Estimating the position of mistracked coil of EMA Data using GMM-based methods

Qiang Fang[1], Jianguo Wei[2], Fang Hu[1], Aijun Li[1], Haibo Wang[3]*

[1] Phonetics Lab., Institute of Linguistics, CASS, China
[2] School of Computer Science, Tianjin University, China
[3] Phonetics Lab., Institute of Ethnology and Anthropology, CASS, China

{fangqiang, hufang, liaj, hbwang}@cass.org.cn, 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 the methods based on Gaussian Mixture Model (GMM). These methods approximate the probability density function of the positions for the concerned coil given the positions of the other coils, then elaborating regression functions by using Minimum Mean Square Error (MMSE) and Maximum Likelihood (ML) methods. The results indicate that: i.) The positions of mistracked coils could be reconstructed from the positions of correctly tracked coils with the RMSE between 1mm and 1.5mm; ii.) The performance can be further improved by incorporating the velocity information in most cases.

Index Terms: EMA, mistracking, GMM, MMSE, ML

1. Introduction

The articulatory data is important for exploring the mechanism of speech production[1], analyzing the behavior of speech therapy, improving the performance of speech recognition[2] and synthesis system[3], and estimating vocal tract configuration from speech signals[4]. Various techniques such as X-ray movie, X-ray microbeam, EMA, ultrasound, and magnetic resonance imaging have been widely applied for these purposes. Compared with other techniques, EMA has high temporal resolution, and does no harm to subjects. This makes EMA the most popular technique for collecting large-scale articulatory database.

However, it always takes great efforts to collecting EMA data. During EMA experiment, coils are glued to the concerned articulators. 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) is available publicly. Recently, we plan to construct a phonetically balanced Chinese EMA database for articulatory-based speech synthesis, and speech inversion. 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, the question comes to whether we can reconstruct the positions of the mistracked coil from those of the others. If the answer is yes, then we can apply machine-learning techniques to estimate the positions of mistracked coils quite accurately. As we know, some of the articulators (such as tongue and jaw, lower lip and jaw) are physiologically connected, and some of the articulators (such as tongue and lips) are functionally associated to fulfill speech tasks. Therefore, it is possible to exploit the correlations between different articulators to estimate the positions of one articulator based on those of the other articulators. Several work have been conducted towards this direction based on an X-Ray microbeam corpus. For example, Roweis[5] 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[6] 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 the present study, we extend Qin's [6] method in the following two aspects: i.) making use of additional information of the Gaussian mixtures; ii.) introducing more articulatory information into the input feature.

The remainder of paper is organized as follows. In section 2, we will give a brief introduction of our EMA corpus and analyze three types of mistracking. In section 3, the methods for estimating the positions of mistracked coils will be described. In section 4, the experiment results based on ground truth data will be illustrated. In section 5, we will give a short summary about current work.

2. Material

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 the 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 RE, LE, and NOSE) serve as the references (shown in **Fig. 1**). Two subjects (1 male and 1 female) are recruited in the EMA experiments.

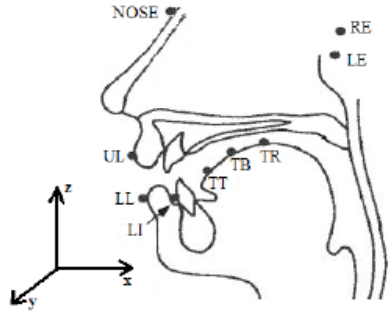

Fig 1. Position of coils in EMA experiment.

Three types of mistracking are discovered in the collected EMA data (as shown in **Fig** 2 and **Fig** 3): i.) abrupt jump of coil position at the beginning and in the middle of utterances; ii.) continuous shifting of coil position at the end of utterances; iii.) coil position beyond the region of vocal tract.

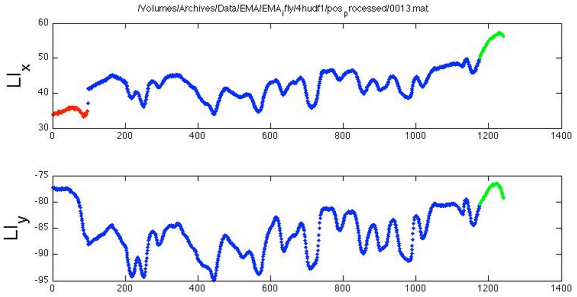

Fig. 2 The 1st (the red curve at the beginning of a utterance in the upper panel) and 2nd (the green curves at the end of an utterance in both lower and upper panels) type of coil mistracking in EMA data.

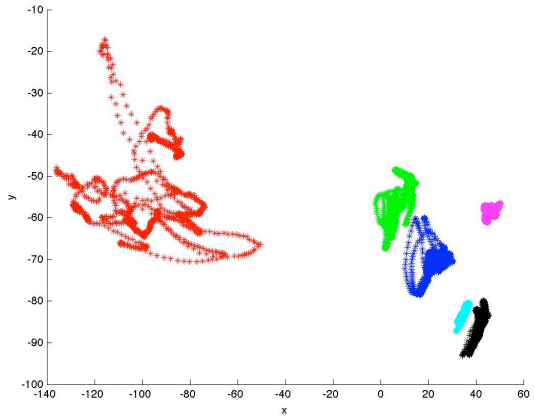

Fig. 3 The 3rd type of coil mistracking. The clouds with different colors stand for TR, TB, TT, LI, LL, and UL, respectively. The coil for TR (denoted by red spots) is beyond the vocal tract.

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. Mistracking occurs most often on one coil at a time and very rarely on multiple articulators. If we discard the whole utterance when mistracking is detected, a number of data could not be used for further studies. This will make us fail to collect a phonetically balanced database for various applications. Nevertheless, the movements of articulators are either physiologically or functionally correlated. Therefore, it is possible to exploit this property and apply machine-learning techniques to reconstruct the positions of mistracked coils from those of the correctly tracked coils.

3. Methods

In this part, we will introduce the methods to estimate the position of the 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 positions 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. In this study, GMM is applied to approximate the conditional probability density function $p(\boldsymbol{y}|\boldsymbol{x})$. Then the mapping function is elaborated by applying MMSE and ML method based on $p(\boldsymbol{y}|\boldsymbol{x})$.

3.1. Conditional 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).

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

$$\boldsymbol{\mu}_k = \left[(\boldsymbol{\mu}_k^x)^T,(\boldsymbol{\mu}_k^y)^T\right]^T \quad (2)$$

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{xx} & \boldsymbol{\Sigma}_k^{xy} \\ \boldsymbol{\Sigma}_k^{yx} & \boldsymbol{\Sigma}_k^{yy} \end{bmatrix} \quad (3)$$

where π_k is the weighting coefficient of the *k-th* mixture, $\boldsymbol{\mu}_k^x$ and $\boldsymbol{\mu}_k^y$ are the mean of source and target vectors of the *k-th* mixture, respectively. $\boldsymbol{\Sigma}_k^{xx}$ and $\boldsymbol{\Sigma}_k^{yy}$ are the covariance matrices of the *k-th* mixture for source and target vectors, respectively. $\boldsymbol{\Sigma}_k^{xy}$ and $\boldsymbol{\Sigma}_k^{yx}$ are the cross-covariance matrices of the *k-th* mixture between source and target vectors, respectively. Then, the probability density function of $\boldsymbol{y}$ given $\boldsymbol{x}$ could be expressed by Eq.4~7

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

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

$$\sum_k^{\mathbf{y}|\mathbf{x}} = \sum_k^{\mathbf{yy}} - \sum_k^{\mathbf{yx}} \left(\sum_k^{\mathbf{xx}}\right)^{-1} \sum_k^{\mathbf{xy}} \quad (6)$$

$$w_k = \frac{\pi_k N(\boldsymbol{x}; \boldsymbol{\mu}_k^{\mathbf{x}}, \sum_k^{\mathbf{xx}})}{\sum_{k=1}^{K} \pi_k N(\boldsymbol{x}; \boldsymbol{\mu}_k^{\mathbf{x}}, \sum_k^{\mathbf{xx}})} \quad (7)$$

3.2. MMSE method

In conventional applications, people usually using 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 $\hat{\boldsymbol{y}}_t$, the target vector could be estimated by using Eq.9.

$$\boldsymbol{y}^* = \int \boldsymbol{y} p(\boldsymbol{y} \mid \boldsymbol{x}) d\boldsymbol{x} = \sum_{k=1}^{K} w_k \boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}} \quad (9)$$

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 $p(\boldsymbol{y}|\boldsymbol{x})$.

3.3. Maximum Likelihood method

It is obvious that Eq.9 only makes use of the weighting coefficients $\boldsymbol{w}_k$ and mean vectors $\boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}}$ of the Gaussian mixtures. Introducing more information may further improve the performance of estimation. ML method, which takes both means and covariance matrices of Gaussian components into account, would be a feasible candidate. Let $Y = \left[\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, ..., \boldsymbol{y}_N^T\right]^T$, $X = \left[\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, ..., \boldsymbol{x}_N^T\right]^T$ be the target and source vector sequence of an utterance, respectively. The equation for ML is as follows.

$$Y^* = \arg\max_{Y} \ln p(Y \mid X) \quad (10)$$

The optimal value of $\boldsymbol{Y}$ could be obtained by using E-M algorithm, where the auxiliary function is formulated as in Eq.11.

$$\begin{aligned} Q(Y, \hat{Y}) &= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{k,i} \ln p(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}}, \sum_k^{\mathbf{y}|\mathbf{x}}) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{k,i} \left(-\frac{1}{2} \boldsymbol{y}_i^T (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1} \boldsymbol{y}_i + \boldsymbol{y}_i^T (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1} \mu_k^{\mathbf{y}} + C\right) \\ &= \sum_{i=1}^{N} \left[-\frac{1}{2} \boldsymbol{y}_i^T \left(\sum_{k=1}^{K} \gamma_{k,i} (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1}\right) \boldsymbol{y}_i + \boldsymbol{y}_i^T \left(\sum_{k=1}^{K} \gamma_{k,i} (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1} \boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}}\right)\right] + \bar{C} \\ &= Y^T D Y + Y^T M + \bar{C} \end{aligned} \quad (11)$$

$$\gamma_{k,i} = \frac{w_k p(\hat{\boldsymbol{y}}_i | \boldsymbol{x}_i; \boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}}, \sum_k^{\mathbf{y}|\mathbf{x}})}{\sum_{k=1}^{K} w_k p(\hat{\boldsymbol{y}}_i | \boldsymbol{x}_i; \boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}}, \sum_k^{\mathbf{y}|\mathbf{x}})} \quad (12)$$

$$D = diag(D_1, D_2, ..., D_N), D_i = \sum_{k=1}^{K} \gamma_{k,i} (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1} \quad (13)$$

$$M = \left[M_1^T, M_2^T, ..., M_n^T\right], M_i = \sum_{k=1}^{K} \gamma_{k,i} (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1} \boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}} \quad (14)$$

Thus, Y could be estimated iteratively by using Eq.15.

$$\hat{Y} = D^{-1} M \quad (15)$$

Namely, the position of target vector at each instant could be iteratively estimated by using Eq.16.

$$\hat{\boldsymbol{y}}_i = \left(\sum_{k=1}^{K} \gamma_{k,i} (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1}\right)^{-1} \sum_{k=1}^{K} \gamma_{k,i} (\sum_k^{\mathbf{y}|\mathbf{x}})^{-1} \boldsymbol{\mu}_k^{\mathbf{y}|\mathbf{x}} \quad (16)$$

If $\sum_k^{\mathbf{y}|\mathbf{x}} = \sum$, which means all the Gaussian components share the same covariance matrix, then the mapping function of ML degrade to a similar result of MMSE.

3.4. Incorporating Dynamic feature

The above ML method tries to improve the performance from the algorithm point of view. In this part, we will make some efforts from the perspective of input feature. As we know, the articulatory data sequence itself contains not only static position information but also dynamic information, e.g. velocity. Therefore, the position vector augmented with velocity information will provide more information of articulatory movements, and may helps to further improve the performance.

Let $X_t = \left[x_t^T, \Delta x_t^T\right]$, $Y_t = \left[y_t^T, \Delta y_t^T\right]$ be the source and target vectors that contain both position and velocity at instant *t*, respectively. Consequently, the corresponding source and target vector trajectories for an utterance are formulated as $X = \left[X_1^T, X_2^T, ..., X_N^T\right]^T$ and $Y = \left[Y_1^T, Y_2^T, ..., Y_N^T\right]^T$, respectively. If the position vector is $\boldsymbol{y} = \left[y_1^T, y_2^T, ..., y_N^T\right]^T$, then the relation between *Y* and $\boldsymbol{y}$, would be:

$$Y = W\boldsymbol{y} \quad (18)$$

where $\boldsymbol{W}$ is the same as the matrix that Tokuda used in parameter trajectory generation for HMM-based speech synthesis[7]. Finally, based on the ML method, the trajectory of the mistracked coils could be reconstructed by using Eq.19.

$$\hat{\boldsymbol{y}} = \left(W^T D W\right)^{-1} W^T M \quad (19)$$

4. Experiment results

300 sentences in set B serve as the training set to train the GMM and derive the mapping function, and the other 32 sentences 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.

4.1. Experiments based on MMSE

The influences of mixture number and coils/articulators identity are investigated by using the MMSE method. The results are shown in Fig 4. The mixture number varies form 8 to 1024. It indicates that the RMSE of all the coils decreases when the number of mixtures increases from 8 to 256, while the RMSE increases when the number of mixtures increases from 256 to 1024. The RMSE of the coil for upper-lip is lowest (about 0.88mm), while the RMSE of

the coil for TT is highest (about 1.34mm). 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. This is comparable with the result reported by Qin on X-ray microbeam data set [6].

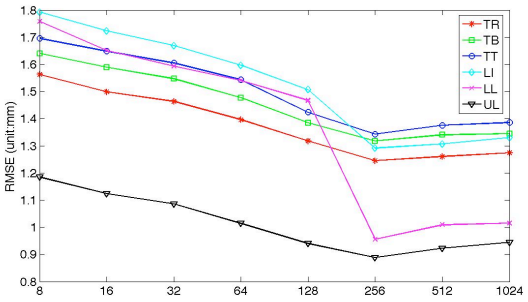

Fig 4. The influences of mixture number and articulator identity on the performance of MMSE.

4.2. Experiments based on ML and ML-dyn

In this part, we test whether the accuracy of the reconstructed positions can be improved by incorporating additional information (covariance matrix/dynamic feature). The GMMs with 256 mixtures are trained to approximate the joint probability density functions for ML and ML-dyn methods. For ML method, only the position vectors of coils are taken as the input feature. While for ML-dyn method, both position and velocity vectors are taken as input feature.

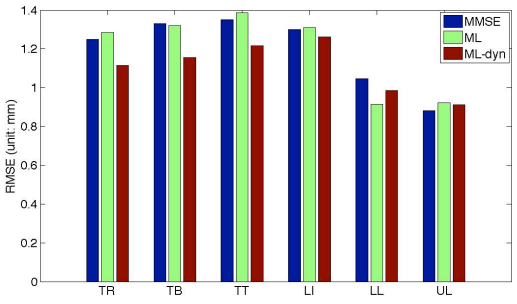

Fig 5. Comparison of the performance of the MMSE, ML, and ML with dynamic information.

For experiment based on ML and ML-dyn methods, where full covariance matrices are used, and the results of MMSE are taken as initial values.

For the results obtained by ML method, the RMSE of TR, TB, TT and UL is about 0.1mm larger than that obtained by MMSE. The RMSE of LI is almost the same as that obtained by MMSE. However, the RMSE obtained by ML is about 0.4mm less than that obtained by MMSE. This indicates that the performance of ML method does not outperform that of MMSE method in the current experiment.

For the results obtained by ML-dyn method, the RMSEs of TR, TB, TT, and LI are less than that obtained by both MMSE and ML methods. The RMSE of LL is larger than that obtained by ML, but still less than that obtained by MMSE method. The RMSE of UL is less than that obtained by ML method, but larger than that obtained by MMSE method. It indicates that the velocity information helps improve the performance of estimation.

5. Conclusion

In this study, we attempt to reconstruct the positions of mistracked coils by using GMM-based mapping functions. To this end, we exam three methods (MMSE, ML, and ML-dyn) and compared the performance of the three methods. It indicates that the performance of ML-dyn is better than that of MMSE and ML in most cases. The RMSEs of the reconstruct coils are about 1.2mm on test set. This result is comparable with the measurement error of EMA machine and the result of Qin [6] on X-ray microbeam dataset. This suggests that the positions of mistracked coils can be reconstructed from the positions of correctly tracked coils. However, the performance of ML is a little bit worse than that of MMSE. This is not as we have expected from theoretical analysis. This may be caused by either insufficient data for estimating the full covariance matrices of Gaussian mixtures or misclassification of correctly tracked data in EMA data.

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] P. Hoole, "On the lingual organization of the German vowel system," *J. Acoust. Soc. Am.,* vol. 106, pp. 1020-1032, 1999.

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

[3] L. Z., K. Richmond, J. Yamagishi, and R. Wang, "Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 17, pp. 1171-1185 2009.

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

[5] S. Roweis, "Data driven production models for speech processing," Ph.D, California Institute of Technology, 1999.

[6] C. Qin and M. A. Carreira-Perpinan, "Reconstructing the full tongue contour from EMA/X-Ray microbeam," in *ICASSP*, 2010.

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," presented at the ICASSP, 2000.