CONSTRUCTING SPEAKER-SPECIFIC ARTICULATORY VOCAL TRACT MODELS FOR TESTING SPEECH MOTOR CONTROL HYPOTHESES

Alfonso Nieto-Castanon[1] and Frank H. Guenther[1,2]
[1]*Department of Cognitive and Neural Systems, Boston University*
[2]*Research Laboratory of Electronics, Massachusetts Institute of Technology*

ABSTRACT

Inter-speaker differences in the articulation of speech sounds are common, presumably due in part to differences in vocal tract morphology. These differences can confound the process of using articulator movement data to evaluate theories of speech motor control. This abstract describes software that allows relatively fast construction of speaker-specific articulatory vocal tract models from midsagittal vocal tract profiles, acquired with Magnetic Resonance Imaging (MRI) while speakers produce a list of approximately 30 phonemes. The resulting vocal tract models will be used in concert with experimental data to test models of speech motor control.

1. INTRODUCTION

Many examples of inter-speaker differences in articulation have been documented in the speech production experimental literature. These differences are likely based, at least in part, on individual differences in vocal tract morphology, and they can confound the process of using movement data to evaluate theories of speech movement control. Computational models of speech movement control that utilize generic, speaker-independent vocal tract models cannot account for inter-speaker differences, regardless of whether or not the control schemes used by the models are accurate reflections of biological control schemes. In contrast, if speaker-specific vocal tract models are utilized within computational motor control models, then some control schemes may be able to account for the inter-speaker differences, thus favoring these control schemes as explanations for the biological processes underlying speech movement control over control schemes that do not account for the differences.

This paper describes a procedure for constructing a speaker-specific vocal tract model from MRI midsagittal vocal tract profiles and corresponding acoustic data. The analysis steps described herein have been implemented in MATLAB software, allowing easy portability to different platforms. Emphasis was placed on extraction of realistic movement degrees of freedom and acoustic output using a minimal number of 2-D MRI scans. An overview of this procedure is provided in Figure 1.

2. PROCESSING THE ARTICULATORY DATA

2.1. MRI image acquisition

2-D MRI midsagittal profiles (4mm slice, 14s acquisition time) were acquired using a T1 Flash scan for a male subject while producing a set of 32 phonemes. This data was processed to obtain a representation of the vocal tract cavity profiles and the principal articulatory degrees of freedom of the individual speaker.

2.2. Preprocessing MRI images

Images are first binarized to allow differentiation of air and tissue regions. The intensity histogram is clustered into 8 representative levels, and the relevant air regions are selected using a flood-fill algorithm with 8-connected foreground starting from a user defined point. The vocal tract air cavity is then automatically segmented into 5 different regions (lips, hard palate, velum, laryngeal region, and tongue body), characterized by a fixed set of equally spaced points defining the contour of each region. The outline of the jaw is also obtained, as well as other external contours, for plotting purposes. This process is user-supervised via a graphical interface to allow on-line corrections, air cavity boundary constructions, etc. in any step of the otherwise automatic segmentation of the raw MRI image into relevant outlines.

2.3. Characterization of articulatory space

To construct a parametric description of articulatory movements, most previous approaches [1,2,3] create a grid in the midsagittal plane and obtain the vocal tract area function from the intersection of this grid with the vocal tract cavity. The major disadvantage of this approach, from a motor control point of view, is that the transformation between the vocal tract articulator movements and the corresponding area functions is not continuous (independent of the resolution of the grid); thus any continuous control scheme acting on the area function parameters (or any set of parameters derived continuously from them) will not produce realistic articulator movements.

We create a parametric definition of the articulator space that is simple, articulatorily meaningful, and continuous with respect to movement of the articulators using a common statistical technique known as principal component analysis (PCA [4,5]). This technique effectively decomposes the covariance of the articulator positions into an ordered set of principal axes, allowing pruning of the less significant axes in order to obtain a minimal basis in the shape space spanning the possible configurations of the different articulators. These principal axes are obtained from the eigenvectors of the covariance matrix **C**, while the covariance along these axes is obtained from the eigenvalues of **C**. Pruning the less significant axes is accomplished by selecting those axes that account for more than 5% to the total covariance.

Prior to the principal component analysis, the effect of jaw movement is extracted by performing a linear regression of all articulator positions over the jaw position (defined as the projection over the first axis from the PCA of the jaw outlines).

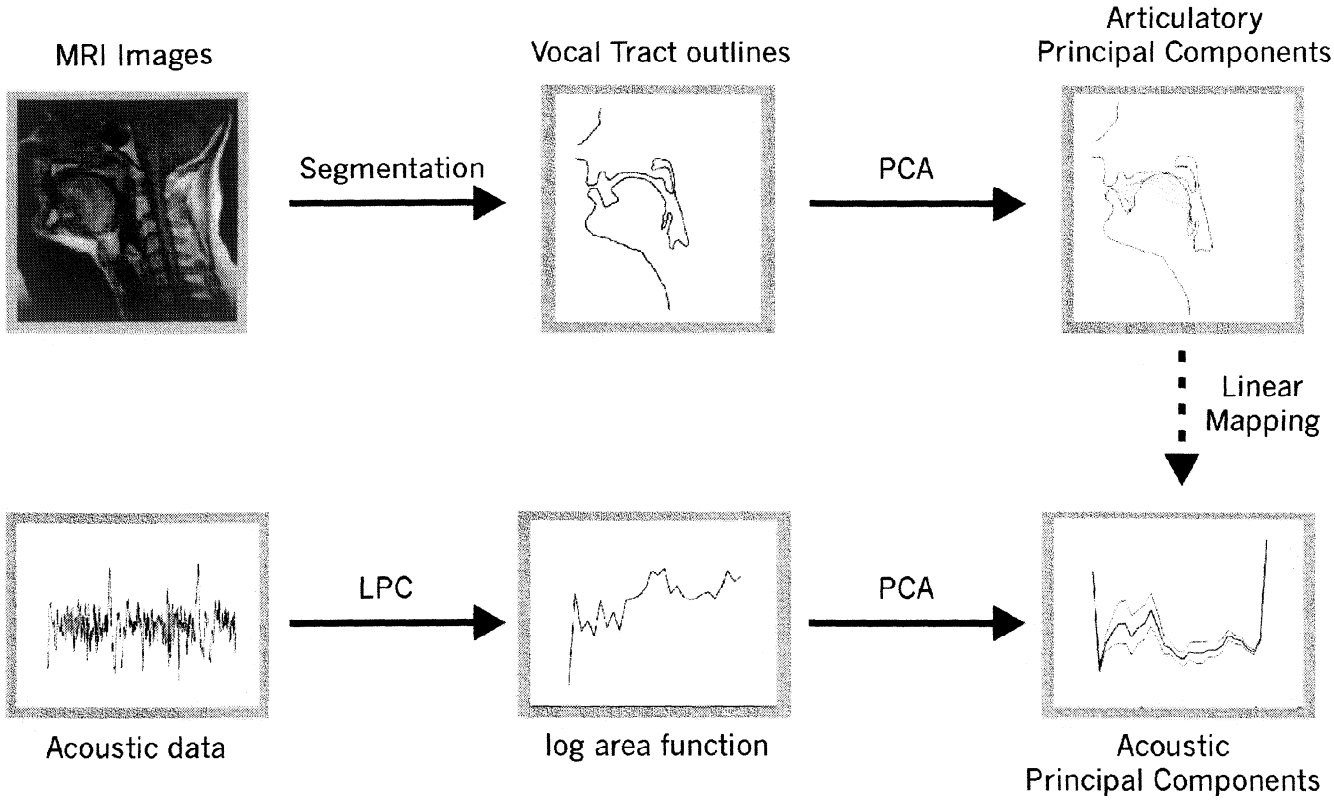

Figure 1. Overview of the processing steps for creation of the speaker-specific vocal tract model.

After this stage, PCA is performed over the different segments of the vocal tract cavity, providing a set of meaningful parameters approximating the primary articulatory degrees of freedom of the individual speaker.

An initial analysis of the raw articulator data (Figure 2) shows that a set of 7 parameters can account for 78% of the total variance of the data, while any single additional parameter would account for less than an additional 5% of the variance. The dimensionality of the relevant degrees of freedom of the vocal tract is consistent with Maeda's parametrization using a similar method [5].

The contribution of this set of articulator parameters (1 for the jaw, 3 for the tongue body, 1 for the lips, 1 for the velum, ,and 1 for the laryngeal region) to the vocal tract configuration is shown in the top half of Figure 3. The jaw component primarily describes the aperture of the lips and raising/lowering of the tongue body without any significant change in its shape; the three tongue components describe the raising/lowering of the apical and dorsal areas of the tongue and its front/back movement; the lip component describes the frontal extension of the lips; the velum component describes the opening/closing of the nasal cavity; and the laryngeal component describes the raising/lowering of the base of the laryngeal region.

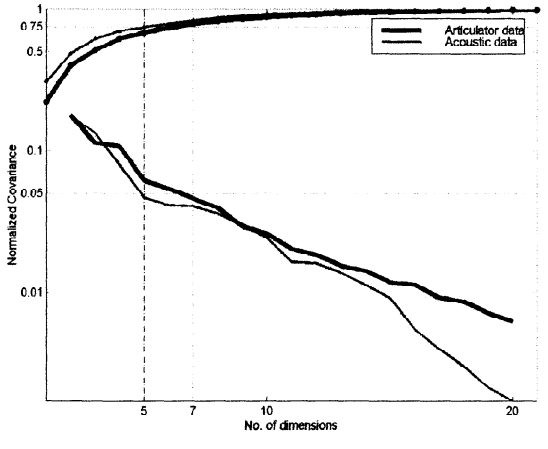

Figure 2. Principal component analysis for the articulatory (thick black line) and acoustic (thin gray line) data. The bottom two curves indicate normalized covariance along a principal axis as a function of the index of the axis. The top two curves indicate the cumulative normalized covariance. The articulatory data analysis results in 7 significant components (with covariance greater than 5% of the total variance), while the acoustic data analysis yields only 5 significant components, consistent with an effective many-to-one mapping between articulatory and acoustic information.

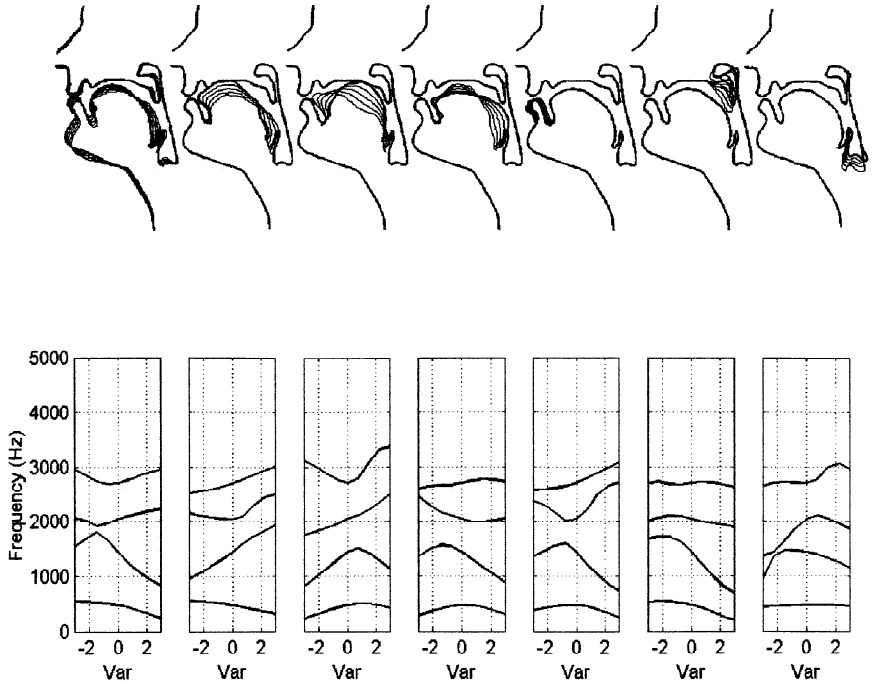

Figure 3. Contribution of principal articulatory components, ranging 3 standard deviations from the mean value, to vocal tract configuration (upper row; thick line represents highest positive deviation); and formants F1 to F4 (lower row) as obtained through the estimated articulatory-to-acoustic mapping. The set of principal components consists of (from left to right) 1 jaw component, 3 tongue components, 1 lip component, 1 velar component, and 1 laryngeal region component.

3. PROCESSING THE ACOUSTIC DATA

3.1. Introduction

In order to produce an acoustic signal with the vocal tract model, an algorithm is needed to convert midsagittal vocal tract profiles into area functions or an equivalent description of the vocal tract transfer function. In previous models, the vocal tract area function was typically approximated from midsagittal widths, either using a difficult to tune elliptical approximation [5] or an elegant but more complex estimation procedure based on multiple 3-D volumetric MRI representations of the vocal tract [6]. Collection of 3-D volumetric data for multiple phonemes is time-consuming and can suffer from problems in determining the location of the teeth, which do not show up on MRI images and thus adversely affect the measured area function.

To avoid these problems, we developed a new procedure for estimating the acoustic signal (see Figure 1). First, acoustic data collected at the start of each scan is used to estimate the vocal tract area function during that scan via linear predictive coding techniques [7]. We then perform a principal component analysis on these area functions to derive a set of "acoustic" principal components. These principal components describe the vocal tract area function as approximated from the acoustic signal. Finally, we use regression techniques to define a linear mapping between these acoustic principal components and the articulatory principal components described in Section 2.3. This linear mapping can then be inverted to transform arbitrary articulator configurations, defined in terms of the 7 articulatory principal components, into corresponding acoustic signals.

3.2. Preprocessing

Acoustic data were collected while the subject was producing phonemes in the MRI machine. Following Wakita [7], we use pre-emphasis (single delay FIR filter, a_1=.95) of the acoustic signal to remove effects due to radiation and the glottal pulse, followed by linear predictive coding (p=26, Fs=22KHz) in order to obtain equivalent vocal tract area functions. Note that these area functions are simply suitable for producing the acoustic signal; they are not necessarily the actual vocal tract area functions.

3.3. Characterization of acoustic space

The variance of the acoustic-derived area functions is again analyzed using PCA in order to find a minimal set of descriptors of the acoustic variability. Figure 2 shows that a set of 5 parameters (pruning those that account for less than .05 normalized variance) is enough to account for 74% of the total variance. This result, when compared to the set of 7 articulatory parameters, is consistent with the idea of an effective many-to-one mapping between articulator and acoustic data, while adding some insight into the degree of this dimensionality reduction.

	CPA	F1	F2	F3	F4
Vowels	E	589	1443	2381	2734
		649	*1505*	*2382*	*2668*
	IY	331	1891	2534	2647
		349	*1757*	*2538*	*2789*
	AE	674	1451	2023	2519
		658	*1403*	*2503*	*2798*
	UH	592	1232	2342	2485
		589	*1321*	*2132*	*2706*
	OO	342	1148	2115	2853
		357	*1240*	*2235*	*2719*
	I	392	1498	1970	2840
		384	*1600*	*2104*	*2712*
	OW	688	1096	2343	3143
		726	*1332*	*2430*	*2973*
	OU	502	1085	2262	2965
		497	*1130*	*2174*	*3028*
	EI	483	1573	1909	2362
		470	*1684*	*2074*	*2694*
Semivowels	RR	463	1324	1672	2827
		487	*1369*	*1864*	*2867*
	RR	671	1119	1652	3003
		626	*1441*	*1970*	*3054*
Sonorant	L	556	1061	2292	2716
		609	*1474*	*1892*	*2493*
	N	324	1439	2169	2470
		526	*1408*	*1845*	*2512*
	M	363	1395	2184	2649
		551	*1414*	*1874*	*2683*
Fricatives	F	562	1510	2422	3013
		444	*1309*	*2496*	*2819*
	V	495	1310	2449	3145
		372	*1287*	*1840*	*3132*
	ZH	374	1419	2307	3440
		420	*1166*	*2181*	*2873*
	SH		957	2390	2946
		283	*1278*	*1785*	*2788*
	S		943	1937	2592
		358	*1162*	*1777*	*2818*
	TH		807	1824	2480
		417	*1154*	*2390*	*2735*

Table 1. Measured formants (first row) and formants obtained using a linear mapping between articulatory and acoustic PCA coefficients to estimate the acoustic signal (second row). The linear mapping was obtained using data pairs corresponding to vowel and semivowel phonemes. Reasonable generalization to sonorant and some fricative consonants is observed.

3.4 Linear mapping between acoustic PCs and articulatory PCs

In order to relate the acoustic and articulatory representations, a linear mapping between both the two sets of PCs was derived using linear regression over the vowel and semivowel data pairs. With this mapping, a set of articulatory PC values (e.g., as specified by a computational model of speech motor control) can be mapped into acoustic PCs, which can in turn be inverted to create an acoustic signal that approximates the signal produced by the modeled speaker when his/her articulators are in the positions described by the articulatory PCs.

Table 1 shows the results of this map in formant space compared to the collected acoustic data. The learning set for the linear map was the set of vowels and semivowels. This mapping generalizes to some extent to sonorant and frivative consonants. Figure 3 shows the contribution of the primary articulatory degrees of freedom to the vocal tract transfer function obtained through the linear articulatory-to-acoustic mapping.

4. CONCLUSION

A procedure to construct speaker-specific vocal tract models from MRI midsagittal images and acoustic data has been described. The principal articulatory degrees of freedom of an individual speaker are efficiently represented with a set of 7 articulatorily meaningful parameters: 1 for the jaw, 3 for the tongue body, 1 for the lips, 1 for the velum, ,and 1 for the laryngeal region. Additional processing allows acoustic signals approximating those of the modeled speaker to be generated given a set of articulator positions. The resulting vocal tract model can be used to help differentiate the effects of speaker-specific vocal tract morphology from the effects of speaker-independent control strategies.

ACKNOWLEDGMENTS

This research was supported by NIDCD grant 1R29DC02852.

REFERENCES

[1] Story, B.H., and Titze, I.R. 1998. Parametrization of vocal tract area functions by empirical orthogonal modes. *Journal of Phonetics*, 26, 223-260.

[2] Story, B.H. Titze, I.R., and Hoffman, E.A. 1996. Vocal Tract Area Functions from Magnetic Resonance Imaging. *Journal of the Acoustical Society of America*, 100, 537-554.

[3] Perrier, P., Boë, L.-J., and Sock, R. 1992. Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: Modeling the transition with two sets of coefficients. *Journal of Speech and Hearing Research*, 35, 53-67.

[4] Strang, G. 1988. *Linear algebra and its applications*, 3rd ed. New York: Academy Press, Inc.

[5] Maeda, S. 1990. Compensatory articulation during speech; Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle and A. Marchal (eds.), *Speech production and Speech modelling* (pp. 131-150). Dordrecht, Holland: Kluwer.

[6] Tiede, M.K., and Yehia, H. 1996. A shape-based approach to vocal tract area function estimation. *Proc. ASA-ASJ 3rd Joint Meeting*, 861-866.

[7] Wakita, H. 1973. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics*, AU-21 (5), 417-427.