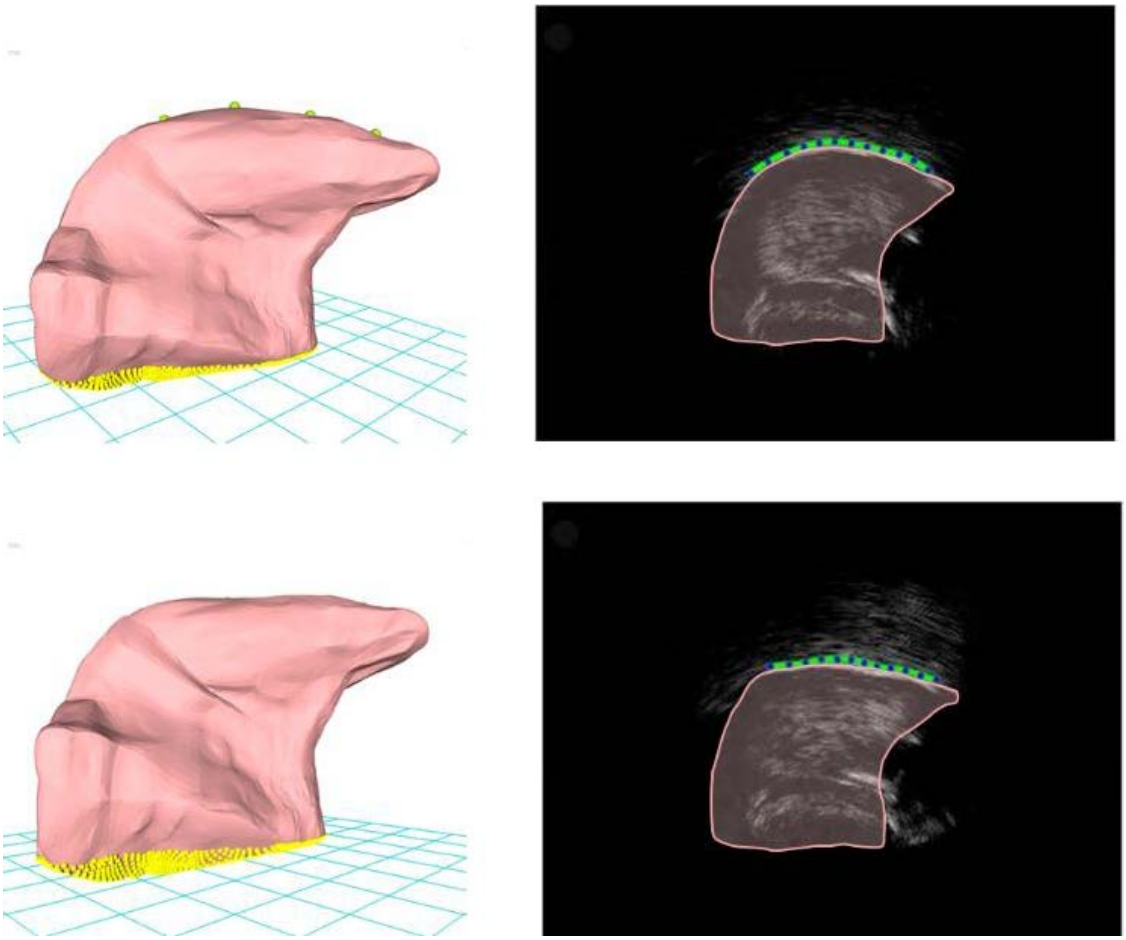


Figure 5-9 Validation for the proposed method for 3D tongue modeling. The left column gives the 3D tongue model, while the right column gives the ultrasound tongue image with tongue extracted (the green lines denote the contour extracted). The midsagittal planes of the 3D tongue model are placed over the ultrasound tongue images in transparency.

5.5 Conclusion

In this section, we first described the techniques used for the 3D tongue modeling, which include model reduction and model warping. Then speckle tracking-based tongue modeling method is given, which transits the displacements of the speckles to drive the 3D tongue model, but, as mentioned earlier, the speckle tracking error will be accumulated. Although real-time modeling can be achieved, but some unrealistic deformation was generated, thus an alternative method is needed to modeling the motion of the tongue.

Contour tracking is more stable to characterize the motion of the tongue, but the registration between the contour in 2D and 3D tongue model is of great difficulty. In this chapter, we explored how to convert this problem to “3D shape retrieve” problem. Firstly, we build a 3D tongue model database, which consists of 1000 different shapes of the tongue. Each different tongue shape is described by the assigned displacements for the constraint nodes. After contour tracking, the similarity is calculated between the curves tracked in the ultrasound tongue image, and the curve on the midsagittal plane in the 3D tongue models. For each time step, by search for the most similar shapes in the tongue base, a more realistic 3D tongue dynamic modeling is achieved.

Chapter 6
Conclusions

6.1 Conclusions

Speech is perhaps the most important human bio-signal, but not all of the characteristics of speech production are fully understood. Ultrasound imaging provides such a tool to improve the interpretation of articulator configurations by providing the real-time human tongue movement. Using ultrasound tongue images, most of the traditional approaches to the analysis of the tongue motion (or deformation) stays in the 2-D dimension, which ignored lots of the motion information. As a new tool for understanding speech production, the proposed 3D tongue motion visualization platform has been developed, based on ultrasound images, using modal analysis and model warping to perform the simulation in real time. We believe this to be the first combination of ultrasound imaging with a 3D tongue model to visualize the motion in real time.

The organization is based on our attempts to model the tongue motion using B-mode ultrasound tongue images in 3D, which can be roughly divided into two main parts: tongue motion tracking part and 3D tongue dynamic modeling part.

Tongue motion tracking: Accurate, robust tongue motion tracking remains a challenging problem for ultrasound sequences of long duration, due to acoustic effects, speckle noise and poor signal-to-noise-ratio (SNR). In this thesis, to characterize the motion of the tongue, different tracking methods were tested, which can be divided into two kinds of methods: speckle tracking and contour tracking. For speckle tracking, deformation registration, optical flow and local invariant feature-based motion tracking methods were tested on the ultrasound tongue image sequence. By tracking the “tissue point” lied on the tongue surface, a comparative study was conducted to evaluate the performance of different speckle tracking methods in the ultrasound tongue images. Moreover, an image similarity-

based speckle tracking re-initialization method is proposed to improve the robustness of the tracking. The tracking results demonstrate that speckle tracking can provide point correspondence information with relatively high robustness, but the tracking error will be accumulated during the tracking processing, and the tracking is sometimes unstable.

Compared to speckle tracking, contour tracking can provide superior performance to follow the surface of the tongue in ultrasound tongue images. However, despite significant research efforts, long duration contour tracking still poses a challenge in ultrasound tongue images. Most previous efforts aimed to extract the contour from single frame, without taking the temporal information into accounts. However, as the deformation of the tongue is physical, prior motion information can be helpful to extract the contours. In this thesis, we tested a temporal regularization method as the prior information, to guide the contour tracking. Based on which, we obtain relatively robust contour tracking in the ultrasound tongue images.

Both speckle tracking and contour tracking have their own advantages and disadvantages. For speckle tracking, the tracking performance is lower as it is not always stable during the whole sequence. However, this kind of method can provide extra points correspondences information, which is vital for the motion-driven 3D tongue modeling. Contour tracking gives better performance to characterize the motion of the tongue. Nevertheless, the contour cannot provide much point correspondence information.

3D tongue dynamic modeling: Based on the motion tracking part, a general framework is presented for the 3D tongue motion modeling. Both speckle tracking-based and contour tracking-based modeling methods were explored in this thesis.

For speckle-tracking-based 3D tongue modeling method: Speckle tracking-based modeling simulated the motion of the tongue in real-time, while retained relative realistic deformation in most cases. However, as the tracking error accumulated during the tracking processing, some unrealistic deformations occurred during our experiment.

For contour-tracking-based 3D tongue modeling method: As contours cannot provide tissue point correspondence information, the registration between the 2D tongue image and the 3D tongue model is a great challenge. In this thesis, by converting the problem into a “3D shape retrieve” problem, we can avoid the registration processing step. The modeling work can be divided into three main modules: 1) 3D tongue shapes database construction; 2)

Contour extraction from the B-mode ultrasound tongue image; 3) Similarity measurement between the contour extracted from 2D ultrasound image and the contours projected from the 3D tongue shapes. Based on this kind of method, a more realistic 3D tongue motion visualization is achieved.

6.2 Perspectives

The study has offered a promising approach to simple but practical 3D tongue motion visualization solution using ultrasound tongue image sequences. However, it also encountered certain limitations, as well as brought to light a number of new ideas, both of which could be addressed in our future work.

- Firstly, for speckle tracking-based modeling method, the registration between 2D points in the ultrasound image and the nodes on the 3D tongue model is manual tweaking, while a more robust and feasible registration method is desired.
- Secondly, for contour-based modeling approach, the MSD error measurement may not be the optimal choice to measure the similarity between curves, and a more specific measurement may need to be developed. Furthermore, there are non-midsagittal motions (or out-plane motions) of the tongue, and employing the motion information from the midsagittal plane only is not enough to generate fully accurate tongue shapes.
- Thirdly, during speech production, the tongue should be volume-preserved. However, although the volume change is small during the motion modeling, a more explicit volume-preserving constraint needs to be added.
- Fourthly, the performance of the tongue motion visualization framework still need to be evaluated quantitatively by making use of other imaging modalities such as MRI and EMA.

- Lastly, we would like to go back to the "Silent Speech Interface" concepts. Presently, the quality of 3D tongue motion modeling is still far from been used to improve the performance of the SSI systems, although the platform developed can provide extra visual information to understand the speech production. But, we believe that, with the advances in 3D ultrasound imaging of the tongue, other imaging modality, image processing techniques and physics-based modeling techniques, 3D tongue motion modeling can provide extra quantitative information, which can be helpful for the SSI systems.

Publications

K. Xu, Y. Yang, A. Jaumard-Hakoun, M. Adda-Decker, A. Amelot, S. Al Kork, L. Crevier-Buchman, P. Chawah, G. Dreyfus, T. Fux, C. Pillot-Loiseau, P. Roussel, M. Stone, B. Denby. 3D tongue motion visualization based on ultrasound image sequences, *The 15th Annual Conference of the International Speech Communication Association (InterSpeech),* Singapore, 2014.

S. Al Kork, D. Ugurca, C. Sahin, P. Chawah, L. Crevier-Buchman, M. Adda-Decker, **K. Xu**, B. Denby, P. Roussel, B. Picart, S. Dupont, F. Tsalakanidou, A. Kitsikidis, F. Maria Dagnino, M. Ott, F. Pozzi, M. Stone, E. Yilmaz, A Multi-Sensor Helmet to Capture Rare Singing, An Intangible Cultural Heritage Study, *The 10th International Seminar on Speech Production (ISSP)*, Cologne, Germany, 2014.

P. Chawah, S. Al Kork, T. Fux, M. Adda-Decker, A. Amelot, N. Audibert, B. Denby, G. Dreyfus, A. Jaumard-Hakoun, C. Pillot-Loiseau, P. Roussel, M. Stone, **K. Xu**, L. Crevier-Buchman, An educational platform to capture, visualize and analyze rare singing. *The 15th Annual Conference of the International Speech Communication Association (InterSpeech)*, Singapore, 2014.

E. Yilmaz, D. Ugurca, C. Sahin, F. M. Dagnino, M. Ott, F. Pozzi, K. Dimitropoulos, F. Tsalakanidou, A. Kitsikidis, S. K. Al Kork, **K. Xu**, B. Denby, P. Roussel, P. Chawah, L. Buchman, M. Adda-Decker, S. Dupont, B. Picart, J. Tilmanne, M. Alivizatou, L. Hadjileontiadis, V. Charisis, A. Glushkova, C. Volioti, A. Manitsaris, E. Hemery, F. Moutarde, N. Grammalidis, Novel 3D Game-like Applications driven by body interactions for learning specific forms of Intangible Cultural Heritage, *The 10th International Conference on Computer Vision Theory and Applications (VISAPP2015)*, Berlin, Germany, 2015.

S. Al Kork, D. Ugurca, C. Sahin, P. Chawah, L. Buchman, M. Adda-Decker, **K. Xu**, B. Denby, P. Roussel, B. Picart, S. Dupont, F. Tsalakanidou, A. Kitsikidis, F. Maria Dagnino, M. Ott, F. Pozzi, M. Stone, E. Yilmaz, A novel human interaction game-like application to learn, perform and evaluate modern contemporary singing: "Human

Beat Box", *The 10th International Conference on Computer Vision Theory and Applications (VISAPP2015)*, Berlin, Germany, 2015.

L. Crevier-Buchman, T. Fux, A. Amelot, S. K. Al Kork, M. Adda-Decker, N. Audibert, P. Chawah, B. Denby, G. Dreyfus, A. Jaumard-Hakoun, P. Roussel, M. Stone, J. Vaissiere, **K. Xu**, and C. Pillot-Loiseau, Acoustic data analysis from multi-sensor capture in rare singing: Cantu in Paghjella case study, International Journal of Heritage in the Digital Era, 4(1), 2015

K. Xu, Y. Yang, A. Jaumard-Hakoun, C. Leboullenger, G. Dreyfus, P. Roussel, M. Stone, B. Denby, Development of a 3D tongue motion visualization platform based on ultrasound image sequences, *The 18th International Congress of Phonetic Sciences (ICPhS, 2015)*, Glasgow, Scotland. (Oral presentation)

A. Jaumard-Hakoun, **K. Xu**, P. Roussel, G. Dreyfus, B. Denby. Tongue contour extraction from ultrasound images based on deep neural network, *The 18th International Congress of Phonetic Sciences (ICPhS, 2015)*, Glasgow, Scotland.

K. Xu, Y. Yang, M. Stone, A. Jaumard-Hakoun, G. Dreyfus, P. Roussel, B. Denby. Robust contour tracking in the ultrasound tongue image sequences, *Clinical Linguistics and Phonetics*. 30(3-5), 2016

X. Wang, Y. Yang, J. Shi, Y. Zeng, **K. Xu**. A novel hybrid mobile malware detection system integrating anomaly detection with misuse detection, *The 6th ACM International Workshop on Mobile Cloud Computing and Services (MSC@MobiCom 2015)*, Paris, France. (Oral presentation)

K. Xu, Y. Yang, C. Leboullenger, P. Roussel, B. Denby. Contour-based 3D tongue motion visualization using the ultrasound image sequences, *41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*. Shanghai, China

K. Xu, T. Gábor Csapó, P. Roussel, B. Denby, A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization. *The Journal of the Acoustical Society of America*. 139 (5), EL154-EL160, 2016

A. Jaumard-Hakoun, **K. Xu**, C. Leboullenger, P. Roussel-Ragot, B. Denby. An articulatory-based singing voice synthesis using tongue and lips imaging. *17th Annual Conference of the International Speech Communication Association (InterSpeech)*. 2016

B. Denby, Z. Gan, Y. Wan, Y. Bao, C. Leboullenger, **K. Xu**, P. Roussel, Y. Yang. Towards a Kinect for the Tongue. *The 4th International Workshop on Biomechanical and Parametric Modeling of Human Anatomy (PMHA)*. Vancouver, Canada, 2016

K. Xu, Y. Yang, A. Jaumard-Hakoun, C. Leboullenger, P. Roussel, B. Denby. Is speckle tracking feasible in ultrasound tongue images? *Submitted to 42th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, USA, 2017

K. Xu, T. Gábor Csapó, P. Roussel, B. Denby. Convolutional neural network-based automatic classification of midsagittal tongue gestures using B-mode ultrasound images. Submitted to *The Journal of the Acoustical Society of America*

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert and J. Brumberg, "Silent speech interface," *Speech Communication,* vol. 52, no. 4, pp. 270-287, 2010.

[2] T. Hueber, E. L. Benaroya, G. Chollet, B. Denby, G. Dreyfus and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication,* vol. 52, no. 4, pp. 288-300, 2010.

[3] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics & Phonetics,* vol. 19, no. 6-7, pp. 455-501, 2005.

[4] N. Shrikanth, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America,* vol. 136, no. 3, pp. 1307-1311, 2014.

[5] "USC-TIMIT," [Online]. Available: http://sail.usc.edu/span/usc-timit/.

[6] J. Westbury, P. Milenkovic, G. Weismer and R. Kent, "X-ray microbeam speech production database," *The Journal of the Acoustical Society of America,* vol. 88, no. S1, p. S56, 1990.

[7] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *Proceedings 5th Seminar of Speech Production*, Kloster Seeon, Bavaria, Germany, 2000.

[8] "MOCHA-TIMIT," [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

[9] Y. S. Akgul, C. Kambhamettu and M. Stone, "Automatic extraction and tracking of the tongue contours," *IEEE Transactions on Medical Imaging,* vol. 18, no. 10, pp. 1035-1045, 1999.

[10] M. Li, C. Kambhamettu and M. Stone, "Automatic contour tracking in ultrasound images," *Clinical Linguistics and Phonetics,* vol. 19, no. 6-7, pp. 545-554, 2005.

[11] A. Roussos, A. Katsamanis and P. Maragos, "Tongue tracking in ultrasound images with active appearance models," in *16th IEEE International Conference on Image Processing*, Cairo, Egypt, 2009.

[12] I. Fasel and J. Berry, "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech," in *20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010.

[13] D. Fabre, T. Hueber, F. Bocquelet and P. Badin, "Tongue Tracking in Ultrasound Images using EigenTongue Decomposition and Artificial Neural Networks," in *InterSpeech*, Dresden, Germany, 2015.

[14] A. H. Jaumard, K. Xu, P. Roussel, G. Dreyfus and B. Denby, "Tongue contour extraction from ultrasound images based on deep neural network," in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland, 2015.

[15] L. Tang, T. Bressmann and G. Hamarneh, "Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves," *Medical image analysis,* pp. 1503-1520, 8 16 2012.

[16] G. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, no. 5786, pp. 504-507, 2006.

[17] M. Stone, "A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data," *The Journal of the Acoustical Society of America,* vol. 87, no. 5, pp. 2207-22017, 1990.

[18] R. W. Tricarico, "Physiological modeling of speech production: Methods for modeling soft-tissue articulators," *The Journal of the Acoustical Society of America,* vol. 97, no. 5, pp. 3085-3098, 1995.

[19] O. Engwall, "Vocal tract modeling in 3D," KTH TMH-QPSR, 1999.

[20] O. Engwall, "A 3d tongue model based on MRI data," in *Interspeech*, Beijing, China, 2000.

[21] O. Engwall, "Using linguopalatal contact patterns to tune a 3d tongue model," in *InterSpeech*, Aalborg, Denmark, 2001.

[22] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," *Speech Communication,* vol. 41, no. 2, pp. 303-329, 2003.

[23] I. Stavness, J. E. Lloyd and S. Fels, "Automatic prediction of tongue muscle activations using a finite element model," *Journal of biomechanics,* vol. 45, no. 16, pp. 2841-2848, 2012.

[24] J. E. Lloyd, I. Stavness et S. Fels, «ArtiSynth: A Fast Interactive Biomechanical Modeling Toolkit Combining Multibody and Finite Element Simulation,» chez *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, Berlin, Germany, Springer Berlin Heidelberg, 2012, pp. 355-394.

[25] Y. Yang, X. Guo, J. Vick, L. G. Torres and T. F. Campbell, "Physics-based deformable tongue visualization," *IEEE Transactions on Visualization and Computer Graphics,* vol. 19, no. 5, pp. 811-823, 2013.

[26] K. Martin, "Introduction to B-mode imaging," in *Diagnostic ultrasound: physics*

and equipment, Cambridge, UK, Cambridge University Press, 2010, pp. 1-22.

[27] M. Tanter and M. Fink, "Ultrafast imaging in biomedical ultrasound," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control,* vol. 61, no. 1, pp. 102-119, 2014.

[28] M. Stone, "Preface to the Special Issue on Ultrasound Imaging of the Tongue," *Clinical Linguistics & Phonetics,* vol. 19, no. 6-7, p. 453–454, 2005.

[29] J. Cleland, J. Scobbie and N. Zharkova, "Insights from ultrasound: Enhancing our understanding of clinical phonetics," *Clinical Linguistics & Phonetics,* vol. 30, no. 3-5, pp. 171-173, 2016.

[30] T. L. Szabo, Diagnostic ultrasound imaging: inside out, Academic Press, 2004.

[31] "Paul_Langevin_wiki," [Online]. Available: https://en.wikipedia.org/wiki/Paul_Langevin.

[32] L. V. Wang and S. Hu, "Photoacoustic tomography: in vivo imaging from organelles to organs," *Science,* vol. 335, no. 6075, pp. 1458-1462, 2012.

[33] S. C. Manion and J. P. Rathmell, Atlas of ultrasound-guided procedures in interventional pain management, New York: Springer-Verlag, 2010.

[34] J. Curie and P. Curie, "Développement, par pression, de l'électricité polaire dans les cristaux hémièdres à faces inclinées," *Comptes Rendus de l'Académie des Sciences,* 1883.

[35] K. Martin, Diagnostic ultrasound: physics and equipment, Cambridge, UK: Cambridge University Press, 2010.

[36] B. Cohen and I. Dinstein, "Motion estimation in noisy ultrasound images by maximum likelihood," in *15th International Conference on Pattern Recognition*, Barcelona, Spain, 2000.

[37] B. Denby, J. Cai, T. Hueber, P. Roussel, G. Dreyfus, L. C. Buchman, C. P. Loiseau, G. Chollet, S. Manitsaris and M. Stone, "Towards a practical silent speech interface based on vocal tract imaging," in *International Seminar on Speech Production*, Montreal, Canada, 2011.

[38] B. H. Amundsen, T. H. Valle, T. Edvardsen, H. Torp, J. Crosby, E. Lyseggen and A. Støylen, "Noninvasive myocardial strain measurement by speckle tracking echocardiography: validation against sonomicrometry and tagged magnetic resonance imaging," *Journal of the American College of Cardiology,* vol. 47, no. 4, pp. 789-793, 2006.

[39] T. Helle-Valle, J. Crosby, T. Edvardsen, E. Lyseggen, B. H. Amundsen, H.-J. Smith and B. D. Rosen, "New noninvasive method for assessment of left ventricular rotation speckle tracking echocardiography," *Circulation,* vol. 112, no. 20, pp. 3149-3156, 2005.

[40] M. Jacob, H. L. LeHouillier, S. Bora, S. McAleavey, D. Dalecki and J. McDonough, "Speckle tracking for the recovery of displacement and velocity information from sequences of ultrasound images of the tongue," in *The 8th International Seminar on Speech Production*, Strasbourg, France, 2008.

[41] J. D'hooge, «Principles and different techniques for speckle tracking,» chez *Myocardial imaging: tissue Doppler and speckle tracking*, Wiley, 2007, pp. 17-25.

[42] C. Kontogeorgakis, M. G. Strintzis, N. Maglaveras and I. Kokkinidis, "Tumor detection in ultrasound b-mode images through motion estimation using a texture detection algorithm," in *Computers in Cardiology*, 1994.

[43] F. Yeung, S. F. Levinson and K. J. Parker, "Multilevel and motion model-based ultrasonic speckle tracking algorithms," *Ultrasound in medicine & biology,* vol. 24, no. 3, pp. 427-441, 1998.

[44] J. F. Krucker, G. L. LeCarpentier, B. J. Fowlkes and P. L. Carson, "Rapid elastic

image registration for 3-D ultrasound," *IEEE Transactions on Medical Imaging,* pp. 1384-1394, 11 21 2002.

[45] T. C. Poon and N. R. Robert, "Three-dimensional extended field-of-view ultrasound," *Ultrasound in medicine & biology,* vol. 32, no. 3, pp. 357-369, 2006.

[46] M. G. Strintzis and I. Kokkinidis, "Maximum likelihood motion estimation in ultrasound image sequences," *IEEE Signal Processing Letters,* vol. 4, no. 6, pp. 156-157, 1997.

[47] B. Cohen and I. Dinstein, "New maximum likelihood motion estimation schemes for noisy ultrasound images," *Pattern Recognition,* vol. 35, no. 2, pp. 455-463, 2002.

[48] B. K. Horn and B. G. Schunck, "Determining optical flow," in *International Society for Optics and Photonics Technical Symposium East*, 1981.

[49] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, Vancouver, BC, Canada, 1981.

[50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,* vol. 60, no. 2, pp. 91-110, 2004.

[51] C. Liu, Y. Jenny and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 5, pp. 978-994, 2011.

[52] Z. Wang, A. Bovik, S. Hamid and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* vol. 13, no. 4, pp. 600-612, 2004.

[53] M. P. Sampat, Z. Wang, S. Gupta, A. Bovik and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions*

on Image Processing, vol. 18, no. 11, pp. 2385-2401, 2009.

[54] A. Karasaridis et E. Simoncelli, «A filter design technique for steerable pyramid image transforms,» chez *IEEE International Conference on Acoustics Speech and Signal Processing*, 1996.

[55] A. A. Wrench et P. Balch, «Towards a 3D Tongue model for parameterising ultrasound data,» chez *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland, 2015.

[56] X. Zhou, X. Huang, J. Duncan et W. Yu, «Active contours with group similarity,» chez *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, 2013.

[57] E. J. Candès, X. Li, Y. Ma and J. Wright, "Robust principal component analysis?," *Journal of the ACM,* vol. 58, no. 3, p. 11, 2011.

[58] Y. Nesterov, "Gradient methods for minimizing composite objective function," CORE Discussion Papers, 2007.

[59] T. G. Csapó and S. M. Lulich, "Error analysis of extracted tongue contours from 2D ultrasound images," in *Sixteenth Annual Conference of the International Speech Communication Association*, Dreston, Germany, 2015.

[60] M. Gérard, J. Ohayon, V. Luboz, P. Perrier and Y. Payan, "Indentation for estimating the human tongue soft tissues constitutive law: application to a 3D biomechanical model," in *Medical Simulation*, Berlin Heidelberg, Springer, 2004, pp. 77-83.

[61] M. G. Choi and H. S. Ko, "Modal warping: Real-time simulation of large rotational deformation and manipulation," *IEEE Transactions on Visualization and Computer Graphics,* vol. 11, no. 1, pp. 91-101, 2005.

[62] Y. Yang, G. Rong, L. Torres and X. Guo, "Real-time hybrid solid simulation:

spectral unification of deformable and rigid materials," *Computer Animation and Virtual Worlds,* vol. 21, no. 3-4, pp. 151-159, 2010.

[63] S. Al Kork, A. Jaumard-Hakoun, M. Adda-Decker, A. Amelot, L. Buchman, P. Chawah, G. Dreyfus, T. Fux, C. Pillot-Loiseau, P. Roussel, M. Stone, K. Xu et B. Denby, "A Multi-Sensor Helmet to Capture Rare Singing, An Intangible Cultural Heritage Study", *International Seminar on Speech Production*, Cologne, Germany, 2014.

[64] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision,* vol. 1, no. 4, pp. 321-331, 1988.