Parametrising Degree of Articulator Movement from Dynamic MRI Data

Zeynab Raeesy, Ladan Baghai-Ravary, John Coleman

Phonetics Laboratory, University of Oxford, UK

{zeynab.raeesy,ladan.baghai-ravary,john.coleman}@phon.ox.ac.uk

Abstract

A new approach is proposed for quantifying the degree of articulator movement within a phoneme as a single scalar value, using vocal tract images captured using dynamic MRI. This indicates the degree of physical movement of the articulators involved in speech production, rather than the acoustic consequences of that movement.

We go on to show that this is a valid method for characterising the overall dynamics of the vocal tract, by demonstrating a coefficient of determination (R^2) of 0.61 between it and a similarly-defined scalar measure of the acoustic dynamics of the signal.

The calculation of the new measure involves production of images showing the exact location of any movement within the vocal tract, and additionally shows this information separately for the initial and final segments of each phoneme.

Finally, we demonstrate that although some sounds may involve *more* movement of the articulators than would be expected from the dynamics of the acoustic signal, it is rare for the degree of articulation derived from the MRI data to be significantly *less* than expected.

Index Terms: degree of articulation, phoneme duration, vocal tract shape, dynamic MRI.

1. Introduction

Some phonemes require greater movement of the vocal tract's articulators than others. There have already been many attempts to analyse the movement of these articulators [1-6] and to apply those movements to speech coding [7], speech synthesis [8, 9] and speech recognition [10, 11]. These methods have invariably attempted to use detailed models of the articulators' positions to predict details of the temporal and spectral aspects of the signal. This can make the analysis much more complicated and error-prone than is necessary in some applications.

This work addresses estimation of the degree of articulatory movement required to produce the speech sounds, rather than analysing the exact movement of any particular articulator. Each time a speaker articulates a phoneme in a different context there are variations in its articulation. We use multiple instances of articulation in different contexts and find an average of normalised measures of the degree of dynamic articulation within those instances. We propose this as a method for quantifying the degree of dynamic articulation involved in producing a particular phoneme.

The paper is structured as follows: a brief introduction to MRI vocal tract imaging is presented in section 1.1, followed by a description of the database in section 1.2. The methodology is explained in sections 1.3 and 1.4. Finally, the results are presented and discussed in section 2, followed by concluding points in section 3.

1.1. Vocal Tract MRI

Magnetic resonance imaging techniques have been commonly used for obtaining models of human articulation and vocal tract shape [12 – 17]. However, the natural articulation rate is considerably faster than the typical acquisition time required by the MRI scanners and different approaches have been suggested in the literature for addressing this problem.

In static MRI imaging, a sustained steady-state articulation is required for the image acquisition to be complete, and it is therefore more useful in capturing steady vowels and fricatives [12]. Real-time MRI captures the images at natural speaking rate, and therefore has to deal with spatial and temporal resolution trade-offs [13, 14]. In dynamic MRI, images are captured and aggregated over a sequence of repetitions of an utterance, and can offer a relatively high spatial and temporal resolution [15, 16, 17].

Dynamic and real-time MRI imaging techniques are both used to capture articulation during running speech; the main difference between the two methods is in the procedure for obtaining and reconstructing the images. In dynamic MRI, the image data are collected from a dynamically articulating subject and the images are reconstructed offline across different repetitions upon completion of the acquisition procedure. In real-time MRI, the MRI movies are captured in real time, while the speaker is articulating. Consequently, real-time MRI is able to give the visual impression of continuous articulation, but details of some very rapid articulatory movements are likely to be missed. The animation constructed offline by dynamic MRI can show the movement of articulators with a very fine degree of temporal and spatial resolution.

The choice of MRI technique depends on the final application and the importance of the image's temporal and spatial resolution.

1.2. MRI Data

We used a database of dynamic MRI movies and their corresponding acoustic data (i.e. audio recordings), collected as the outcome of a previous research project [17]. The images were collected from 20 native speakers of British English (10 male and 10 female). The MRI device used was a 1.5 Tesla MRI unit[1]. Due to the long exposure time required for capturing each image, only a relatively small number of phones were selected for inclusion in this database. These are listed in IPA notation in Table 1.

Table 1. *Phone list (IPA).*

[ˈɑː]	[ə]	[ˈʌ]	[ˈɔː]
[d]	[f]	[ɪ]	[ˈɪ]
[ˈiː]	[n]	[s]	

To acquire the images, the subjects lay 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 images was captured at intervals set according to the metronome rate (approximate image temporal resolution of 0.05s). The images are mid-sagittal images from the bottom of the subjects' neck

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

up to the top of the head. Figures 1 and 2 are the first and last one-thirds of the average image for a single speaker's articulation of the phone [ˈɔː].

Acoustic 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 out using signal processing techniques, as in the original project [17]. Because of the recording procedure, both the acoustic waveforms and MRI images are very similar from one instance to the next, allowing a simple pixel-by-pixel comparison between images.

For the work described here, six speakers were selected on the basis of the relatively high quality of their data.

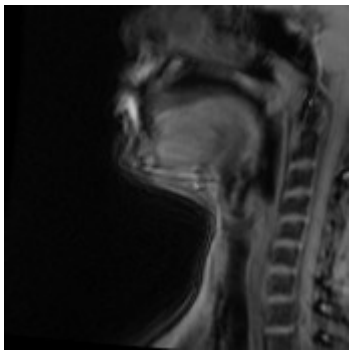

Figure 1: *Averaged image for the first one-third of the phone* [ˈɔː] *for a single speaker*

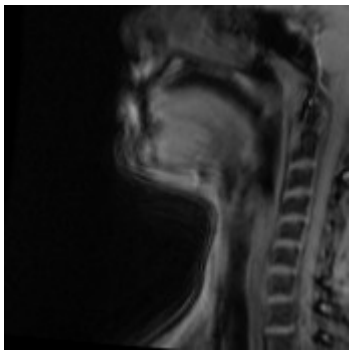

Figure 2: *Averaged image for the final one-third of the phone* [ˈɔː] *for a single speaker*

1.3. Alignment

Although the transcriptions of the speech signals in the MRI database are known, neither the acoustic signals nor the visual (MRI) data are aligned in time with the transcriptions. To align the audio recordings with the transcriptions, and thus determine the phoneme boundaries in the acoustic data, we developed a set of acoustic HMMs [18] and used Viterbi alignment to label the signal.

Due to the levels of electromagnetic and acoustic noise inside the scanner, the collected speech data is relatively poor in quality even after the noise cancellation. In addition, neither the quantity nor the diversity of the acoustic MRI data were sufficient to train an HMM-based alignment system. Therefore, the models were trained on a separate "clean" corpus, supplemented with the speech from the MRI database. The clean corpus contains speech recordings collected for training British English acoustic models in a previous project [19]. The MRI audio recordings were also included to provide data that is more representative of the characteristics of the target signal.

The trained alignment system was used to force-align the MRI acoustic data with the transcriptions. A subset of the data was also aligned manually, and the automatic labels were compared against the manual labels. 93% of the automatic labels agreed to within 50ms of the manual ones (81% agreed to better than 20ms).

The initial frame of the image sequence was then located manually. The video and audio signals were synchronised by identifying subjectively plausible start-times for the respective signals. The locations of individual phonemes were derived from the acoustic data, with respect to the start-times.

1.4. Degree of Articulation

1.4.1. MRI Data

For this work, we took images covering each one-third duration of the respective phones, and averaged them separately over multiple instances of the relevant phone. We divided the durations into thirds on the assumption that the widely-accepted 3-state phoneme models used in ASR would be sufficiently detailed to capture articulatory movement as well as acoustics.

By taking the difference between the averaged MRI images of the first, middle and final thirds of each phone we quantified the amount of articulatory movement within an instance of a phone. The average magnitudes of the differences between first and middle, and between middle and last, thirds of the phone [ˈiː] are shown in Figure 3 and 5.

The overall degree of articulation of each phone is quantified by summing the intensity levels indicated in images such as Figure 3 and 5. These figures capture the articulatory movements in the first and final parts of each phone, respectively.

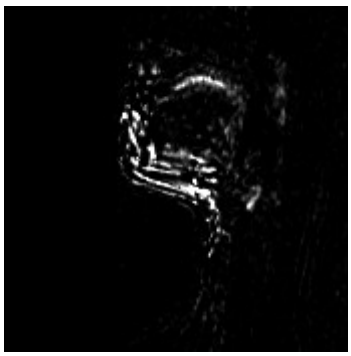

Figure 3: *Image showing the average articulatory movement between the first and middle thirds of the phone* [ˈiː] *for a single speaker. Most movement is observed in the back, near the hard palate, and the lower jaw.*

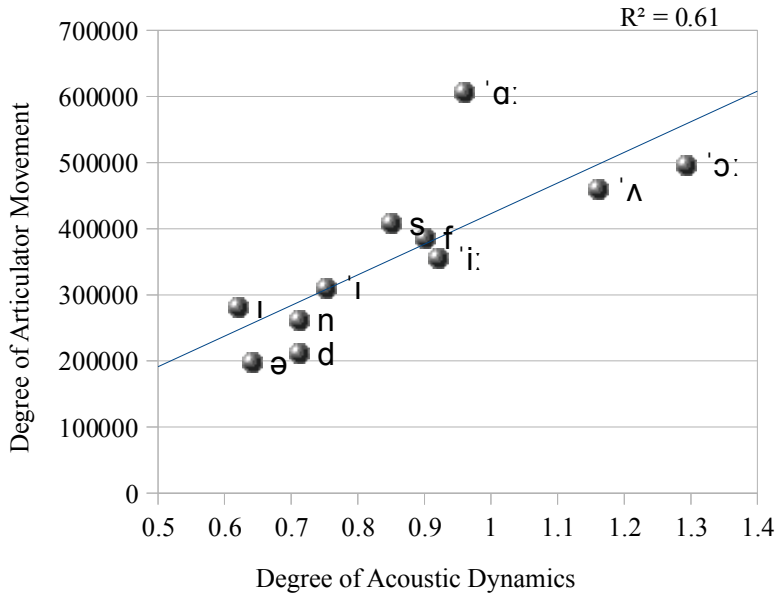

Figure 4: *Phone articulation measures averaged over six speakers: the degree of dynamics in the MRI data plotted against that in the acoustic signal.*

Figure 5: *Image showing the average articulatory movement between the middle and final thirds of the phone* [ˈiː] *for a single speaker. The jaw movement is comparable with that in Figure 3, but the movement of the tongue is now at the front of the mouth.*

We then normalised these 'degree of articulation' images to suppress variability due to the size of the image and the speakers' vocal tract. These normalised measures indicate the articulator movement involved in production of each phone, and the measures from the different speakers can be averaged to give an overall measure of 'degree of articulator movement' for first and final parts of each phone.

1.4.2. Acoustic Data

For the acoustic data, a measure of acoustic dynamics was devised, again by comparing the first, middle, and final thirds of each labelled phone. In this case, a modified symmetrical form of the Itakura-Saito distance was calculated between the acoustic signals for the same regions as in the MRI analysis. A single Itakura-Saito distance was calculated based on the whole of each third of the respective phone. The Itakura-Saito distance was chosen for its mathematical simplicity and well-documented behaviour.

These acoustic parameters were, like the MRI-based measure of articulator movement above, averaged over all instances of each phone, spoken by an individual speaker, and their range normalised separately for each speaker. These normalised values were then averaged across all speakers to provide a single measure of 'degree of acoustic dynamics' associated with the first and final parts of each phone.

2. Results and Discussion

To illustrate the relationship between articulator movement and acoustic dynamics, the measures described in Sections 1.4.1. and 1.4.2. were plotted in Figure 4. This figure combines the data for first and final parts of the respective phones, by summing them; this is in order to capture the total degree of dynamics within each phone.

Table 2. *Locations of visually significant movement for first and final parts of each phone for a single speaker. Parentheses indicate very slight movement.*

Phone (IPA)	First Part	Final Part
[ˈɑː]	**Jaw**	**Tip of tongue & jaw**
[ə]	-	-
[ˈʌ]	**Tip of tongue** (& jaw)	(jaw)
[ˈɔː]	**Tip & back of tongue** (& jaw)	**Tip of tongue & jaw**
[d]	-	-
[f]	**Lips** (& jaw)	**Lips** (& jaw)
[ɪ]	-	(Tip of tongue & jaw)
[ˈɪ]	**Front edge of tongue** (& jaw)	(Jaw)
[ˈiː]	**Back of hard palate & jaw**	**Front of hard palate & jaw**
[n]	**Tip of tongue** (& jaw)	**Velum & root of tongue** (& jaw)
[s]	**Tip of tongue** (& jaw)	(Tip of tongue & jaw)

There is a clear correlation between articulator movement and acoustic dynamics, with a coefficient of determination (R^2) of 0.61, which is highly significant ($t=3.75$; $p<0.01$). One phone [ˈɑː] has significantly greater articulator movement than would be expected from the associated acoustic dynamics, but the others are very clearly and directly related.

The full set of 'degree of articulation' images of a single speaker (corresponding to Figure 3 and 5 for all the phones in Table 1) were examined, and regions of high movement which they revealed were noted.

The results are summarised in Table 2. These subjective observations were made before the objective results in Figure 4 were available, but are qualitatively in close agreement with them.

From these observations it appears that although most of the phones in the table would normally be considered "steady" sounds, only three [ə], [ɪ] and [d] involve very little or no movement of the articulators during their production. This lack of articulation of unstressed vowels [ə] and [ɪ] may be because the short duration of the sounds obtained from the aligned acoustic data. The shorter durations lead to fewer frames being associated with the phone articulation and, not much variation can be observed in short sequences of their corresponding image frames, which can be very few in number because of the relatively low frame rate of the MRI data.

The jaw movement observed in most of the phones may be related to the *manner of articulation* of the phones, or it could be the result of surrounding phone articulations..

3. Conclusions

In this work, we propose a novel method for observing the relationship between the physical movement of the speech articulators and the acoustic productions. Instead of looking at the shape of the vocal tract and area functions, we proposed using the degree of articulation for characterizing the acoustics.

Our results suggest that although the details of phoneme articulation can be speaker-specific, a strong pattern of correlation can be observed between the degree of spectral variation (acoustic dynamics) and the amount of movement in the vocal tract. This correlation indicates that the measure of 'degree of articulator movement' described in this paper, is indeed a valid method for characterising the overall dynamics of the vocal tract.

There are many possible uses for this novel parameter – primarily in research areas concerned with identifying the relative importance of articulatory movement during the production of different sounds.

4. References

[1] Kaburagi, T., Honda, M., "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory–acoustic database", Proc. ICSLP, pp. 433–436, 1998.

[2] Shiga, Y., King, S., "Accurate spectral envelope estimation for articulation-to-speech synthesis", Proc. ISCA Speech Synthesis Workshop, pp. 19–24, 2004.

[3] Hiroya, S., Honda, M., "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model", IEEE Trans. Speech Audio Processing, vol. 12, issue 2, pp. 175–185 , 2004.

[4] Kello, C.T., Plaut, D.C., "A neural network model of the articulatory–acoustic forward mapping trained on recordings of articulatory parameters", J. Acoust. Soc. Am., vol. 116, issue 4, pp. 2354–2364, 2004.

[5] Nakamura, K., Toda, T., Nankaku, Y., Tokuda, K., "On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum". Proc. ICASSP, pp. 93–96, 2006.

[6] Toda T., Black, A.W., Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model", Speech Communication, vol. 50, issue 3, pp. 215–227, 2008.

[7] Schroeter, J., Sondhi, M.M., "Speech coding based on physiological models of speech production", In: Furui, S., Sondhi, M.M. (Eds.), Advances in Speech Signal Processing, Marcel Dekker, pp. 231–268, 1992.

[8] Sondhi, M.M., "Articulatory modeling: a possible role in concatenative text-to-speech synthesis", Proc. IEEE Workshop on Speech Synthesis, pp. 73–78, 2002.

[9] Story, B.H., "Speech synthesis by mapping articulator movement patterns to a shape-based area function model of the vocal tract (A)", J. Acoust. Soc. Am., vol. 109, issue 5, pp. 2444–2445, 2001.

[10] Wrench, A.A., Richmond, K., "Continuous speech recognition using articulatory data", Proc. ICSLP, pp. 145–148, 2000.

[11] Frankel, J., Richmond, K., King, S., Taylor, P., "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces". Proc. ICSLP, vol. 4, pp. 254–257, 2000.

[12] Baer, T., Gore, J.C., Gracco, L.C., Nye. P.W., "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels". J. Acoust. Soc. Am., vol. 90, issue 2, pp. 799–828, 1991.

[13] Demolin, D., Hassid, S., Metens, T., Soquet, A., "Real-time MRI and articulatory coordination in speech", C. R. Biol., vol. 325, issue 4, pp. 547–556, 2002.

[14] Narayanan, S.S., Nayak, K.S., Lee, S., Sethy, A., Byrd, D., "An approach to real-time magnetic resonance imaging for speech production", J. Acoust. Soc. Am., vol. 115, issue 4, pp. 1771–1776, 2004.

[15] Foldvik, A.K., Kristiansen, U., Kværness, J., "A time-evolving three dimensional vocal tract model by means of magnetic resonance imaging (MRI)". Proc. Eurospeech, pp. 557–560 , 1993.

[16] Shadle, C.H., Mohammad, M., Carter, J.N., and Jackson, P.J.B., "Dynamic magnetic resonance imaging: new tools for speech research.", Proc. ICPhS, pp. 623– 626, 1999.

[17] Alvey, C., Orphanidou, C., Coleman, J., McIntyre, A., Golding, S., Kochanski, G., "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, vol. 3, issue 5, pp. 457–464, 2008.

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

[19] Loukina, A., Kochanski, G., Shih, C., Keane, E, "Rhythm measures with language independent segmentation. Proc. Interspeech, pp. 1531–1534, 2009.