RECONSTRUCTION OF VOCAL TRACT BASED ON MULTI-SOURCE IMAGE INFORMATION

Song Wang[1], Shen Liu[1], Jianguo Wei[1,], Qiang Fang[3] and Jianwu Dang[1, 2]*

[1]Tianjin University, Tianjin, China
E-mail: yuwenqingsun@126.com; tobyliu33@hotmail.com; jianguo@tju.edu.cn
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Phonetics laboratory, Chinese Academy of Social Sciences, China
E-mail: jianguo@tju.edu.cn

ABSTRACT

At present, a variety of instruments recording the articulation have its own pros and cons. None is able to record the data containing all the information of articulators. For example, ultrasound system can obtain main surface information of the tongue, but the images are noisy and cannot record tongue tip under some cases. While the EMA system can precisely record trajectory data of the key points associated with attached sensors on the tongue surface. Therefore, we use EMA and ultrasound simultaneously as a complementary. In this paper, we will use the ultrasound system and the EMA system to record the tongue's movement. We obtain the ultrasound images and the synchronous audio by the ultrasound system; the EMA system is used to collect the EMA data and the synchronous audio. We register and match the ultrasound images and the EMA data by the audio files. And we integrate spatially the ultrasound images and the EMA data of each time point.

***Index Terms*—** Ultrasound images, EMA, tongue, vocal tract, alignment

1. INTRODUCTION

As with X-ray, Computed Tomography (CT), magnetic resonance imaging (MRI), ultrasound [1], [2] imaging technology is one of the four major medical imaging techniques. The four instruments recording the articulation have its own advantages and disadvantages. None is able to record the data containing all the information of articulator. Ultrasound imaging technology has the advantage of convenient, safe, fast, real-time, which is widely used in clinic. However, due to the particularity of the imaging mechanism, ultrasound images have serious speckle noise and provide limited information of the subject's articulator [3]. While Electromagnetic Articulography (EMA) [4], [5] data store the precise location of the sensor information but lack the complete information of the surface of the tongue.

*Corresponding author

In the experiment, EMA system and ultrasound system are simultaneously used as a complementary to record the tongue's movement. We obtain the ultrasound images and the synchronous audio by the ultrasound system; the EMA system is used to collect the EMA data and the synchronous audio. We register and match the ultrasound images and the EMA data by the audio files. Then we integrate spatially the ultrasound images and the EMA data of each time point.

The paper is organized as followed. In Section 2, we introduce the acquisition system, including the hardware and software system. Section 3 describes the speech corpus and databases recording. We analyze process and integrate the ultrasound and EMA data in section 4. The conclusions are given in Section 5.

2. ACQUISITION SYSTEM

Before the experiment with recording databases, we build the environment and connect the machines for the acquisition system. The main machines of the system are shown in Fig.1.

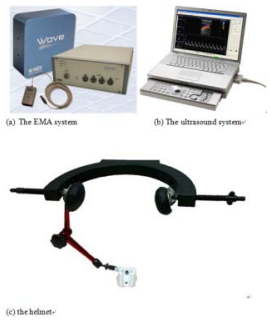


Fig.1. (a) the EMA system (b) the ultrasound system (c) the helmet

The acquisition system consists of:

- Ultrasound system: a Terason T3000 ultrasound system with a 8MC3 micro-convex transducer, providing ultrasound images of the subject's tongue
- EMA system: the WAVE system of Northern Digital Inc with 8 channels, the WAVE system contains: Field Generator, Mounting Arm, System Control Unit,

System Interface Unit, Micro-sensor, Audio Synchronization cable

- Microphone: Studio Project Cs5 condenser microphone
- Helmet: stabilization helmet for ultrasound probe

2.1. Hardware

In the experiment, the subject's head will unavoidably move. So we need a helmet to fix the probe to the subject's chin. So far, some groups are working on the ultrasound based system ESPCI [6], HATS [7], Palatron [8]. Our acquisition system makes use of a helmet, which is used for stabilizing the ultrasound probe. As far as we concerned, if the head is fixed too tightly, the subject will feel uncomfortable and then will be tired in the recording process. If the head is fixed too loosely, when the subject speak, his/her head will inevitably move relative to the probe. If so, we will get invalid data. Because what we want to obtain is the image of a tongue in the sagittal plane. The helmet ensures that the probe of the ultrasound system and the subject remain relatively static. Fig.2 shows our acquisition system in the experiment.

Fig.2. our acquisition system

In order to obtain the information of the tongue tip and tongue back, the EMA Field Generator is placed at the side of the head of the experimenter.

The system can keep the ultrasound probe fixed to the jaw and EMA aside with the head without restricting head movement. Although the helmet makes the ultrasound probe static relative to the head, we find wearing the helmet is heavy and makes the subject tired for long recording in the experiment.

2.2. Software

We would like to obtain three sets of data from the experiments: ultrasound images, EMA data, and audio files. The ultrasound image acquisition program developed based on the SDK supplied by the Terason Ultrasound System. We add some functions, such as File storage, information display and so on. We process the audio and image streams in parallel using the multithreading programming techniques [6]. Each image is named with timestamp to synchronize with audio file and EMA data.

The EMA data includes the trajectory of the point and audio files provided by the WAVE system. In the experiment, we use two computers to run the programs.

3. DATABASE ACQUISITION

3.1 Experiment

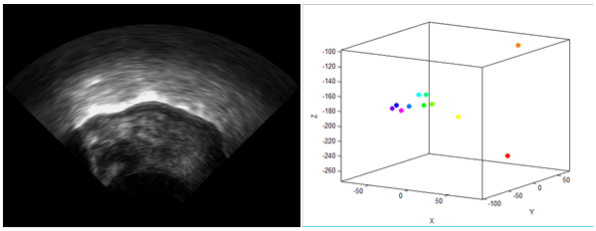
At first step, we open the WAVE system and adjust the location of field generator roughly. Then we start attach sensors for the subject. All sensors should be disinfected with alcohol and pasted to 11 points of the subject (left ear, right ear, nose, tongue tip, tongue middle, tongue dorsum, frontal side of tongue, lateral side of tongue, jaw, lower lip, upper lip) to collect articulatory data. After sensor attachment procedure, we detect sensor status and adjust their position in the magnetic field slightly. In the next step, the ultrasound program was started and was used for checking whether ultrasound probe is working properly. Then the subject wears the helmet and smears glue on the ultrasound probe and adjust the helmet to fix the probe under the jaw tightly. According to the live images on the monitor, the image parameters (including image size, image depth, gain, compression ratio, noise suppression, etc.) are adjusted, in order to get the ultrasound images with clear tongue contours and a more stable and higher acquisition frequency.

The corpus we use is "863 speech synthesis database", in which we select 100 sentences from the corpus. Reading every sentence is about 8-15 seconds. The whole recording process lasts approximately 20-30 minutes.

3.2 Data Structure

The ultrasound system can provide image and audio files. The images are bitmap, 8bits. The resolution is 640×480 pixels. The image stream is 60fps. The audio file is 8 bits and the sample rate is 44.1 kHz. We named each images with a timestamp from the system current time accurate to the millisecond.

The EMA system can provide '.raw' files to store articulatory data and '.wav' file to store speech sound. We use the two files to get the sensors' trajectory and corresponding audio. The frequency of the sensors movement is 100 Hz. As the same with the ultrasound system, the audio file is 8 bits and the sample rate is 44.1 kHz. The ultrasound images and the EMA data are shown in Fig.3.



(a) Ultrasound images (b) The EMA data

Fig.3. (a) the ultrasound image (b) the EMA data about the sensors' trajectory

4. DATA ANALSIS

4.1 Synchronization

Two groups of audio files are recorded during one acquisition. One group of the audio files is synchronized with the ultrasound images and the other group is synchronized with EMA data. The two microphones record the same sound, hence theoretically the two audio file should be the same and at least they have the same length. So the two sets of audio files can completely register. By the two audio files, the ultrasound images and EMA data can match on the time axis. Fig.4 shows the waveform of the audio files of a set of Chinese vowels (/a/, /o/, /e/, /i/, /u/, /ü/). The waveform is drawn by software *Pratt.*

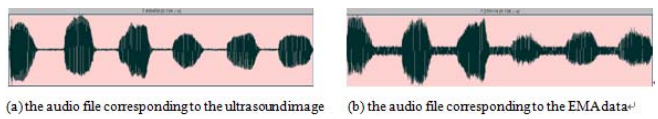


Fig.4. (a) and (b) are the audio files corresponding to the ultrasound images and the EMA data, respectively.

4.2 Alignment in the space

The EMA data stores the 5D information including the sensor position and sensor rotation while the ultrasound image stores the pixel values of the 2D plane in a matrix. The ultrasound image is the image of the midsagittal plane of the tongue (head). We should find the corresponding plane if we want to combine the two different sets of data. In order to find the corresponding plane, we should also find the midsagittal plane of head for the EMA data.

4.2.1 Finding the common plane

We find the head midsagittal plane for the EMA data using three points (left ear, right ear and nose) as the reference to determine a plane and remove the movement of the head.

We set the point of left ear as Point *A*, right ear as point *B* and the nose as point *C*. We can get the midsagittal plane of the head by the three point *A, B, C*. Point *A* and point *B* compose the vector $\overrightarrow{AB}$ which is perpendicular to the midsagittal plane. Point *C* is on the plane.

We can get the coordinate of point *A, B, C* from the EMA data. Then we can obtain the equation of the midsagittal plane:

$$ax + by + cz + d = 0 \tag{1}$$

where *a, b, c, d* is the constant, which are determined by the coordinate of point *A, B, C.*

We set the intersection of the plane and the connection of the two ears as the original point of the plane-coordinate system. The direction from the original point to the point of the nose is the direction of X axis. The direction of Y axis is the direction which is perpendicular to the X axis and pointing down from nose to tongue.

After defining the midsagittal plane of the EMA data and the coordinate system, we project other EMA points on the plane and get the plane coordinates. Given a point $\vec{P}(X,Y,Z)$ and point $\vec{P}_0(X_0,Y_0,Z_0)$ is the $\vec{P}$'s projection on the sagittal plane. We can calculate $\vec{P}_0$ by the following formula.

$$\vec{P}_0 = \vec{P} + [at, bt, ct] \tag{2}$$

Here, *a, b, c* is the constant in the equation of the midsagittal plane. And t is defined by:

$$t = -(aX + bY + cZ + d)/(a^2 + b^2 + c^2) \tag{3}$$

Where *X, Y, Z* is the coordinate of point $\vec{P}$ and *a, b, c*, d are the constant in the equation of the midsagittal plane.

4.2.2 Coordinate transformation

After we get the corresponding plane and the coordinate of the EMA points on the plane, we transform the EMA coordinate to the ultrasound image coordinate by three steps which are scaling, rotation and translation.

Given $\vec{P}_0$ is the point on the EMA coordinate and $\vec{P}$ is the transformed point on the ultrasound image coordinate.

$$\vec{P} = s \times \left(\vec{P}_0 \times r\right) + t \tag{4}$$

Where *s, r* and *t* are the parameters correspond to the values of scaling, rotation and translation. And *s* is a constant that represents enlargement or reduction during the coordinate transformation; *r* is a degree which represents the angle of rotation; *t* is a vector which represents the amount and direction of translation.

In the coordinate of EMA data, the unit length is 1mm. In order to calculate the value of the scaling parameter, we need to know the unit length of the coordinate system of the ultrasound image. When we configure the parameters of the ultrasound image, we set the depth (the height of the sector) 8cm as is shown in Fig.5.

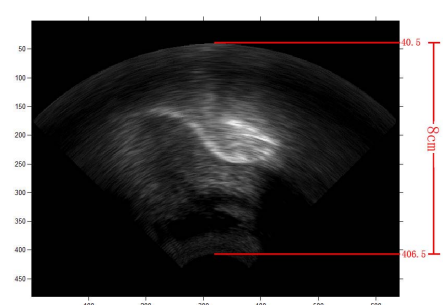


Fig.5. the height of the sector in the ultrasound image

Hence, in this way we can calculate the real unit length of the ultrasound image coordinate system which is 0.219mm, so the scaling parameter *s* is 4.575 in the coordinate transformation.

For each image, we need to align the four EMA points (tongue tip, tongue middle, tongue dorsum and jaw) onto the contour (the bright line) in the ultrasound image. Adjust the parameters of rotation and translation manually in order to finish the alignment. Fig.6 shows the alignment result of 6 vowels (a, o, e, i, u, ü).

Table 1. The value of scaling, rotation and translation

Vowel	*s*	*r*	*t*
a	4.575	-26°	[-160,85]
o	4.575	-10°	[-100,-30]
e	4.575	-14°	[-140,-10]
i	4.575	-20°	[-160,75]
u	4.575	-15°	[-135,40]
ü	4.575	-20°	[-155,98]

Table 1 shows the value of the parameters (*s, r and t*) corresponding to each vowel.

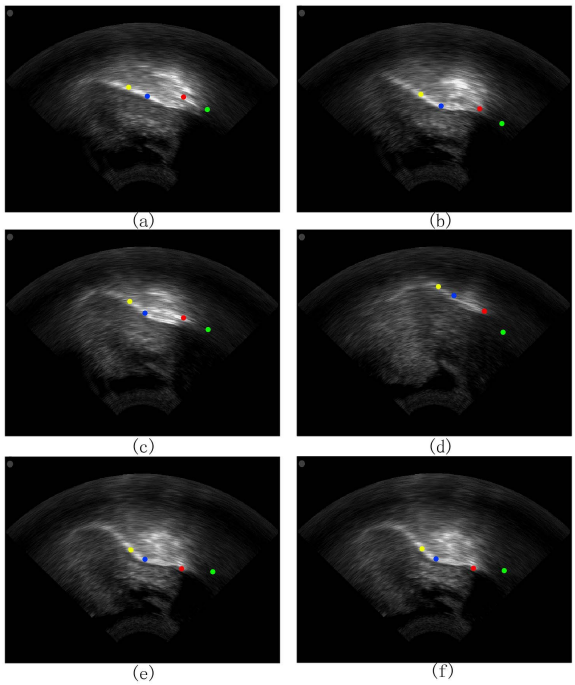


Fig.6. (a), (b), (c), (d), (e), (f) represent the results of alignment for the six vowels (a, o, e, i, u, ü), respectively. The yellow, blue, red, green point represents tongue dorsum, tongue middle, tongue tip, and jaw, respectively.

5. CONCLUSION

In this paper, we introduce our acquisition system and use the system to record the ultrasound images and EMA data. With the ultrasound images and EMA data, we analyze and fusion them together. During the data combination, we synchronize them by the two audio files and get the alignment image with both EMA information and ultrasound image. Although we only get six sets of the parameters (s*, r and t)* for the six vowels manually, we can use the method in the further study about the automatic alignment of EMA data and ultrasound image. In the future, we will evaluate the performance of our system with more datasets and experiments conditions. The system could be applied for vocal tact visualization.

6. ACKNOWLEDGMENT

This work is sponsored by the national natural science foundation of china under contract No. 61175016. And it is supported by Major State Basic Research Development Program of China (973 Program, Contract No. 2013CB329305). This study was also supported in part by Microsoft Research Asia under contract No. FY11-RES-OPP-008.

7. REFERENCES

[1] Stone, M., Sonies, B., Shawker, T., Weiss, G., Nadel, L., 1983. Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system, *Journal of Phonetics*, 11, pp. 207-218.

[2] Thomas, H., Elie-Laurent B., Gérard, C., Bruce, D., Gérard, D, Maureen, S., "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips", *Speech Communication*, Vol. 52, Iss. 4, pp. 288-300, April, 2010.

[3] Boisvert, J., Gobbi, D., Vikal, S., Rohling, R., Fichtinger, G., and Abolmaesumi, P., "An open-source solution for interactive acquisition, processing and transfer of interventional ultrasound images," in [*Workshop on Systems and Architectures for Computer Assisted Interventions*], MICCAI 2008 (2008).

[4] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., and Jackson, M., "Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements", *J. Acoust.* Soc. Am. 92, pp. 3078-3096, 1992.

[5] Hoole, P., Nguyen, N., 1999. Electromagnetic articulography in coarticulation research, in Hardcastle, W.H., Hewlitt, N. Eds., Coarticulation: Theory, Data and Techniques, pp. 260-269, Cambridge University Press, 1999.

[6] Florescu, V-M., Crevier-Buchman, L., Denby, B., Hueber, T., Colazo-Simon, A., Pillot-Loiseau, C., Roussel, P. Gendrot, C., Quattrochi, S. (2010), "Silent *vs* Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface", *Proceedings of Interspeech (Makuari, Japan)*, pp. 450-453.

[7] Stone, M., and Davis, E. (1995), "A Head and Transducer Support System for Making Ultrasound Images of Tongue/Jaw Movement," *Journal of the Acoustical Society of America*, 1995 98 (6), pp. 3107-3112.

[8] Mielke, J. Baker, A., Archangeli, D., Racy, S. 2005, Palatron: A technique for aligning ultrasound images of the tongue and palate. *In Siddiqi, D., Tucker, B.V.* (eds.), Coyote Papers 14, 97-108.