The DKU-JNU-EMA Electromagnetic Articulography Database on Mandarin and Chinese Dialects with Tandem Feature based Acoustic-to-Articulatory Inversion

Zexin Cai[1], Xiaoyi Qin[2], Danwei Cai[1], Ming Li[1], Xinzhong Liu[3], Haibin Zhong[4]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China
[2]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China
[3]Chinese Language and Literature Department, Jinan University, Guangzhou, China
[4]Jiangsu Jinling Science and Technology Group Limited

ming.li369@dukekunshan.edu.cn, ozwgb@jnu.edu.cn

Abstract

This paper presents the acquisition of the Duke Kunshan University Jinan University Electromagnetic Articulography (DKU-JNU-EMA) database in terms of aligned acoustics and articulatory data on Mandarin and Chinese dialects. This database currently includes data from multiple individuals in Mandarin and three Chinese dialects, namely Cantonese, Hakka, Teochew. There are 2-7 native speakers for each language or dialect. Acoustic data is obtained by one head-mounted close talk microphone while articulatory data is obtained by the NDI electromagnetic articulography wave research system. The DKU-JNU-EMA database is now in preparation for public release to help advance research in areas of acoustic-to-articulatory inversion, speech production, dialect recognition, and experimental phonetics. Along with the database, we propose an acoustic-to-articulatory inversion baseline using deep neural networks. Moreover, we show that by concatenating the dimension reduced phoneme posterior probability feature with MFCC features at the feature level as tandem feature, the inversion system performance is enhanced.

Index Terms: deep neural network, electromagnetic articulography, acoustic-to-articulatory inversion, phoneme posterior probability, tandem feature

1. Introduction

Speech signals, produced by human vocal tract and speech production organs, carry various types of information such as lexical contents and paralinguistic characteristics, e.g. language, speaker, emotion, age, gender, etc. With the progress and refinement in articulography [1, 2], three-dimensional (3D) modeling of the human speech production system has become available. Such technique provides accurately aligned acoustic signals and articulatory trajectories, which can be useful for research works in a variety of areas, such as speech production [3], speech recognition [4], speaker recognition [5], emotion recognition [6], speech synthesis [7, 8], etc.

The articulography speech research system helps capture the configurations of the articulators, that are, the locations and the movements of lips, tongue, velum, etc. The speech signals are also recorded simultaneously. These collected data is known as the electromagnetic articulography (EMA) data. Several well-known free and publicly available corpora for EMA data are mentioned in [9]. The publicly available EMA corpus MNGU0 [10] contains 1354 utterances from a single native British English speaker, while another public corpus [11] is a multi-channel/multi-speaker English database. However, the resources on Mandarin and Chinese dialects are very limited. In order to provide more EMA resources on Mandarin and Chinese dialects, we collected the multi-speaker/multi-dialect Duke Kunshan University Jinan University Electromagnetic Articulography (DKU-JNU-EMA) database, which is under final preparation for public release. Furthermore, the database may potentially promote the acoustic-articulatory researches on Asian languages since it contains Mandarin and three different Chinese dialects, and there are 2-7 native speakers for each language or dialect. The database was collected in Jinan University, China. Basically, we use the NDI wave research system to record the movements of the articulators in the midsagittal plane. The DKU-JNU-EMA database consists of over 3000 utterances in four different kinds of reading materials. The reading materials cover all phonemes in Mandarin, Hakka, Teochew and Cantonese. We believe that the DKU-JNU-EMA database can help advance research in areas like speech production [3], acoustic-to-articulatory inversion [12, 13, 14], dialect recognition, and experimental phonetics.

Along with the DKU-JNU-EMA database, this paper implements an acoustic-to-articulatory inversion baseline with one subject's data from the database. Acoustic-to-articulatory inversion is a technique that determines the articulatory trajectories from speech signals and has made substantial progress recently by adopting deep neural network for the inversion [13]. Following the methods in [14] , we employ the deep neural networks (DNN) structure to model and predict the tract variable trajectories. Conventionally, the network input is acoustic features extracted from speech signals, and output is the synchronized articulatory coordinates. In this study, motivated by the tandem feature concept in [10] , we first use a DNN based ASR acoustic model, trained on the publicly available HKUST database [15] , to obtain the phoneme posterior probabilities (PPP). Then we apply principal component analysis (PCA) on PPP features for dimension reduction and concatenate with MFCC together as a kind of tandem features for the subsequent modeling. We show that the phonetic level information introduced by the PPP features can enhance the inversion performance.

This paper is organized as follows. In Section 2, we will describe the details of the DKU-JNU-EMA database. Section 3 presents acoustic-to-articulatory inversion baseline. The experimental results are discussed in Section 4 while conclusions are provided in Section 5.

978-1-5386-5627-3/18/$31.00 ©2018 IEEE

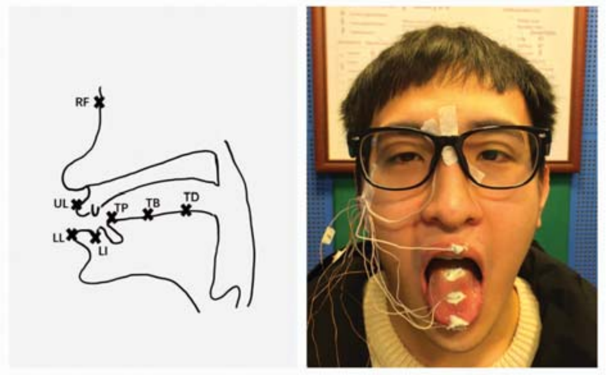

Figure 1: *Left: Position of sensors in the DKU-JNU-EMA database, Right: The EMA recording setup*

2. DKU-JNU-EMA database

In this section, we describe the new electromagnetic articulography database containing Mandarin, Cantonese, Hakka and Teochew languages produced by multiple speakers. The DKU-JNU-EMA database contains about 10.66 hours of recording.

2.1. Data collection setup

We use the NDI electromagnetic articulography speech research system to capture the real-time tract variable trajectories, as shown in the figure 1 and table 1. Subjects were asked to place six sensors in mouth and one at the bridge of nose as a reference point. Table 1 shows the locations of six sensors, that are, upper lip, lower lip, lower incisor, tongue tip, tongue body, and tongue dorsum [9] . In addition, subjects wear a head-mounted close talk microphone to record the speech signal simultaneously. We also used a palate probe to perform the palate tracing. The morphological shapes of hard palate could be useful in speech production and speaker recognition.

Table 1: *Location of the sensors*

Location	Label
Upper lip	UL
Tongue tip	TP
Lower lip	LL
Tongue body	TB
Lower incisor	LI
Tongue dorsum	TD
Nose bridge	RF

2.2. Database composition

The DKU-JNU-EMA database includes data from Mandarin and three different Chinese dialects. For each language or dialect, subjects were required to record 4 sessions of utterances as follows:

- Sentence session: subjects read complete sentences or short texts.
- Consonant session: for each given consonant, subjects read related words composed by the specific consonant.
- Vowel session: for each given vowel, subjects read related words composed by the specific vowel.
- Tone session: for each given word, subjects read words with every tone of that language or dialect.

Each language and dialect has a reference alphabet, and the phonetically balanced texts and sentences selected for recording. The reading materials can be found in the database. Unfortunately, there are no utterances in Consonant session, Vowel session and Tone session for Mandarin due to the lack of Mandarin alphabet reading material. Similarly, the lack of phonetically balanced sentences for Hakka and Teochew leads to two empty Sentence sessions. For each language or dialect, each utterance was recorded once by every subject. However, several unqualified recordings were disregarded.

As shown in table 2 , The Mandarin database contains 7 subjects (4 male, 3 female), and up to 2100 utterances in the Sentence session. Cantonese database has 5 subjects (2 male, 3 female), 113 Sentence utterances, 55 Tone utterances, 119 Consonant utterances, 361 Vowel utterances. Hakka database has 2 male subjects, recording 12 Tone utterances, 34 Consonant utterances and 146 Vowel utterances. Teochew database has 1 male subject and 1 female subject recorded 20 Tone utterances, 46 Consonant utterances and 195 Vowel utterances. In addition, each subject has recorded hard palate trace shapes multiple times.

Table 2: *The data composition of the DKU-JNU-EMA database*

	Mandarin	Cantonese	Hakka	Teochew
male:female	4:3	2:3	2:0	1:1
sentences	2100	113	\	\
tone	\	55	12	20
consonant	\	119	34	46
vowel	\	361	146	195
palate trace	68	41	18	18

Each recorded utterance has both articulatory and acoustic data. The acoustic signal is recorded by a head-mounted MEMS microphone at 22kHz sample rate. In the articulatory domain, NDI EMA speech research system captures 5D data of each sensor with a sample frequency of 100 Hz. The 5D data is composed of quaternion rotation parameters and X, Y, Z coordinates in the reference three dimensional space.

3. Acoustic-to-articulatory inversion methods

In this section, we introduce our deep neural network and tandem feature based acoustic-to-articulatory inversion deep neural network baseline.

3.1. Data pre-processing

Each speech utterance is downsampled from 22kHz to 16kHz. Then we apply an energy based voice active detection (VAD) module to remove the silence parts in the audio. However, to preserve phoneme contexts, we keep 50ms silence before and after each speech segment.

The raw EMA data of seven sensors, one for reference, provides six individual measurements in 5D layout. Typically, the lips and tongue have little movements in the Z axis(left and right). We select the X and Y coordinates (back/front and up/down respectively) in the midsagittal plane for further analysis. Consequently, 12 dimensional coordinate vector is derived from 6 sensors. We then, normalize the 12 dimensional coordinate vector by subtracting the corresponding global mean from each dimension, and then dividing by each dimension's global standard deviation. Like other electromagnetic articulograph speech research system [16], there are mis-tracking points in the collected EMA data due to some anomalies in our device's performance. Typically, the mis-tracking points show as NAN in the EMA data. However, the average mis-tracking rate is 0.85%, which is low. Hence we can simply use the interpolated values for the missing point.

3.2. Mel Frequency Cepstral Coefficient(MFCC)

This study uses MFCC as the basic acoustic feature in our acoustic-to-articulatory inversion system. First, we extract 13 dimensional (13D) MFCC features for each utterance, with cepstral mean subtraction and variance normalization. Since the tract variable trajectories are continues, context information is beneficial for the inversion system. Therefore, a second order delta features are appended to the 13D MFCC features to obtain the 39 dimensional MFCC features [17].

3.3. Tandem acoustic feature

Here we propose to employ tandem feature [7] in our DNN acoustic-to-articulatory inversion system. Figure 2 shows the main procedure in generating tandem features. In speech recognition, MFCC features from the entire context window are fed into a DNN acoustic model to generate the phoneme posterior probabilities (PPP). Actually, the PPP is a vector of probabilities, each element on each senone triphone tied state. In automatic speech recognition, conventionally, the PPP would go directly to an Hidden Markov Model(HMM) decoder to find the word sequence, but instead, we use them as the phonetic features. We believe this phonetic level feature could potentially enhance the inversion system performance since it provides additional information from the phonetic point of view.

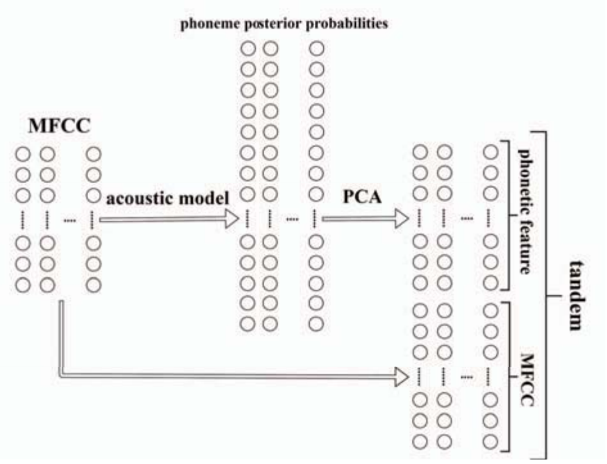

Figure 2: *The procedure of extracting tandem feature*

The dimension of the PPP feature vector is the same as the size of the output layer in the speech recognition acoustic model, which is 6500 in our experiments. Since the dimensionality of PPP is too high for the acoustic-to-articulatory inversion system, principal component analysis (PCA) is applied on PPP. Then the dimension reduced PPP is concatenated with the MFCC features together as a kind of tandem feature. We believe the tandem feature including both the acoustic and phonetic characteristic is able to improve the inversion performance.

In this paper, the Mandarin acoustic model for generating PPP is trained from the HKUST database([1]). HKUST Mandarin Telephone Transcript Data contains 200 hours of Mandarin Chinese conversational telephone speech from Mandarin speakers in mainland China.

3.4. Deep neural networks (DNN) setup

The proposed system adopts a 4 layers DNN setup, with 300 nodes for each layer. We use RELU as the activation function and Adam optimizer [18] for stochastic optimization.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}(e_i - t_i)^2} \tag{1}$$

Root mean-squared error (RMSE) is widely used for measuring the performance of acoustic-to-articulatory inversion systems. Hence, we also use it as the loss function in the training procedure. RMSE is defined as equation 1 , where e_i is the predicted tract variable and t_i is the groundtruth tract variable.

4. Experimental results

We choose one subset named LY from the Mandarin database to demonstrate the system performance. We use the first 260 utterances for training, and the remaining 40 utterances for testing. 10ms frame shift is adopted in extracting MFCC features which matches with the 100Hz sample rate of the EMA system. In this way, we are able to align the tandem features with the articulatory trajectories.

We employ an Mandarin acoustic model learnt from the HKUST database to derive the phoneme posterior probability(PPP) vector. Then PCA is applied to the PPP features to reduce its dimensionality to 40. In this work, we compares three different features listed below for acoustic-to-articulatory inversion.

- MFCC with a context window of 11 frames
- line spectral frequency (LSF) [19] with a context window of 11 frames
- Tandem feature with a context window of 11 frames. Since the PPP already has contexts information, the concatenate feature consist of 5 frames of MFCC before current frame, tandem feature (current frame), and 5 frames of MFCC after.

As shown in the following equation, we also use the average correlation to evaluate the system performance.

$$r = \frac{1}{N}\sum_{i} corrcoef\,(e_i, t_i) \tag{2}$$

N denotes the dimensionality of the EMA groundtruth data, which is 12 in this paper. e_i is the predict trace of dimension i and t_i is the actual measured trace.

In order to make sure it's not accidental that the tandem feature system outperforms MFCC and LSF, the DNN inversion system is trained and tested 10 times with random initial parameters for each kind of input features. The results on different features are shown in the table 3.

[1] https://catalog.ldc.upenn.edu/LDC2005S15
https://catalog.ldc.upenn.edu/LDC2005T32

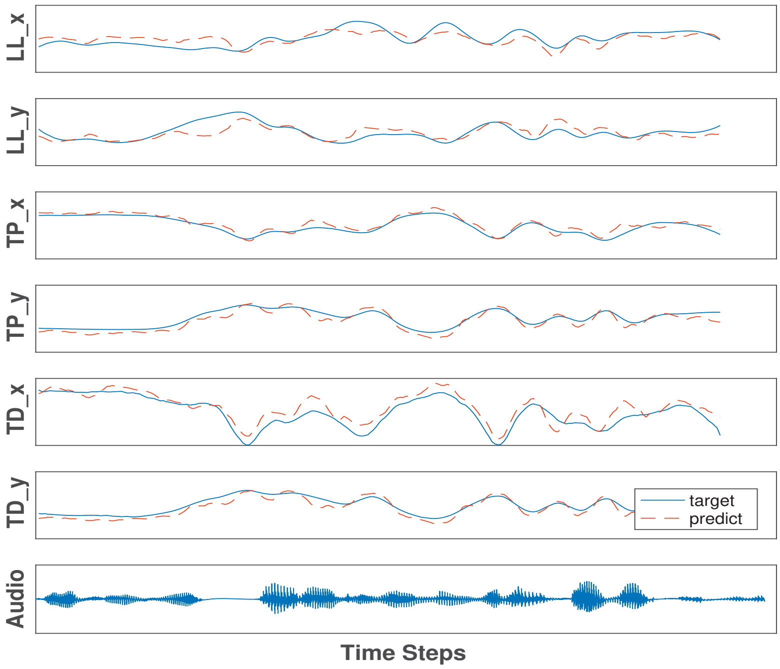

Figure 3: *Predicted and groundtruth trajectories of the proposed inversion system on LL, TP and TD sensors.*

Table 3: *RMSE and correlation with three different features*

Feature	r	RMSE		
		MEAN	MAX	MIN
MFCC	0.78551	0.54698	0.5522	0.5404
LSF	0.78435	0.548024	0.5559	0.54464
Tandem	0.79518	0.53441	0.5386	0.5301

From table 3, we can observe that tandem feature outperforms MFCC and LSF in terms of both RMSE and average correlation. Hence, the introduction of phonetic level tandem feature enhances the inversion performance.

Figure 3 shows the predicted and groundtruth trajectories of 3 articulators (lower lip, tongue tip, lower incisor). The predicted tract variable trajectories are well aligned with the real measured ones in our acoustic-to-articulatory inversion system.

5. Conclusions

This paper presents a new Electromagnetic Articulography database called DKU-JNU-EMA. Different from the existing publicly available EMA database [9, 2, 20] , the proposed DKU-JNU-EMA database focus on Mandarin and three widely used Chinese dialects, each has data from multiple speakers. Releasing DKU-JNU-EMA database as a public and free resource to the community could potentially benefit the research works in related areas.

We also provide an acoustic-to-articulatory inversion system baseline with 300 utterances from a single speaker in the DKU-JNU-EMA database. Experimental results show that by concatenating the phonetic level PPP feature with the acoustic level MFCC feature together as the inputs, the inversion system performance is enhanced in every experiment. However, the numerical differences are small among three different inversion systems. It is likely that the limited training data restricts the performance of tandem feature system.

6. Acknowledgements

This research was funded in part by the National Natural Science Foundation of China (61773413), the National Social Science Fund of China (14BYY038), the Natural Science Foundation of Guangzhou City (201707010363), the Science and Technology Development Foundation of Guangdong Province (2017B090901045), the National Key Research and Development Program (2016YFC0103905).

7. References

[1] P. Badin and A. Serrurier, “Three-dimensional modeling of speech organs : Articulatory data and models,” *Ieice Technical Report*, vol. 106, no. 177, pp. 29–34, 2006.

[2] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, “A multimodal real-time mri articulatory corpus for speech research,” in *INTERSPEECH*, 2011, pp. 837–840.

[3] S. Hiroya and M. Honda, “Estimation of articulatory movements from speech acoustics using an hmm-based speech production model,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[4] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, and A. Bezmaman, “Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 jhu summer workshop,” in

IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, pp. IV–621 – IV–624.

[5] M. Li, J. Kim, A. Lammert, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," *Computer speech & language*, vol. 36, pp. 196–211, 2016.

[6] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *INTERSPEECH*, 2005, pp. 497–500.

[7] Z. Ling, 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, no. 6, pp. 1171–1185, 2009.

[8] A. Toutios and S. Narayanan, "Articulatory synthesis of french connected speech from ema data," pp. 2738–2742, 2013.

[9] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus." *Interspeech*, pp. 1505–1508, 2011.

[10] M. Li, L. Liu, W. Cai, and W. Liu, "Generalized i-vector representation with phonetic tokenizations and tandem features for both text independent and text dependent speaker verification," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 207–215, 2016.

[11] A. A. Wrench and W. J. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Seminar on Speech Production, Kloster Seeon*, 2000, pp. 131–132.

[12] S. Hiroya and M. Honda, "Speaker adaptation method for acoustic-to-articulatory inversion using an hmm-based speech production model," *Ieice Transactions on Information & Systems*, vol. 87, no. 5, pp. págs. 1071–1078, 2004.

[13] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *INTERSPEECH*, 2006, pp. 577–580.

[14] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4450–4454.

[15] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "Hkust/mts: a very large scale mandarin telephone speech corpus," in *International Conference on Chinese Spoken Language Processing*, 2006, pp. 724–735.

[16] J. S. Boyle, D. Williamson, R. Cederwall, M. Fiorino, J. Hnilo, J. Olson, T. Phillips, G. Potter, and S. Xie, "Numerical instabilities and three-dimensional electromagnetic articulography," *Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3941–9, 2012.

[17] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *Ttps*, vol. 2, 2010.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[19] M. Ghorbandoost, A. Sayadiyan, M. Ahangar, H. Sheikhzadeh, A. S. Shahrebabaki, and J. Amini, "Voice conversion based on feature combination with limited training data," *Speech Communication*, vol. 67, no. 67, pp. 113–128, 2015.

[20] A. Wrench, "The mocha-timit articulatory database," 1999, http://www/cstr.ed.ac.uk/artic/mocha.html.