NORMALIZATION OF ARTICULATORY DATA THROUGH PROCRUSTES TRANSFORMATIONS AND ANALYSIS-BY-SYNTHESIS

Daniel Felps, Sandesh Aryal, and Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University
{dlfelps,sandesh,rgutier}@cse.tamu.edu

ABSTRACT

We describe and compare three methods that can be used to normalize articulatory data across speakers. The methods seek to explain systematic anatomical differences between a source and target speaker without modifying the articulatory velocities of the source speaker. The first method is the classical Procrustes transform, which allows for a global translation, rotation, and scaling of articulator positions. We present an extension to the Procrustes transform that allows independent translations of each articulator. The additional parameters provide a 35% increase in articulatory similarity between pairs of speakers when compared to classical Procrustes. The proposed extension is finally coupled with a data-driven articulatory synthesizer in an analysis-by-synthesis loop to select model parameters that best explain the predicted acoustic (rather than articulatory) differences. This normalization method is able to increase acoustic similarity between source and the target speaker by 34%. However, it also reduces articulatory similarity by 22%, which suggest that improvements in acoustic similarity do not necessarily require an increase in articulatory similarity.

Index Terms– analysis-by-synthesis**,** articulatory synthesis, speaker normalization

1 INTRODUCTION

Paralinguistic information in the speech signal, such as cues to the speaker's gender, dialect, emotional state, and age, are generally treated as a source of noise in automatic speech recognition. A number of acoustic *normalization* techniques have been proposed to reduce the impact of these "noise" sources while maximizing the discrimination of linguistic information (i.e. phones) [1]. Gender-specific acoustic differences are well documented, e.g., females have higher fundamental frequencies and an enlarged F1-F2 vowel space. Many of these differences, however, are rooted in the human anatomy. Female vocal folds are smaller, which cause them to vibrate at a higher frequency, and the enlarged F1-F2 vowel space has been conjectured to be a compensation for the reduced spectral sampling resulting from greater distances between harmonics [2].

Techniques for articulatory normalization are rarer, but seem timely given the recent interest in using articulatory information to boost the performance of automatic speech recognition systems [3-6]. In this article, we compare three methods for articulatory normalization: the Procrustes transform (global translation, rotation, and scaling), an extension of the Procrustes transform with independent translation components for each articulator, and an analysis-by-synthesis method that seeks to minimize acoustic rather than articulatory differences.

Relation to prior work. Our Procrustes transform is similar to the one used by Geng and Mooshammer [7] to investigate vowel production strategies independently of speaker-specific vocal tract anatomy. The authors presented a qualitative evaluation on a few vowels; in contrast, we evaluate the normalization method objectively in terms of improvements in articulatory and acoustic similarity between speakers. Our analysis-by-synthesis approach is also related to that of McGowan and Cushing [8], but has several key differences. First, we use a data-driven concatenative articulatory synthesizer [9], as opposed to one based on a physical model of the vocal tract. Second, our approach is fully automated and allows a direct comparison of articulators from two speakers (i.e., as opposed to requiring an intermediate vocal tract model.) Finally, our study includes consonants and diphthongs and is validated on a large number of speaker pairings.

2 PRIOR WORK

Techniques for articulatory normalization can be found in research on speech kinematics. For instance, Hashi et al. [10] identified palatal height as a systematic source of variation that could be accounted for by first scaling the articulatory data to a common size then expressing the tongue relative to the palate, and the lips relative to one another. In an investigation of the different ways to produce [r], Westbury et al. [11] expressed tongue pellet positions as ordered triple angles to generalize the shape of the tongue. Simpson [12] described a weighted palatal normalization method to account for gender-specific differences of articulatory space. These methods aim to reduce articulatory variability within phone categories while preserving differences across categories [10].

These approaches to articulatory normalization, however, are in contrast with auditory-based theories of speech production, which argue that speakers aim for auditory (rather than articulatory) targets because the "same" vowels can be created using different articulatory gestures. A few studies have taken this view into consideration [8, 13] to perform articulatory inversion through analysis-by-synthesis. McGowan and Cushing [8] sought to find the static parameters of an articulatory synthesizer (vocal-tract anatomies and postural settings) such that synthesized vowels matched those of a target speaker. The study showed that an articulatory normalization step was necessary to account for anatomical differences between the target speaker and the standard vocal tract model used by the articulatory synthesizer. In contrast, Hiroya and Honda [13] focused on differences in articulatory dynamics across speakers. For this purpose, the authors developed a production model consisting of HMMs of articulatory parameters for each phoneme, and state-specific linear mappings for articulatory-to-acoustic conversion. To adapt the inversion model to a new speaker, the authors used an analysis-by-synthesis loop to

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

adjust the acoustic-to-articulatory mapping parameters so the output matched acoustic recordings from the new speaker.

3 METHODS

Assume a database of articulatory recordings from a source speaker $S = [s_1\ s_2 \dots s_n]$, where s_i is an articulator vector containing the (x, y) position of P articulatory markers in a given speech frame, and a parallel[1] database from a target speaker T. Assume that each articulator vector is organized as $[x_1, y_1, x_2, y_2, \dots, x_P, y_P]$, where x denotes anteroposterior position (front to back) and y denotes dorsoventral (top to bottom) position. We seek to develop a transformation that adapts articulatory data from the source speaker to the target speaker; the resulting normalized source will be denoted by $\hat{S}$. The remainder of this section presents three methods of articulatory normalization.

3.1 Classical Procrustes

The first method considered in this study is the classical Procrustes (CP) transform. CP is a geometric transform that includes a global translation (c_x, c_y), scaling (b), and rotation (θ). For $P = 2$ articulators, the transform can be expressed as:

$$\begin{bmatrix} \hat{x}_1 \\ \hat{y}_1 \\ \hat{x}_2 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta & & \\ -sin\theta & cos\theta & & \\ & & cos\theta & -sin\theta \\ & & -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \end{bmatrix} b + \begin{bmatrix} c_x \\ c_y \\ c_x \\ c_y \end{bmatrix} \quad (1)$$

where $S = [x_1, y_1, x_2, y_2]$ are the coordinates of the two articulators and $\hat{S} = [\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2]$ are the normalized coordinates. The goal is to find parameters $\{\theta, b, c_x, c_y\}$ that minimize the L2 norm ($\|\cdot\|$) between the target and normalized source articulators:

$$art_{err}(T, \hat{S}) = \frac{1}{n}\sum_{k=1}^{n} \|t_k - \hat{s}_k\| \quad (2)$$

CP has several useful properties: the parameters have a clear geometrical interpretation; it can be applied to any point on the (x, y) plane, which allows subsequent normalization of palatal outlines; and it preserves the relative anatomical shape of the source speaker. CP also has a unique global minimum in the *articulatory* domain, when limited to a two-dimensional translation, scaling, and rotation.

3.2 Extended Procrustes

The CP transform is suitable when the relative position of markers remains constant, which is not generally the case across speakers or even for different recording sessions of the same speaker. To address this issue, we propose an extension of the CP transform that allows local translations, but maintains a global rotation and scale factor in the x/y plane. This extended Procrustes (EP) transform adds $2P - 2$ parameters to the CP transform, which becomes (for $P = 2$ articulators):

$$\begin{bmatrix} \hat{x}_1 \\ \hat{y}_1 \\ \hat{x}_2 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta & & \\ -sin\theta & cos\theta & & \\ & & cos\theta & -sin\theta \\ & & -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \end{bmatrix} b + \begin{bmatrix} c_{1x} \\ c_{1y} \\ c_{2x} \\ c_{2y} \end{bmatrix} \quad (3)$$

The EP solution can achieve a lower articulatory error (equation (2)) than CP, but it also has certain disadvantages. First, the generality of the transform is lost—it is no longer defined for the entire x/y plane but only for each specific marker. Second, an analytical solution does not exist; instead, model parameters must be optimized through an iterative technique. Our implementation uses the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton optimization procedure, a generalization of the secant method for multidimensional problems. Though the EP transform has a single optimal solution, the solver may converge to a local minimum. To discourage this from happening, model parameters are initialized at the CP solution.

3.3 Analysis-by-synthesis

Unlike the previous two methods, which minimize the *articulatory distance* between two speakers, the analysis-by-synthesis (ABS) approach seeks to find the parameters of an EP transform that minimize the expected *acoustic distance*. Our approach uses an articulatory-to-acoustic mapping similar to the data-driven concatenative articulatory synthesis procedure of Kaburagi and Honda [9].

Consider a database containing synchronized recordings of articulatory $U = \{u_1, u_2 \dots u_T\}$ and acoustic frames $V = \{v_1, v_2 \dots v_T\}$, where u_i is an *2P*-dimensional vector containing articulator positions, and v_i is an *N*-dimensional vector containing the corresponding acoustic features (i.e., MFCCs). In data-driven concatenative articulatory synthesis, one seeks to estimate the acoustic features $\hat{v}$ for a novel articulatory vector u by combining units from the database (U, V). Assuming that each articulatory channel has been previously autoscaled to $N(0,1)$, the squared distance from u to each of the units in the database is:

$$e_i = (u - u_i)^T (u - u_i) \quad (4)$$

At this point, and following Kaburagi and Honda [9], we estimate the acoustic vector $\hat{v}$ by finding the *M* lowest squared distances e_j and computing the weighted sum over the corresponding acoustic vectors v_j:

$$\hat{v} = \sum_{j=1}^{M} w_j v_j \quad (5)$$

where $w_j \propto e_j^{-2}$ subject to the constraint $\sum w_j = 1$. To improve the run time of the ABS algorithm, we reduce the database to 10% of its original size using k-means clustering—new units are created from the cluster centers of the combined articulatory and acoustic vectors. The weighted sum is then performed over the entire reduced database to obtain the acoustic estimate $\hat{v}$:

$$\hat{v} = \sum_{\forall i} w_i v_i \quad (6)$$

with w_i subject to the previous constraints. The objective function then becomes a measure of acoustic error:

[1] Further, assume that both datasets have been aligned at the frame level, e.g., via dynamic time warping [14].

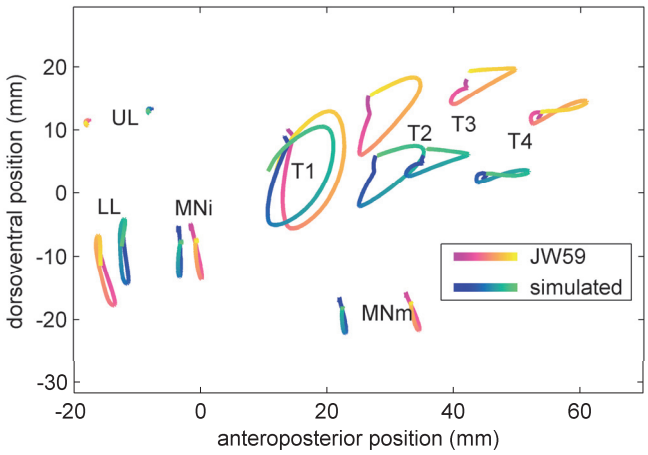

Figure 1. The first experiment uses a known transformation of speaker JW59. Here is his production of /sowd/ and the one created using the parameters of the simulated speaker. Both the EP and ABS methods recovered the simulated parameters. The XRMB database includes markers for the upper lip (UL), lower lip (LL), mandibular incisor (MNi), mandibular molar (MNm), and four positions of the tongue (T1-4).

$$acc_{err}(T,\hat{S}) = \frac{1}{n}\sum_{k=1}^{n}\|AS\{t_k\} - AS\{\hat{s}_k\}\| \quad (7)$$

where $AS\{\cdot\}$ represents the articulatory synthesis function, which takes a $2P$-dimensional articulator vector and returns an N-dimensional acoustic vector. As before, we used the BFGS procedure to optimize the EP parameters.

4 EXPERIMENTS

We compared the three normalization methods using the Wisconsin X-ray Microbeam (XRMB) Speech Production Database [15], which captures articulatory movements of eight sagittal articulatory positions (i.e. upper lip, lower lip, mandibular incisor, mandibular molar, and four tongue locations). Ten native Wisconsin speakers[2] were selected: five male and five female without mistakes during the VCV (TP013) or CVC (TP016) task. The selected tasks were combined and then split into training and test sets according to Table 1 to ensure a similar number of CVC and VCV tasks. For a given source speaker, target speaker, and method of normalization, the normalized source $\hat{S}$ was evaluated in terms of the articulatory and acoustic errors in equations (2) and (7), respectively. The final error values in each case was the average of two scenarios: training model parameters on set A and testing on set B; training on set B and testing on set A (refer to Table 1).

The three normalization methods were validated on two experiments. The first experiment established proof-of-concept by estimating the normalization transform for a simulated speaker (created from JW59) using known rotation (10 degrees), scaling (0.9), and translation (16 displacements randomly drawn from the normal distribution $\mathcal{N}(0,10mm)$; as a comparison, the four tongue pellets span 40 mm). This allowed us to compare the adapted articulators $\hat{S}$ against the (known) ground truth, see Figure 1. The second experiment tested the normalization methods on every source/target combination from the ten speakers (90 cases excluding same-same pairings).

5 RESULTS

Results are presented as relative improvements with respect to the distance between target and source speakers:

$$\text{art}_{\text{imp}}(T,\hat{S}) = \frac{art_{err}(T,S) - art_{err}(T,\hat{S})}{art_{err}(T,S)} \quad (8)$$

$$\text{acc}_{\text{imp}}(T,\hat{S}) = \frac{acc_{err}(T,S) - acc_{err}(T,\hat{S})}{acc_{err}(T,S)} \quad (9)$$

These measures were deemed to be more reliable than art_{err} and acc_{err} since baseline scores vary from speaker to speaker.

Articulatory and acoustic scores from the first experiment are shown in Figure 2. Both EP and ABS recovered the solution within the tolerances of the optimization algorithm. In contrast, CP provides only a 20% improvement in articulatory error, since independent translation parameters are needed in order to account for changes in the relative position of pellets. Interestingly, results with the CP transform show that small improvements in the articulatory similarity (20%) can produce considerable improvements in acoustic similarity (60%).

The second experiment exposes the inherent difficulties of articulatory normalization when working with real data. Results are shown in Figure 3. A multiple comparison test using Tukey's honestly significant difference (HSD) criterion found all pairwise comparisons to be significantly different. EP outperforms CP in both measures, attaining the largest articulatory improvement (59%) and the second most acoustic improvement (22%). The benefit of the added translation parameters in EP with respect to CP is evident by an increase in articulatory improvement from 24% to 59%. ABS obtained a 34% acoustic improvement although its articulatory score was 22% worse than baseline (i.e., initial distance between source and target speaker).

Table 1. The contents of task TP013 and TP016 were merged and split to create two balanced sets containing VCVs and CVCs.

Set A		Set B	
CVC	VCV	CVC	VCV
side	uhka	sewed	uhfa
seed	uhma	sod	uhra
sued	uhta	sawed	uhzha
sid	uhza	sad	uhva
surd	uhcha	said	uhha
sud	uhba	soid	uhsha
sowd	uhla	sood	uhya
sayed	uhga		uhwa
	uhpa		uhna
	uhja		uhda
			uhsa

[2] JW11, JW12, JW14, JW26, JW35, JW39, JW40, JW53, JW54, and JW59

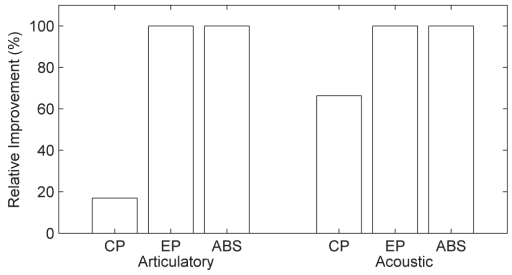

Figure 2. Relative improvement of the normalized simulated speaker over the baseline indicates that the optimization function is able to estimate the exact parameters of a known transform using EP and ABS. CP was not expected to capture the complexity of the transform since it does not possess enough degrees of freedom.

We also analyzed results with respect to the choice of gender for the source and target speaker. For this purpose, we partitioned results for each of the six conditions in Figure 3 into four categories—FF, FM, MF, and MM (e.g. FM: female source, male target) and performed a two-way analysis of variance. Results are summarized in Table 2. In terms of articulatory scores, gender effects were observed for CP and ABS, but not for EP. CP showed a primary effect for target gender: scores for female targets were significantly higher than those for male targets (XF>XM). An interaction was also present, though not in the direction that was predicted: cross-gender scores for FM pairings were significantly higher than same-gender scores MM; this was most likely caused by a larger baseline difference in the FM condition than MM, which allowed for greater gains to be made through normalization. ABS also displayed a significant interaction (FF > MF). In terms of acoustic scores, gender effects were present for CP and EP, but not for ABS. Both source gender and target gender had a significant effect for CP; in both cases, female was the higher scoring gender (FX>MX and XF>XM). EP showed significantly higher scores when the source gender was female (FX>MX).

6 DISCUSSION

We have presented three methods for articulatory normalization that are based on two geometric transforms (classical vs. extended Procrustes) and two alternative objective functions (articulatory vs. acoustic similarity). Our results show that the extended Procrustes transform can account for a wider range of speaker differences than classical Procrustes. When viewed in terms of articulatory similarity, the extended Procrustes transform is not affected by the gender of the source or target speaker. By incorporating an analysis-by-synthesis loop, we have also shown that the extended Procrustes transform can be used to find articulatory normalizations that optimize acoustic (rather than articulatory) similarity. This result is also shown to be unrelated to the gender of the source or target speaker.

In all differences due to gender factors (e.g. CP on articulatory scores), females were found to be in the higher scoring group. One explanation for this result is that females may have been more careful and consistent in their production. Several studies have found females to be more intelligible than males [16]. Lindblom [17] explains this hypothesis from the perspective of energy conservation—since male articulatory gestures consume more energy to compensate for their larger vocal tracts, they are also more lax in their articulatory targets. A more recent study of gender differences in vocal tract dynamics [18] shows that male diphthongs are created using greater articulatory excursions at higher articulatory speeds, but did not find significant differences in articulatory accuracy across genders.

By coupling a data-driven articulatory synthesis in analysis-by-synthesis loop, the method allowed us to optimize any articulatory transform to maximize acoustic similarity. This technique significantly increased acoustic similarity compared to a method that aimed towards maximizing the articulatory similarity alone. However, the improvements in acoustic similarity came with a reduction in articulatory similarity. The results suggested that increases in acoustic similarity do not necessarily require corresponding increases in articulatory similarity. A possible explanation for this finding is the well-known many-to-one relationship [19, 20] between articulators and their acoustic observations.

A potential future direction for this research involves a multi-objective optimization that simultaneously minimizes articulatory and acoustic errors. We also plan to investigate whether solutions trained on VCV and CVC tasks generalize to continuous utterances or whether continuous utterances can be used for training.

7 ACKNOWLEDGEMENTS

This work was supported by NSF award 0713205 and the SMART scholarship program at the Department of Defense.

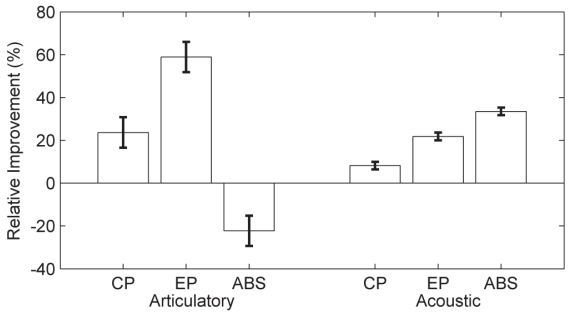

Figure 3. Means and confidence intervals for each condition. The gains are modest compared to the first experiment; one result is worse than the baseline in terms of articulatory error (though it offers the most acoustic improvement).

Table 2. Significance results for the 6 two-way ANOVA tests that examined source and target gender effects within each method. EP was the only method that did not show gender effects in the articulatory score. ABS was the only method that did not show gender effects in the acoustic score. Blank cells were not significant.

		Source gender	Target gender	Interaction
Articulatory scores	CP		p<0.001	p<0.05
	EP			
	ABS			p<0.05
Acoustic scores	CP	p<0.001	p<0.001	
	EP	p<0.01		
	ABS			

8 REFERENCES

[1] Johnson, K., "Speaker Normalization in Speech Perception," in *The Handbook of Speech Perception*, ed: Blackwell Publishing Ltd, 2008, pp. 363-389.

[2] Ryalls, J. H. and Lieberman, P., "Fundamental frequency and vowel perception," *The Journal of the Acoustical Society of America,* vol. 72, pp. 1631-1634, 1982.

[3] Arora, R. and Livescu, K., "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *ICASSP*, Vancouver, BC, Canada, 2013, pp. 7135 - 7139.

[4] Ghosh, P. and Narayanan, S., "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America,* vol. 130, p. EL251, 2011.

[5] King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M., "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America,* vol. 121, pp. 723-742, 2007.

[6] Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., and Goldstein, L., "Robust word recognition using articulatory trajectories and Gestures," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[7] Geng, C. and Mooshammer, C., "How to stretch and shrink vowel systems: Results from a vowel normalization procedure," *The Journal of the Acoustical Society of America,* vol. 125, p. 3278, 2009.

[8] McGowan, R. S. and Cushing, S., "Vocal tract normalization for midsagittal articulatory recovery with analysis-by-synthesis," *The Journal of the Acoustical Society of America,* vol. 106, pp. 1090-1105, 1999.

[9] Kaburagi, T. and Honda, M., "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *ICSLP*, Sydney, Australia, 1998, pp. 433-436.

[10] Hashi, M., Westbury, J. R., and Honda, K., "Vowel posture normalization," *The Journal of the Acoustical Society of America,* vol. 104, pp. 2426-2437, 1998.

[11] Westbury, J. R., Hashi, M., and J. Lindstrom, M., "Differences among speakers in lingual articulation for American English /r/," *Speech Communication,* vol. 26, pp. 203-226, 1998.

[12] Simpson, A. P., "Gender-specific articulatory-acoustic relations in vowel sequences," *Journal of Phonetics,* vol. 30, pp. 417-435, 2002.

[13] Hiroya, S. and Honda, M., "Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model," *IEICE TRANSACTIONS on Information and Systems,* vol. 87, pp. 1071-1078, 2004.

[14] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 26, pp. 43-49, 1978.

[15] Westbury, J. R., "X-Ray Microbeam Speech Production Database User's Handbook Version 1.0," Waisman Center on Mental Retardation & Human Development, University of Wisconsin, Madison, WI1994.

[16] Bradlow, A., Torretta, G., and Pisoni, D., "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication,* vol. 20, pp. 255-272, 1996.

[17] Lindblom, B., "Economy of speech gestures," in *The production of speech*, ed: Springer, 1983, pp. 217-245.

[18] Simpson, A., "Dynamic consequences of differences in male and female vocal tract dimensions," *The Journal of the Acoustical Society of America,* vol. 109, p. 2153, 2001.

[19] Gay, T., Lindblom, B., and Lubker, J., "Production of bite-block vowels: Acoustic equivalence by selective compensation," *The Journal of the Acoustical Society of America,* vol. 69, pp. 802-810, 1981.

[20] Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W., "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *The Journal of the Acoustical Society of America,* vol. 63, pp. 1535-1555, 1978.