Reconstruction of mistracked articulatory trajectories

Qiang Fang[1], *Jianguo Wei*[2*], *Fang Hu*[1]

[1] Phonetics Lab., Institute of Linguistics, CASS, China
[2] School of Computer Science, Tianjin University, China
{fangqiang, hufang}@cass.org.edu, jianguo@tju.edu.cn

Abstract

Kinematic articulatory data are important for researches of speech production, articulatory speech synthesis, robust speech recognition, and speech inversion. Electromagnetic Articulograph (EMA) is a widely used instrument for collecting kinematic articulatory data. However, in EMA experiment, one or more coils attached to articulators are possible to be mistracked due to various reasons. To make full use of the EMA data, we attempt to reconstruct the location of mistracked coils with Gaussian Mixture Model (GMM) regression method. In this paper, we explore how additional information (spectrum, articulatory velocity, etc.) affects the performance of the proposed method. The result indicates that acoustic feature (MFCC) is the most effective additional features that improve the reconstruction performance.

Index Terms: EMA, GMM, MMSE

10.21437/Interspeech.2014-194

1. Introduction

Kinematic articulatory data plays more and more important roles in the field of exploring the mechanism of speech production[1], analyzing the behavior of speech therapy, improving the performance of speech recognition[2] and synthesis[3], and estimating vocal tract configuration from speech signals[4]. X-ray microbeam and EMA are the most popular equipments applied to build kinematic arrticulatory corpus for the above purposes.

Collecting kinematic articulatory data is much more difficult than recording acoustic data. When recording the kinematic articulatory data with EMA or X-ray microbeam, coils are glued to the articulators of concern. It makes subjects very uncomfortable, and some of the coil may fall over the articulators in the recording process. Because of these, for the moment, only a British English database (MOCHA)[5] and Wisconsin X-ray microbeam database[6] is publicly available. In EMA experiments, coils are possible to be mistracked due to various reasons[4]. Since it is not easy to get kinematic articulatory data, it is better to make full use of the collected data. Thus, we attempt to reconstruct the mistracked portion of the articulatory data.

It is well known that the articulators (such as tongue and jaw, lower lip and jaw) are either physiologically connected, or (such as tongue and lips) functionally associated to fulfill speech tasks. Therefore, it is possible to exploit the correlation between different articulators to reconstruct the position of one articulator based on those of the other articulators. Several works have been conducted towards this direction based on an X-Ray microbeam corpus. For example, Roweis[7] proposed a method which learned a low-dimensional manifold to represent the data and intersected the manifold with the constraints provided by the measured values. Qin [8] applied the GMM-based MMSE method to estimated missing data sequence of articulation recorded by using X-ray microbeam. Both of these two methods obtained good results. In previous work we implement GMM-regression method to reconstruct the mistracked articulatory trajectories based on the position of the correctly tracked coils.

However, it is not difficult to notice that we have simultaneous information (such as the velocity, acceleration of coils and the corresponding acoustic features) at hand. Previous work only made use of parts of the information. Some studies show that coils' position could be estimate from acoustic signal with RMS of about 2.5mm [4, 9, 10]. It is possible that some of the information, which is highly correlated with the articulatory configuration, will further constraint the result reconstructed from the positions of other coils.

Hence, in the study, we attempt to investigate whether the performance of estimating the position of mistracked coils can be improved by introducing synchronous information.

2. Material

2.1. Experiment setup

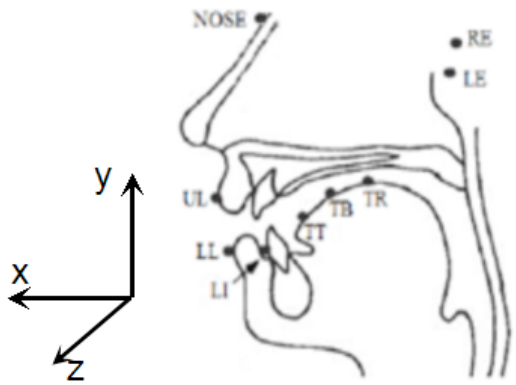

Fig 1. Positions of coils in EMA experiment.

Currently, we are constructing a Chinese kinematic articulatory database for articulatory-based speech synthesis, speech-to-articulatory inversion, and other applications. 400 phonetically balanced Chinese sentences are selected to serve as the recording scripts. In EMA experiment, coils are attached to Tongue Rear (TR), Tongue Blade (TB), Tongue Tip (TT), Lower Incisor (LI), Lower Lip (LL) and Upper Lip (UL), respectively. Another 3 coils (attached to the process behind Right and Left Ears – RE and LE, and NOSE) serve as the references (shown in **Fig. 1**). Two subjects (1 male and 1 female) are recruited in the EMA experiments. The acoustic signal and articulatory data are recorded simultaneous. The sampling frequencies are 16,000Hz for acoustic signal and 200Hz for the articulatory signal.

2.2. Preprocessing

2.2.1. Data smoothing

The original articulatory trajectories recorded by EMA are usually contaminated by noise. To alleviate the influence of noise, here, Savitzky-Golay filter is implemented to smooth the articulatory trajectories. Among all the articulators, the tongue tip moves fastest to form the approximation and closure in the front part of the oral cavity with hard palate. The coil which records tongue tip movement should contain more meaningful high frequency information. Therefore, the data smoothing should not make too much distortion in the high frequency range. Fig.2 gives examples of the raw and smoothed trajectory of the coil attached to tongue tip, while Fig.3 shows the corresponding result of Fourier analysis. The result indicates that the smoothed trajectory less the undesired fluctuation of the position of coil while introduce few distortion in frequency domain.

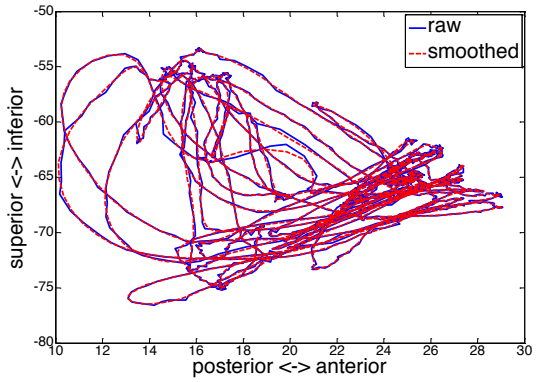

Fig. 2 Examples of the raw and smoothed (with span of 9 and 3rd order polynomial model for Savitzky-Golay filter) trajectory of tongue tip.

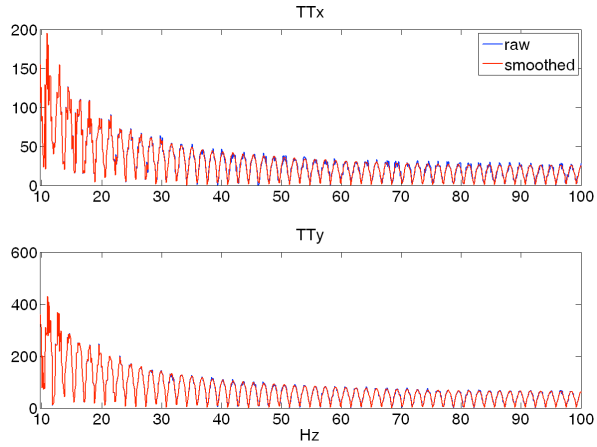

Fig. 3 Fourier analysis of TTx and TTy of raw and smoothed trajectory of an utterance.

2.2.2. Anomaly detection

Three types of mistracking are discovered in the collected EMA data (as shown in **Fig.**4 and **Fig**.5): i.) abrupt jump of coil position at the beginning and in the middle of utterances; ii.) continuous drifting of coil position at the end of utterances; iii.) coil position beyond the region of vocal tract.

To extract the correctly tracked EMA data, we estimate the mean and covariance matrix for each coil based on the whole data set. Then, outliers are detected by using 4 times standard deviation (std.). The samples within 4 times std. are classified as correctly tracked coils, while the others are classified as mistracked coils. Finally, coil mistracking is detected in 68 utterances, which are 16% of the total utterance.

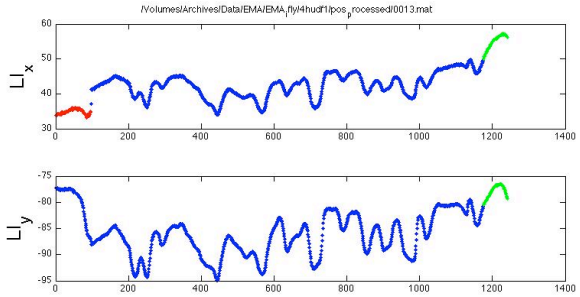

Fig. 4 The 1st (the red curve at the beginning of a utterance in the upper panel) and 2nd (the green curves at the end of an utterance in both lower and upper panels) type of coil mistracking in EMA data.

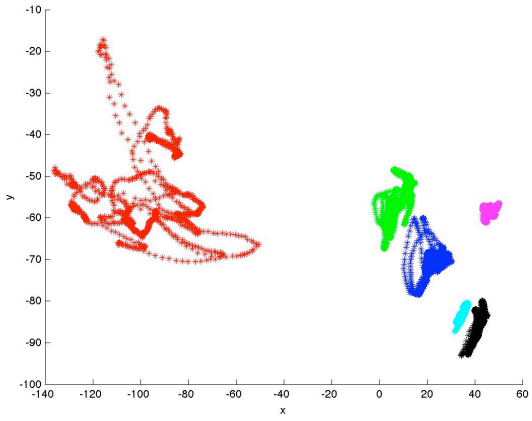

Fig. 5 The 3rd type of coil mistracking. The clouds with different colors stand for TR, TB, TT, LI, LL, and UL, respectively. The coil for TR (denoted by red spots) is beyond the vocal tract.

3. Methods

In this part, we will introduce the method that reconstructs the position of mistracked coil. Let $\boldsymbol{y}_t$ be the target vector that stands for the position of the concerned coil at instant t, and $\boldsymbol{x}_t$ be the source vector that stands for the position of the other coils at instant t. In our case, the collected EMA data could be divided into 3 sets: A={$\boldsymbol{x}_t$, $\boldsymbol{y}_t$ |both $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are correct}; B={$\boldsymbol{x}_t$, $\boldsymbol{y}_t$|$\boldsymbol{y}_t$ is problematic, while $\boldsymbol{x}_t$ is correct}; C={$\boldsymbol{x}_t$, $\boldsymbol{y}_t$|both $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are problematic}. Thus, a mapping function, $\boldsymbol{y}=f(\boldsymbol{x})$, could be trained and evaluated on set A, and the target vector $\hat{\boldsymbol{y}}_t$ of mistracked coil in set B could be reconstructed by using the trained mapping function. For the samples in data C, it is possible to estimate the position of the mistracked coils by utilizing the synchronous acoustic feature. This will not be discussed in the current work. In this study, GMM is applied to approximate joint probability density function $p(\boldsymbol{x}, \boldsymbol{y})$. Then the conditional probability density function $p(\boldsymbol{y}|\boldsymbol{x})$ calculated from $p(\boldsymbol{x}, \boldsymbol{y})$. Finaly, the mapping function are calculated by applying MMSE criterion based on $p(\boldsymbol{y}|\boldsymbol{x})$.

3.1. Joint probability density function

Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are the source and target vectors, respectively. The joint probability density function $p(\boldsymbol{x}, \boldsymbol{y})$ could be approximated by GMM (shown in Eq.1~3).

$$p(\boldsymbol{x},\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x},\boldsymbol{y}) \quad (1)$$

$$= \left[()^T, ()^T \right]^T \quad (2)$$

$$= \begin{bmatrix} & \\ & \end{bmatrix} \quad (3)$$

where π_k is the weighting coefficient of the *k-th* mixture, and are the mean of source and target vectors of the *k-th* mixture, respectively. and are the covariance matrices of the *k-th* mixture for source and target vectors, respectively. and are the cross-covariance matrices of the *k-th* mixture between source and target vectors, respectively.

3.2. Conditional probability density function

Then, the probability density function of $\boldsymbol{y}$ given $\boldsymbol{x}$ could be expressed by Eq.4~7

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{K} w_k N(\boldsymbol{y}|\boldsymbol{x}) \quad (4)$$

$$= + \sum_k^{yx} \left(\sum_k^{xx} \right)^{-1} \left(\boldsymbol{x} - \right) \quad (5)$$

$$= \left) \quad (6)$$

$$w_k = \frac{\pi_k N(\boldsymbol{x})}{\sum_{k=1}^{K} \pi_k N(\boldsymbol{x})} \quad (7)$$

3.3. Minimum Mean Square Criterion

In conventional applications, people usually use MMSE criterion to estimate target vector.

$$\boldsymbol{y}^* = \arg\min_{\hat{\boldsymbol{y}}} E\left[(\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}}) \right] \quad (8)$$

By taking derivative on $\hat{\boldsymbol{y}}$, the target vector could be estimated by using Eq.9.

$$\boldsymbol{y}^* = \int \boldsymbol{y} p(\boldsymbol{y} \mid \boldsymbol{x}) d\boldsymbol{x} = \sum_{k=1}^{K} w \quad (9)$$

It means that the estimated target vector is a weighted sum of the mixtures' mean vectors, and the weighting coefficient of a specific mean vector is the corresponding weighting coefficient in $p(\boldsymbol{y}|\boldsymbol{x})$.

4. Experiments

300 sentences in set B serve as the training set to train the GMM and derive the mapping function, and the other 32 sentenc*es* in set B serve as the testing set. To evaluate the performance of above methods, we black out the trajectory of one coil over the entire utterance, and estimate their positions given the positions of the remaining coils. Then, the estimated positions are compared with the corresponding ground truth. With reference to our previous work[11], GMM with 256 mixtures are trained for the joint probability density function, and the corresponding conditional probability density functions are calculated for each coil based on the joint probability density function.

4.1. Input feature

4.1.1. Dynamic articulatory feature

The articulatory data sequence itself contains not only static position information but also dynamic information, e.g. velocity, acceleration. Therefore, the position vector augmented with velocity information will provide more information of articulatory movements, and may helps to further improve the performance. Let y_t denotes the positions of coils at time instant *t*. Then, the velocity can be calculated according to Eq.10.

$$\begin{aligned} \Delta y_t &= (y_{t+1} - y_{t-1})/2 \quad \text{if} \quad 1 < t < N \\ \Delta y_t &= (y_{t+1} - y_m)/2 \quad \text{if} \quad t = 1 \\ \Delta y_t &= (y_m - y_{t-1})/2 \quad \text{if} \quad t = N \end{aligned} \quad (10)$$

where **N** is the length of the utterance, and y_m is the mean positions of the coils of the utterance.

4.1.2. Acoustic feature

In addition to the articulatory information, acoustic signals are recorded simultaneously in the EMA experiment. A number of studies in the field of speech inversion have proven that the trajectory of articulators could be estimated from acoustic signal with high accuracy[4, 9, 10]. Hence, the performance of reconstruction is possible to be improved by introducing acoustic features to the feature vectors. In the current work, Mel-Frequency Cepstral Coefficients are extracted from acoustic signal (hamming window, frame length = 25ms, frame shift = 5ms), and incorporated into the input feature to train the GMM-regression model.

4.2. Results

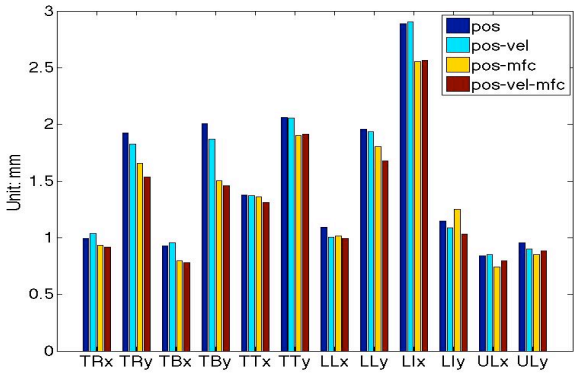

Fig. 6 Comparison of the performance based on different features.

Eq.9 are implemented for each coil based on the input features (position: 'pos'; position and velocity: 'pos-vel'; position and corresponding mfcc: 'pos-mfc'; position, velocity, and mfcc: 'pos-vel-mfc). The results are shown in Fig.6.

It shows the performance of GMM-regression by using different features. The RMSE of the coil for UL is lowest, while the RMSE of the coil for LI is highest. The RMSEs of other coils are in between. For the coils on tongue (TR, TB, and TT), the position of TB is easiest to estimate, while the position of TT is the most difficult to estimate. This is comparable with the result reported by Qin on X-ray microbeam data set[8].

If the velocity of coils are introduced as an additional input feature, the performance could be improve a little in most case. If the synchronous acoustic feature (MFCC) is introduced as an additional input feature, the performance could be improved about 0.5mm at most for TBy, and least for TTx. T-test indicates that the improvements are statistically significant ($p<0.05$) in for most of the coils. If both velocity of coils and acoustic features are introduced as additional input features, the performance can be further improved in most cases. And this improvement is also statistically significant ($p<0.05$). From the above analysis, it is obvious that introducing more input feature, such as velocity and MFCC will improve the performance in most cases.

5. Conclusions

In this study, we attempt to reconstruct the position of mistracked coil by using GMM-regression. To this end, we utilize different input features to estimate the position of mistracked coil. The RMSE of the coil for UL is lowest, while the RMSE of the coil for LI is highest. The RMSEs of other coils are in between. For the coils on tongue (TR, TB, and TT), the position of TR is easiest to estimate, while the position of TT is the most difficult to estimate. When introducing additional input feature, the result demonstrates that the information carried by the acoustic feature is more critical than the velocity of the correctly tracked coils. If only introducing an addition feature into the input feature set, it obvious that the performance is better by introducing acoustic feature (MFCC) than by introducing velocity information of the coils' velocity.

6. Acknowledgements

This study is partly supported by Key project of NSFC (No. 61233009), NSFC Project (No. 60975081, No. 61175016), and Innovation Project of Chinese Academy of Social Sciences.

7. References

[1] Hoole, P., *On the lingual organization of the German vowel system.* J. Acoust. Soc. Am., 1999. **106**: p. 1020-1032.

[2] Markov, K., J. Dang, and S. Nakamura, *Integration of Articulatory and Spectrum Features based on the Hybrid HMM/BN Modeling Framework,.* Speech Communication, 2006. **48**(2): p. 161-175.

[3] Ling, Z., et al., *Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis.* IEEE Transactions on Audio, Speech and Language Processing, 2009. **17**(6): p. 1171-1185

[4] Richmond, K., *Estimating articulatory parameters from the acoustic speech signal*, 2001, University of Edinburgh.

[5] Wrench, A.A. *A multi-channel/multi-speaker articulatory database for continuous speech recognition research*. in *Phonus 5*. 2000. Saarbrucken: Institute of Phonetics.

[6] Westbury, J.R., *X-Ray Microbeam Speech Production Database User's Handbook*. 1994, UW-Madison.

[7] Roweis, S., *Data driven production models for speech processing*, 1999, California Institute of Technology.

[8] Qin, C. and M.Á. Carreira-Perpiñán. *Estimating missing data sequences in X-ray microbeam recordings*. in *InterSpeech2010*.

[9] Hiroya, S. and M. Honda, *Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model.* IEEE Transactions on Speech and Audio Processing, 2004. **12**(2): p. 175-185.

[10] Ling, Z.H., K. Richmond, and J. Yamagishi, *An Analysis of HMM-based prediction of articulatory movements.* Speech Communication, 2010. **52**(10): p. 834-846.

[11] Fang, Q., et al. *Estimating the position of mistracked coil of EMA Data using GMM - based methods*. in *APSIPA2013*. 2013.