ELSEVIER

Speech Communication 38 (2002) 201–209

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

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

Changsheng Yang [a], 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., 5-A-12, Baltimore, MD 21201, USA

Received 23 August 2000; received in revised form 1 January 2001; accepted 18 June 2001

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 0.5 mm.

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

1. 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° in 10 s. This method is suitable for sustained vowels and consonants, but 10 s 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 5–6 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

[*] Corresponding author. Fax: +1-410-706-0193.
E-mail address: mstone@umaryland.edu (M. Stone).

0167-6393/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(01)00053-X

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, 1979; 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 eight ultrasound images (2D), five coronal and three 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).

2. Method

2.1. 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 T_y and T_x, respectively. The lengths of T_y and T_x 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 DP 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. Fig. 1(a) shows an example of DP matching and its constraint window. A possible path $P = p_1, p_2, \ldots, p_K$ is a sequence of K points which begins from $p_1 = (1, 1)$, end at point $p_K = (T_x, T_y)$. 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:

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

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