Analysis of normal and infrequent articulation based on comparison of simulation and observation

Akikazu Nishikido, Jianwu Dang

Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi city, Ishikawa 923-1292, Japan

{a-nishi,jdang}@jaist.ac.jp

Abstract. *To reveal a novel constraint for inverse estimation from speech sound to articulation, we focus on the difference between normal speech and ventriloquism, where the latter is supposed to use some different articulatory manners and/or places that are not used in normal speech. For inverse estimation of normal speech, the ventriloquism, infrequent articulation, should be excluded because it is a local minimum. To do so, we generated all possible articulation using a physiological articulatory model and analyzed them using principal components analysis. Moreover, the generated articulation was compared with articulatory observation. As a result, it was found that vowel /e/ had a normal articulation within covariance ellipse corresponding to three times standard deviation, and infrequent articulation scattered around the ellipse. Furthermore, the difference between the normal articulation and the infrequent one was also investigated for the tongue deformation and the muscle force.*

1. Introduction

The inverse estimation of articulation from speech sound faces an essential problem, the one-to-many problem. To reduce the problem, Atal et al. (1978) introduced a spatial constrain in their data set. Suzuki et al. (1998) used a search technique in a codebook of articulatory-acoustic vector pairs on estimating articulatory movements, which dynamic constraints were involved. In contrast, Dang and Honda (2002) combined the morphological, dynamic and physiological constraint with the inverse estimation from speech sound to articulatory movements by using a physiological articulatory model. As other approach, Hiroya and Honda (2004) used a Hidden Markov Model-based speech production model for estimating articulatory movements from speech sound.

The one-to-many problem is reduced by introducing the above constraints but it is far from being solved. For example, Ogawa et al. (2000) speculated that articulatory manner and place for the same phoneme differed between normal speech and ventriloquism. In both case, all the spatial, dynamic and physiological constraints are satisfied, which have been used in past studies. This situation requires us to reveal some novel constraints for the inverse estimation.

Comparing between normal speech and ventriloquism, it is supposed that many articulatory configurations, which are not used in normal speech, are able to generate the same phoneme within an identical category in acoustic space. The articulatory configurations, which seldom appear in normal speech, are referred to as infrequent articulation, while the one used in normal speech are denoted as normal articulation. If this assumption is confirmed, the existence situation of the infrequent articulation can be used as prior knowledge to exclude the infrequent ones from the inverse estimation for normal speech so that the one-to-many problem can further be reduced. Note that if no special explanation, articulation indicates the normal one.

This study attempts to reveal a novel constraint for inverse estimation from speech sound to articulation by using the prior knowledge on normal and infrequent articulation. For this purpose, we examine all possible articulation and distinguish infrequent articulation from normal ones in a statistical standpoint. To do so, possible articulation is generated using a physiological articulatory model with the forces uniformly distributed in muscle force space (Dang and Honda, 2004). Then, the tongue shapes for vowels are analyzed articulatory-acoustically using the simulation data sets and compared with articlatory observation. In addition, we examine difference between the normal articulation and infrequent one based on the tongue deformation and the muscle force.

2. Analysis of distribution of articulation for vowels

2.1. Generation of articulatory-acoustic data set

The articulatory-acoustic data set was generated using the physiological articulatory model (Dang and Honda, 2004). To obtain the articulatory data, 28 muscle sets with two or three tongue muscles were used to drive the tongue to generate articulatory movements. Forces with eight levels were used to activate each tongue muscle. For the jaw movement, the force with five levels was applied for the jaw opening muscle group and two levels for the jaw closing muscle group. Degree of deformation of the tongue between forces for each muscle was nearly uniform. However, the generated articulatory distribution was not uniform because some configurations can be produced by many muscle combinations. Articulatory data are represented using 17 observation points on the tongue surface in the midsagittal plane.

Based on the articulatory movement of the model, a time-varying vocal tract width was obtained by summing the widths in the midsagittal and parasagittal planes. After combining a lip tube with the articulatory model, a time-varying cross-sectional area function of the vocal tract was obtained using an improved $\alpha - \beta$ model (Dang and Honda, 2002). Based on the cross-sectional area function, the resonant peaks for the vocal tract were calculated using a transmission line model.

Since the model movement was controlled in muscle space, a lot of generated vocal tract shapes and their resonant peaks did not belong to any vowel category. In other words, many resonant peaks were not vowel formant in general meaning. Projecting all the resonant peaks, the ones fallen into a vowel category would be treated as vowels. The resonant peaks denote formants. The first and second formants are used in the analysis of this study.

2.2. Analysis of distribution of articulation

The aim of analysis is to clarify whether the infrequent articulation exists or not for the vowel categories of five Japanese vowels. The criterion of the vowel category is defined as a region around the typical frequency values of the first formant (F1) and the second formant (F2) within 10% (Nakagawa et al., 1982). As a result, the data used this analysis acoustically belongs to a certain vowel category. The number of data within each category was 7190 for /a/, 590 for /i/, 5163 for /u/, 5345 for /e/ and 1777 for /o/, respectively.

In the articulatory space, we investigated the distribution of the articulation of the data that were within the vowel category. To make the simulation data compatible with the observation obtained by the electromagnetic midsagittal articulographic (EMMA) data (Okadome and Honda, 2001), four observation points were selected on the tongue model surface. In the model coordinate, these points are the node 11, 9, 7 and 5 from the tongue root. The points for each category were subtracted average and were analyzed using principal component analysis (PCA). For all vowels, the first two principal components (PCs) explain more than 80% of the variance. The PCs for vowel /e/ are shown in Fig. 1. The large ellipse represents covariance ellipse corresponding to three times standard deviation. A large area with dense data appears in the center of distribution and several local areas with sparse data around the large area distribute. In the local areas, the local area out of the ellipse corresponds to a cluster for clustering the PCs. To show the density of the distribution clearly, a histogram is calculated based on the PCs. The occurrence is normalized for setting the maximum occurrence to be 1. Fig. 2 shows a surface plot of the histogram. The primary area has a large peak and indicates the high density, while the sub areas have small peaks. Vowel /i/ and /o/ showed a similar tendency. This implies that the distribution of articulation within a vowel category consists of primary area with high density and sub areas with low density. The former may correspond to the normal speech and the latter may be the infrequent articulation. On the other hand, for vowel /a/ and /u/, existence of the sub area with low density was not clear.

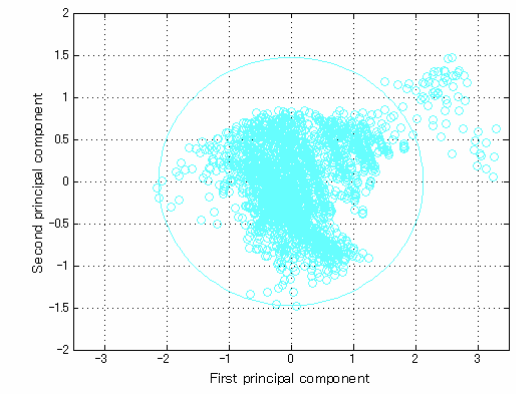

Fig. 1. Distribution of the first two PCs for the vowel /e/. The large ellipse represents the covariance ellipse.

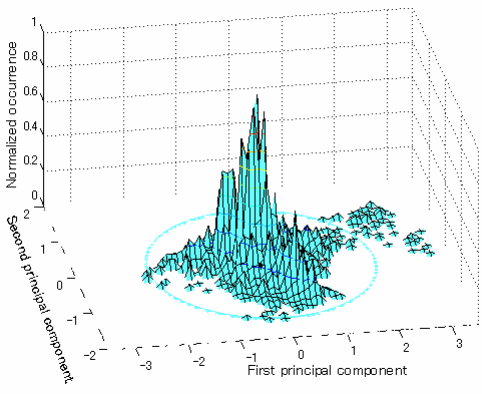

Fig. 2. Histogram of the first two PCs for the vowel /e/. The large ellipse represents the covariance ellipse

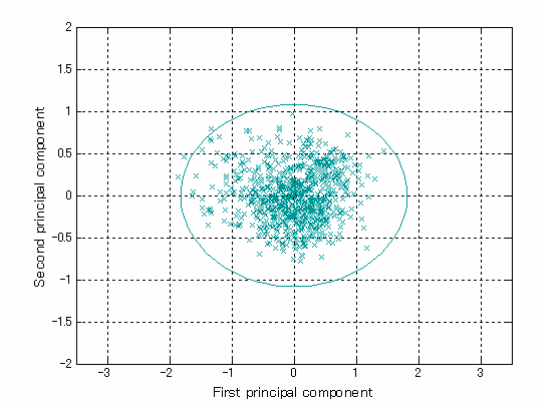

Fig. 3. Distribution of the first two PCs of the vowel /e/ for the observation. The large ellipse represents the covariance ellipse.

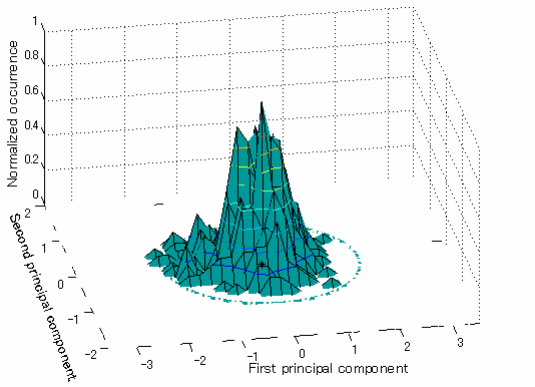

Fig. 4. Histogram of the first two PCs of the vowel /e/ for the observation. The large ellipse represents the covariance ellipse.

2.3. Comparison between simulation and observation

Using PCA, we also analyzed the observation data, four tongue points (T1, T2, T3 and T4), obtained by EMMA system in NTT (Okadome and Honda, 2001). Before PCA, the data from subject TM were transformed to match the shape of the tongue and the palate of the model of the simulation. PCs of the transformed data were calculated for five Japanese vowels. The number of data is 1624 for /a/, 1111 for /i/, 799 for /u/, 762 for /e/ and 1220 for /o/, respectively. The first two PCs explain more than 80% of the variance. The PCs of the vowel /e/ for the observation are shown in Fig. 3. The large ellipse represents covariance ellipse corresponding to three times standard deviation. Comparing with the distribution of the simulation, there are not distinguishing sub areas. For the observation, a histogram is calculated using the same way as that for the simulation, and shown in Fig. 4. There is only one primary area with a large peak. For other vowels, there is a primary area with a large peak or several large peaks and there is not sub area.

To investigate the relationship between the simulated data and the observed data, the simulated data were projected to the first two PCs space of the observed data. For vowel /e/, the first two PCs of the observed data and the projection of the simulated data are shown in Fig. 5. The small circles represent the simulated data and the crosses represent the observed data. The ellipse with solid line represents covariance ellipse corresponding to three times standard deviation for the simulated data, while the ellipse with dash-dotted line represents the covariance ellipse for the observed data. For the simulated data, some sub areas located outside both the covariance ellipses for the observation and simulation. Histogram of the PCs of the observed data and the projected data were calculated separately. Both histograms are normalized by each maximum occurrence, and shown in Fig. 6. For both histograms, location of the peak almost overlaps. For other subject, comparison between the simulated data and the observed data showed the same tendency. Accordingly, the primary area with high density can be considered to correspond to the articulation used in normal speech.

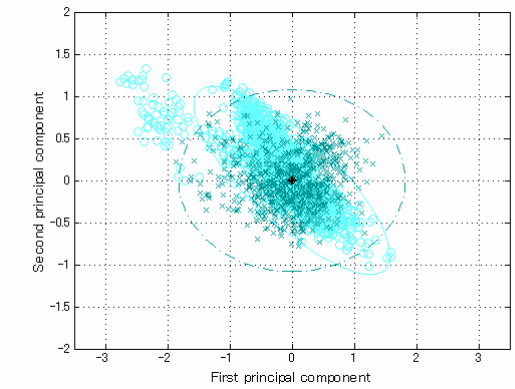

Fig. 5. Distribution of the first two PCs of the vowel /e/ for observed data and the projection of the simulated data. The small circles represent the simulated data and the crosses represent the observed data. The large ellipse with solid line represents covariance ellipse for the simulated data and the large ellipse with dash-dotted line shows the covariance ellipse for the observed data

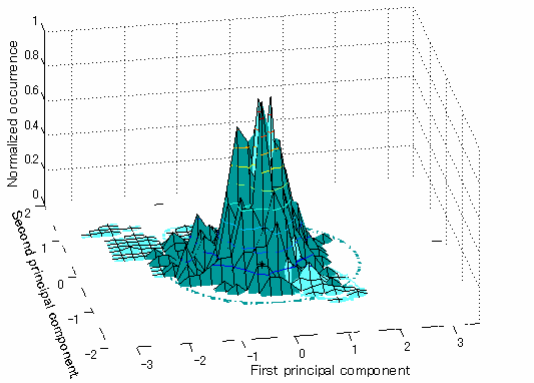

Fig. 6. Histogram of the first two PCs of the vowel /e/ for observed data and the projection of the simulated data. The large ellipse with solid line represents covariance ellipse for the simulated data and the large ellipse with dash-dotted line shows the covariance ellipse for the observed data.

Moreover, this suggests a possibility of the existence of the sub areas with low density that may be corresponding to infrequent articulation. For the vowel /i/ and /o/, the sub areas of the projected data do not always locate outside of the ellipse for the observation.

3. Examination of normal and infrequent articulation

To reveal a novel constraint for inverse estimation, we examine difference between the primary area with high density and the sub areas with low density. As shown in Fig. 1, the sub areas are located far from the central part. It implies that infrequent articulation performs more complicated the tongue shape than normal articulation. It is can be reasonably considered that infrequent articulation is generated using larger muscle force or more complicated activation patterns than that for normal articulation, because large deformation of the tongue shape needs more efforts in general. For this reason, we investigated the degree of deformation of the tongue shape and muscle force using the first two PCs for the simulated data, where this paper just treats with vowel /e/. Deformation of the tongue shape is represented by the mean distance of the simulated shape to the initial shape over 17 observation points on the tongue surface in the midsagittal plane. Muscle force is represented by the sum of force used to activate each muscle for each articulation. The tongue deformation and muscle force are shown in Fig. 7 and Fig. 8, respectively, by the first two PCs. Dark red represents larger value and dark blue represents smaller value. The large ellipse shows covariance ellipse corresponding to three times standard deviation.

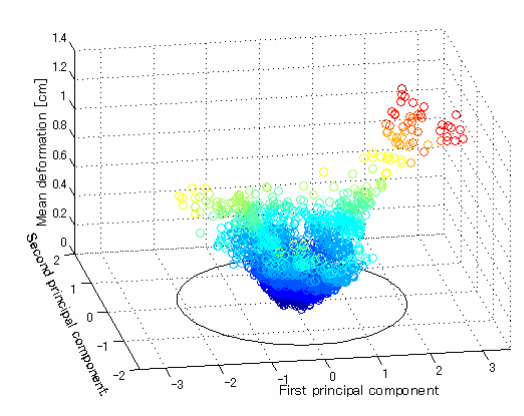

Fig. 7. The tongue deformation [cm] vs the first two PCs of the simulated data for the vowel /e/. Dark red represents larger deformation and dark blue represents smaller deformation. The large ellipse represents the covariance ellipse.

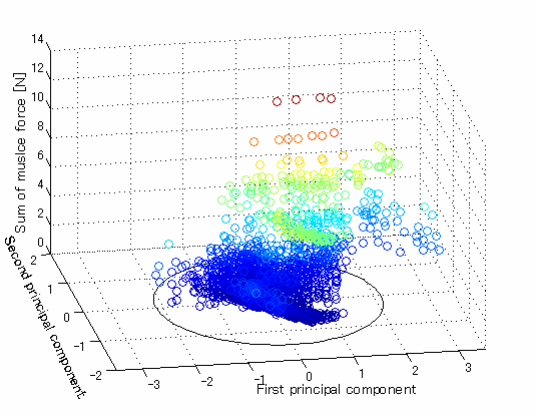

Fig. 8. The muscle force [N] vs the first two PCs of simulated data for the vowel /e/. Dark red represents larger force and dark blue represents smaller force. The large ellipse represents the covariance ellipse.

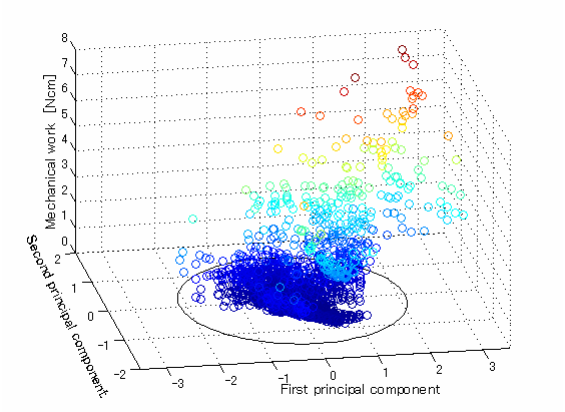

Fig. 9. Mechanical work [Ncm] vs the first two PCs of the simulated data for the vowel /e/. Dark red represents larger work and dark blue represents smaller work. The large ellipse represents the covariance ellipse.

As seen in Fig. 7, within the covariance ellipse, the articulation in the central part has smaller deformation and the ones removed from the central part have the larger deformation. Then, the largest deformation locates in the sub area out of the ellipse and data with larger deformation gather in the sub area. This indicates infrequent articulation perform larger deformation than deformation of normal articulation. In contrast, for the muscle force in Fig. 8, the largest force locates not in the sub area but in the ellipse. However, data in the sub area are larger force than mean force. Thus, difference between the primary area and the sub area is considered to be represented clearly by the product of the tongue deformation and muscle force. The product corresponds to mechanical work. The mechanical work for the first two PCs of the vowel /e/ is shown in Fig. 9. As seen in the left of Fig. 9, the difference between inside and outside the ellipse looks more clearly. It implies that the energy such as mechanical work used in the normal articulation is smaller than that used in the

infrequent articulation. Thus, it is possible that the energy is useful as the criterion to distinguish between normal articulation and infrequent one.

4. Conclusions

To reveal a novel constraint for inverse estimation from speech sound to articulation, we examined all possible articulation and sorted infrequent articulation according to the distribution of the articulation. The tongue shapes and its resonance characteristics were analyzed using the simulated data sets by PCA and then compared with articulatory observation. As a result, for the vowel /e/, the articulation has a primary area with high density within covariance ellipse corresponding to three times standard deviation, relative to normal articulation, and sub areas with low density out of the ellipse concerned with the infrequent articulation. In addition, the difference between normal and infrequent articulation was investigated for paying attention to the tongue deformation and muscle force. Analysis using mechanical work derived from the deformation and muscle force indicated that energy for generating a normal articulation is smaller than that for infrequent articulation. In the feature study, we are going to use a statistical model reflected the difference between the normal articulation and the infrequent articulation and apply it in the inverse estimation.

Acknowledgements

The authors especially thank NTT communication science laboratories for permitting us to share the articulatory data. This research is conducted as a program for the "21st Century COE Program" by Ministry of Education, Culture, Sports, Science and Technology and is supported in part by Grant-in-Aid for Scientific Research of Japan (No. 17300182). This research is also supported in part by the MSRA/IJARC project.

References

Atal, S., Chang, J., Mathews, J. and Tukey, W. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America,* 63 (5), pages 1535-1555, 1978.

Dang, J. and Honda, K. Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. *Journal of Phonetics,* 30 (3), pages 511-532, 2002.

Dang, J. and Honda, K. Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America,* 115 (2), pages 853-870, 2004.

Hiroya, S. and Honda, M. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Speech and Audio Processing,* 12 (2), pages 175-185, 2004.

Nakagawa, T., Saito, S. and Yoshino, T. Tonal difference limens for second formant frequencies of synthesized Japanese vowels. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo,* 16, pages 81-88, 1982.

Ogawa, Y., Uemi, N. and Ifukube, T. Speech production process in ventriloquism of the phonemes where the place of articulation is lips. IEICE Technical Report, H2000-37, pages 1-8, 2000. (in Japanese)

Okadome, T. and Honda, M. Generation of articulatory movements by using a kinematic triphone model. *The Journal of the Acoustical Society of America,* 110 (1), pages 453-463, 2001.

Suzuki, S., Okadome, T. and Honda, M. Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints. In *Proceedings of the5th International Conference on Spoken Language Processing*, pages 2251-2254, 1998.