International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:2, 2007

Comparison of Fricative Vocal Tract Transfer Functions Derived using Two Different Segmentation Techniques

K. S. Subari, C. H. Shadle, A. Barney and R. I. Damper

***Abstract*—**The acoustic and articulatory properties of fricative speech sounds are being studied using magnetic resonance imaging (MRI) and acoustic recordings from a single subject. Area functions were derived from a complete set of axial and coronal MR slices using two different methods: the Mermelstein technique and the Blum transform. Area functions derived from the two techniques were shown to differ significantly in some cases. Such differences will lead to different acoustic predictions and it is important to know which is the more accurate. The vocal tract acoustic transfer function (VTTF) was derived from these area functions for each fricative and compared with measured speech signals for the same fricative and same subject. The VTTFs for /f/ in two vowel contexts and the corresponding acoustic spectra are derived here; the Blum transform appears to show a better match between prediction and measurement than the Mermelstein technique.

***Keywords*—**Area functions, fricatives, vocal tract transfer function, MRI, speech.

I. Introduction

LIMITED knowledge regarding fricative speech sounds and their production mechanisms, in comparison to vowels and stop consonants, constrains the task of accurately synthesizing, classifying and/or identifying them. This in itself hinders the progress of, for instance, advanced recognition systems and articulatory speech synthesizers, as fricatives play a considerable role in speech.

Our approach was to obtain articulatory and acoustic data during fricative productions. We then used the articulatory data to predict acoustic output, and compared that to the actual speech spectra with the ultimate goal of studying the effect of vowel context on the fricatives. The articulatory data consisted of volumetric magnetic resonance imaging (MRI) data, from which we determined the 3D vocal tract shape. We then used two different techniques of parameterizing the 3D shape as an area function. We sought an area function because a wide variety of models exist with which to compute the acoustic transfer function from the area function.

Manuscript received November 5th, 2004. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK.

K. S. Subari is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK (phone: +44 (0)2380-594882; fax: +44(0)2380-595499; e-mail: kss01r@ecs.soton.ac.uk)

Christine. H. Shadle is with Haskins Laboratories, 270 Crown St., New Haven, CT 06511, USA (email: shadle@haskins.yale.edu).

A. Barney is with the Institute of Sound and Vibration Research, University of Southampton, SO17 1BJ (e-mail: ab3@soton.ac.uk)

R. I. Damper is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK (e-mail: rid@ecs.soton.ac.uk).

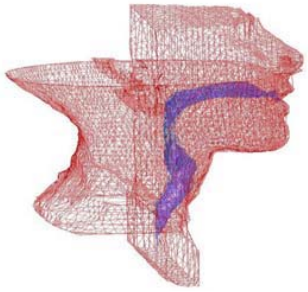

Fig. 1: Complete 3D model of vocal tract consisting of vertical and axial MR image slices.

In order to generate an area function consisting of cross-sectional areas along the tract, we needed to select locations along the tract's axis at which to slice the 3D shape and determine the area at that slice. Two methods were used, both originally developed for use on 2D midsagittal tract outline: the Mermelstein technique [8], and the Blum transform [2] as applied by Goldstein [6].

We wish to understand any observed differences between the area functions generated by the two vocal tract segmentation methods and whether or not these lead to differences in the acoustic predictions. If so, which prediction is the more accurate?

II. Method

The MR image processing mainly used a software package known as 3D-Doctor™[1]. 3D speech production models consisting of an outline of the head, neck and vocal tract were generated from the MR images (Fig. 1). Because MRI cannot distinguish air from bone (i.e. teeth), certain image sequences needed additional processing to determine the vocal tract airway in the oral cavity. This process is described in detail in the following sections. Following this, the vocal tract was perpendicularly sliced along its main axis at angles which depended on the aforementioned segmentation techniques.

A. Acquisition of Image Corpus

MR images were acquired for female subject CHS, who speaks American English, using the spin echo technique for good image resolution. Full volume scans were taken in the axial and coronal planes for each token. The complete image corpus consisted of the tokens [(a)f], [(i)f], [(u)f], [(a)s], [(a)θ] and [(a)ʃ]. The vowels in the parentheses indicate that the images were acquired while the subject was sustaining the fricative in that particular vowel context.

The thickness of the slices was 5 mm in the axial plane and 4 mm in the coronal plane. Both sets of volume scans consist of 25-30 slices, ranging from the larynx area up to the hard palate for the axial scans, and from the lips to the back of the pharyngeal wall for the coronal scans. Refer to [9] for full details.

Prior to the MR imaging sessions, dental impressions of the subject had been taken and made into casts for electropalatography palates. The casts were replicated using silicone rubber which was sliced in the coronal direction from the front at 4 mm thicknesses (to adapt to the coronal images). Subsequently, each slice was scanned into the computer for editing using a document scanner. The borders of each slice were manually outlined using 3D-Doctor, and matched to the corresponding image slice by associating it with the shape of the gums and the tongue against the teeth. This procedure was done prior to outlining the vocal tract boundary.

B. Image Processing

The image corpus was processed using 3D-Doctor. A manual trace of the boundary between the subject's profile and the background, and the vocal tract airway was made separately on each image slice. The outlines from each slice were then automatically connected to render 3D models of the subject's head and the vocal tract, two for each token. These axial and coronal 3D models were merged to create a single 3D model (as in Fig. 1). The ears, nose, and chin act as landmarks to assist in aligning the axial and coronal models correctly. The 3D models were used to generate vocal tract area functions (AFs). The process is described in detail in Section III. From the area functions the corresponding vocal tract transfer functions (VTTFs) were derived.

C. Generating the VTTFs using ACTRA

The AFs were processed to produce VTTFs using ACTRA [7], based on an earlier program called VOAC [4] and originally developed from an exhaust system modeling program. In ACTRA, geometric elements characterized by type, area, length and hydraulic radius, are concatenated to model the vocal tract following the assumptions of [5]. The effect of sudden area expansions is modeled by calculating equivalent end corrections. The program is capable of modeling the effect of side-branches to the main tract, allowing a full implementation of the Blum segmentation technique to be achieved. Additionally, any point in the tract can be used as the location of an acoustic source, enabling fricative sounds to be realistically modelled.

To derive the acoustic transfer function, the positive- and negative-going acoustic pressure wave components in the sections of tract both anterior and posterior to the source location are considered. Full details of the calculation process may be found in [7].

For the models described here, the pressure source is assumed to be located at a distance of 5 or 10 mm downstream of the constriction area depending on the AF. Note that ACTRA is based on a plane-wave model of acoustic transmission and therefore is only valid below the cut-on frequency of the first higher-order acoustic mode, about 4 kHz. We show VTTFs only up to 2 kHz in order to highlight the differences in the pole and zero locations for the different AFs. Future work may run VTTFs of up to 5 kHz.

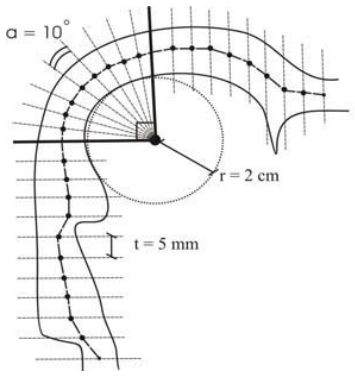

Fig. 2: Grid implementation according to the Mermelstein technique. The mid-line was found by connecting the mid-points of adjacent slices.

III. Grid-Line Implementation

A. The Mermelstein Technique

Mermelstein [8] developed the basis for an articulatory model of speech production. Variables were used to describe articulatory positions with respect to the jaw during speech. The tongue body was presented as a circle with a moving center. A fixed radius of 2 cm for the circle was found to be a good match from the X-ray tracings [3].

The vocal tract bend was segmented radially by grid-lines 10° apart converging at the center of the tongue circle, and the remainder of the tract was segmented with parallel grid lines 5 mm apart, vertically in the oral region, and horizontally in the laryngeal region. After reading the area of each plane at the grid line, the mid-line (or longitudinal axis) of the vocal tract was determined by connecting the center points of adjacent planes. In cases where the mid-line was not normal to the plane, the difference in angle was computed and the corresponding area was multiplied by a cosine factor (Fig. 2).

In this technique, any branches off the main tract are considered as a (sudden) increment in area, if at that location the branch area is connected to the plane in which the area reading is taken. However, in cases where the branches form completely separate areas from the main tract (such as the pyriform sinuses), only the area reading of the main tract is taken and the areas of the branches are discarded.

B. The Blum Transform

Blum [2] introduced the medial axis to define a systematic method of describing biological shapes. The technique was later adapted by Goldstein [6]. The technique involves the use of circles that are fitted neatly into the tract outline, such that the borders of the circle are tangent at two points of the tract.

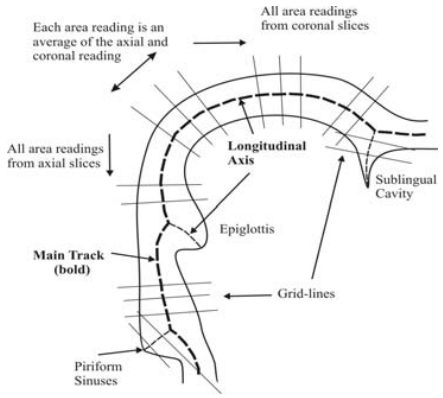

Fig. 3: Grid implementations according to the Blum transform. Assume the grid-lines down the main tract are at 5 mm intervals. There are three side-branches found in this example: sublingual cavity, epiglottis and pyriforms.

The mid-line is identified by connecting the centre-point of each closely-spaced circle. Applying this technique takes into account any side-branches that branch off the main tract. Subsequently, the areas of these side-branches are also computed for a plane normal to that of the side-branch (Fig. 3).

IV. Results

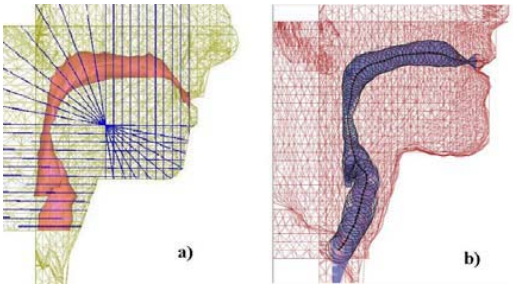

Fig. 4: Mid-sagittal view of subject uttering a) [(a)s] with tract sliced according to the Mermelstein technique and b) [(a)θ], with mid-line computed according to the Blum transform.

A. Segmentation

Fig. 4(a) shows an example of the vocal tract outline for the utterance [(a)s] segmented according to the Mermelstein technique. The grid-lines are integrally defined in this technique. Fig. 4(b) shows an example of the longitudinal axis of the tract while the subject was uttering [(a)θ] defined according to the Blum transform. This axis, consisting of the center-points of the circles, acts as the normal to the grid-lines, placed at 5mm intervals, where each area reading is taken.

B. Area Functions

It was observed that the Mermelstein technique frequently gave area readings that were substantially lower in comparison to the Blum technique, specifically in the laryngeal area of the tract. However, in a few cases, the AFs showed similar area readings throughout the vocal tract. Examples of both cases will be discussed here.

Fig. 5(a) shows the AFs for the token [(i)f] for the Mermelstein technique, and Figs. 5(b)-(c), for the Blum technique. Fig. 5(c) shows the areas of any separate branches from the Blum technique. In this example, the two techniques produced similar AF shapes in the main tract, although one may notice the difference in the overall vocal tract length.

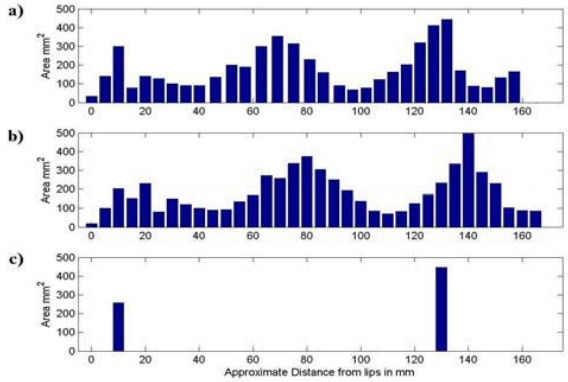

Fig. 5: AFs for [(i)f] from a) Mermelstein technique, b) Blum transform, c) branches from Blum transform.

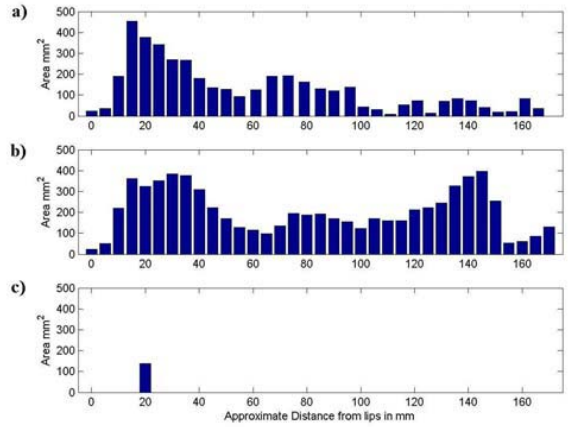

Fig. 6: AFs for [(u)f] from a) Mermelstein technique, b) Blum transform, c) branches from Blum transform.

Fig. 6(a)-(c) shows the AFs for the token [(u)f]. In this case, the Mermelstein technique did not capture as much area in the laryngeal region as the Blum technique. This occurrence was observed in 4 out of the 6 AFs we derived.

C. Vocal Tract Transfer Functions

Fig. 7(a) shows the VTTF generated by ACTRA from the AFs of the token [(i)f] by both techniques. The similarities in the AFs seen in Fig. 5 are reflected in the similar VTTFs. It is important to note that the VTTFs from the Blum data incorporate areas of the side-branches into their computation.

Fig. 7(b) shows the VTTFs for the AFs of the token [(u)f]. Corresponding to the differences in the AFs, these transfer functions show clear differences in shape and in the locations of the resonances, especially at higher frequencies.

V. Discussion

We have identified three reasons why the Blum transform produces area functions that differ from the Mermelstein technique. 1) The longitudinal axis of the vocal tract is irregular; placing vertical/horizontal grid-lines 5 mm apart in the oral and laryngeal region of the tract results in discrepancies in the true length measurement of the tract, hence the differences in the total lengths between the two AFs. 2) The Mermelstein technique determines the longitudinal axis after grid placement in order to compute the area. This means that the position of the subject could compromise the data, especially if the subject was not in exactly the same position for image acquisitions of the different tokens. 3) Areas of any side-branches off the main tract are modeled as separate tubes in the Blum technique, but in the Mermelstein technique, they are either included in the area of the corresponding plane or else neglected depending on the connectivity of the regions.

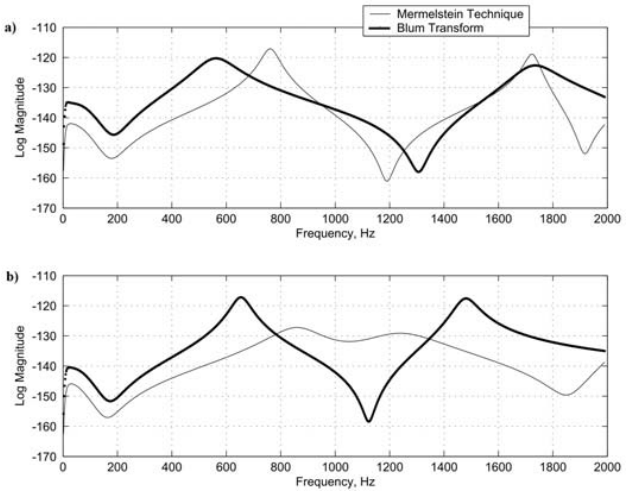

Fig. 7: VTTFs for the tokens a) [(i)f], pressure source at 5 mm from lips and, b) [(u)f], pressure source at 10 mm from lips, for the AFs shown in Figs. 5 and 6 respectively.

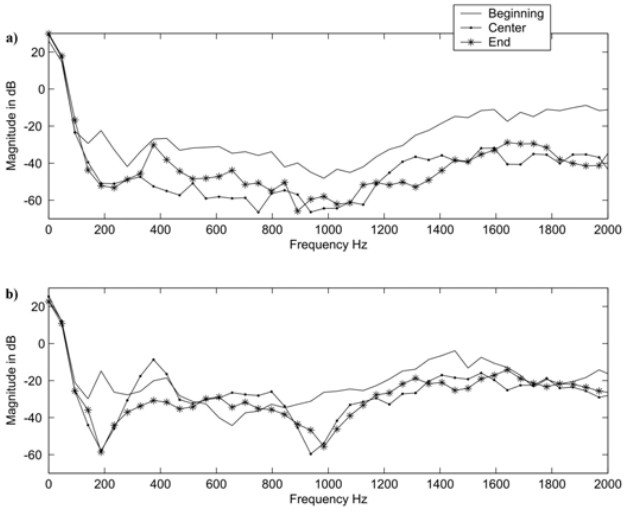

Fig. 8: PSDs for /f/ in the vowel contexts a) /i/ and b) /u/ measured at three separate locations within a sustained token.

It is clear from Fig. 7 that differences in the AFs generated from the two techniques can be expected to translate into differences in acoustic behaviour. We need to consider which technique gives the most realistic approximation to the measured sound output for these fricatives.

Fig. 8 shows the power spectral densities (PSDs) of the fricative portion of the tokens [ifi] and [ufu] respectively, uttered by subject CHS while lying down. The fricative portion was sustained for approx. 3-4 seconds. The PSD was calculated at three separate locations along the fricative steady-state: beginning, center, and end, with approximately 170 ms of signal in each segment. Bear in mind that the PSDs of the measured speech signal are the product of the vocal tract filter with an acoustic source spectrum, thus features in the PSDs may be influenced by features of the source. They are not pure acoustic transfer functions; hence comparisons with the VTTFs should be made with caution.

The VTTFs derived from the AFs and the measured spectra show similarity in the existence of a trough at 200 Hz. As for both VTTFs for [(i)f] and the Blum VTTF for [(u)f], the measured spectra show a tendency towards a pattern with two distinct peaks. For [ifi] these lie at approximately 400 Hz and 1600 Hz, giving a better comparison to the Blum-derived VTTF especially for lower frequencies. For [ufu] the peak at 400 Hz is rather lower than that in either VTTF while the upper peak at approximately 1500 Hz corresponds best with the Blum-derived VTTF. For [ufu] the trough at about 1000 Hz also corresponds well to the Blum-derived VTTF.

We conclude that differences in area function do result in differences in acoustic behaviour and that using the Blum technique generally offers a slightly more realistic approximation of the area function for use in predicting the vocal tract acoustics of fricatives.

References

[1] Able Software Corp., `http://www.ablesw.com/3d-doctor/`.
[2] H. Blum, "Biological shape and visual science: Part 1," *Journal of Theoretical Biology*, vol. 38, pp. 205-287, 1973.
[3] C. Coker and O. Fujimura, "Model for specification of the vocal tract area function," *Journal of the Acoustical Society of America*, vol. 40, pp. 1271, 1966 (abstract).
[4] P. O. A. L. Davies, R. S. McGowan and C. H. Shadle, "Practical flow duct acoustics applied to the vocal tract," in *Vocal Fold Physiology: Frontiers in Basic Science*, 1st ed. R. Titze, Ed. Singular Publishers, San Diego, CA, 1993, pp. 93-142.
[5] J. L. Flanagan and L. Cherry, "Excitation of vocal tract synthesizers," *Journal of the Acoustical Society of America*, vol. 45, no. 3, pp. 764-769, 1969.
[6] U. G. Goldstein, "An Articulatory Model for the Vocal Tracts of Growing Children," PhD dissertation, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1980.
[7] P. J. B. Jackson, "Characterisation of Plosive, Fricative and Aspiration Components in Speech Production," PhD dissertation, Department of Electronics and Computer Science, University of Southampton, UK, 2000.
[8] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070-1082, 1973.

[9] C. H. Shadle, M. Tiede, S. Masaki, Y. Shimada and I. Fujimoto, "An MRI study of the effects of vowel context on fricatives," *Proceedings of the Institute of Acoustics*, vol. 18, no. 9, pp. 187-193, 1996.