SPEAKER-SPECIFIC TYPICAL VOCAL TRACT SHAPES OBTAINED USING DYNAMIC MRI

Zeynab Raeesy

Phonetics Laboratory, University of Oxford, UK
zeynab.raeesy@phon.ox.ac.uk

ABSTRACT

Biomedical imaging techniques are applied to observe the hidden process of human articulation. We propose a new approach for obtaining a *typical* image for a phoneme's articulation. We use vocal tract images captured using dynamic MRI, and by considering image spatial characteristics (intensities), we calculate the average of different instances of phoneme articulation. The obtained image presents a typical phoneme articulation for each speaker, acquired during running speech. We investigate the comparability of our results with what is suggested in the literature.

Keywords: typical articulation, vocal tract shape, dynamic MRI

1. INTRODUCTION

Understanding human articulation is an important part of human speech research. Biomedical imaging techniques have been commonly used for obtaining models of human articulation. The early images of human articulation were mostly captured by x-ray technologies (x-ray radiography [4, 11] and cinefluroscopy/x-ray films [6, 9]). These preliminary images and movies played an important role in observing vocal tract shape and articulation.

More recent studies use CT [7, 12] and magnetic resonance imaging (MRI) technique for capturing vocal tract shape [2, 3, 5, 10]. In static MRI imaging, a sustained articulation for a period of time is required for the image acquisition to be complete, and is therefore more useful in capturing vowels and consonants that can be sustained over time. Dynamic and real-time MRI imaging techniques have been used to capture articulation during running speech. Dynamic MRI images are captured and aggregated over a sequence of repetitions of speech, and are relatively high-resolution. Real-time MRI captures the images at natural speaking rate, and therefore has to deal with spatial and temporal resolution trade-offs.

The imaging techniques have been used for observing different aspects of articulation, among which we only focus on obtaining images that depict a phoneme's articulation. In prior work, the obtained image model for articulation of phonemes has been either the static image resulted from holding the production of a sound, or a single instance image selected manually from a range of images of a particular phoneme articulation. In general, the obtained model depicts only one position and sometimes one moment of phoneme articulation, and the variability resulting from articulating phoneme in different contexts and with different speaking rate is ignored.

In this paper, we propose a method for obtaining typical phoneme articulation image for each individual speaker. For each phoneme, we specify a set of articulation instances that are different depending on the context and speaking rate. Having a range of various articulation instances for each phoneme, we produce a representative image that can demonstrate the typical phoneme articulation of a speaker.

2. MRI DATABASE

We used a database of dynamic MRI movies and their corresponding acoustic data (i.e. audio recordings), collected as the output of a previous research project [1]. The images were collected from 20 British native speakers (10 males and 10 females). The MRI device used was a 1.5 Tesla MRI unit.[1]To acquire the images, the subjects laid in the MRI scanner and articulated a set of utterances repeatedly 20 times. The timing of the repetitions of the utterance was governed by a metronome, and the speakers spoke in time to the metronome beats.

For each utterance, a sequence of 68 mid-sagittal images were captured at intervals set according to the metronome rate. Figure 2 (a) and (b) are two sample MRI negatives from the database. The "head wrap" in the images at the top of the speaker's head is an artefact caused by scanning the area beyond the "field of view" of the image. These areas are aliased back to the image and result in the wrap-around.

Audio signals were simultaneously recorded during the image acquisition by a non-magnetic gradient microphone that was fitted inside the scanner approximately 5 cm from the subject's mouth.

To achieve a better SNR, the scanner noise was cancelled from the signal by signal processing techniques. The repetitions are generally very similar.

2.1. Audio alignment

Although the transcriptions of the speech signals in the MRI database are known, the signals are not aligned in time with the transcriptions. To align the audio and transcriptions and determine the phone boundaries in the acoustic data, we developed an HTK ASR system [13]. Due to the inevitable noisy conditions inside the scanner and the narrow bandwidth of the microphone, the collected speech data are relatively poor in quality even after the noise cancellation. In addition, the available data were not sufficient to train an ASR system. Therefore, the models were trained on a separate "clean" corpus, in addition to the speech from the MRI database. The clean corpus contains speech recordings collected for training British English acoustic models in a previous project [8]. The models were retrained with MRI audio data to be adapted to the characteristics of the target acoustics. The trained ASR system was used to force-align the MRI audio data with the transcriptions.

2.2. Alignment of images and audio

Each sequence of images, despite being acquired over multiple repetitions, only depicts one articulation instance of the target phrase, and thus needs to be handled with care.

The main step in image alignment is to map subsequences of images from the sequence to the segments of audio signal. In other words, this step involves specifying windows of sequential frames representing the articulation of each phone. The number of sequential images, i.e. window length, is calculated from the duration information of the transcription-aligned audio data (equation 1):

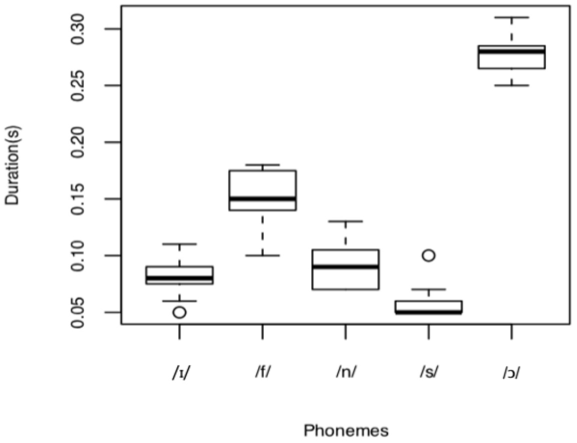
$$(1) \quad winlen_{s_p} = K \frac{dur_{s_p}}{\sum_{i=1}^{P} dur_{s_i}}$$

where P is the number of phonemes in a phrase, dur_{s_i} is the duration of the phoneme s_i, and K is the number of frames in the phrase movie.

The transcription-aligned speech signals, however, represent the audio data recorded over the entire duration of image acquisition, consisting of nearly 20 repetitions of the same phrase. Consequently, there are multiple instances of the phoneme repeated over the entire utterance. To achieve a single value for each phoneme duration, the average duration of different occurrences of the phoneme in the utterance is considered. Note that if a phoneme occurs in two different positions in one phrase, the durations at each position is calculated independently.

To check the validity of this approach, we calculated the median and deviation of phoneme durations for different utterances. The results suggested that maximum duration deviation from median for each phoneme in each utterance is mostly an order of milliseconds. The boxplot in Figure 1 illustrates the results for the utterance "enforce". In this example, the maximum deviation from the median is almost 50 ms (/f/).

Figure 1: Variation of phoneme durations in multiple repetitions of phrase "enforce" in a single utterance.



For each speaker, all the obtained image subsequences for a particular phoneme across different utterances are merged into one set of articulation frames for that phoneme. An alternative approach would be to select only the middle frame in the image window of each phoneme (in each utterance) to be added to the set of articulation frames of that phoneme. We only investigate the first approach in this work, and leave the latter for future work.

3. TYPICAL PHONEME ARTICULATIONS

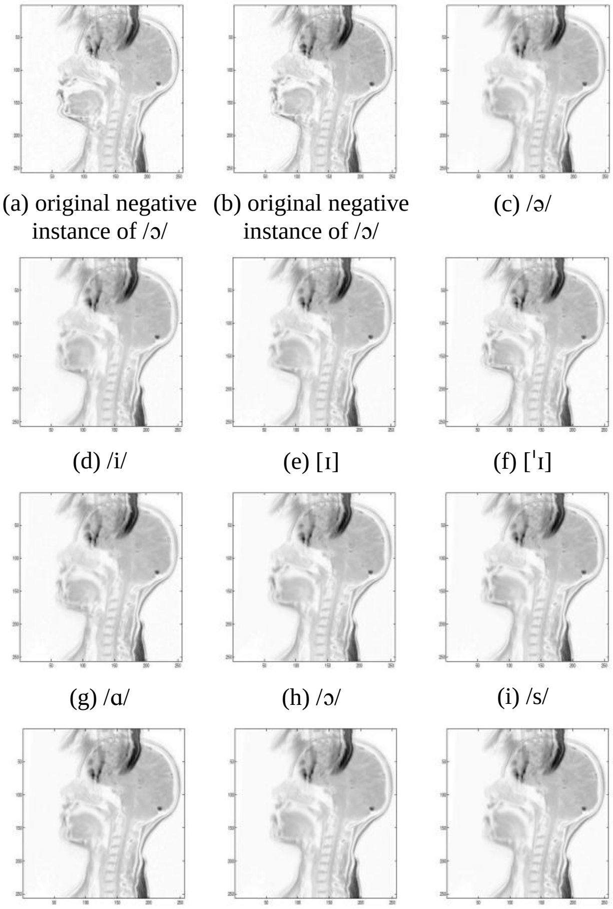
To obtain a representative image for each phoneme, the intensity properties of the digital image are used. The MRI images are 256×256 pixels, where each pixel in the image matrix has a value between 0-255 in grey scale. The vocal tract shape is visible in the negative images as a white cavity shaped by the contrast of the colours in the boundaries of the vocal tract tissues and the air. The image representing a typical phoneme articulation for each speaker is calculated by pixel-by-pixel averaging over all the images corresponding to that particular phoneme, obtained during image alignment process described in section 2.2. The result is an image depicting the average position of articulators for a particular phoneme. Each tissue in the average image may appear in different shades of grey due to the overlapping of tissue and cavity areas in different images (appearing as a blurring around the edges).

4. RESULTS AND DISCUSSION

Figure 2 (c)-(h) shows the average images of 6 vowels articulated by a single subject. The mean articulatory configuration in the figures is more or less in agreement with the expected vocal tract shapes for each sound. For example, in articulation of mid- central vowel /ə/, the vocal tract shape has a relatively uniform cross-sectional width in both the upper and lower airways, as has been observed in many previous studies. In contrast, the vocal tract cavity has a horn-shape with narrow opening in the up- per airway and a wide gap in the pharynx for vowel /i/. A similar pattern can be observed for vowels [ɪ] and [ˈɪ], where a narrow constriction is expected in the upper airway. The tongue body in the latter two images is relatively front, and a wider aperture is visible in the pharynx. For open vowel /ɑ/, a reverse configuration is observed with a relatively wide opening between the tongue body and palate in the front, and a narrower airway at the pharynx. The articulatory model in the average image of vowel /ɔ/ shows a constriction at the back of the throat between the tongue back and the soft palate.

The configurations observed in average images representing the production of consonants (Figure 2 (i)-(l)) also agree with prior work. In the image for phoneme /s/, a constriction of the tongue tip is clearly visible. For the alveolar plosives /d/ and /n/ a similar configuration can be observed, with the tip of the tongue contacting the alveolar ridge, and with the velum slightly lowered creating an entrance into the nasal cavity. The average image for /f/ can be distinguished from the rest of the images in this set considering the low position of the tongue tip. The distance between the upper and lower lips is smaller in /f/ compared to /d/, /s/ and /n/ as the lower lip is constricted towards the upper teeth.

Figure 2: Two randomly selected original MRI negatives from the window of articulation images of /ɔ/ ((a) and (b)), and average articulation images ((c) - (l)) obtained across 20 utterances of a female speaker.



(a) original negative instance of /ɔ/ (b) original negative instance of /ɔ/ (c) /ə/

(d) /i/ (e) [ɪ] (f) [ˈɪ]

(g) /ɑ/ (h) /ɔ/ (i) /s/

(j) /f/ (k) /d/ (l) /n/

This approach can be used to image articulatory configurations such as the hold phase of plosives, which is hardly possible with other methods such as static MRI. Another application of this technique would be to show the amount of variation in each sound which could inform theories of coarticulatory resistance. Moreover, it would be interesting to show how much each articulation (in different vowel contexts) differs from the averaged articulation. However, measuring the variation requires quantifying the image through segmentation which is a nontrivial task. In order to have more quantifying characteristics of vocal tract such as average vocal tract width and the associated variations, initially we need to analyse the images, which has been the focus of this work. We obtained typical phoneme articulation images of two other speakers from the database and observed consistent results. However, we leave inter-speaker typical articulation comparisons for future work.

5. CONCLUSIONS

A typical phoneme articulation for each speaker should be representative of articulation in different context with different rhythm and speed. Despite the many models of articulation obtained by imaging techniques in the literature, most of them depict single instances of articulation rather than a general and typical model. In this work, we obtained a typical speaker-specific phoneme articulation model by calculating an average of the articulation instances for each phoneme.

6. ACKNOWLEDGEMENTS

The author is grateful to Professor John Coleman for his insightful comments on this work.

7. REFERENCES

[1] Alvey, C., Orphanidou, C., Coleman, J., McIntyre, A., Golding, S., Kochanski, G. 2008. Image quality in non-gated vs. gated reconstruction of tongue motion using Magnetic Resonance Imaging: A comparison using automated image processing. *Int. J. Comp. Assisted Radiography and Surgery* 3(5), 457-464.

[2] Baer, T., Gore, J.C., Gracco, L.C., Nye. P.W.1991. Analysis of vocal tract shape and dimensions using Magnetic Resonance Imaging: Vowels. *J. Acoust. Soc. Am.* 90(2), 799-828.

[3] Demolin, D., Hassid, S., Metens, T., Soquet, A. 2002. Real-Time MRI and articulatory coordination in speech. *C. R. Biol*. 325(4), 547-556.

[4] Fant, G. 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton.

[5] Foldvik, A.K., Kristiansen, U., Kværness, J. 1993. A time-evolving three dimensional vocal tract model by means of Magnetic Resonance Imaging (MRI). *Proc. 3rd Eurospeech*, 557-560.

[6] Heinz, J.M., Stevens, K.N. 1964. On the derivation of area functions and acoustic spectra from cineradiographic films of speech (A). *J. Acoust. Soc. Am*. 36(5), 1037-1038.

[7] Kiritani, S., Tateno, Y., Iinuma, T., Sawashima, M. 1977. Computed tomography of the vocal tract. In: Sawashima, M., Cooper, F. (eds), *Dynamic Aspects of Speech Production*. Tokyo: University of Tokyo Press, 203-206.

[8] Loukina, A., Kochanski, G., Shih, C., Keane, E. 2009. Rhythm measures with language independent segmentation. *Proc. 10th Interspeech* Brighton, 1531-1534.

[9] Munhall, K.G., Vatikiotis-Bateson, E., Tohkura, Y. 1998. X-Ray film database for speech research. *J. Acoust. Soc. Am*. 95(5), 2822-2822.

[10] Narayanan, S.S., Nayak, K.S., Lee, S., Sethy, A., Byrd, D. 2004. An approach to real-Time Magnetic Resonance Imaging for speech production. *J. Acoust. Soc. Am*. 115(4), 1771-1776.

[11] Sundberg, J. 1969. On the problem of obtaining area functions from lateral x-ray pictures of the vocal tract. *STL QPSR* Stockholm, 43-45.

[12] Sunderberg, J., Johansson, C., Wilbrandb, H., Ytterbergh, C. 1987. From sagittal distance to area: A study of transverse vocal tract cross-sectional area. *Phonetica* 44, 76-90.

[13] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. 2006. *The HTK Book. http://htk.eng.cam.ac.uk/*

[1] Signa HDx, GE Medical Systems, Milwaukee, WI.