Is average RMSE appropriate for evaluating acoustic-to-articulatory inversion?

Qiang Fang
Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
Email: fangqiang@cass.org.cn

Abstract Acoustic-to-articulatory inversion has potential application in number of fields. For decades, average root mean square error and Pearson correlation coefficient are the most prevalent quantities adopted to evaluate the performance of acoustic-to-articulatory inversion. Various inversion methods have been developed to less the average root mean square error, but very few studies explored whether the average root mean square error is appropriate for evaluating and comparing the performance of different inversion methods. In this study, we attempt to tackle this issue by comparing not only the average root mean square error but also channel root mean square error of each articulatory channel, and the root mean square error of the critical and non-critical portions of each articulatory channel for methods within and between different groups. It is found that: i) the root mean square error of each articulatory channel, and the root mean square error of the critical and non-critical portions of each articulatory channel decrease while the average root mean square error decrease if the AAI methods belong to the same group; ii) exceptions are found if the inversion methods belong to different categories; iii) the average root mean square error is dominated by that of non-critical portions of articulatory channels. This suggests that new methods, which pay more attention to the performance of acoustic-to-articulatory inversion on critical articulators and facilitate the comparison of performance of methods belonging to different categories, should be developed in the future.

1. Introduction

Movements of articulators are slow and smooth compared to corresponding acoustic features. Hence, it naturally has advantages in applications which requires short-term steady features, such as speech synthesis, coding, and recognition. Though articulatory movement information is important, collecting articulatory movement data is not so easy as collecting acoustic signal data. It always requires some types of special instruments, such as ElectroMegnetic Articulograph (EMA), Ultrasound, MRI etc. But none of these instruments can be directly incorporated into the applications of flexible speech synthesis, speech recognition, and speech coding at site. Hence, the features of articulatory movements should be inferred from corresponding acoustic features, which is called acoustic-to-articulatory inversion (AAI).

To tackle the AAI issue, for decades, a number of works has been conducted based on parallel acoustic-articulatory databases. In this case, the AAI is formulated as regression tasks, which produce an output form associated input based on models trained on input-output data pair. For this purpose, a number of studies have been carried out. Various statistical models have been applied to the task of AAI, such as multilayer perceptron (MLP) [1] [2] [3], mixture density network[3], trajectory Gaussian Mixture Model[4], HMM-based speech production model[5], trajectory HMM [6], deep forward trajectory density neural network[7], bidirectional long-short term memory RNN [8] [9]. And the effectiveness of the input acoustic features has been extensively studied. Qin [10] explored the effects of choosing different popular acoustic features (LPC, LSF, FBANK, MFCC, LPCC, PLP, RASTA-PLP) with and without dynamic features, different short-time window lengths , and different levels of smoothing of the acoustic temporal trajectories on the performance of acoustic-to-articulatory inversion with MLP. Ghosh [11] used mutual information as the criterion to rank the MFCC and their derivative according to the information on different articulatory features in acoustic-to-articulatory inversion. Some studies tried to incorporate phoneme information to enhance the performance of acoustic-to-articulatory inversion [12, 13]. To evaluate the AAI performance, the average Root Mean Square Error (a-RMSE) and Pearson correlation coefficient of all the measured articulatory channels are adopted in most of the previous studies.

Nonetheless, whether the a-RMSE is appropriate for evaluating the performance of AAI has not been extensively investigated. To our knowledge, proper quantities for evaluating AAI performance should satisfy the following requirements at least: i) the RMSEs of each channel should decrease consistently while the a-RMSE decreases; ii) the RMSE of critical portion of movements of all critical channels should decrease consistently while the a-RMSE decreases (see Section 4.2 for definition). In this study, we make effects to shed light on these two issues by comparing the results obtained by four widely used AAI methods.

2. Dataset

2.1. The MOCHA database

The publicly available multichannel articulatory database (MOCHA), released by the Centre for Speech Technology Research, University of Edinburgh, is used in this study. In MOCHA database, the 460 British TIMIT sentences were uttered by two subjects, fsew0 and msak0. Four data streams were recorded: the waveform (16 kHz sample rate, with 16-bit precision for quantification) together with laryngograph, electropalatograph (EPG), and EMA data. The waveform signal and articulatory information were synchronized and output to a computer simultaneously.

The EMA was used to retrieve the movement of articulators. For this purpose, coils were attached to upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD) and velum (V) to track their positions while sentences were uttered. Each of these coils provided horizontal (x) and vertical (y) coordinates in the midsagittal plane. Finally, 14 channels of articulatory information were recorded in total. Additional coils were attached to the nose bridge and the upper incisor to server as references. The movement of coils attached to the articulators were sampled with the sampling frequency of 500 Hz.

2.2. Data processing

Before feeding synchronized acoustic-articulatory data to train and evaluate different AAI methods, some preprocessing procedure need be conducted. Firstly, silences at the beginning and end of each speech and corresponding EMA files are omitted, since articulators can possess any status at those silent part.

The speech signals are transformed to MFCC parameters (12 mel-cepstral + C_o) with the setting (Hamming window with the length corresponding to 25ms speech signals, frame shift with the length corresponding to 10ms speech signals, and 26 channels for filter-bank analysis) by using HTK speech recognition toolkit. A context window of 11 frames are used to organize acoustic feature sequence into input sequence of model for AAI.

The EMA data are bidirectionally filtered with a lowpass filter (10-order finite impulse filter with cutoff frequency of 20Hz), first forward then backward, to remove the high-frequency artifacts and avoid phase distortion. Then, the trajectory of 460 utterances' means is smoothed by a Savitzky-Golay filter (with the order of 5 and frame-size of 121) to obtain slow varying means of utterance. At last, the articulatory data of each utterance are normalize with the reference to the mean of each utterance [14]. To match the frame frequency of acoustic feature, the EMA data is down-sampled 100 frames per second. The EMA frames corresponds to the 6th acoustic feature frame in the input sequence (a sequence of 11 frames of acoustic features) are extracted to formulate the target outputs for AAI.

Subject fsew0's data is used in this study. The set of 460 utterance is divided into 3 subsets with no overlap: a training set with 370 utterances, a validation set with 45 utterances, and a testing set with 45 utterances.

3. Methods

3.1. Gaussian mixture model

Let $\boldsymbol{x}_t$, $\boldsymbol{y}_t$, $\Delta\boldsymbol{y}_t$, and $\Delta\Delta\boldsymbol{y}_t$ denote the acoustic parameter, the position, velocity, and acceleration of articulators at instance *t*, respectively Then, $\boldsymbol{X_t} = [\boldsymbol{x}_{t-c}^T, \dots, \boldsymbol{x}_t^T, \dots, \boldsymbol{x}_{t+c}^T\,]^T$, $\boldsymbol{Y_t} = [\boldsymbol{y}_t^T, \Delta\boldsymbol{y}_t^T, \Delta\Delta\boldsymbol{y}_t^T\,]^T$ are the contextual acoustic parameter vector and articulatory parameter vector used to train a joint gaussian mixture model (GMM). The joint probability density function (PDF) $p(\boldsymbol{X}_t,\, \boldsymbol{Y}_t)$ is formulated as :

$$p(\boldsymbol{X}_t,\, \boldsymbol{Y}_t) = \sum_{k=1}^{M} \pi_k \mathcal{N}(\boldsymbol{X}_t,\, \boldsymbol{Y}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

$$\boldsymbol{\mu}_k = \left[\boldsymbol{\mu}_k^{X^T}, \boldsymbol{\mu}_k^{Y^T}\right]^T \tag{2}$$

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{XX} & \boldsymbol{\Sigma}_k^{XY} \\ \boldsymbol{\Sigma}_k^{YX} & \boldsymbol{\Sigma}_k^{YY} \end{bmatrix} \tag{3}$$

where π_k is the probability that the k^{th} distribution $\mathcal{N}(\boldsymbol{X}_t,\, \boldsymbol{Y}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is used to generate sample $(\boldsymbol{X}_t,\, \boldsymbol{Y}_t)$. $\boldsymbol{\mu}_k^{X}$ and $\boldsymbol{\mu}_k^{Y}$ are the mean of acoustic and articulatory parameter vectors of the k^{th}component, respectively. $\boldsymbol{\Sigma}_k^{XX}$ and $\boldsymbol{\Sigma}_k^{YY}$ are the covariance matrices of the k^{th} component of the acoustic and articulatory vectors, respectively. $\boldsymbol{\Sigma}_k^{XY}$ is the cross-covariance matrices of the k^{th} component distribution between acoustic and articulatory parameter vectors, respectively. Then, the probability density function of $\boldsymbol{Y}_t$ given $\boldsymbol{X}_t$ could be expressed by Eq.4~7

$$p(\, \boldsymbol{Y}_t | \boldsymbol{X}_t) = \sum_{k=1}^{M} w_k \mathcal{N}(\boldsymbol{Y}_t | \boldsymbol{X}_t; \boldsymbol{\mu}_k^{Y|X}, \boldsymbol{\Sigma}_k^{Y|X}) \tag{4}$$

$$\boldsymbol{\mu}_k^{Y|X} = \boldsymbol{\mu}_k^{Y} + \boldsymbol{\Sigma}_k^{YX}\left(\boldsymbol{\Sigma}_k^{XX}\right)^{-1}\left(\boldsymbol{X} - \boldsymbol{\mu}_k^{X}\right) \tag{5}$$

$$\boldsymbol{\Sigma}_k^{Y|X} = \boldsymbol{\Sigma}_k^{YY} - \boldsymbol{\Sigma}_k^{YX}\left(\boldsymbol{\Sigma}_k^{XX}\right)^{-1}\boldsymbol{\Sigma}_k^{XY} \tag{6}$$

$$w_k = \frac{\pi_k \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_k^{X}, \boldsymbol{\Sigma}_k^{X})}{\sum_{k=1}^{M} \pi_k \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_k^{X}, \boldsymbol{\Sigma}_k^{X})} \tag{7}$$

3.1.1 Minimum mean square error estimation

Given an acoustic parameter vector, the articulatory parameter vector can be determined by Eq. (8) if Minimum Mean Square Error (MMSE) criterion is taken.

$$\widehat{\boldsymbol{Y}_t} = \sum_{k=1}^{M} w_k \boldsymbol{\mu}_k^{Y|X_t} \tag{8}$$

The articulatory parameters estimated with MMSE are noisy since only the mean $\boldsymbol{\mu}_k^{Y|X_t}$ and the w_k of the PDF of the current frame is taken into account, and the information of the PDF of neighbor frame are omitted. This drawback can be overcome by using a trajectory model.

3.1.2 Trajectory estimation

The acoustic and articulatory vector trajectories of an utterance can be formulated as $\boldsymbol{X} = [\boldsymbol{X}_1^T, \dots, \boldsymbol{X}_t^T, \dots, \boldsymbol{X}_N^T]^T$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \dots, \boldsymbol{Y}_t^T, \dots, \boldsymbol{Y}_N^T]^T$. If the articulatory position vector is denoted by $\boldsymbol{y} = [\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_t^T, \dots, \boldsymbol{y}_N^T]$, the relation between $\boldsymbol{Y}$ and $\boldsymbol{y}$, would be:

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{9}$$

where $\boldsymbol{W}$ is the same as the matrix that Tokuda et al. used in parameter trajectory generation for HMM-based speech synthesis[15]. For a given acoustic parameter sequence $\boldsymbol{X}$, the articulatory position sequence $\boldsymbol{y}$ can be estimated by using Maximum Likelihood Parameter Generation (MLGP) method:

$$\widehat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{Y}|\boldsymbol{X}) \tag{10}$$

Where

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{m} P(\boldsymbol{Y}|\boldsymbol{m}, \boldsymbol{X}) P(\boldsymbol{m}|\boldsymbol{X}) \tag{11}$$

A sequence of the mixture component indices $[m_1, \dots, m_t, \dots, m_N]$ is denoted as $\boldsymbol{m}$.

$$Q\left(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}\right) = \sum_{all\ \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{Y}) \boldsymbol{log} P(\boldsymbol{m}, \boldsymbol{Y}|\boldsymbol{X}) \quad = \\ -\frac{1}{2}\widehat{\boldsymbol{y}}^T \boldsymbol{W}^T \overline{\boldsymbol{D}^{(Y)^{-1}}} \boldsymbol{W}\widehat{\boldsymbol{y}} + \widehat{\boldsymbol{y}}^T \boldsymbol{W}^T \overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}} + \boldsymbol{K} \tag{12}$$

Where

$$\overline{\boldsymbol{D}^{(Y)^{-1}}} = diag\left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}}, \overline{\boldsymbol{D}_2^{(Y)^{-1}}}, \dots, \overline{\boldsymbol{D}_t^{(Y)^{-1}}}, \dots, \overline{\boldsymbol{D}_T^{(Y)^{-1}}}\right] \tag{13}$$

$$\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_{\boldsymbol{1}}^{(Y)^{-1}}\boldsymbol{E}_{1}^{(Y)}}, \overline{\boldsymbol{D}_{2}^{(Y)^{-1}}\boldsymbol{E}_{2}^{(Y)}}, \ldots, \overline{\boldsymbol{D}_{t}^{(Y)^{-1}}\boldsymbol{E}_{t}^{(Y)}}, \ldots, \overline{\boldsymbol{D}_{T}^{(Y)^{-1}}\boldsymbol{E}_{T}^{(Y)}}\right] \quad (14)$$

$$\overline{\boldsymbol{D}_{\boldsymbol{t}}^{(Y)^{-1}}} = \sum_{\boldsymbol{m}=\boldsymbol{1}}^{\boldsymbol{M}} \boldsymbol{\gamma}_{\boldsymbol{m},\boldsymbol{t}}^{(\boldsymbol{Z})} \boldsymbol{D}_{\boldsymbol{m}}^{(Y)^{-1}} \quad (15)$$

$$\overline{\boldsymbol{D}_{\boldsymbol{t}}^{(Y)^{-1}}\boldsymbol{E}_{\boldsymbol{t}}^{(Y)}} = \sum_{\boldsymbol{m}=\boldsymbol{1}}^{\boldsymbol{M}} \boldsymbol{\gamma}_{m,t}^{(\boldsymbol{Z})} \boldsymbol{D}_{m}^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)} \quad (16)$$

$$\boldsymbol{\gamma}_{m,t}^{(\boldsymbol{Z})} = P\left(m \middle| \boldsymbol{X}_{t}, \boldsymbol{Y}_{t}, \boldsymbol{\lambda}^{(z)}\right) \quad (17)$$

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^{\boldsymbol{T}}\overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\right)^{-1} \boldsymbol{W}^{\boldsymbol{T}}\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} \quad (18)$$

The articulatory trajectory can be obtained by iteratively minimize the Q function with Eq.13~18

3.2. Neural network-based methods

3.2.1 Multilayer perceptron

MLP is a class of feedforward neural network. It consists of, at least, three layers of nodes: an input layer, a hidden layer, and an output layer. The nodes in hidden layers usually use nonlinear activation functions to transform their inputs into their outputs. And the nodes in the output layer usually use nonlinear activation functions for classification task, and linear activation for regression task. MLP can be trained using backpropagation method [16]. MLP can be used to approximate any nonlinear function with appropriate parameter setting. Hence, MLP has been implemented to various classification and regression tasks.

In this study, we use an MLP with 2 hidden layers with sigmoid neurons to estimate articulatory trajectory. And each hidden layer has 300 neurons.

3.2.2 Deep bidirectional LSTM RNN

MLP can only use the input information of current frame, although this can flaw can be made up to some extent by using contextual feature frame. To consider the contextual information in a more natural way, recurrent connections can be inserting into a feedforward neural network, which is called recurrent neural network (RNN). RNN is able to remember previous inputs and allow them to persist in the network internal states with recurrent connections. Therefore, RNN can map the history of previous input vectors to each output vector.

Nevertheless, articulatory parameter at current time correlates with the acoustic parameter at current time as well as those in the past and in the future. Hence, it is desirable to incorporate the future acoustic context for acoustic-to-articulatory inversion. A Bidirectional RNN (BRNN) computes both forward hidden sequence $\vec{h}$, and backward hidden sequence $\overleftarrow{h}$. The outputs from both the forward and backward pass are combined together to serve as the input to the output layer. So, it is able to access past and future context by processing data in both directions.

Unfortunately, because of the vanishing gradient problem, RNNs or BRNNs can only access a limited range of context. Long short term memory (LSTM) [17], which consists of a input gate, a forget gate, a outputs gate, and a cell memory, is a solution to solve the vanishing gradient problem in RNN. Bidirectional LSTM (BLSTM) can be implemented by replace the normal neuron cells in hidden layers with LSTM cells.

In this study, the Deep Bidirectional LSTM (DBLSTM) network consists of four hidden layers, in which the inputs are connected to 2 feedforward neural networks with sigmoid activation function, then feed to 2 BLSTM layers. Each hidden layer has 300 neurons.

4. RESULTS

Here, we will present the results obtained by the four methods described Section 3. In this study, we attempt to explore: i) whether the RMSE of each channel decrease consistently while the a-RMSE decreases; ii) whether the RMSEs of critical portion of all the channels decrease consistently while the a-RMSE decreases. So, the models' parameters are mostly adopted from published papers, and are not extensively explored to make models achieve their best performance. When a model achieves comparable performance reported in corresponding literature, the parameter searching procedure for the model is stopped. Eq.19 and Eq.20 are adopted for calculating the root mean square error (RMSE) of each channel and the a-RMSE of all the channels.

$$E(j) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_{i,j} - y_{i,j}\right)^{2}} \quad (19)$$

$$E_{avg} = \frac{1}{N_c}\sum_{i=1}^{N_c} E(i) \quad (20)$$

where $y_{i,j}$ is the ground truth of the *j*th channel of the *i*th sample, $\hat{y}_{i,j}$ is the corresponding estimate, and *Nc* is the number of articulatory channels.

4.1. RMSE of each channel

Table 1. The average RMSE 1 obtained by four different inversion methods (unit: mm).

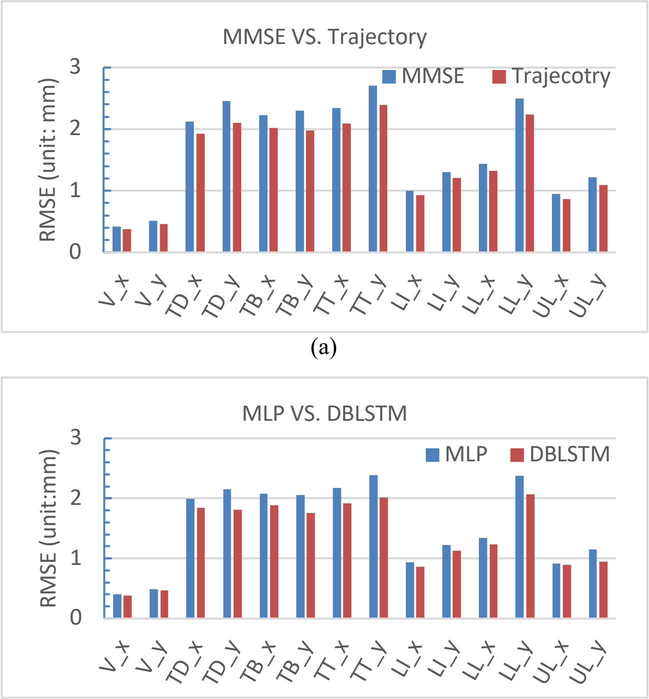
	MMSE	*Trajectory*	*MLP*	*DBLSTM*
Avg.	1.68	1.50	1.56	1.37

Table 1. presents the a-RMSEs obtained by the four methods, namely GMM-based MMSE (MMSE), GMM-based Trajectory (Trajectory), MLP, and DBLSTM. The a-RMSEs obtained by MMSE, Trajectory, and MLP methods are consistent with the results reported by Sudhakar[18], Toda[4], and Richmond[14] on the same database, respectively. As for the performance of DBLSTM methods for AAI, it is difficult to compare our work with previous works quantitatively since few results have been reported on MOCHA database. Fortunately, Liu[9] and Zhu[19] found that the a-RMSEs of DBLSTM are smaller that of MLP on MNGU0 database. Similar phenomenon is also observed in our experiments, which qualitatively proves that the results achieved by DBLSTM is reasonable. The above preliminary evaluation indicates that our experiment reproduced the results reported in previous studies. In addition, the obtained a-RMSEs are in the order that: DBLSTM < Trajectory < MLP < MMSE.

Then, the four methods are grouped into two categories according to the criterions adopted to train and predict articulatory status. Among them, MMSE and Trajectory methods, which estimate the joint acoustic-articulatory PDF with maximum likelihood criterion and predict articulatory status based on the conditional PDF derived from the estimated joint PDF, are grouped into one category, while

MLP and DBLSTM, which estimate the model parameters based on least mean square error criterion and predict articulatory status directly from input acoustic parameters, are grouped into the other group.

Firstly, it is of interest to know whether the RMSE of each channel decreases consistently while the a-RMSE decreases for the methods belonging to the same group. Fig 1(a) and Fig 1(b) presents the results of the two groups methods. As shown in Fig 1(a), all the channel-RMSEs of Trajectory method are smaller than those of MMSE method. Similar phenomenon is observed for MLP and DBLSTM methods. This indicates that the RMSE of each channel decreases consistently while the a-RMSE decreases if the AAI methods are in the same group.

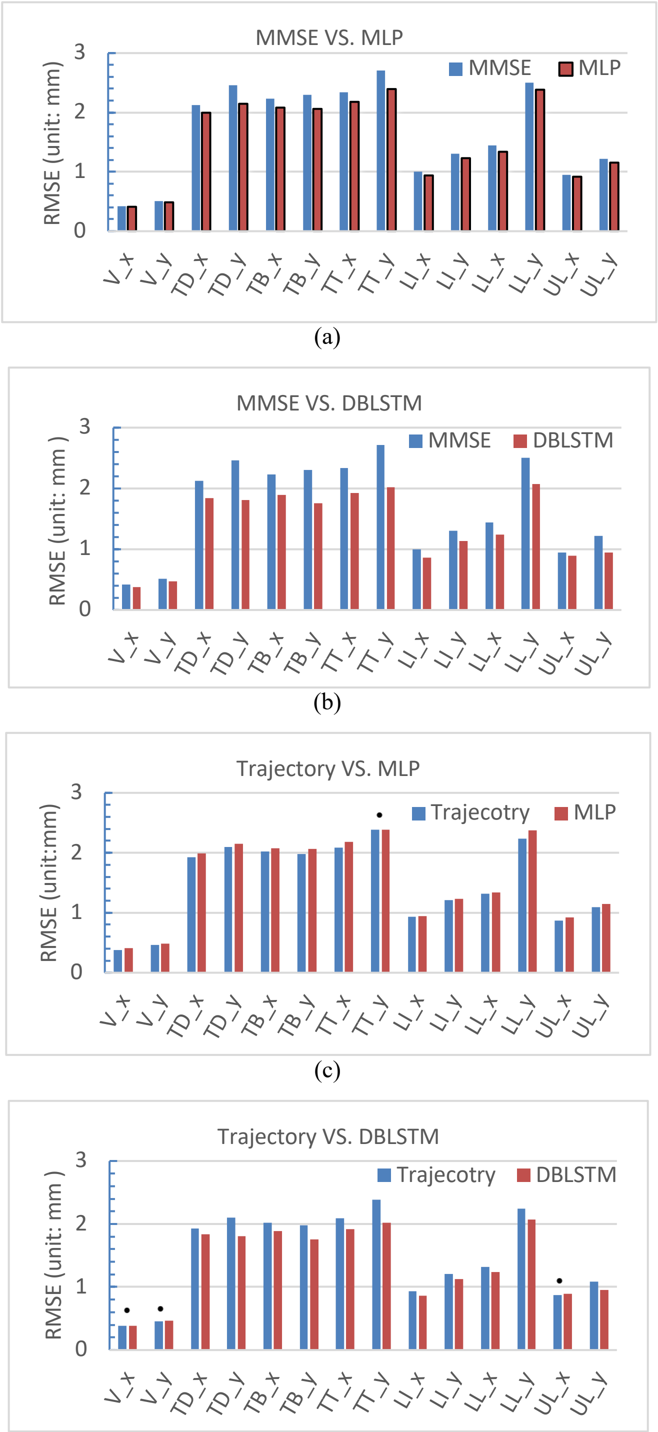


(a)

(b)

Fig. 1 Comparison of the RMSE of each channel obtained by methods in the same group. (a) Comparison of channel RMSEs obtained MMSE and Trajectory method (b) Comparison of RMSEs obtained by MLP and DBLSTM method.

Secondly, it is of interest to know whether the RMSE of each channel decreases consistently while the a-RMSE decrease for the methods belonging to different groups. Fig. 2 presents the results of the comparison of the channel-RMSEs between the methods belong to different groups. As shown Fig 2(a) and Fig 2(b), when comparison is made between the channel-RMSEs obtained by MMSE and by MLP/DBLSTM, it is observed that the RMSE of each channel decrease consistently while the a-RMSE decreases. However, exceptions are found when comparison is made between the channel-RMSEs obtained by Trajectory and by MLP/DBLSTM (TT_y in Fig 2(c), and V_x, V_y, UL_x in in Fig 2(d), denoted by black filled circles). It indicates that the channel-RMSEs do not necessarily decrease while the a-RMSE decreases if the AAI methods belong to different groups.



(a)

(b)

(c)

(d)

Fig. 2 Comparison of the RMSE of each channel obtained by methods in different groups. (a) Comparison of channel RMSEs obtained by MMSE and MLP methods (b) Comparison of channel RMSE obtained by MMSE and DBLSTM methods (c) Comparison of channel RMSEs obtained by Trajectory and MLP methods (d) Comparison of channel RMSEs obtained by Trajectory and DBLSTM methods.

4.2 RMSEs of critical and non-critical portions

Theoretically, different articulators play different roles for producing specific speech sounds. For example, the position tongue tip is required to be in a very restricted area to form closure at the alveolar to produce speech sound /t/ and /d/, while the tongue body and lips are allowed to take positions with large variation. Accordingly, the tongue tip is the critical articulator and the other articulators are non-critical articulators when producing sound /t/ and /d/. In continuous speech, for each articulatory channel, it serves as critical articulator in some portions (call critical portion) while serves as non-critical articulator in other portions (call non-critical portions) according to the identities of uttered phonemes. To take an further insight on the performance of the four methods on critical and non-critical portions of articulator channels, we annotate the articulators at each time stamp as critical/noncritical articulators with reference to the segmental information offered in MOCHA database and the criterion used by Papcun [1] and Okadome [20].

Table 2 presents the result of average RMSE of critical (a-crt -RMSE) and noncritical (a-ncrt-RMSE) articulators obtained by the four methods. For all the AAI methods, the a-crt-RMSEs are in the order of DBLSTM < Trajectory < MLP < MMSE, and the a-ncrt-RMSEs are also in the order of DBLSTM < Trajectory < MLP < MMSE. But the a-ncrt-RMSEs are nearly 25% larger than the a-crt-RMSEs. If comparing the results in Table 2 with the results in Table 1, one can find that the a-RMSEs are dominated by the a-ncrt-RMSEs.

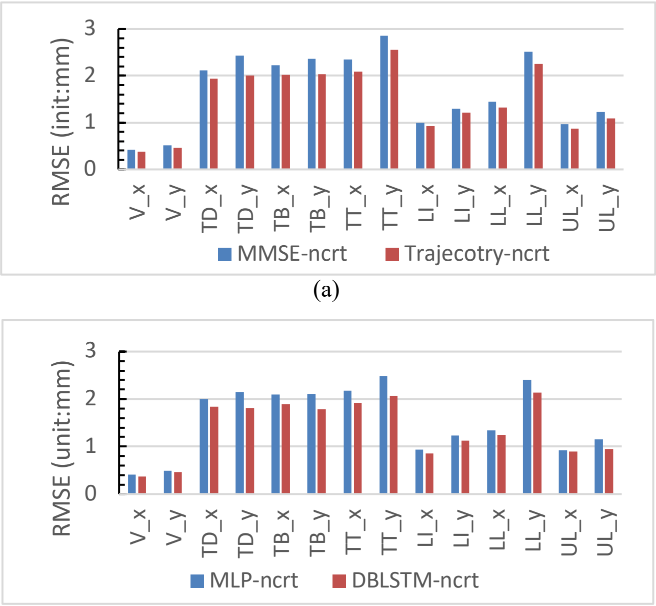
Table 2. The average RMSEs of critical (denoted by crt) and noncritical (denoted by ncrt) portions of articulatory channels.

	MMSE		*Trajectory*		*MLP*		*DBLSTM*	
	ncrt	crt	ncrt	crt	ncrt	crt	ncrt	crt
Avg.	1.69	2.26	1.51	2.00	1.57	2.04	1.39	1.69

4.2.1 RMSEs of non-critical portions

As discussed in Section 4.1, it is of interest to know whether the RMSE of non-critical portions of each channel decreases consistently while the a-RMSE decreases for the methods belonging to the same group. Fig 3(a) and Fig 3(b) presents the results of the two groups methods. As shown in Fig 3(a), all the ncrt-RMSEs obtained by Trajectory method are smaller than those obtained by MMSE method. Similar phenomenon is observed for MLP and DBLSTM methods. This indicates that the RMSE of non-critical portions of each channel decreases consistently while the a-RMSE decreases if the AAI methods belong to the same group.

Furthermore, it is of interest to know whether the ncrt-RMSE of each channel decreases consistently while the a-RMSE decreases for the methods belonging to different groups. Fig 4 presents the results of the comparison of the ncrt-RMSE of each articulatory channel between the methods belong to different groups. As shown Fig 4(a) and Fig 4(b), when comparison is made between the ncrt-RMSE of each channel obtained by MMSE and by MLP/DBLSTM, it is observed that the nct-RMSE of each channel decreases consistently while the a-RMSE decreases. However, exceptions are found when comparison is made between the ncrt-RMSE of each channel obtained by Trajectory and by MLP/DBLSTM (TT_y in Fig 4(c), and V_x, V_y, UL_x in in Fig 4(d), denoted by black filled circles). It indicates that the ncrt-RMSE of each channel does not necessarily decrease while the a-RMSE decreases if the AAI methods belong to different groups.



(a)

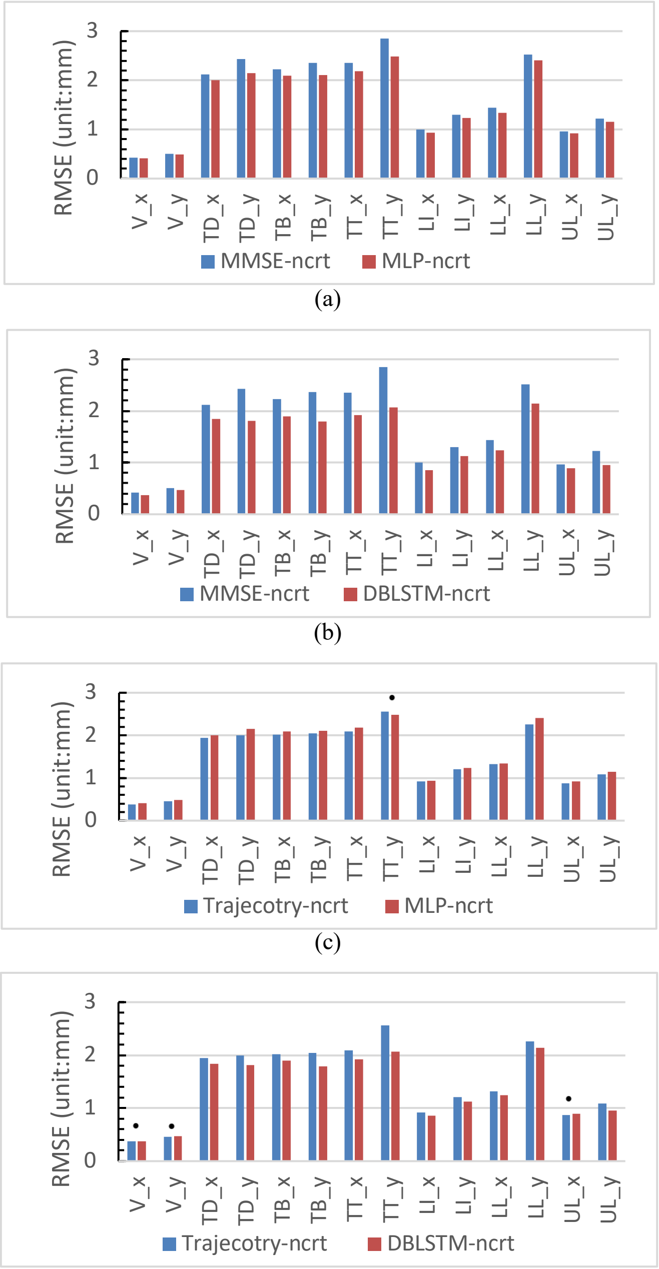
(b)

Fig. 3 Comparison of the RMSE of non-critical portions of each channel obtained by methods belonging to the same group. (a) Comparison the RMSEs of non-critical channels obtained MMSE and Trajectory method; (b) Comparison of the RMSEs of non-critical channels obtained by MLP and DBLSTM method.

4.2.2 RMSE of critical portions

Since the position of critical articulators determine the identity of the utter speech sound, it is of interest to know whether the RMSE of critical portions of each channel decreases consistently while the a-RMSE decreases for the methods belonging to the same group. Fig 5(a) and Fig 5(b) presents the results of the two groups methods. As shown in Fig 5(a), all the crt-RMSEs obtained by Trajectory method are smaller than those obtained by MMSE method. Similar phenomenon is observed for MLP and DBLSTM methods. This indicates that the RMSE of critical portions of each channel decreases consistently while the a-RMSE decreases if the AAI methods belong to the same group.

Furthermore, it is of interest to know whether the crt-RMSE of each channel decreases consistently while the a-RMSE decreases for the methods belonging to different groups. Fig 6 presents the results of the comparison of the crt-RMSE of each articulatory channel between the methods belong to different groups. As shown Fig 6(a) and Fig 6(b), when comparison is made between the crt-RMSE of each channel obtained by MMSE and by MLP/DBLSTM, and by Trajectory and by DBLSTM, it is observed that the crt-RMSE of each channel decreases consistently while the a-RMSE decreases. However, exceptions are found when comparison is made between the ncrt-RMSE of each channel obtained by Trajectory and by MLP (TD_y, LL_y in Fig 6(c), denoted by black filled circles). It indicates that the crt-RMSE of each channel does not necessarily decrease while the a-RMSE decreases if the AAI methods belong to different groups.



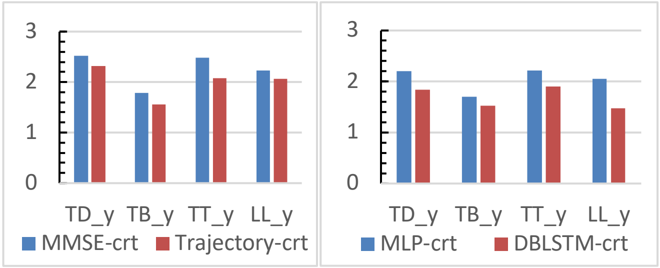
(a)

(b)

(c)

(d)

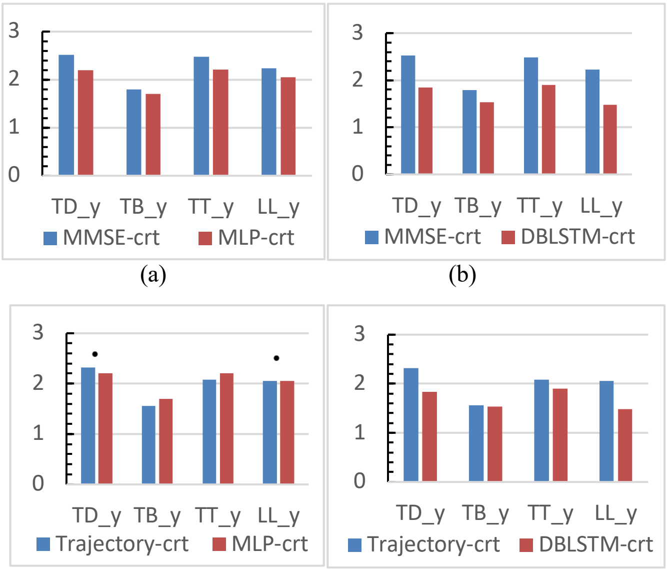
Fig. 4 Comparison of the RMSE of non-critical portions of each channel obtained by methods in different groups. (a) Comparison of the RMSEs of non-critical portions obtained by MMSE and MLP methods; (b) Comparison of the RMSEs of non-critical portions obtained by MMSE and DBLSTM methods; (c) Comparison of the RMSEs of non-critical portions obtained by Trajectory and MLP methods; (d) Comparison of the RMSEs of non-critical portions obtained by Trajectory and DBLSTM methods.



(a)

(b)

Fig. 5 Comparison of the RMSE of critical portions of each channel obtained by methods belonging to the same group. (a) Comparison the RMSEs of critical portions obtained MMSE and Trajectory method; (b) Comparison of the RMSEs of critical portions obtained by MLP and DBLSTM method.



(a)

(b)

(c)

(d)

Fig. 6 Comparison of the RMSE of critical portions of each channel obtained by methods in different groups. (a) Comparison of the RMSEs of critical portions obtained by MMSE and MLP methods; (b) Comparison of the RMSEs of critical portions obtained by MMSE and DBLSTM methods; (c) Comparison of the RMSEs of critical portions obtained by Trajectory and MLP methods; (d) Comparison of the RMSEs of critical portions obtained by Trajectory and DBLSTM methods.

5. Conclusion

In this study, preliminary analysis is conducted on the performance of different AAI methods, which roughly belong to two different categories. It is found that the RMSE, crt-RMSE, and ncrt-RMSE of each articulatory channel decrease while the a-RMSE decreases if the AAI methods belong to same category. While some exceptions are found if the AAI methods belong to different categories. This indicates that a-RMSE maybe proper for comparing the performance of methods belonging to the same categories, but not appropriate for comparing the performance of methods belonging to different categories. Therefore, when comparing the performance of methods from different categories with a-

RMSE, one should keep in mind that the decrease of a-RMSE doesn't necessarily mean the decrease other important measures. Besides, it's found that a-RMSEs are dominated by ncrt-RMSEs, and the crt-RMSEs are about 25% larger than the ncrt-RMSEs. This suggests that new methods, which pay more attention to the performance of AAI on critical articulators and facilitate the comparison of performance of inversion methods belonging to different categories, should be developed in the future.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science-Foundation of China (No.61573254), Key Project of National Social Science Foundation of China (No.15ZDB103), Advanced Innovation Center for Language Resource and Intelligence (KYR17005), CASS Innovation Project "Articulatory model for pronunciation training", and the programs of Beijing Municipal Administration of Hospital Clinical Medicine Development of Special Funding Support (Code. XMLX201848).

REFERENCES

[1]. Papcun, G., et al., *Inferring articulation and recognising gestures from acoustics with a neural network trained on X-ray microbeam data.* J. Acoust. Soc. Am., 1992. **92**(2): p. 688–700.

[2]. Qin, C. and M.A. Carreira-Perpinan, *A Comparison of Acoustic Features for Articulatory Inversion*, in *InterSpeech2007*. 2007: Antwerp. p. 2469-2472.

[3]. Richmond, K., *A trajectory mixture density network for the acoustic-articulatory inversion mapping*, in *InterSpeech2006*. 2006. p. 577–580.

[4]. Toda, T., A.W. Black, and K. Tokuda, *Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model.* Speech Communication, 2008. **50**: p. 215–227.

[5]. 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.

[6]. Zhang, L. and S. Renals, *Acoustic-articulatory modeling with the trajectory HMM.* IEEE Signal Process Letter, 2008. **15**: p. 245–248.

[7]. Uría, B., et al., *Deep Architectures for Articulatory Inversion*, in *InterSpeech2012*. 2012.

[8]. Zhu, P., L. Xie, and Y. Chen, *Articulatory Movement Prediction Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks andWord/Phone Embeddings*, in *Interspeech2016*. 2016: San Francisco.

[9]. Liu, P., et al., *A DEEP RECURRENT APPROACH FOR ACOUSTIC-TO-ARTICULATORY INVERSION*, in *ICASSP 2015*. 2015. p. 4450-4454.

[10]. Qin, C., et al. *Predicting tongue shapes from a few landmark locations*. in *Interspeech*. 2008.

[11]. Ghosh, P.K. and S. Narayanan, *Information Theoretic Acoustic Feature Selection for Acoustic-to-Articulatory Inversion*, in *Interspeech2013*. 2013: Lyon. p. 3177-3181.

[12]. Xie, X., X. Liu, and L. Wang, *Deep Neural Network Based Acoustic-to-articulatory Inversion Using Phone Sequence Information*, in *Interspeech2016*. 2016: San Francisco. p. 1497-1501.

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

[14]. Richmond, K., *Estimating articulatory parameters from the acoustic speech signal*. 2002, University of Edinburgh.

[15]. Tokuda, K., et al., *Speech parameter generation algorithms for HMM-based speech synthesis*, in *ICASSP*. 2000. p. 1315-1318.

[16]. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning internal representations by error propagation*, in *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*, A. Collins and E.E. Smith, Editors. 1988, Morgan Kaufmann Publishers, Inc. p. 399-421.

[17]. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural Computation, 1997. **9**(8): p. 1735-1780.

[18]. Sudhakar, P. and P.K. Ghosh. *Sparse smoothing of articulatory features from Gaussian mixture model based acoustic-to-articulatory inversion: Benefit to speech recognition*. in *Interspeech2014*. 2014. Singapore.

[19]. Zhu, P., L. Xie, and Y. Chen, *Articulatory Movement Prediction Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks andWord/Phone Embeddings*, in *Interspeech2015*. 2015. p. 2192-2196.

[20]. Okadome, T. and M. Honda, *Generation of articulatory movements by using a kinematic triphone model.* J. Acoust. Soc. Am., 2001. **110**(1): p. 453-463.