# Dynamic programming method for temporal registration of three dimensional tongue surface motion from multiple utterances

Changsheng Yang [a] and Maureen Stone [b,*]

[a] Symantec Corporation, 525 Butler Farm Road, Suite 106, Hampton, VA 23666, USA

[b] Department of Oral and Craniofacial Biological Sciences, Department of Orthodontics, University of Maryland Dental School, 666 W. Baltimore St. Baltimore, MD,21201, USA

* Correspondence author:
Maureen Stone, Ph.D.
**Email: mstone@umaryland.edu**

Phonetic symbols: I, B, D, A

I as in “hit”.

B as in “hat”.

A as in “father”.

D is the schwa, found in most minimally stressed syllables of English, such as “tomato” (tDmeitou).

The number of pages: 18

The number of figures: 6

**Abstract**

This study proposes a new method to reconstruct three dimensional (3D) tongue surface motion during speech using only a few sections of the tongue measured with ultrasound imaging. Reconstruction of static 3D tongue surfaces has been reported. This is the first report for reconstruction of 3D tongue surface motion using ultrasound imaging. To temporally align data from multiple scan locations, a Dynamic Programming (DP) algorithm was used to line up the tokens collected from different repetitions by using the acoustic signals recorded simultaneously with the ultrasound images. Reconstruction error was evaluated by using a pseudo-motion measurement of known 3D tongue shapes. The average error was 0.39 mm, which was within the ultrasound measurement error, of 0.5mm.

*Keywords:* Reconstruction of 3D tongue surface; 3D tongue surface motion; Ultrasound imaging; Dynamic programming

## Introduction

Ultrasound imaging has been used to assess three dimensional (3D) tongue surface shapes of English consonants and vowels (Stone and Lundberg, 1996). By this method, a series of static 2D contours was spatially aligned to reconstruct a detailed 3D tongue surface. A special transducer collected 60 ultrasound scans in a polar sweep of 60 degrees in 10 seconds. This method is suitable for sustained vowels and consonants, but 10 seconds is too slow to collect tongue motion. Lundberg and Stone (1999) used the same data sets (Stone and Lundberg, 1996) to determine a minimal number of slices, or optimized sparse set, for reconstruction of 3D static tongue surfaces without significantly reducing the reconstruction quality. The results showed that 5 to 6 coronal slices were adequate to reconstruct 3D tongue surfaces, i.e., the 3D tongue surface could be reconstructed by collecting a few 2D tongue contours at the optimized locations.

In order to reconstruct 3D tongue surface motion during speech we must collect multiple 2D data sets at different scan locations. Any single data set, which is a sequence of 2D tongue contours, contains the 2D tongue motion at that specific location. The premise is that a number of 2D data sets can be combined later into a single 3D tongue motion by spatial and temporal alignment.

The spatial position of different scans can be aligned using the pre-measured location of the transducer. For the temporal alignment we must consider that subjects vary in speaking rate and articulation for multiple repetitions. Speaking rate differences are even more likely when the repetitions are separated in time by other speech materials, as is the case with multiple ultrasound data sets. A time-warping algorithm is needed to align temporal variations in multiple repetitions. In automatic speech recognition, to eliminate the effect of large variation in the speaking rates and inter and intra-speaker variation, the dynamic programming (DP) algorithm has been used successfully (Itakura, 1975; Sakoe and Chiba, 1978; Rabiner and Juang, 1993; Ney and Ortmanns, 1999). The DP algorithm finds the optimal time registration between two repetitions based on the minimum total distance measure of the acoustic feature. Dang et al. (1997) used an X-ray microbeam system to measure the position of 8 points on the tongue surface during speech. Five metal pellets glued along the mid-sagittal tongue and three metal pellets glued on the para-sagittal tongue (1 cm apart from the mid-sagittal) were tracked separately by the system. Time differences between the data sets were synchronized by using spectrograms of the speech signals. Strik and Boves (1991) applied the DP algorithm to time-alignment and averaging of repeated physiological signals to improve the signal-to-noise ratio. The result showed that the DP algorithm was able to correct the timing differences among the repetitions.

In this study, ultrasound imaging is used to reconstruct 3D tongue motion during normal speech using 8 ultrasound images (2D), 5 coronal and 3 sagittal slices which were collected at different scan angles. The different slices were aligned manually on the computer using pre-measured information as to transducer location. The sagittal slices also were used to retrieve tongue tip information. For each section, the acoustic signal was recorded simultaneously with the ultrasound images. To temporally align data from multiple scans, a DP algorithm based on Rabiner and Juang (1993) was used on the acoustic signals to line up the tokens collected from different repetitions. Reconstruction error of the proposed method was evaluated with the 3D tongue shapes of Lundberg and Stone (1999).

## I. Method

**I.1** Dynamic Programming Algorithm

Consider two patterns, a reference pattern ($S_R$) and a test pattern ($S_T$). Both patterns are represented by a sequence of feature vectors extracted from speech signals. The lengths of the two patterns are *Ty* and *Tx*, respectively. The length of *Ty* and *Tx* may be different. The frames of the two patterns define a grid of $T_x$ times $T_y$ points. It is a time matching and normalization problem to reduce the effect of speaking rate and articulation variation for different repetitions. The dynamic programming algorithm is used to efficiently find the optimal time matching for the two patterns. A constraint window is used to limit the search path within a reasonable region. Figure 1(a) shows an example of DP matching and its constraint window. A possible path $P=p_1, p_2, \ldots, p_\kappa$ is a sequence of $K$ points which begins from $p_1=(1,1)$, end at point $p_\kappa=(T_x, T_y)$. The total distance between the two patterns $S_T$ and $S_R$ for a given path $P$ is the weighted sum of the local distance at each point $p_k$:

$$D_P(S_T, S_R) = \sum_{k=1}^{K} d(p_k) w(k). \qquad (1)$$

Where $p_k=(k_x, k_y)$ is the possible point within the constraint window in Fig.1. The weighting coefficient $w(k)$ is defined in Fig. 1(b). Each sub-path is constricted by the five step sequences. The slope weighting coefficients control the distribution of the local distance for each path. The symmetrical form DP-matching is used because it is reported to give better performance (Sakoe and Chiba, 1978). The local distance $d(p_k)$ is the distance between the feature vector of frame $k_x$ of $S_T$ and that of frame $k_y$ of $S_R$. Here we use LPC cepstrum coefficients $c_i$ as the feature vector. The distance $d$ is defined as:

$$d = \sum_{n=1}^{L} (c_n - c_n')^2. \qquad (2)$$

This distance $d$ is thus a measure of the spectral distances and reflects articulation differences.

The optimal path from (1,1) to $(T_x, T_y)$ is the path that minimizes the equation $D_P(S_T, S_R)$. The optimal path represents a time matching between the reference pattern and the test pattern. The speech signals are thus lined up according to time matching.

<Figure 1 about here>

The DP algorithm is summarized as following:

1. Initialization:

$$D(1,1) = d(1,1). \qquad (3)$$

2. Recursive computation
   For each point $(i_x, i_y)$

$$1 \le i_x \le T_x, \quad 1 \le i_y \le T_y,$$

within the constraint window, calculate equation (4):

$$D(i_x, i_y) = \min \begin{Bmatrix} D(i_x - 3, i_y - 1) + 2d(i_x - 2, i_y) + d(i_x - 1, i_y) + d(i_x, i_y), \\ D(i_x - 2, i_y - 1) + 2d(i_x - 1, i_y) + d(i_x, i_y), \\ D(i_x - 1, i_y - 1) + 2d(i_x, i_y), \\ D(i_x - 1, i_y - 2) + 2d(i_x, i_y - 1) + d(i_x, i_y), \\ D(i_x - 1, i_y - 3) + 2d(i_x, i_y - 2) + d(i_x, i_y - 1) + d(i_x, i_y). \end{Bmatrix} \quad (4)$$

3. Termination:

$$D_p(\mathrm{S_T}, \mathrm{S_R}) = D(T_x, T_y). \quad (5)$$

The optimal path can be determined from the search steps recorded during the recursive procedure.

**I.2** Data Acquisition and Analysis

A head and transducer support system HATS (Stone and Davis, 1995) was used to collect ultrasound images. A female native English speaker served as subject. The speech token was the sentence "It ran a lot". As mentioned before, to reconstruct 3D tongue surface shape optimally from ultrasound images, 5 to 6 cross-sections are needed to collect the ultrasound images at different scan angles (Lundberg and Stone, 1999). In the present report, 5 cross-sections and 3 sagittal (one mid-sagittal, two para-sagittal) sections, each 10mm apart, were collected. The sampling rate of ultrasound is 28 scans per second. The ultrasound images during speech are recorded on videotape at a rate of 30 frames per second. Two scans per second are repeated. Most speech movements that are generated by muscle contractions have a bandwidth below 15 Hz (Perkell et al., 1992), so ultrasound is able to capture most parts of the tongue. It is slow however for measuring raising and lowering movements of the tongue tip, which has velocities in the range of 80 cm/s (Perkell, 1969).

The acoustic signal was recorded on the audio channel of the videotape. The VTR data were grabbed by an SGI computer and saved as movie files. A media-conversion program was then used to convert the video signals to TIFF image sequence files, and to convert the audio signals to WAV format files. Tongue surface contours were extracted from the TIFF image sequences by a software system called "tonTrak" (Akgul, Khambamettu, and Stone, 1999) and were saved as x-y coordinate points. Figure 2(a) shows an example of an ultrasound tongue image and overlaid surface contour. Figure 2(b) shows the mid-sagittal tongue surface contour sequence of the sentence. The sentence begins at the right end of the sequence. The tongue tip is on the right. The sentence chosen has two phonemes (/t/ and /l/), which might be too rapid for adequate capture by ultrasound. Figure 2(b), however, indicates that tip elevation was captured, at least in part, by the data.

On the videotape, a lateral view of the speaker's face, including the transducer, was inserted at the lower left corner of the ultrasound image (Fig. 2(a)). This small window was used to measure the position and angle of the transducer (see Fig. 3).

<Figure 2 about here>

<Figure 3 about here>

The acoustic signal was sampled at a sampling frequency of 11025 Hz, and was analyzed by the LPC method. The LPC analysis order was 12 using a Hamming window. The analysis window was 20 ms with a shift of 5 ms. The first 12 cepstrum coefficients ($c_1$ - $c_{12}$), which were used as the feature vectors, were converted from LPC coefficients (Rabiner and Schafer, 1978).

**I.3** Time Alignment

The dynamic programming algorithm was applied to time-align ultrasound samples at different scan angles using the acoustic signal. The acoustic signals were calculated as LPC cepstrum vector sequences at a rate of 200 frames per second. For each repetition, the line-up region was restricted to the region between the first pitch of /i/ and the last /t/. The token of the mid-sagittal section was selected as the reference pattern. All other tokens were aligned with the reference pattern. Speaking rates for the eight repetitions were all different. The longest duration, which was the mid-sagittal section, was 1.54 s, and the shortest was 1.05 s. Figure 4 shows an example of the non-linear time alignment. In the figure, the top and bottom are the sound spectrogram of the reference and one of the other tokens. Vertical lines on the sound spectrograms indicate the boundaries of the phonemes. Lines in the middle graph indicate the time registration for the two patterns. The time interval of the lines is 30 ms on the reference side. It can be seen that the phonemes are well matched at their boundaries.

<Figure 4 about here>

**I.4** 4D Tongue Surface Reconstruction

The acoustic sequence was sampled at a rate of 200 frames per second. It is higher than the sampling rate of ultrasound imaging which is 30 frames per second. After the time alignment, the acoustic sequences were re-sampled at the same rate as the image signal to convert them to the matched indexes for image frames. The time alignment based on the DP algorithm was not a linear function; temporal compression and stretching were needed on the test repetitions and their related 2D tongue contour sequences. For temporal compression, two tongue contour frames were replaced by a single frame which was made by linear interpolation of the two frames. For temporal stretching, an additional tongue contour frame was created by interpolating linearly the contours before and after it. In this report, we use a linear algorithm to interpolate tongue contours where temporal compression or stretching is necessary. Visual examination showed that it was a good approximation. More sophisticated algorithms need to be studied to improve the reconstruction quality, and will depend on further research on tongue dynamics.

After time alignment, all the image tokens were aligned temporally and had the same frames as the reference pattern. In this experiment, the maximum number of deleted frames for one sequence was 8, and that of inserted frames was 21. The eight tongue contour sequences were reconstructed into a 3D surface sequence by spatially orienting the 2D contours at each time frame. The spatial orientation of each coronal contour was determined relative to the mid-sagittal contour by the angle of the transducer and the mid-sagittal contour. All three sagittal sections were used to check if the spatial positions of the coronal contours were determined properly. The three sagittal contours were also used to reconstruct the tongue tip beyond the anterior-most coronal slice. The anterior points in the sagittal contours were connected with interpolating splines to create a surface.

A linear moving average algorithm (3, 5 and 7 points) was applied in the study to reduce articulation variations and errors of the tongue surface in both spatial and temporal domains. According to the evaluation experiment, the five-point moving average was the best at smoothing the 3D tongue surface sequence. After spatial and temporal alignments, each 2D coronal contour (digitized to 100 points) was smoothed. Then the algorithm was applied along the frames of the sequence.

**II. Evaluation Experiment**

It is important to validate quantitatively a measurement technique. However, there are no 3D tongue surface data that can be collected without reconstruction. In order to evaluate the proposed method, an evaluation experiment was performed using the 3D tongue surface shape data collected previously by Stone and Lundberg (1996). The static data were used because the 2D and 3D shapes were known and temporal reconstruction of them, based on the DP algorithm, could be evaluated. A sentence "She wants Ray loose" was designed as the speech material, because all the static tongue shapes of the phonemes were found in the static data set. The naturally occurring temporal variation found in multiple repetitions was simulated as follows. Six repetitions of the sentence, uttered by a native English speaker, were recorded on an audio recorder. The acoustic signal was sampled at a sampling frequency of 11,025 Hz. The beginning and end points of each repetition were determined manually using a sound-editor program. A reference or "true" tongue surface motion was based on the longest acoustic repetition, and a "reconstructed" tongue surface motion was based on the other 5 acoustic repetitions.

The "true" tongue motion, was made by aligning the static 3D tongue shapes to the onset of each phoneme of the longest acoustic repetition, (the onsets were determined manually). Tongue shapes between phonemes were reconstructed by linear interpolation at a rate of 30 frames per second, to reflect the timing of the acoustic signal.

The "reconstructed" tongue motion, was reconstructed from five 2D coronal tongue contour sequences. One contour sequence incorporated one of the five acoustic data sets and one of the ultrasound slice location data sets (10, 18, 24, 32, 38 degrees). The slice number indicates the transducer's angle during the scan; the chosen sets were the ones determined to have minimum reconstruction error (see Lundberg and Stone, 1999). Motion sequences for each of the 5 coronal locations were created in the same manner as for the "true" tongue motion, but using 2D contour instead of 3D surfaces. In this way, five 2D coronal tongue contour sequences were produced. These sequences represented 2D tongue surface shapes at different scan positions and phonated by different

repetitions. The DP algorithm was applied to the five repetitions to align them respectively with the reference repetition. Then the same procedure in **I.4** was applied to make the "reconstructed" 3D tongue surface motion.

The 3D static tongue surface data were all coronal scans. No sagittal data were used in the evaluation experiment, so the tongue tip shape beyond the anterior-most coronal slice could not be evaluated.

Ideally, if all the phoneme boundaries matched by DP were the same as those marked manually, and all the tongue motion variations of the six repetitions were linear, the "true" and the "reconstructed" tongue motions would be identical. As is well known, speaking rate variations are non-linear in real speech. So the "reconstructed" tongue motion may be different from the "true" tongue motion. The error measure, the distance between the "true" 3D tongue shape sequence and the "reconstructed" sequence, was calculated point by point. Over the whole sentence, the average error, maximum error, and standard deviation (SD) error were 0.41 mm, 2.50 mm and 0.58, respectively. The average error was within the ultrasound measurement error, 0.5mm. Errors (mismatched frames) occurred when the tongue contour slices of the test pattern were not aligned to the same slices of the reference pattern. Large errors occurred at those mismatched frames where the duration of a phoneme was very short and the adjacent tongue shapes were very different. This would be improved if the sampling rate of ultrasound image could be increased. The maximum error occurred at the frame between /r/ and /e/. The tongue for /r/ was widely and deeply grooved in the back. For /e/, the back tongue had a shallow, narrow groove. Temporally, all mismatches were within one image frame, i.e. less than 33 ms, except for one case which was off by two frames. To reduce the mismatch, each repetition was cut into two sections at the onset of /r/. The DP algorithm was performed on the two sections. The errors were reduced to 0.39 mm, 2.17 mm, and 0.54 mm, respectively. The option of manually picking an intermediate alignment point can be considered in the 3D tongue motion reconstruction system to reduce mismatches of the acoustic signal.

This experiment used static tongue shapes to simulate tongue motion by linear interpolation. In real speech, the tongue shape of extreme phonemes (target) may be reached or not, depending on speaking rate. In the experiment, we assumed that the target shape would always be reached. The duration of /w/ to /A/ varied from 58 ms to 111 ms, and that of /r/ to /e/ varied from 56 ms to 126 ms within the six repetitions. The target tongue shape changed rapidly within 2 image frames for the shortest phoneme. The variation of speaking rate may not only affect the velocity of tongue movement, but also affect tongue shape in forming the target phoneme. These factors were not controlled specifically in the experiment. Their effects probably increased the error.

**III Results and Discussion**

The 3D tongue surface sequence of the sentence "It ran a lot" was reconstructed in the experiment. The HATS system (Stone and Davis, 1995) keeps the subject's head and the transducer positions unchanged during the scan section. The spatial position of the coronal contours was checked against the three sagittal sections. The transducer position and angle measured in Fig. 3 determine the 3D position of each scan. Minor adjustment of 3D positions was manually made for the coronal sequences of no more than 2.4 mm.

Figure 5(a) shows the three sagittal and five coronal tongue contours used to reconstruct the tongue surface of /I/. Splines were drawn to form a boundary at the tip. The dark lines in Fig. 5(b) are the five coronal and the tip contours. The contours were bi-linearly interpolated and rendered to form a 3D tongue surface. Two grid planes were drawn with a 1 cm unit to provide a 3D perspective of the surface. The key features of /I/, a deep groove near the root and an arched front, were preserved very well.

The reconstructed 3D tongue surfaces of extreme phonemes in the sentence are shown in Fig. 6. These surfaces are the result of the DP alignment of the 5 contour sequences, based entirely on the acoustic features. The number at the upper left corner of each tongue surface is the frame number. Phonetic symbols are printed at the left on each image. Except for the tongue tip, the main features of 3D tongue shape of the extreme phonemes are well preserved in the figure. The deep groove became a shallow slope from /I/ (frame 3) to /r/ (frame 13). The constriction position changed from the back vowel /A/ (frame 41) to the consonant /t/ (frame 49).

The tongue tip shape reconstructed from the three sagittal images extended the surface beyond the coronal images. Coronal slices, by their nature, cannot measure the tongue orthogonal to the plane. Although the sampling rate was low (28 frame per second), the tongue tip movement is visible in the sagittal plane. In Fig. 2(b), we can see tongue tip raising for phonemes /t/, /n/ and /l/. It is not clearly reflected in Fig. 6, because there is not enough detail to reconstruct the tongue tip. It should be noted that ultrasound does not pass beyond the jaw and floor of the mouth, or due to air beneath the tip, the very tip of the tongue may not be visualized for some phonemes even in sagittal slices. Further improvement is needed for this technique.

<Figure 5 about here>

<Figure 6 about here>

**Conclusions**

To measure and model 3D tongue motion, is the goal of our research. This experiment provided a way to gather detailed information about the performance of DP algorithm used for 3D tongue motion reconstruction. It also gave us an approximate evaluation of the proposed method. The study established a reasonable method to reconstruct 3D tongue surface movement during speech using only a few sections of ultrasound images. The spatial position of the coronal contours was checked against the three sagittal sections, and the result showed that the relative shape of the coronal tongue changed consistently with the sagittal changes. Tongue surface reconstruction quality can be improved with a finely tuned surface smoothing algorithm. The 3D tongue surface sequence can be converted to a movie file containing the acoustic signal of the reference token. The movie facilitates observation of tongue motion during speech. The method provides a visualization tool to investigate tongue movement during speech. In future work, a 3D motion model is expected to extract tongue shape features from the 3D tongue motion data.

**Acknowledgment**

This work was supported by NIH Research Grant No. R01 DC 01758.

## References


Akgul, Y., Khambamettu, C., and Stone, M., 1999. Extraction and tracking of the tongue surface from ultrasound image sequences. IEEE Trans. on Medical Imaging. 18(10), 1035-1045.

Dang, J., Honda, K., and Tohkura, Y., 1997. 3-D observation of tongue articulatory movement for Chinese vowels. Technical Report of IEICE (Japan), SP97-119, 9-16.

Itakura,F., 1975. Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. Speech signal process. ASSP-23, 67-72.

Lundberg, A. and Stone, M., 1999. Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data. J. Acoust. Soc. Am. 106(5), 2858-2867.

Perkell, J., 1969. Physiology of speech production: Results and implications of a quantitative cineradiographic analysis. Research Monograph No. 53, MIT, Cambridge, MA.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I. and Jackson, M., 1992. Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. J. Acoust. Soc. Am. 92(6), 3078-3096.

Ney, H. and Ortmanns, S., 1999. Dynamic programming search for continuous speech recognition. IEEE Signal Processing, 16(5), 64-83.

Rabiner, L. and Schafer, R., 1978. Digital processing of speech signals, Prentice-Hall Inc., New Jersey.

Rabiner, L. and Juang, B.-H., 1993. Fundamentals of speech recognition, PTR Prentice-Hall Inc., New Jersey.

Sakoe, H. and Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. ASSP-26, 43-49.

Sakoe, H., 1979. Two-level DP matching – a dynamic programming-based pattern matching algorithm for connected word recognition. IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 588-595.

Stone, M. and Davis, E.P., 1995. A head and transducer support system for making ultrasound images of tongue/jaw movement. J. Acoust. Soc. Am. 98(6), 3107-3112.

Stone, M. and Lundberg, A., 1996. Three-dimensional tongue surface shapes of English consonants and vowels. J. Acoust. Soc. Am. 99(6), 3728-3737.

Strik, H. and Boves, L., 1991. A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech. J. Phonetics, 19, 367-378.


**Figure 1.** An illustration of the dynamic programming algorithm. (a) The time warping function and the constrain window, (b) the five possible steps and the weighting coefficients.

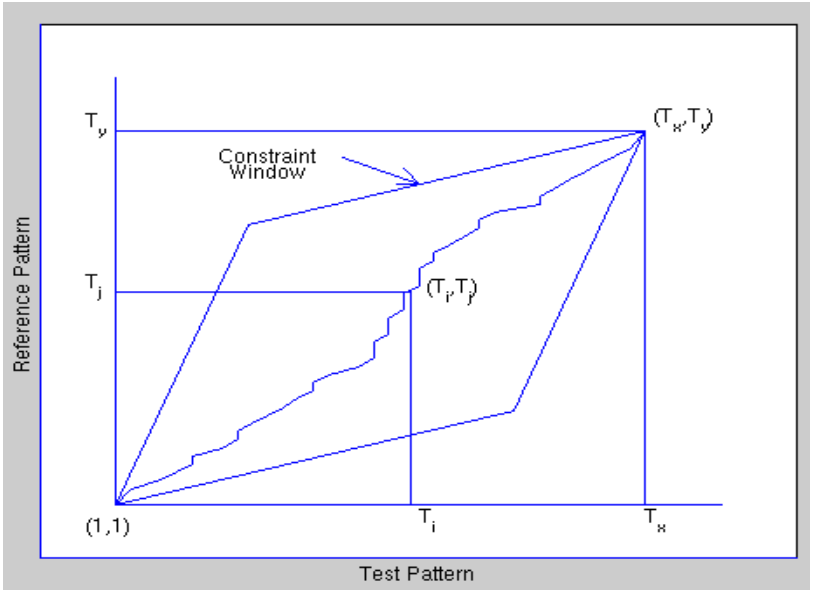
**Figure 2.** (a) A mid-sagittal ultrasound image of the tongue and its tongue surface contour. (b) The mid-sagittal tongue contour sequence of the sentence “It ran a lot.”

**Figure 3.** The little window inserted at the lower-left corner of the ultrasound image. The lines indicate how to measure the transducer position and angle.

**Figure 4.** An example of time alignment based on the DP matching. Top figure is the sound spectrum graph of the reference pattern, and bottom is that of the test pattern. The lines in the middle window represent the time alignment of the two patterns.

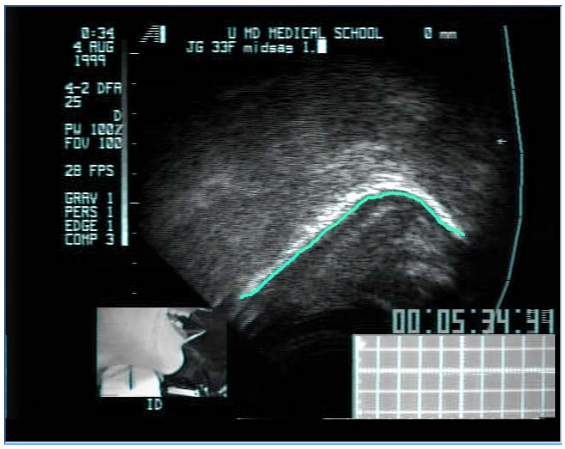
**Figure 5.** Contours used to reconstruct the 3D tongue surface of /I/ (a), and the reconstructed tongue surface (b).

**Figure 6.** 3D tongue surfaces of the extreme phonemes in the sentence “It ran a lot.”

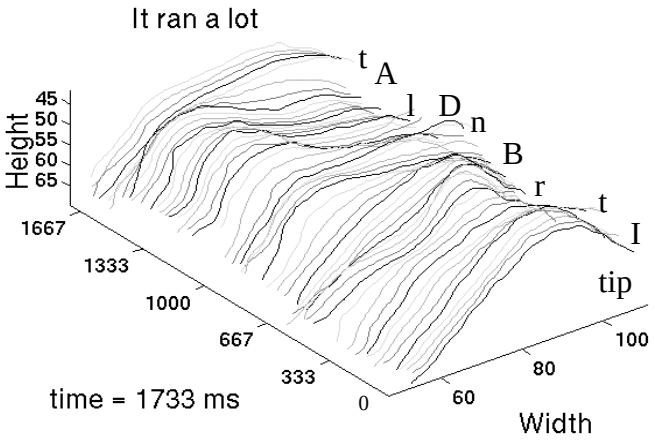


(a) (b)

Figure 1


0:34
4 AUG
1999
U MD MEDICAL SCHOOL
JG 33F midsag 1.
0 mm
4-2 DFR
25
D
PW 100%
FOV 100
28 FPS
GRAY 1
PERS 1
EDGE 1
COMP 3
00:05:34:99
ID

(a)


It ran a lot
Height
45
50
55
60
65
1667
1333
1000
667
333
0
time = 1733 ms
60
80
100
Width
tip

(b)

Figure 2

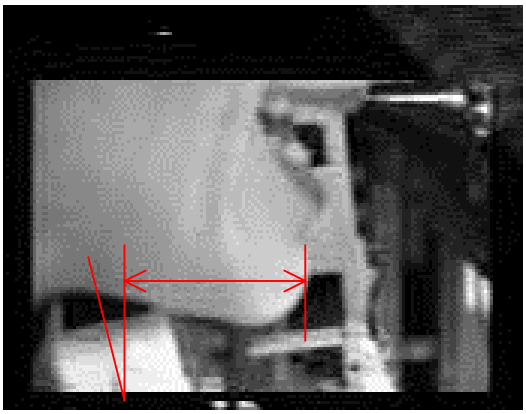


Figure 3

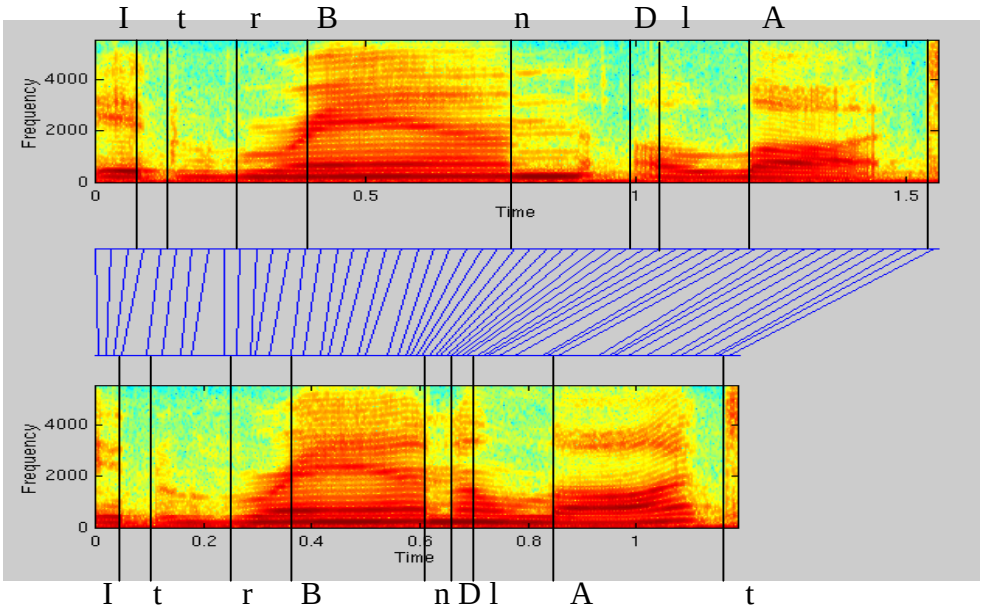
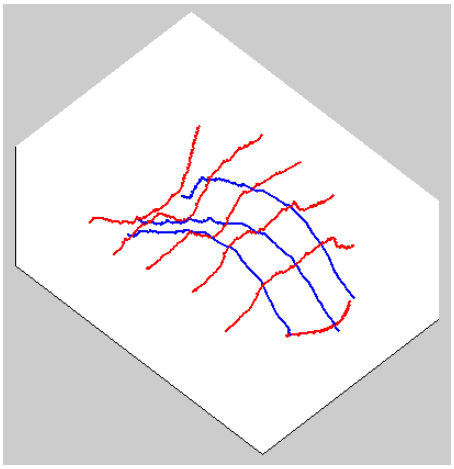


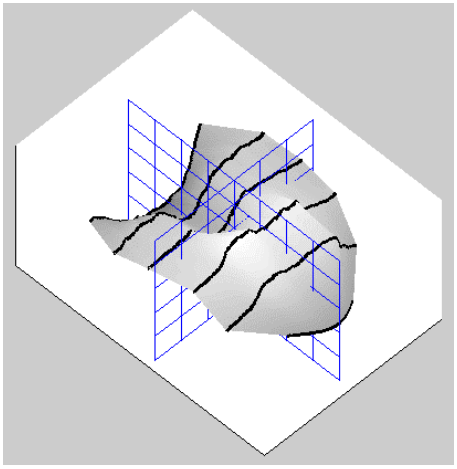

Figure 4



(a)

(b)

Figure 5



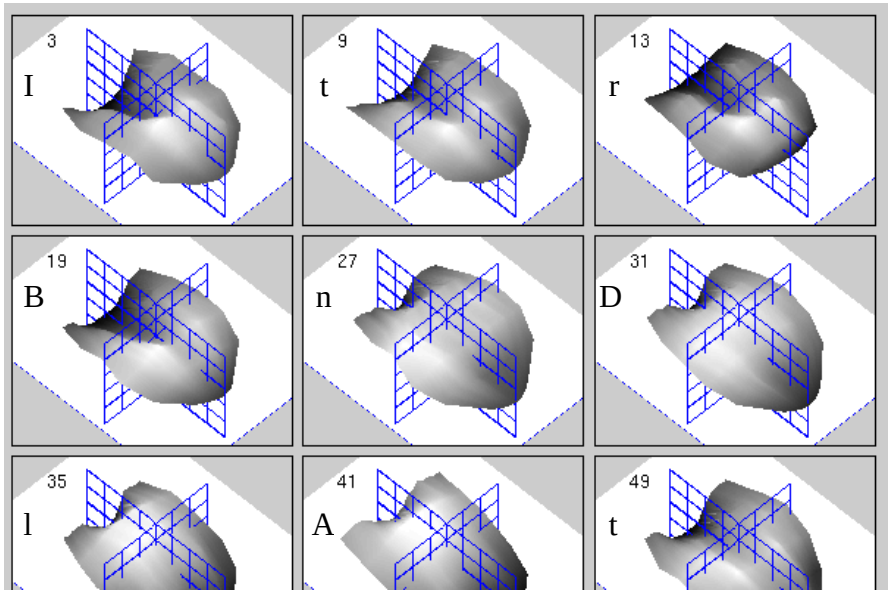


Figure 6