



Predicting Unseen Articulations from Multi-speaker Articulatory Models

G. Ananthakrishnan¹, Pierre Badin², Julián Andrés Valdés Vargas², Olov Engwall¹

¹Centre for Speech Technology, KTH (Royal Institute of Technology), Stockholm, Sweden

²GIPSA-Lab (Département Parole & Cognition / ICP),
UMR 5216, CNRS - Grenoble University, France

agopal@kth.se, Pierre.Badin@gipsa-lab.grenoble-inp.fr,

Julian-Andres.Valdes-Vargas@gipsa-lab.inpg.fr, engwall@kth.se

Abstract

In order to study inter-speaker variability, this work aims to assess the generalization capabilities of data-based multi-speaker articulatory models. We use various three-mode factor analysis techniques to model the variations of midsagittal vocal tract contours obtained from MRI images for three French speakers articulating 73 vowels and consonants. Articulations of a given speaker for phonemes not present in the training set are then predicted by inversion of the models from measurements of these phonemes articulated by the other subjects. On the average, the prediction RMSE was 5.25 mm for tongue contours, and 3.3 mm for 2D midsagittal vocal tract distances. Besides, this study has established a methodology to determine the optimal number of factors for such models.

Index Terms: Factor analysis, Multi-speaker Articulatory Model

1. Introduction

An important aspect of vocal tract articulatory modeling, from the point of view of both speech production and its effect on acoustic perception, is the variability between different speakers. While it is obvious that size and general conformation of the articulators are speaker-specific, it is not clear which differences exist between strategies adopted by different speakers to produce the same phoneme. It is thus of high interest to study the interdependence between articulatory control and size/conformation to be able to deal with the inter-speaker normalization problem. Several studies based on measurements using Electromagnetic Articulography (EMA) or Magnetic Resonance Imaging (MRI) have tried to elucidate this problem. Harshman *et al.* [1] performed a three-mode factor analysis on vowel productions of different speakers of American English using the Parallel Factor analysis (PARAFAC) algorithm. They concluded that the tongue postures could be decomposed into two factors which explain the variations in the production by different speakers up to a multiplication constant. They also showed how factor analysis did not merely explain the variations in a statistical sense but also provided interesting interpretative capabilities. Johnson *et al.* [2] provided a more extensive theoretical and conceptual basis for studying speaker variability in the American English vowel system. They used canonical discriminant analysis to study differences between speakers under several conditions such as speaking rate, tense/lax, dialect, gender, palate shape etc. and cast serious questions on the universal articulatory hypothesis. Hoole *et al.* [3, 4] performed PARAFAC on the German vowel system and showed that the variability for different speakers can be modeled by two factors. However it is not clear whether the same number of parameters can be applied to consonants as well. Geng and Mooshamer [5] showed a method for normalizing vowel systems both in the acoustic and articulatory domains so as to minimize the variability between different speakers with respect to a normal

tongue using a generalized Procrustes analysis. They showed that the discriminability between different vowels in the articulatory space improved by up to 90% on normalized articulatory measurements, thus lending some credibility to the universal articulation theory.

The present study attempts to make a similar analysis as that of Hoole [4], but extended to consonants. Studies (such as [6]) have suggested a higher number of factors to model the variations in tongue shapes for consonants. More importantly, this paper tries to establish a paradigm where an unknown (or unmeasured) articulation of a given speaker can be predicted using a model made from measurements of several speakers (including the given speaker) articulating a partial set phonemes and the corresponding phoneme articulated by the other speakers.

2. Techniques for Three-mode Factor Analysis

Three mode-analysis is an extension of two-mode factor analysis methods like Principal Component Analysis (PCA). Here the three modes (or ways) of variation are across articulations (phonemes), articulators (tongue posture) and speakers (subjects). The following are the factor analysis methods we use.

2.1. Two-mode factor analysis

Two-mode factor analysis has been used to analyze the different modes of variation in several studies before [6, 7]. Consider articulatory measurements from speaker s , which consists of a row vector of measurements $\bar{x}_p(n : 1 \leq n \leq N)$ for articulation $p : 1 \leq p \leq P$ such that $X_s = [\bar{x}_1^T \dots \bar{x}_p^T \dots \bar{x}_P^T]^T$. X_s is decomposed into a set of control parameters $\Pi_s^{[P \times F]}$ (predictors or scores which control the variations in articulations) and posture model parameters $C_s^{[N \times F]}$, (factor loadings which explain the correlation between the different articulators) by the following equation

$$X_s = \Pi_s * C_s^T + \gamma_s \quad (1)$$

where F are the number of factors and γ_s is the residual error. Since we want to account for maximum variability in the tongue postures we perform a PCA to effect the decomposition for each individual speaker $s : 1 \leq s \leq S$. Typically, X_s is mean subtracted, which means that $\sum_{p=1}^P \bar{x}_p(n) = 0 \forall n$.

2.2. Inversion

Knowing the posture model parameters C_s , suppose one wants to find the control parameters π_p for a new articulation \bar{x}_p (not included in the initial data), we can do so as follows

$$\pi_p = \bar{x}_p * [C_s^T]^\dagger \quad (2)$$

where $[\cdot]^\dagger$ is the pseudo-inverse of the matrix. We call this process inversion.

2.3. PARAFAC

PARAFAC is a three-way factor analysis approach which is often used to derive parsimonious representations that explain variations in 3-mode data. The decomposition is realized as

$$X_s = \Pi * \Phi_s * C^T + \gamma_s \quad (3)$$

where Φ_s is a diagonal matrix made from the s^{th} row of matrix Φ which are the loading factors for the speaker variability. Π and C are speaker independent and the inter-speaker variability is modeled as a scaling for every factor by Φ .

2.4. Tucker Decomposition

Tucker3 [8] also called three-mode PCA, is an extension of PCA to three modes of variation. Here every data point can be modeled as

$$X_{psn} = \sum_{a=1}^{F_1} \sum_{b=1}^{F_2} \sum_{c=1}^{F_3} \Pi_{pa} \Phi_{sb} C_{nc} G_{abc} \quad (4)$$

where $X^{[P \times N \times S]}$ is the 3-mode data and G is called the core matrix which contains the factor loadings for all three modes of variation, while matrices Π , Φ and C are the control parameters (predictors, scores) for variations along articulation, speaker and articulator points respectively. Here the number of factors for each mode can be fixed independently to F_1 , F_2 and F_3 respectively. However, the solutions for Tucker3 are not unique, unlike those for PARAFAC. In this study, factor F_2 is fixed to be the number of speakers S and factor $F_1 = F_3 = F$, in order to simplify the comparison between different methods for the same number of factors.

2.5. Two-Level PCA

A more conventional approach has also been considered for the current problem of generalization in order to carry out inversion in a three-mode decomposition. At the first level the three dimensional matrix, X , is unfolded to a two dimensional matrix $\hat{X}^{[P \times N \times S]} = [X_1 X_2 \dots X_s \dots X_S]$ and decomposed at the first level using PCA as follows

$$\hat{X} = \Pi * \Psi^T + \gamma_1 \quad (5)$$

where $\Pi^{[P \times F]}$ are the universal control parameters for all the speakers and γ_1 are the residual errors. $\Psi^{[N \times S \times F]} = [\Psi_1 \dots \Psi_s \dots \Psi_S]^T$ contains the concatenated loading factors of the articulators for all the speakers. At the second level, Ψ is unfolded further and decomposed into speaker specific control parameters $\hat{\Psi}^{[F \times S \times F]}$ and universal loading factors $C^{[N \times F]}$ for the different articulators.

$$\hat{\Psi} = \hat{\Phi} * C^T + \gamma_2 \quad (6)$$

where $\hat{\Psi}^{[F \times S \times N]} = [\Psi_1^T \dots \Psi_s^T \dots \Psi_S^T]^T$ and γ_2 is the residual at the second level. $\hat{\Phi}$ contains $\Phi_s^{[F \times F]}$, the speaker-specific adaptation matrix.

It is now possible to perform inversion when data $\overline{x_{s\rho}}$ is not available from speaker s in the following way:

$$\pi_\rho = [\overline{x_{1\rho}} \dots \overline{x_{(s-1)\rho}} \overline{x_{(s+1)\rho}} \dots \overline{x_{S\rho}}] * \widehat{\Psi}_{-s}^\dagger \quad (7)$$

where $\widehat{\Psi}_{-s}$ is the same as $\hat{\Psi}$ without the component Ψ_s for speaker s .

2.6. Inversion by Expectation Maximization

If data for a particular articulation is unavailable for a certain speaker, then it is still possible to construct the three-way model estimated through Expectation Maximization (EM) [9]. Since all the factor analysis methods deal with estimating the covariance matrix, EM algorithm allows this estimation even when part of the data is missing. The model learnt from the rest of the data can then be used to predict the missing values.

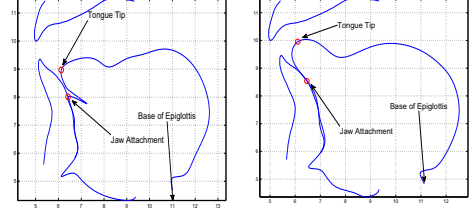


Figure 1: Examples of ‘as visible’ tongue contours for speaker ‘pb’: left /a/ where the sub-lingual cavity is visible; right /fa/ where the sub-lingual cavity is not visible

We used the MatlabTM implementation of PCA, the three-mode analysis implementation of PARAFAC and Tucker3 by Andersson and Bro [10] and used our own implementation of the two-level PCA.

3. Data

The data consisted of midsagittal MRI images of three French speakers who sustained articulations of 73 different phonemes for 16 seconds each. The reconstructed images had a resolution of about 1 mm in the midsagittal plane. There were two male and one female speaker all of them without marked dialectal accents. The corpus consisted of the 10 French oral vowels /i, y, u, e, ø, o, ε, œ, ɔ, a/, the 3 nasal vowels /ā, ē, ē/ articulated without any context and the 10 consonants /f, k, l, m, n, p, b, s, j, t/ articulated in symmetric VCV context of six vowels, namely /a, e, ε, i, o, u/. The contours of tongue, jaw, palate, epiglottis, velum, pharynx and lips were hand-traced. For some tongue contours, the sub-lingual cavity was observable, while for others it was hidden. In order to be consistent in modeling all the tongue postures, the tongue tip and the point where the tongue is attached to the lower jaw was detected as shown in Figure 1. In the articulations for which no sub-lingual cavity was visible (as decided by the expert), the contour from the ‘jaw attachment’ to the tongue-tip was considered as missing data. The contour from the base of the epiglottis to the tongue tip is defined by 150 equidistant points while the contour from the tongue-tip to the jaw attachment is defined by 50 equidistant points giving a total of 200 articulator points to model what we call ‘Tongue Contours’ for every articulation ($N = 200$). We also used an automated method to extract the 2-dimensional (2D) midsagittal vocal tract distance function, which measures the distance from the lower contour, consisting of lower lips, lower jaw, tongue and epiglottis to the upper contour consisting of the upper lip, palate, velum and pharynx wall respectively. This distance is measured perpendicular to the mid-point between the two contours. For each articulation, the closest points on the upper contour to every point in the lower contour is calculated. The corresponding mid-points are fit into an 8th order polynomial curve to give an approximate mid-point curve along the vocal tract. This curve is sampled uniformly into 50 ($N = 50$) points for which normal lines are found at every point. The distance between the intersection of each such normal onto the upper and lower contours are the measured ‘2D vocal tract function’. Figure 2 illustrates the results of this automated method for the three speakers.

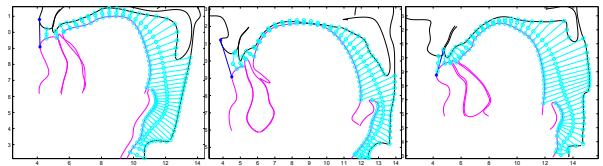


Figure 2: The 2D vocal tract functions for the speakers, from left to right ‘pb’, ‘hl’ and ‘yl’ pronouncing the the phoneme /i/

4. Experiments and Results

In order to have a realistic starting point, we compared 3-mode articulatory modeling using the different methods listed in section 2. Since the goal was not necessarily to explain the factors that contributed to the variations, but to find a suitable multi-speaker model explaining the inter-speaker variation, we tested the methods for different number of factors. Figures 3 and 4 show the Root Mean Square value (RMS error) for the model residual in each of the methods for a varying number of factors. Models for individual speakers are also fitted to tongue contours ‘as visible’ (by skipping the sub-lingual cavity where it is hidden) in order to verify whether applying the EM algorithm on the missing sub-lingual cavities helps in reducing the error. It was found that the error is reduced slightly only for a higher numbers of factors (> 5).

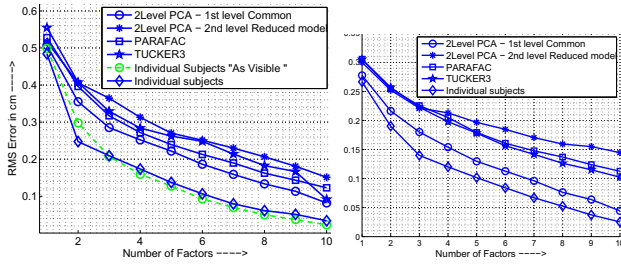


Figure 3: Comparison of the RMS errors (cm) for the different three-way methods (discussed in section 2) over 13 vowel articulations. Left: for the tongue contours. Right: for the 2D vocal tract functions.

4.1. Comparison of the three-way analysis methods

Using PARAFAC analysis on tongue contours measured with MRI, Hoole *et al.* [4] found that only two factors lead to an RMS error of 2.3 mm and explained over 87% of the variance for a set of 9 speakers and 7 vowels. In our first experiment we modeled only the 13 vowels as shown in Figure 3 in order to make a fair comparison. Using a two factor PARAFAC model, the average midsagittal tongue contour reconstruction error over our three speakers was 3.9 mm while the error for the 2D vocal tract function was 2.5 mm accounting for a variance of 71% and 66% respectively. While the error for the 2D vocal tract function is comparable to Hoole’s study, the error for the tongue contour including the sub-lingual cavity is much larger.

When we extended the modeling to all the phonemes, including consonants, shown in Figure 4, we obtained an average RMS error of 2.3 mm for the three speakers with individual 5 factor models for each speaker. This is again higher than what was reported in the literature. Badin *et al.* [6] reported an average RMS error of 1.6 mm for the tongue surface using 5 factors while Engwall [7] obtained an RMS error of 1.3 mm using 6 factors. On the other hand, the 2D vocal tract function (which includes the lips) gave an average error of 1.58 mm for the three speakers, which is in the same range as the other reported errors. Along with consonants, a two factor PARAFAC gives an average error of 5 mm. The three-mode methods, in general, require twice as many factors as individual speaker models to get approximately the same error. The first level of unfolded PCA seems to perform the best compared to all the other methods even though the improvement is not statistically significant ($p > 0.1$). However, the number of parameters that need to be estimated using a 1st level unfolded PCA is much higher than for PARAFAC or TUCKER3.

4.2. Generalizations

The second set of experiments is aimed at generalizing the multi-speaker articulatory models that have been learnt, to pre-

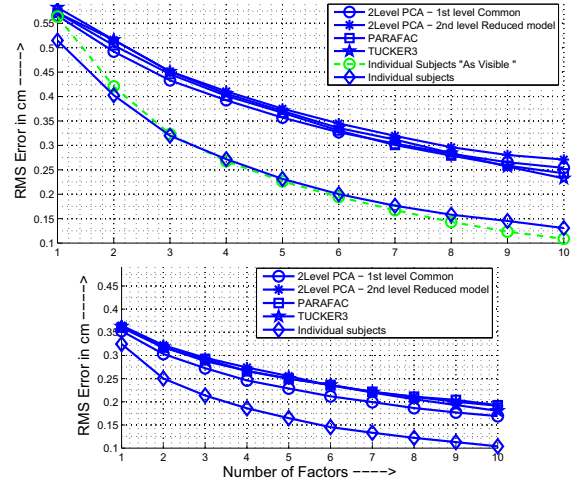


Figure 4: Comparison of the RMS errors (cm) for the different three-way methods (discussed in section 2) over 73 articulations. Left: for the tongue contours. Right: for the 2D vocal tract functions.

dict unknown articulations for one of the speakers. For this experiment, we have excluded one articulation from one speaker and used the rest of the data to try to predict that articulation. We used this leave-one-out approach for every speaker and every articulation thereby trying to establish the over-all generalization capabilities of these methods. For the PARAFAC and TUCKER3 methods, the missing articulation from one speaker is predicted using the EM algorithm as explained in section 2.6. For the unfolded PCA, the missing data was predicted from both the EM algorithm and using inversion (Equation 7). For the inversion methods, the model is built using only the articulations of the common phonemes for all three speakers and then inverted using only the articulations of two of the speakers for the said phoneme. The universal control parameters, π_ρ , and the posture model parameters, Ψ_s , are then used to predict the articulation made the remaining speaker for the said phoneme using Equation 5. In order to make a fair comparison, we have also included the case where the individual models are made from articulations of all phonemes except the said one and then the control parameters are estimated through inversion as in Equation 2.

Another approach towards generalization is finding a linear transformation from the control parameters estimated on individual speaker models. In this method, models are made for individual speakers with whatever articulations are known for the speaker. For the articulations of the common phonemes between all the speakers, a ‘Linear Transform’ matrix is estimated (using linear regression) between the control parameters of the different speakers. The control parameters for the unknown articulation of one of the speakers is then estimated using this transformation matrix. Finally, ‘Direct Linear Transformation’ between tongue postures for different speakers is also estimated and used to predict tongue postures for unknown articulations.

Figure 5 displays the comparative performances of the different methods for this generalization. The optimal performance for all the methods is clearly reached when the number of factors is ‘three’. The linear transformation on control parameters (predictors) seems to be least affected by the number of factors used for modeling. While TUCKER3 modeling gave the best results for generalization of midsagittal tongue contours, the first-level inversion of the unfolded PCA gave the best results for the 2D vocal tract function measurements. However, there are no statistically significant differences between the performances of the different methods.

Figure 6 shows the generalization to unknown articulations

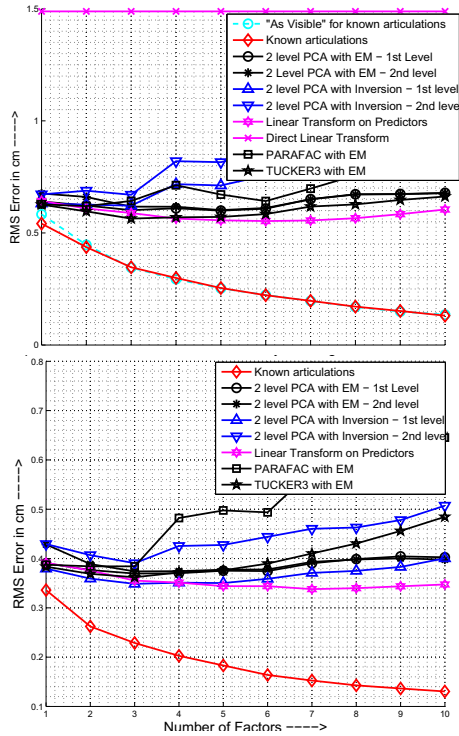


Figure 5: The comparative performance in terms of RMS error (cm) of the generalization properties of the three-way models, with a leave-one-out per speaker experiment. Top: ‘tongue contours’. Bottom: 2D vocal tract functions.

of /i/ and /a/ by speakers ‘yl’ and ‘pb’ respectively. Although the RMS error is much larger than when the articulation data is known, one can see that the shape is preserved and is distinguishable. The figures also show the estimated sub-lingual cavity even if it was hidden in the original contour.

5. Conclusion and Future Work

We performed a 3-mode analysis to study inter-speaker variability in articulations of different French phonemes with an MRI database of 3 speakers. As far as we are aware, this is the first such study which includes both vowels and consonants. The primary focus of the paper was to establish a framework for predicting the articulation of one speaker from the other speakers along with a model of other articulations from that speaker. The experiments that were carried out on this data showed that such a kind of generalization is possible with an RMS error of 5.25 mm for midsagittal tongue contours (including the sub-lingual contour) and 3.3 mm for the 2D vocal tract function.

Since studies in the past have not usually assessed the generalization properties of the articulatory models, the number of factors used in these models was decided arbitrarily, although based on expert knowledge (e.g. Harshman *et al.* [1] considered 2 parameters). The present study shows clearly that 3 parameters are optimal for establishing a model capable of generalization to unseen data. Though this choice may not be suitable to model all kinds of variations for all languages, our approach offers a more reliable and objective methodology for deciding a particular number of factors. Another contribution of this paper has been the means to model unseen data, such as the sub-lingual cavity of the tongue. The method to extract the 2D vocal tract function is completely automated for different speakers without having to specify a grid as has been done in other previous studies.

Future work is to be directed at finding the minimum amount of articulation data that is required for a particular

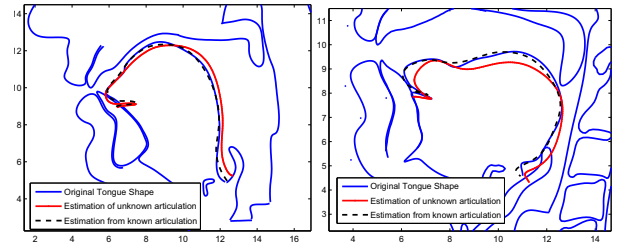


Figure 6: Prediction of the tongue posture using 1st level inversion when the original data is unknown. Left: Speaker ‘yl’ articulating /i/. The RMS error when the data is known and unknown are 1.21 mm and 3.12 mm respectively. Note that the sub-lingual cavity that was hidden in the original contour is predicted by the method. Right: Speaker ‘pb’ articulating /a/. The RMS error when the data is known and unknown are 1.68 mm and 5.23 mm respectively.

speaker in order to be able to generalize the model to all other articulations. This may also be affected by the number of speakers used to make this universal model. The methods discussed in this paper can be employed to clarify the relations between speaker variability due to the shape and size as against variability due to articulatory strategies. Another important aspect that needs to be looked into is whether this generalization reduces the discriminability between the different articulations for the particular speaker. Only such an analysis would be able to validate the generalization method suggested in this paper.

6. Acknowledgements

We sincerely thank L. Lamalle (IRM 3T, CHU, Grenoble, France) for performing the MRI recordings for subjects YL and HL, and S. Masaki, S. Takano, I. Fujimoto, and Y. Shimada (ATR, Kyoto, Japan) for subject ‘pb’. This work has been partially supported by the French ANR-08-EMER-001-02 grant (project ARTIS) and the Swedish Research Council project 80449001, Computer-Animated Language Teachers.

7. References

- [1] Harshman, R., Ladefoged, P., and Goldstein, L., “Factor analysis of tongue shapes,” *J. Acoust. Soc. Am.*, 62(3):693–707, 1977.
- [2] Johnson, K., Ladefoged, P., and Lindau, M., “Individual differences in vowel production,” *J. Acoust. Soc. Am.*, 94(2):701–714, 1993.
- [3] Hoole, P., “On the lingual organization of the German vowel system,” *J. Acoust. Soc. Am.*, 106(2):1020–1032, 1999.
- [4] Hoole, P., Wismüller, A., Leinsinger, G., Kroos, C., Geumann, A., and Inoue, M., “Analysis of tongue configuration in multi-speaker, multi-volume MRI data,” in *Proceedings of the 5th Seminar on Speech production: Models and Data*, 157–160, 2000.
- [5] Geng, C. and Mooshammer, C., “How to stretch and shrink vowel systems: Results from a vowel normalization procedure,” *J. Acoust. Soc. Am.*, 125(5):3278–3288, 2009.
- [6] Badin, P. and Serrurier, A., “Three-dimensional linear modeling of tongue: Articulatory data and models,” 395–40, 2006.
- [7] Engwall, O., “A 3D Tongue Model Based On MRI Data,” in *Proc. Int. Conf. on Spoken Language Processing*, 901–904, 2000.
- [8] Tucker, L., “Some mathematical notes on three-mode factor analysis,” *Psychometrika*, 31(3):279–311, 1966.
- [9] Bilmes, J., “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” *Int. Computer Science Institute*, 4:1–13, 1998.
- [10] Andersson, C. A. and Bro, R., “The N-way Toolbox for MATLAB,” *Chemometrics and Intelligent Laboratory Systems*, 52(1):1–4, 2000.