Improving the performance of acoustic-to-articulatory inversion by removing the training loss of noncritical portions of articulatory channels dynamically

Qiang Fang

Institute of Linguistics, Chinese Academy of Social Sciences

fangqiang@cass.org.cn

Abstract

For decades, average Root Mean Square Error (RMSE) over all the articulatory channels is one of the most prevalent cost functions for training statistical models for the task of acoustic-to-articulatory inversion (AAI). One of the underlying assumptions is that the samples of all the articulatory channels used for training are balanced and play the same role in AAI. However, this is not true from speech production point view. In this study, at each time instant, each articulatory channel is classified to be critical or noncritical according to their roles in the formation of constrictions along the vocal tract when producing speech sound. It is found that the training set is dominated by the samples of noncritical articulatory channels. To deal with the unbalanced dataset problem, several Bi-LSTM networks are trained by removing the of noncritical portions of each articulatory channels if the training errors are less than some dynamic threshold. The results indicate that the average RMSE over all the articulatory channels, the average RMSE over the critical articulators, and the average RMSE over the noncritical articulators can be reduced significantly by the proposed method.

Index Terms: acoustic-to-articulatory inversion, critical /noncritical articulatory channel, RMSE, Bi-LSTM

1. Introduction

The status of articulators reflects important characteristics of speech. Its variation is relative slow and smooth in contrast to acoustic features. This makes articulator trajectories themselves have potential applications in flexible speech synthesis, speech coding, speech recognition, and speech animation. Though articulatory movement information is important, collecting articulatory movement data is not so easy as collecting acoustic signals. It always requires some types of special instruments, such as EMA, Ultrasound, MRI etc., which are difficult to be used in real application. Hence, to incorporate articulatory parameters into real application, it is better to infer articulatory movements from corresponding acoustic signals.

To tackle this issue, for decades, numerous studies have been conducted based on synchronously recorded acoustic-articulatory database, where AAI was formulated as regression problems. All of the studies endeavored to design better features or devise better models so as to achieve better performances. The effects of choosing different popular acoustic features (LPC, LSF, FBANK, MFCC, LPCC, PLP, RASTA-PLP) [1] , and the effects of with/without dynamic features, different window lengths and different levels of smoothing of the acoustic temporal trajectories, and with/without phoneme information have been investigated[1]. While other studies attempted to implement various statistical models for the task of AAI. In literature, MLP[2], mixture density neural network[3], GMM[4], trajectory HMM[5], deep forward trajectory density neural network[6]， bidirectional LSTM RNN[7] have been applied to AAI. Most of the studies adopted average RMSE over all the articulatory channels as the training loss of statistical models, where samples from critical and noncritical articulatory channels are treated equally.

From speech production point of view, gestures define phonemes. Gestures consist of the formation and release of constrictions in the vocal tract and are defined in terms of task dynamics. The movement of articulators produces the formation and release of constrictions. Two phonemes will contrast if they differ in gestural composition[8]. In articulatory phonology, one important aspect of task dynamics is that it is the motion of tract variables characterized dynamically. A tract variable characterizes a dimension of vocal tract constriction. Within some periods, some particular articulators are involved in specific gesture and are more critical than other articulators. Consequently, consistent movement trajectories will be observed for those more critical articulators within those periods. Evidences had been found in real articulatory data. Papcun *et al.* [2] found that the some specific articulatory channel of consonants in various vowel-consonant-vowel sequences illustrated more steady movement pattern than other articulatory channels. The articulatory channel which has steady movement patterns are called critical articulator while the others are called noncritical articulatory channel. Discriminating critical and noncritical articulatory channel is widely accepted in the fields of speech production, especially for articulatory trajectory modelling task [9].

Nevertheless, most of the studies of AAI adopted the average RMSE over all the articulatory channels to train regression models. They blurred the roles of critical and non-critical articulators, and ignored the natural distribution of data for critical/noncritical articulatory channels as well as its effects on the performance of AAI.

In this study, at first, each articulatory channel is classified as critical/noncritical according to its role in the formation of constrictions along the vocal tract when producing speech sound. Next, a loss function that deliberately weights samples is devised to account for the unbalance nature of the examples of critical/noncritical articulators when training a Bi-LSTM network. At last, the performance of the proposed method, in terms of the average RMSE over all channels, average RMSE of critical/noncritical channels, are presented.

2. Dataset

2.1. The MOCHA database

MOCHA database is adopted in this study. In MOCHA database, 460 British TIMIT sentences were uttered by two subjects, fsew0 and msak0. And four data streams were recorded: the waveform (16 kHz sampling rate, with 16-bit precision) together with laryngograph, electropalatograph, and ElectroMagnetic Articulograph (EMA) data. The waveform signal and articulatory information are synchronized and output to a computer simultaneously.

The EMA was used to retrieve the movement of articulators. For this purpose, coils were attached to the upper lip(UL), lower lip(LL), lower incisor(LI), tongue tip(TT), tongue body(TB), tongue dorsum(TD) and velum(V) to track their states when speech was producing. Each coil provided the x- and y-coordinates in the midsagittal plane. In total, 14 channels (V_x, V_y, TD_x, TD_y, TB_x, TB_y, TT_x, TT_y, LI_x, LI_y, LL_x, LL_y, UL_x, UL_y) of articulatory information were recorded. Two additional coils were attached to the nose bridge and the upper incisor to serve as the references. The movement of coils attached to the articulators in the midsagittal plane were sampled with the sampling rate of 500 Hz.

In addition, phoneme identities and corresponding segmental information are offered in the MOCHA dataset by force alignment, which make us able to determine whether a specific articulatory channel is critical or non-critical with in a particular period.

2.2. Data processing

Before feeding the synchronized acoustic-articulatory data to train and evaluate inversion models, some pre-processing procedure are necessary. Firstly, silences at the beginning and end of each speech utterance and corresponding EMA file are omitted, since articulators can take any status in those silent parts.

2.2.1. Acoustic feature

The acoustic signals are transformed to MFCC parameters (12 mel-cepstral+log energy), with the setting of 25ms Hamming window, 10ms frame shift, and 26 channels, by HTK. The acoustic feature vector for inversion at instant t consists of the MFCC coefficients of instant t and those of the preceding and following five consecutive acoustic frames. That is the MFCC coefficients of 11 consecutive acoustic frames ($x_{t-5}, \ldots, x_t, \ldots, x_{t+5}$) are used for constructing the inverse mapping from acoustic features to articulatory configuration.

2.2.2. Articulatory feature

The EMA data are bidirectionally filtered with a lowpass FIR filter, first forward then backward, to avoid the phase distortion. The filter used here is a 10-order finite impulse filter with the cutoff frequency of 20Hz. The slow varying mean of each utterance are obtained by smoothing the trajectory of 460 utterance's means with a Savitzky-Golay filter with the order of 5 and frame size of 121. The mean of each utterance is moved to the global mean by subtracting the difference between the smoothed mean of each utterance and the global mean. Finally, the EMA data is down-sampled to match the frame rate of acoustic feature.

2.2.3. Critical/non-critical articulator

Different articulator plays different role when producing specific speech sounds. For example, tongue tip is critical when producing speech sound /t/, while the tongue body and lips are allowed to take configurations with relatively large variations. Therefore, at the instant of producing sound /t/, tongue tip is the critical articulator and the others are noncritical articulators. In the case of EMA data, it is mainly the vertical movements of articulators cause the formation of constriction in vocal tract. As a result, the vertical direction of an articulator is thought to be critical in this study.

The MOCHA database provides the annotation of segmental information, where the phone identities as well as their start and end instants are offered by force alignment. The phoneme identities, corresponding IPA, and the corresponding critical articulatory channel are shown in Table 1 with reference to the criterions used by Papcun [2] and Okadome [9]. Based on the annotation result, for each articulatory channel at each instant, the critical articulatory channel is marked with 1, while the noncritical articulatory channel is marked with 0. For diphthongs and triphthongs, none of the channels are treated to be critical since the boundaries between the component vowels of diphthong/triphthong are unknown.

Table 1: *Phonemes and corresponding critical articulatory channel in the MOHCA database*.

Transcription	IPA	critical articulatory
@	[ə] in "was"	TD_y
@@	[ɜ] in "thirty"	TD_y
a	[æ] in "Nancy"	TD_y
aa	[ɑ] in "hard"	TD_y
b	[b] in "by"	LL_y
ch	[tʃ] in "chair"	TB_y
d	[d] in "hard"	TT_y
dh	[ð] in "this"	TT_y
e	[ə] in "offensive"	TB_y
f	[f] in "for"	LL_y
g	[g] in "grade	TD_y
h	[h] in "hard"	TD_y
i	[ɪ] in "this"	TB_y
ii	[iː] in "she"	TB_y
iy	[i] in "worry"	TB_y
jh	[dʒ] in "Jane"	TB_y
k	[k] in "work"	TD_y
l	[l] in "lily"	TB_y
m	[m] in "may"	LL_y
n	[n] in "Jane"	TT_y
ng	[ŋ] in "thing"	TD_y
o	[ɒ] in "on"	TD_y
oo	[ɔ] in "more"	TD_y, LL_y
p	[p] in "petrol"	LL_y
r	[r] in "bright"	TB_y
s	[s] in "this"	TT_y
sh	[ʃ] in "she"	TB_y
t	[t] in "thirty"	TT_y
u	[ʊ] in "woolen"	TD_y
uh	[ʌ] in "money"	TD_y
uu	[u] in "jewels"	TD_y, LL_y
v	[v] in "thieves"	LL_y
w	[w] in "work"	TD_y
y	[j] in "yell"	TB_y
z	[z] in "was"	TT_y
zh	[ʒ] in "pleasure"	TB_y

It is not difficult to figure out that an articulatory channel does not always play a critical role in the whole speech utterance since a speech utterance consists of different phonemes that usually have different critical articulatory channels. Therefore, in each articulatory channel, only parts of it are critical for the phonemes in the utterance. And they are called Critical Articulatory Portions (CAP), while the other parts are called Non-Critical Articulatory Portions (NCAP).

The data of subject fsew0 is used in this study. And the set of 460 utterance is divided into three separate subsets: a training

set with 370 utterances, a validation set with 45 utterances, and a testing set with another 45 utterances.

3. Bi-LSTM neural network

Bi-LSTM is a type of recurrent neural network (RNN), which incorporates LSTM units to overcome the limitation of the inability of learning long-range context dependencies and utilize both forward and backward context information of data stream. In Bi-LISTM, forward sequence $\vec{h}$ and backward sequence $\overleftarrow{h}$ are computed by iterating in the forward direction from instant 1 to T, and in the backward direction form T to 1, respectively. After that, the forward state $\vec{h}_t$ and backward state $\overleftarrow{h}_t$ are concatenated and fed to the next layer for further processing. Bi-LSTM networks have been successfully used in many tasks, such as speech synthesis[10], speech recognition [11]. It is well known that articulatory state correlate with the past as well as the future acoustic features. Zhu et al. [12] applied a Bi-LSTM neural network for the task of AAI and achieved the best performance on MNGU0 database. Hence, in this study, a Bi-LSTM is applied to the task of acoustic-to-articulatory inversion. The Bi-LSTM network used in this study contains four hidden layers, in which the inputs are connected to 2 feedforward layers with 'Relu' activation function, then fed to 2 BLSTM layers. And each hidden layer contains 300 neurons.

4. Loss functions

As shown in Table 1, for each phoneme, only one channel is regarded as the critical articulatory channel in most cases. As a consequence, the numbers of examples for critical and noncritical channel are unbalanced. The number of examples for noncritical articulatory channels is about 13 times of those for critical articulator channels.

In addition, the critical and noncritical articulatory channels are with the different importance in speech sound production. When producing a specific speech sound, the coordinates of critical articulatory channels should lie in a restricted area, while the coordinates noncritical articulatory channels are allowed to fall in a much larger region. To keep the phoneme identities unchanged from articulation point of view, the errors in critical articulatory channel should be much smaller than that of the noncritical articulatory channel.

In previous studies, average RMSE, $RMSE_{avg}$, is taken as the loss function for training DNN and Bi-LSTM. Eq.1 and Eq.2 are the routines for calculating the RMSE of each articulatory channel and the average RMSE over all the articulatory channels.

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

$$RMSE_{avg} = \frac{1}{N_c}\sum_{j=1}^{N_c} RMSE(j) \quad (2)$$

where $y_{j,i}$ is the true articulatory state of i^{th} sample of the j-th articulatory channel $\hat{y}_{j,i}$ is the corresponding estimate, N_s is the number of samples of the j^{th} articulatory channel, and N_c is the number of articulatory channels. One can see that the estimation errors of critical and noncritical articulatory channels are treated as of the same importance when training models. The error of critical articulatory channel is possible to be overwhelmed by that of the noncritical articulatory channels since the number of training examples of noncritical articulatory channels are much more than those of the critical articulatory channels.

To deal with these issues, we introduce a loss function as shown in Eq. 3-6.

$$RMSE_t(i) = \sqrt{\frac{1}{N_s}\sum_{t=1}^{N_s}\left(\hat{y}_{i,t} - y_{i,t}\right)^2 I(\mathrm{i}, \left(\hat{y}_{i,t} - y_{i,t}\right)^2)} \quad (3)$$

$$I(i,x) = \begin{cases} 0 & x < Th_i \ \ and\ m\ (i,t) = 0 \\ 1 & \text{else} \end{cases} \quad (4)$$

$$m\ (i,t) = \begin{cases} 0 & noncritical\ at\ instant\ t \\ 1 & critical\ at\ instant\ t \end{cases} \quad (5)$$

$$RMSE_{avg} = \frac{1}{N_c}\sum_{i=1}^{N_c} RMSE_t(i) \quad (6)$$

where $I(i,x)$ is an index function, and Th_i is the dynamic threshold of the i^{th} articulatory channel. In contrast to the traditional RMSE function, an index function is incorporated into the calculation of RMSE. The threshold, Th_i, in the index function is decided by the statistic information of the loss of the CAP of the i^{th} articulatory channel, if the channel acts as critical articulatory channel in some periods while acts as noncritical articulatory channel in other periods. Otherwise, Th_i is decided by the statistical information of loss of all the data of that channel. The purpose of introducing the index function is to deliberately discard the loss of some training example so as to alleviate the issues mentioned above. In this study, experiments are conducted when Th_i is the minimal, mean, and maximal error, respectively.

5. Results

Table 2 presents the RMSE per articulatory channel and the average RMSE over all the channels obtained by the models trained with the proposed loss function. Loss-G denotes the RMSE produced by the model trained with loss from all the training data. Loss-Min, Loss-mean, and Loss-Max denote the RMSEs obtained by the models trained by discarding the loss of NCAP that are less than the dynamically defined minimal, mean, and maximal loss of CAP of each articulatory channel. Loss-crt denotes the RMSE obtained by the model trained with the examples of CAP of each articulatory channel only.

Table 2: *The average RMSE obtained by different training paradigm (unit: mm).*

	Loss-G	*Loss-Min*	*Loss-Mean*	*Loss-Max*	*Loss-crt*
V_x	0.38	***0.38***	***0.33***	***0.36***	0.45
V_y	0.47	***0.45***	***0.41***	***0.45***	0.53
TD_x	1.84	***1.52***	***1.22***	***1.40***	2.10
TD_y	1.81	***1.63***	2.10	2.13	2.09
TB_x	1.89	***1.55***	***1.28***	***1.46***	2,27
TB_y	1.76	***1.67***	2.05	2.13	2.48
TT_x	1.92	***1.60***	***1.34***	***1.52***	2.41
TT_y	2.02	***1.85***	2.35	2.47	2.61
LI_x	0.86	***0.79***	***0.68***	***0.78***	0.98
LI_y	1.13	***1.01***	***0.91***	***0.98***	1.26
LL_x	1.24	***1.08***	***0.93***	***1.06***	1.39
LL_y	2.07	***1.81***	2.83	3.17	3.65
UL_x	0.89	***0.84***	***0.75***	***0.86***	0.94
UL_y	0.95	***0.88***	***0.76***	***0.88***	1.19
Avg.	1.37	***1.22***	***1.28***	1.56	1.73

If we take a look at the last row of Table 2, one can see that: i.) the model trained with Loss-Min achieves the best performance; ii.) the model trained with Loss-Mean is the second-best; iii.) the performance of these two models are better than that of the model trained with Loss-G; iv.) the performance of models trained with Loss-Max and Loss-crt are

worse than that of the model trained with Loss-G. This indicates that the performance of the model, in the sense of average RMSE over all the articulatory channels, could be improved by deliberately discarding part of the training examples of noncritical articulatory channels.

In addition, we take a look at the effects of training a model with the proposed method. It is found that: i.) the RMSEs of all the 14 articulatory channel decrease if the loss of the examples of NCAP less than the minimal loss of CAP are discarded; ii.) the RMSEs of 10 out of the 14 articulatory channels decrease if the loss of the examples of NCAP less than the mean loss of CAP are discarded, while the RMSE of channel TD_y, TB_y, TT_y, and LL_y increase; iii.) the RMSEs all of the 14 articulatory channels increase if the model is trained with loss of examples of CAP only.

Table 3: *The RMSE of CAP obtained by different training paradigm (unit: mm).*

	Loss-G	*Loss-Min*	*Loss-Mean*	*Loss-Max*	*Loss-crt*
TD_y	1.84	***1.38***	***1.05***	***0.95***	***1.04***
TB_y	1.53	***1.39***	***1.12***	***1.11***	***1.02***
TT_y	1.90	***1.42***	***1.14***	***1.05***	***1.14***
LL_y	1.48	***1.32***	***1.04***	***0.91***	***1.01***
Avg.	1.69	***1.38***	***1.09***	***1.01***	***1.05***

Table 4: *The RMSE of NCAP obtained by different training paradigm (unit: mm).*

	Loss-G	*Loss-Min*	*Loss-Mean*	*Loss-Max*	*Loss-crt*
V_x	0.37	0.38	***0.33***	***0.36***	0.45
V_y	0.47	***0.45***	***0.41***	***0.44***	0.53
TD_x	1.84	***1.53***	***1.23***	***1.40***	2.10
TD_y	1.81	***1.74***	2.41	2.44	2.38
TB_x	1.90	***1.56***	***1.28***	***1.47***	2.26
TB_y	1.79	***1.70***	2.14	2.22	2.60
TT_x	1.92	***1.60***	***1.35***	***1.52***	2.40
TT_y	2.07	***2.05***	2.82	3.00	3.15
LI_x	0.86	***0.78***	***0.68***	***0.78***	0.98
LI_y	1.13	***1.01***	***0.90***	***0.98***	1.26
LL_x	1.24	***1.08***	***0.94***	***1.05***	1.39
LL_y	2.14	***1.86***	2.98	3.34	3.85
UL_x	0.89	***0.84***	***0.75***	***0.86***	0.94
UL_y	0.95	***0.88***	***0.76***	***0.88***	1.20
Avg.	1.39	***1.24***	1.39	1.48	1.82

Table 3 and 4 present the corresponding RMSE of CAP and NCAP of the articulatory channels obtained by the models trained with the proposed loss function, respectively. Comparing the result in the 1st column and that of the 2nd column in Table 3 and 4, one can see that the RMSEs of both CAP and NCAP decrease when the model is trained by discarding the loss from NCAP whose training error is less than minimal loss of the CAP.

Comparing the result in the 1st column and that of the 3rd column in Table 3 and 4, one can notice that the RMSEs of CAP and NCAP in most of the articulatory channels decrease, except NCAP in channel TD_y, TB_y, TT_y, and LL_y. Similar results are found when comparing results in the 1st column and that of the 4th column in Table 3 and 4.

Comparing the result in the 1st column and that of the 5th column in Table 3 and 4, it is found that the RMSEs of NCAP of all the articulatory channels increase, while the RMSEs of CAP of all the articulatory channels decrease.

Then, we take a look on the result of channel TD_y, TB_y, TT_y, and LL_y, which sometimes act as noncritical articulatory channel while act as critical articulatory channels at other instants according to the phoneme it produced. It is found that: i.) the RMSEs of the NCAP in channel TD_y, TB_y, TT_y, and LL_y decrease at first, then increase drastically (shown in Table 4); ii.) the RMSEs of CAP decrease with the increase of the dynamic threshold (shown in Table 3). This indicates that discarding the loss of examples from NCAP dose reduce the prediction error of CAP, while takes the risk of increasing the prediction error of NCAP if the articulator channel acts as critical articulatory channels in some periods while acts as noncritical articulatory channels in other periods.

6. Conclusions

In this study, we attempt to incorporate the knowledge of speech production into acoustic-to-articulatory inversion. At each instant when producing speech sound, each articulatory channel is classified as critical/noncritical channel according to their roles in the formation of constrictions along the vocal tract when producing speech sound. It is found that training set is dominated by the examples of noncritical articulatory channels. To deal with the unbalanced dataset problem, several Bi-LSTM networks are trained by discarding the loss of examples from NCAP when the corresponding losses are less than the dynamic threshold of their corresponding CAP. The results indicate that the average RMSE over all the articulatory channels, the average RMSE over the CAP, and the average RMSE over the NCAP could be reduced by discarding the loss of examples from NCAP if they are less than the minimum error of the corresponding critical articulatory channels when training a Bi-LSTM network.

7. Acknowledgements

This work was supported by the National Natural Science-Foundation of China (No.61573254, No.61977049), Advanced Innovation Center for Language Resource and Intelligence (KYR17005), National Major Social Sciences Foundation of China (15ZDB103), Innovation Program of Chinese Academy of Social Science, and the programs of Beijing Municipal Administration of Hospital Clinical Medicine Development of Special Funding Support (Code. XMLX201848),

8. References

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

[2] 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.

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

[6] Uria, B., et al., Deep Architectures for Articulatory Inversion, in InterSpeech2012. 2012.

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

[8] Browman, C.P. and L. Goldstein, *Articulatory Phonology: An Overview*. Haskins Laboratories Status Report on Speech Research, 1992. **SR-111/112**: p. 23–42.

[9] 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.
[10] Fan, Y., et al., TTS Synthesis with Bidirectional LSTM basedRecurrent Neural Networks, in InterSpeech2014. 2014: Singapore. p. 1964–1968.
[11] Graves, A., N. Jaitly., and A. Mohamed, *Hybrid speech recognition with Deep Bidirectional LSTM*. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013.
[12] Zhu, P., L. Xie, and Y. Chen, Articulatory Movement Prediction Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks and Word/Phone Embeddings, in Interspeech2015. 2015. p. 2192–2196.

[This paper was published at Interspeech 2020]