Construction of Midsagittal Vocal Tract Videos from CT, Ultrasound, and Motion Capture Data

Toshifumi Masuda s1120204 Supervised by Professor Ian Wilson

Abstract

This paper describes the creation of a movie file of an English speech sample from the Speech Accent Archive. The movie was created by combining ultrasound movies of the tongue, CT images of the vocal tract, and Vicon motion capture data from the skull and jaw. This movie shows all parts of the vocal tract that are related to speech and were recorded successfully. The movie is beneficial in that it places the ultrasound tongue movie into the vocal tract context. When the subject pronounces t, the ultrasound tongue line clearly makes contact with the alveolar ridge from the CT image. The movie is anticipated to be helpful to English as a Second Language learners in the acquisition of pronunciation, and future research should confirm the degree to which it is useful.

1 Introduction

With a constantly increasing number of people who are using foreign languages in daily life, the number of people trying to acquire pronunciation skills has also increased. Precise pronunciation of languages is often important for communication because the wrong pronunciation can cause misunderstanding.

The use of an ultrasound machine has been shown to be helpful in improving pronunciation [1]. It can provide real-time images of both the tongue tip and the tongue root simultaneously, so people can understand easily their movement. It helps to review and check tongue movement, which is the most important of the speech articulators, and this equipment can provide training to make articulation more precise [2].

Ultrasound is non-invasive and safe for the human body as it is used on pregnant women to view the fetus. Without using this equipment, it is possible to see the tongue tip using a mirror, but without ultrasound, the only way to see the tongue root moving during speech is through the use of x-ray, which is now illegal for research in most countries. As is now well known, X-ray radiation causes some problems for the human body, and if it was used for tongue viewing, the human brain would suffer.

Although the cost of an ultrasound machine has been decreasing rapidly, it is still too expensive to purchase and use personally and so most pronunciation learners have not yet tried this method of learning.

The purpose of this research is to develop movies that help learners understand all the parts of the tongue moving in the ultrasound images, and that show the tongue moving in the context of the rest of the vocal tract. In an ultrasound image alone, the surface of the tongue is visible, but the images are not easy to interpret, especially for people who are not used to seeing them. People can recognize the tongue tip and root easily, but it is difficult for people to understand where the tongue is in relation to other parts of the vocal tract (e.g., the hyoid bone, the mandible, and the palate). Using ultrasound images alone, it is probably difficult for learners to practice pronunciation because they cannot see the jaw, lips and other articulators that are moving, even if the images are real-time. Furthermore, it is anticipated that this research will also benefit phonetics researchers who are investigating speech using ultrasound. The palate contour has the potential for disambiguation of the tongue surface, registration of images within and across subjects, and phonetic calculations [3].

Finally, it is also expected that the movies generated from this research will also be beneficial in the teaching and learning of phonetics. This paper describes, in detail, the steps in making a speech movie using a CT image, motion capture data, and ultrasound data.

In past research, a model of tongue movement was developed by Stone [3], who used two distinct measurement techniques: ultrasound imaging of the tongue and tracking of five pellets on the tongue surface using x-ray microbeam. One purpose of her research was to study the relationship between tongue and jaw movement. In other research, the palate contour was studied by using an ultrasound machine [4]. In that research, water placed in the mouth was used to express the palate in the ultrasound display. One limitation with both of the above studies is that they both involved putting some object inside the mouth. Although subjects could still speak, the tongue may have been moving unnaturally. An experiment involving the tongue in its natural context would perhaps get more realistic tongue tracking data.

2 Method

2.1 Subject & Apparatus

The subject whose data was used to create the movie is a Caucasian who was born and grew up in Ontario, Canada. He speaks English as a first language and Japanese as a second language.

In the data collection, a CT scanner (H1 in Table 1), an ultrasound machine with a probe (H2 and H3), motion capture optical tracking equipment (H4), a computer as a recorder (H5), and audio and video signal converters (H6 and H7) were used. A video camera (H8) was used to record the side view images.

Hardware List	
H1	CT Scanner - Imaging Sciences International iCAT
H2	Ultrasound Machine - Toshiba Famio 8 (SSA-530A)
H3	Ultrasound Probe - Toshiba (PVQ-381A)
H4	Motion Capture Optical Tracker - Vicon MX40
H5	Computer - Apple MacBook Pro (Mac OS X 10.4.9)
H6	Digital Video Converter - Canopus (ADVC-110)
H7	Pre-Amplifier - Audio-Technica (AT-MA2)
H8	Video Camera - JVC GZ-HD7

Table 1: Hardware used in data collection

2.2 Procedure

2.2.1 Data Collection

The CT data were collected in 2005 at the University of British Columbia when the subject was 38 years old and had been living in Canada for 5 years. H1 was used to capture 3 DICOM files: the subject holding the articulation for [ŋ], [k], and [t].

The ultrasound data and the motion capture data were collected in 2007 at the University of Aizu when the subject was 40 years old and had been living in Japan for about one and a half years.

Movie-editing software, iMovie (S1 in Table 2) was used to record the ultrasound images by using H6 and H7 to convert video and audio analog signals to digital. The resulting digital signal was recorded on H5 in DV format. For the ultrasound data collection, 11 markers were put on the subject's face, on the ultrasound probe, and on the glasses (GLU, GLR, GLC, GLL, NOS, JAW, PRU, PRD, CLU, CLD in Figure 1, as well as TOT, a tooth marker). The tooth marker was put on the subject's front upper tooth before the first trial, and this data was collected without the subject speaking. The stimuli paragraph was selected from the Speech Accent Archive [6], which is a website containing many sample audio files of people from various language backgrounds speaking the same paragraph. This paragraph includes almost all

Software List	
S1	Apple iMovie (6.0.3)
S2	Vicon Workstation
S3	OsiriX (2.7.5, 32bit)
S4	ImageJ (10.2)
S5	Apple Final Cut Express (3.5.1)
S6	Digital Image Converter (v3.7b2)
S7	Iconico Screen Protractor
S8	GIMP (2.4.3)
S9	Iconico Screen Compass
S10	The Math Works MATLAB (7.4.0)
S11	Adobe Photoshop CS3 Extended (10.0)
S12	Quick Time Player Pro (7.3.0)

Table 2: Software used in data manipulation

English phonemes and consonant clusters, in various positions in the syllable. In the CT image, the mandible, the tooth, and the nasal spine (a point immediately below the nose) are stable when the subject is resting. Those places could be defined to use when overlaying the ultrasound and CT images. Though there was no stable vocal tract reference point in the ultrasound images, the probe position was visible and could be tracked externally using H4. H4 was also used to record the tooth position before the first speaking trial. By calculating the relationship between the tooth and probe in two dimensions, the tooth position in the ultrasound image could be found. H4 recorded the motion of markers, and the motion capture software, Vicon Workstation (S2) exported x, y and z coordinates in CSM format. The video camera (H8) recorded lip movement from the side in high-definition DV format. Table 3 shows each type and format of the data collected.

Anatomy	Data Type	Format
Skull	CT	JPEG
Mandible	CT	JPEG
	Motion Capture	CSM
Tongue	Ultrasound	DV
Tooth	CT	JPEG
	Motion Capture	CSM
Nose	Motion Capture	CSM
	Motion Capture	CSM
TMJ	CT	JPEG
Lips	CT	JPEG
	Video	DV
	Motion Capture	CSM

Table 3: Type and format of collected data

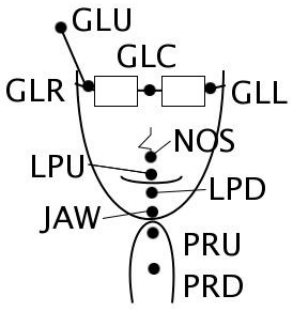

Figure 1: Marker positions

2.3 Data Manipulation

2.3.1 CT Image Preparation

In this research, the DICOM file for [k] sound was used because it had the clearest image of the nasal spine, a good reference point when using the NOS motion capture marker to superimpose the ultrasound image. In the [ŋ] sound DICOM file, the tongue was touching the palate and velum. The mandible and other low parts were ambiguous. As for the [t] sound, the mouth was not open very far. So to overlay and synchronize data, the [k] file was selected.

OsiriX (S3) was used to manipulate the CT DICOM data and export a midsagittal image of the head. Because the three axes (midsagittal, axial, and coronal) of the original DICOM files were slightly offset, S3 was used to rotate and move the three coordinates. In Multi Planar Reformat (MPR) two dimensional mode, the three axes could be moved, and an exact midsagittal image could be obtained (see Figure 2). The midsagittal line was determined from the front tooth, the spinal cord and the inferior nasal concha in the axial view. Then, the midsagittal image was exported as a JPEG file (see Figure 3).

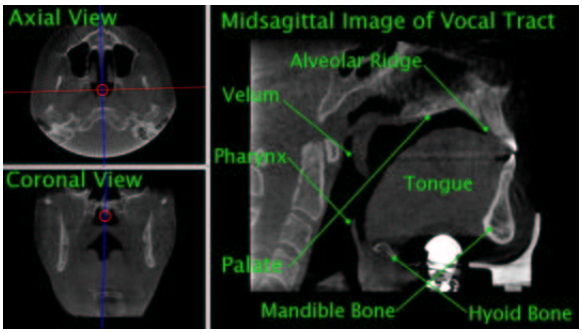

Figure 2: Three image views of the CT data

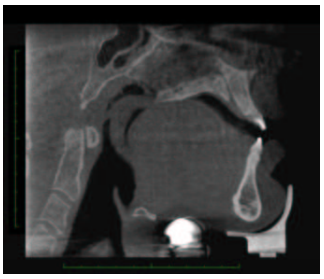

Figure 3: Exported CT image

The image-editing software, ImageJ (S4) was used to edit the image. Because the ultrasound data would be superimposed on the midsagittal CT image, the tongue part of the CT image had to be erased. Also, other parts in the midsagittal image were not used because their movement was not recorded in the data collection. Thus, the tongue, velum, epiglottis, hyoid bone, and the ultrasound probe was were erased from the midsagittal image (see Figure 4).

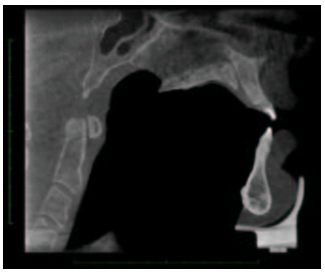

Figure 4: Edited CT image

2.3.2 Motion Capture Data Preparation and Evaluation

Although the data from both the motion capture and the CT scan were three dimensional, this data had to be reduced to two dimensions (the y (front-back) and z (up-down) axes, ignoring the x (right-left) axis) because the ultrasound image is only two dimensions.

There were three types of data collection using motion capture: the tooth trial had 13 markers, each of the 4 normal speech trials had 12 markers, and the probe measurement trial had 3 markers.

The angle between two lines (GLC-NOS and PRU-PRD) was calculated to evaluate these experiment data and select one (out of the four speaking trials) that had the least head movement relative to the ultrasound probe (see Table 4). Though the ultrasound probe was fixed on a microphone stand, and the subject was leaning his

head against a wall to limit the amount of probe movement relative to the head [2], the angle of the probe could have changed during the experiment, because the chin was pushing against the probe during speech. In the evaluation, if the angle between these lines changed much, the ultrasound data would not be reliable. The third trial had the lowest range of probe rotation (1.31 °), so that trial was used to make the movie.

Data	Max	Min	Range
1st	32.49 °	31.05 °	1.44 °
2nd	32.54 °	31.03 °	1.51 °
3rd	32.61 °	31.30 °	1.31 °
4th	32.66 °	31.28 °	1.38 °

Table 4: Angle between GLC-NOS and PRU-PRD

2.3.3 Ultrasound Data Preparation

The movie-editing software, Final Cut Express (S5) was used to export the sequential images from the ultrasound movie at 30 frames/second. The duration of the third trial was approximately 22 seconds, and 661 images (740×480) were obtained.

In using S1 to digitize the ultrasound video, a pixel aspect ratio (PAS) of 0.9:1 (DV NTSC) was used. However, not using square pixels resulted in a stretching of the scale of the ultrasound images in one dimension. Thus, the ultrasound images had to be corrected from 0.9:1 to 1:1 (square pixels). In fact, the images were converted from 740×480 to 656×480 by using Digital Image Converter (S6).

Since the tooth position was visible in both the ultrasound and the CT data, and fixed, it was defined in the ultrasound image to superimpose on the midsagittal. The probe position in the ultrasound image was clearly visible and stable, and so PRT, PRU, and PRD marker locations could be calculated. The PRU and PRD markers were in a relatively fixed geometrical relationship with the tooth marker.

According to the probe measurement trial, the angle between the global vertical PRT line and PRT-PRU was 145.79 °, and the angle between the former line and PRT-PRD was 161.38 °. These two lines were drawn using the screen angle measurement software, Screen Protractor (S7) and the image-editing software, GIMP (S8). The distances between PRT and PRU, and PRT and PRD, were 36.89 mm and 56.16 mm, respectively. Thus, the positions of PRU and PRD were found. In the speech trial, the angle between the global vertical PRU line and PRU-TOT was 24.48 °, and that between the global PRD line and PRD-TOT was 20.42 °. And the PRU-TOT and the PRD-TOT were respectively 75.69 mm and 97.79 mm. Then TOT was found by using the screen radius measurement software, Screen Compass (S9). At a result, the tooth marker was situated at x=496 and y=273 in the 656×480 image format.

To help to superimpose the ultrasound image on the CT image, 4 pixels around x=496 and y=273 in the image were marked in red. Although the probe was fixed on a microphone stand during data collection, the angle changed slightly over time. To correct for this probe movement, MATLAB (S10) was used to calculate the probe rotation for each frame and each ultrasound image was rotated by the respective angle using the MATLAB Image Processing Toolbox. Finally, the rotated ultrasound images (777×615) were edited as shown in Figure 5.

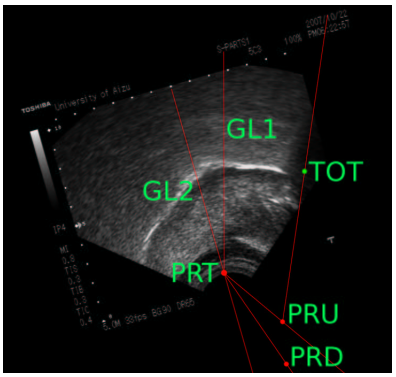

Figure 5: Geometrical relationship between probe and tooth

2.4 Results of Overlaying Images

The three types of prepared data were used to overlay the ultrasound and CT images. This overlay was done using image-editing software, Photoshop (S11). A diamond-shaped tongue image was cropped from the 661 ultrasound images, and superimposed on the midsagittal CT image, as seen in Figure 6. Then, each superimposed image was imported to S5, and then all images were exported at 30 frames/second as a 22-second movie file in MOV format.

3 Discussion

For pronunciation learners who want to improve their English pronunciation, but who do not have an opportunity to use ultrasound for speech training, the movie file produced here is helpful in understanding how one native English speaker's tongue moves during speech. The movie places the ultrasound tongue images in the context of the vocal tract. Learners can see the movement

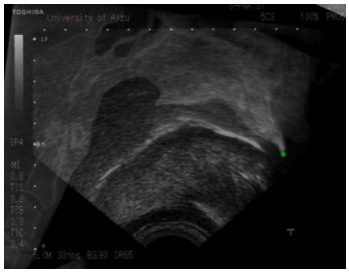

Figure 6: Ultrasound tongue overlaid on CT scan

of the tongue root relative to the tongue tip and use this knowledge with proprioceptive knowledge of articulator positions when they are practicing pronunciation.

An advantage of this movie is that it shows the degree and location of tongue contact with the palate, including the alveolar ridge. This type of information is not easily obtained in natural speech. For /t/, the tongue was clearly touching the alveolar ridge, and for /r/, the tongue was retracted. This type of information is anticipated to be helpful for learners producing English sounds.

However, this movie also shows some unnatural movement. In some frames, the tongue appears to penetrate the palate. There are two reasons for this. First, the tooth location was not necessarily correct in all of ultrasound frame images. Although the probe was rotating slightly during speech (at most 1.31 °), the tooth location was assumed to be fixed in the ultrasound image. This assumption was necessary to simplify the data manipulation process, but it created some inaccuracy in the resultant movie. Second, the CT image was a still image, even though the soft palate could be moving and could be pushed by the tongue. It was not possible to track the soft palate in this research because of a lack of data collection apparatus. Thus, some frames show the tongue penetrating the soft palate.

Because the motion capture data allowed the ultrasound images to be superimposed on the midsagittal CT image, the resultant movie could show a detailed view of inside the mouth without using invasive equipment. This movie is also useful for ultrasound phonetics researchers to understand the effect on the ultrasound images when the tongue makes contact with the palate.

During data collection, the movement of the velum and the hyoid bone were not recorded because they could not be seen using the available apparatus. Therefore, in the CT midsagittal image preparation, all of these parts were ignored, even though they are also moving during speech. The ultrasound images were affected by shadows from the hyoid bone and the mandible. If the movements of these two parts were tracked, a better movie showing the relationship between the shadows on the ultrasound and the movements of the hyoid bone and mandible could be made. Although one jaw marker was used during data collection, an additional jaw marker was needed to calculate jaw rotation. The motion of the jaw could be assumed to simply rotate around the temporomandibular joint (TMJ), but in reality jaw movement is more complex (especially during non-speech activities such as chewing).

Although a microphone stand was holding the ultrasound probe during speech, it could still rotate (in the midsagittal plane) somewhat. The PRU and PRD markers showed its movement. Over all 4 trials, the range of probe rotation was 2.93 °. In the first, second, third and fourth trials, the maximum angles of rotation were 1.44 °, 1.51 °, 1.31 °, and 1.38 °, respectively. As a result, the probe position was not accurate enough to use the movie for measurement purposes, but the accuracy is high enough to use the movie for pedagogical purposes.

The most challenging step in this research project was finding the tooth location in the ultrasound images. Using motion capture data, the tooth location in the ultrasound image was calculated mathematically, but it did not always correspond to the tooth in the overlaid CT midsagittal image. Thus, it was challenging to overlay the images such that the tongue made contact with the palate for stop sounds, but did not penetrate the palate during those sounds.

4 Conclusions and Future Work

This research project involved the contruction of a midsagittal vocal tract movie by merging CT, ultrasound, and motion capture data, and the ultrasound images were successfully superimposed on the midsagittal CT image by using motion capture data, including the tooth location.

In future work, adding another jaw marker during data collection and expressing actual jaw movement would improve the movie. With two jaw markers, the degree of rotation of the jaw relative to the global vertical line could be calculated. Although the upper lip marker (LPU) was used, this lip was moving independently of jaw movement, so it was not useable in this case. To express jaw movement as precisely as possible, another jaw marker should be used.

In other future work, a comparison should be made of three videos: ultrasound tongue only, ultrasound tongue and CT palate, and ultrasound tongue, CT palate, and video face image. It would be valuable to determine which video is most useful for learning pronunciation. Sometimes having too much information can make the

learning process more difficult. In Figure 7, A shows the ultrasound and palate video, and B shows the ultrasound, palate, and face video.

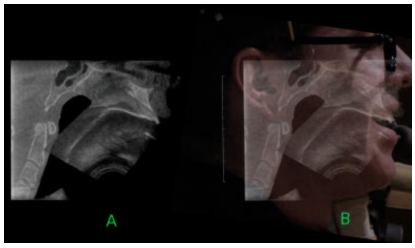

Figure 7: Frames from the two completed movies

Acknowledgements

I thank to the following people. Professor Ian Wilson who was my supervisor of the research, and he worked as a subject in the project. Thanks to him, I got valuable and unusual data: CT DICOM file, and also had opportunities to manipulate various kinds of data. Also he taught and help me to write this thesis kindly. Next, the faculty members of UBIC in the Univ. of Aizu. In the experiment, we used the expensive machine, the motion capture optical tracking system, for many hours. They supported us to go smoothly. Last, Mr. Naoya Horiguchi. He was a sophomore student who was working in the CLR Phonetics Laboratory. He helped me to set up the data collection, use sophisticated software applications and manipulate some data. I would like to show my appreciation for them.

References

[1] B. Gick, B. M. Bermhardt, P. Bacsfalvi, and I. Wilson, "Ultrasound Imaging Applications in Second Language Acquisition," *Phonology and Second Language Acquisition*. Amesterdam: John Benjamins, in press, pp. 309-322.

[2] I. Wilson and B. Gick, "Ultrasound Technology and Second Language Acquisition Research," *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conf. (GASLA 2006)*, Cascadilla Press, 2006, pp. 148-152.

[3] M. Stone, "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," *J. of the Acoustical Society of America*, May 1990, pp. 2207-2217.

[4] M. A. Epstein and M. Stone, "The Tongue Stops Here: Ultrasound Imaging of the Palate," *J. of the Acoustical Society of America*, Oct. 2005, pp. 2128-2131.

[5] J. Aho, "A Quick Guide to Digital Video Resolution and Aspect Ratio Conversions," Dec. 2007; http://lipas.uwasa.fi/ f76998/video/conversion/.

[6] S. H. Weinberger, "The Speech Accent Archive," Nov. 2007; http://accent.gmu.edu/index.php.

Appendix 1

This is the paragraph from the Speech Accent Archive that was used as stimuli in the data collection [6]. There are many samples from speakers of various first languages and backgrounds at that website. It is helpful for speech learners to compare their pronunciation. The paragraph was chosen to contribute to that learning experience.

> Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Appendix 2

This is the inventory of sounds from the paragraph. The numbers in parentheses indicate how many times each sound exists in the paragraph.

Consonants		Vowels	Clusters	
Initial	Final		Initial	Final
k(3)	z(5)	i(12)	pl(2)	sk(1)
t(3)	l(4)	ɑ(4)	st(4)	ŋz(2)
ð(5)	ŋ(1)	ɛ(4)	bɹ(2)	ks(1)
θ(3)	θ(1)	æ(8)	fɹ(3)	nz(2)
w(5)	m(2)	ɪ(11)	sp(1)	bz(1)
s(2)	ɹ(3)	ʌ(2)	sn(3)	nd(3)
f(3)	v(3)	ə(9)	sl(1)	dz(1)
tʃ(1)	ʃ(1)	u(4)	bl(1)	gz(1)
n(1)	k(4)	o(3)	sm(1)	
b(3)	b(1)	aɪ(1)	sk(1)	
l(1)	d(2)	eɪ(5)	θɹ(1)	
ʃ(2)	g(2)	ɔ(3)	tɹ(1)	
d(2)	n(4)	ɚ(5)		
ɹ(1)	p(1)	ɔɪ(1)		
g(1)	t(1)			
m(1)				
h(1)				