INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Combining language corpora in a Japanese electromagnetic articulography database for acoustic-to-articulatory inversion

Tianfang Yan[1], Kikuo Maekawa[2], Yukiko Nota[2], Masayuki Hirata[134]

[1]Department of Neurological Diagnosis and Restoration, Osaka University Graduate School of Medicine, Suita, Osaka, Japan
[2]National Institute for Japanese Language and Linguistics, Japan
[3]Department of Neurosurgery, Osaka University Graduate School of Medicine, Suita, Osaka, Japan
[4]Global Center for Medical Engineering and Informatics, Osaka University, Suita, Osaka, Japan
tianfang_yan@ndr.med.osaka-u.ac.jp

Abstract

This paper presents an electromagnetic articulography database of Japanese sentences. The database includes aligned acoustics and articulatory data from seven males and three females, with a total of five recorded hours. The database is now in preparation for public release to foster research in areas of acoustic-to-articulatory inversion, brain-machine interface communication systems, artificial speech synthesis, and dialect recognition. Moreover, based on this database we established an acoustic-to-articulatory inversion system using a deep, bidirectional, long short-term memory recurrent neural network structure. The results showed that, for the Japanese language, adding English corpora to the training was not beneficial for this speaker-independent model.

Index Terms: electromagnetic articulography, acoustic-to-articulatory inversion, bi-LSTM

1. Introduction

Brain machine interfaces (BMI), which act as communicative bridges between severely paralyzed patients and external robot-assisted equipment, have attracted increasing attention from academic researchers and the public for their recent encouraging brain science advancements. One of the most amazing achievements is a speech neuroprosthetic technology by Anumanchipalli et al.[1], which utilized a two-step method to first decode cortical signals into representations of articulatory movements, and then to transform the representations into audible speech. However, one challenge to this work is how to track intermediate articulatory representations between the neural and acoustic signals in clinical settings, because it is difficult for patients to record articulatory trajectories with electrocorticography (ECoG) equipment. One strategy is to use recent advances in acoustic-to-articulatory inversion (AAI) to estimate reliable natural vocal kinematic trajectories from audio recordings alone[2, 3].

AAI is a technique that infers articulatory kinematic trajectories (AKTs) from speech signals only[4]. The inferred AKTs have contributed to the improvement of text-to-speech synthesis[5, 6], speech recognition[2, 7], and automatic detection of speech production deficits in Parkinson's disease[8]. Speaker-independent articulatory reconstruction is also essential for BMI applications. Following the method of Liu et al.[9], we employed a deep, bidirectional, long short-term memory (bi-LSTM) recurrent neural network structure to model and predict speaker-independent articulatory trajectories. Generally, the network transforms the input of acoustic features to the output of synchronized articulatory coordinates.

Articulatory corpora are inevitable for training such a model. Electromagnetic articulography (EMA) is the most widely used point-tracking-based technique for the study of speech production[10]. EMA data help capture the configurations of the continuous movements and the locations of the lips, tongue and jaw, with high spatial resolution. Such a technique provides accurate aligned acoustic signals, and inside the vocal AKTs use sensors attached to the tongue, lips, and jaw[11, 12].

Some researchers have established their publicly available EMA databases, such as the USC-TIMIT[13], the MOCHA-TIMIT[14], the TORGO[15], the EMA-MAE[16], the mngu0[17], the Haskins[18] databases in English, the MSPKA[19] in Italian, the DKU-JNU-EMA[20] in Chinese, the Mandarin-Tibetan speech corpus[21] in Tibetan, and the database of Norwegian speech sounds[22] in Norwegian. However, there is no such EMA corpora dataset in the Japanese language.

In this study, we establish an EMA database in the Japanese language for the first time. The data were obtained from seven males and three females, for a total of five recorded hours. Then, we used this database to train a bi-LSTM-based AAI model. We believe this database may play an important initial role in AAI for the Japanese language, as well as further improving the BMI communication system, artificial speech synthesis, and dialect recognition.

2. EMA database

2.1. Data collection

We named this the OU-EMA database (OU is short for Osaka University). We used the NDI WAVE electromagnetic articulography system[23] to track real-time inside vocal trajectories. Participants were asked to attach seven recording sensors inside their mouths, with one reference electrode placed at the bridge of the nose. Table 1 and Figure 1 show the sensor locations: tongue tip (TT), tongue body (TB), tongue dorsum (TD), upper lip (UL), lower lip (LL), and lower incisor (LI). In addition, a microphone (Marantz Professional MPM-1000) was used to record the speech signal.

10.21437/Interspeech.2023-36

Table 1: *Sensor locations.*

Location	Label	Location	Label
Tongue dorsum	TD	Lower incisor	LI
Tongue body	TB	Upper lip	UL
Tongue tip	TT	Lower lip	LL
Upper incisor	UI	Nose root	RF

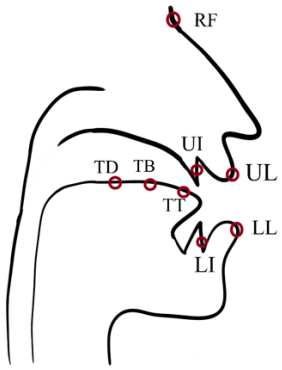

Figure 1: *Sensor locations.*

2.2. Data composition

This database included up to 4050 sentence utterances (303 min in total) from 10 subjects (7 male, 3 female). Participants were asked to read complete sentences from the ATR503[24], a set of 503 short sentences designed to include the main connected speech processes in the Japanese language. The acoustic signal was recorded by a microphone at 22050 Hz sample rate. In the articulatory signal, the NDI WAVE system monitored X, Y, Z coordinate data for each sensor at a sampling frequency of 100 Hz. Each recorded utterance included both articulatory and acoustic data. We checked the recording quality for every sentence and excluded defective ones. For the mis-tracked points in the data, we used the function 'fillmissing' in MATLAB 2020b [25] to fill the interpolated values. The estimated mis-tracked rate is about 0.35% of this dataset. The reading materials were integrated into the database. The forced alignment for each speech audio file-transcription pair was performed using the Julius[26], an open-source speech recognition engine for Japanese that captures the beginning and ending times of each phoneme.

3. Acoustic-to-articulatory inversion

3.1. Data preprocessing

We used midsagittal plane data (X as back/front and Y as the up/down coordinates of the EMA signals) from six sensors for training. The 'UI' sensor was excluded from the analysis because the sensor on the upper incisor was not moved relative to the reference sensor. Consequently, 12 dimensional coordinate vectors were derived. We smoothed each articulatory trajectory using a low-pass filter with a cutoff frequency of 20 Hz. Then, the articulatory vectors were normalized by subtracting the mean over 60 previous and subsequent recordings from each speaker, and then dividing by the speaker-specific standard deviation.

The acoustic waves were downsampled from 20050Hz to 16000Hz. We used the first thirteen dimensional mel-frequency cepstral coefficient (MFCCs) features for each utterance. The offset silences were removed based on the transcript labels. The thirteen MFCCs were normalized for each speaker and were used as input, with a window size of 25 ms and stride of 10 ms. Then, second-order delta features were added to obtain the 39 dimensional MFCC features[27]. We also added 10 context windows: the five previous and five subsequent frames, as in [28].

3.2. Training

We used a bidirectional recurrent neural network architecture similar to Liu's[9], with a convolutional layer as a low-pass filter. We used the root mean-squared error (RMSE) as the loss function in the training procedure for measuring the performance of AAI systems. We used the Adam optimizer with early stopping on the validation set (learning rate 0.001, batch size 10, patience 5). The weights of the low-pass filter were fixed with N = 50 to give a transition band of 0.08. The convolution had one channel, a stride of 1, and padding such that the output had the same size as the input. All training processes consisted of the following steps: (1) validate on a subset of the speakers (10%), (2) test on one speaker, and (3) train on the rest (90%).

3.3. Test results

In order to test the quality of the OU-EMA database in training the AAI model, we compared the articulatory reconstruction results with an already-established EMA database, the Haskins[18]. We used a speaker-independent setting: nine speakers' data for training and one speaker's data for testing. The articulatory reconstruction was evaluated by the RMSE and Pearson's correlation (r). Table 2 compares the results of the OU-EMA database with the Haskins in the same AAI system. Figure 2 shows the results of the predicted and target articulatory trajectories of the AAI system for all recording sensors. The inferred kinematic trajectories were well aligned with the real measured ones.

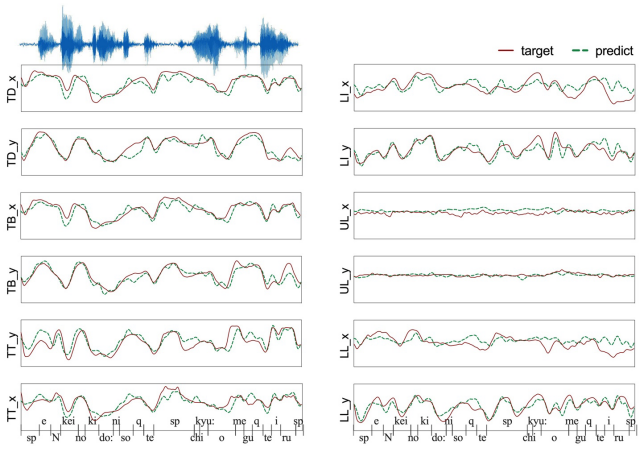

Figure 2: *Predicted and target articulatory trajectories of the AAI system. The sentence is 'En kei no ki do: ni so'!te chi kyu: o me gu'!te i ru'. (sp: silent period) Translation of the sentence: revolving around the earth in a circular orbit.*

Table 2: *Articulatory reconstruction results. (*n=normalized. Smaller is better for RMSE.)*

Database	RMSE	n-RMSE	r
OU-EMA	2.961	0.724	0.704
Haskins	2.958	0.625	0.752

4. Merging languages corpora test

To answer the question: Do we have to set up a new database in the Japanese language, or is it feasible to invert one based on another language, we compared the results of single, cross, and merging corpora. For the single corpus, the training and testing data were from the same database. For the cross corpus, we trained in one language and tested the model in the other language. For the merging corpus, the model was trained on speakers from both languages.

As shown in Table 3, the reconstruction results for the cross corpus were worse than those for the single-corpus condition, showing that it is important to establish a new database in a specific language. In addition, the results of the merging corpus were not better than those of the single corpus, suggesting that adding corpora from another language to the training is not beneficial to the speaker-independent model. This result is similar to Parrot et al.[29], who reported that reconstruction measures were not improved with the addition of different corpora. This indicates that although the AAI system is language-independent, the differences in pronunciation patterns that exist in different languages lead to potentially worse articulatory reconstructions when they are used together.

Table 3: *Articulatory reconstruction results.*

	Training	Testing	RMSE	n-RMSE	r
Single-corpora	OU-EMA	OU-EMA	2.961	0.724	0.704
	Haskins	Haskins	2.958	0.625	0.752
Cross-corpora	OU-EMA	Haskins	3.741	0.790	0.568
	Haskins	OU-EMA	3.659	0.895	0.492
Merging-corpora	Both	OU-EMA	3.180	0.778	0.667
		Haskins	2.899	0.612	0.755

5. Conclusions

This paper introduces the OU-EMA database, a new electromagnetic articulography database in Japanese. This database could potentially benefit research in BMI and related areas. We also established an acoustic-to-articulatory inversion system based on this OU-EMA database. The results showed a relatively good performance by the AAI system in the estimation of articulatory trajectories from only an acoustic signal. Moreover, we suggest it is valuable to establish a new EMA database in a specific language.

6. Ethics statement

The Institutional Review Boards at both the National Institute for Japanese Language and Linguistics (NINJAL) and Osaka University Hospital (22059) approved this study for human experiments.

7. Acknowledgements

The work reported in this article was supported by the National Institute for Japanese Language and Linguistics, Joint Resource-use Projects (B): Electromagnetic Articulography Based Acoustic-to-Articulatory Inversion Research; JST SPRING (Grant Number JPMJSP2138); IPBS Grant-in-Aid for Education and Research 2022. We thank Mark Tiede for sharing his MATLAB tools: MVIEW and Takeru Kuratomi for providing the microphone.

8. References

1. Anumanchipalli, G.K., J. Chartier, and E.F. Chang, *Speech synthesis from neural decoding of spoken sentences.* Nature, 2019. **568**(7753): p. 493-498.
2. Mitra, V., et al., *Joint Modeling of Articulatory and Acoustic Spaces for Continuous Speech Recognition Tasks.* 2017 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp), 2017: p. 5205-5209.
3. Afshan, A. and P.K. Ghosh, *Improved subject-independent acoustic-to-articulatory inversion.* Speech Communication, 2015. **66**: p. 1-16.
4. Richmond, K., *A Trajectory Mixture Density Network for the Acoustic-Articulatory Inversion Mapping.* Interspeech 2006 and 9th International Conference on Spoken Language Processing, Vols 1-5, 2006: p. 577-580.
5. Cao, B.M., et al., *Integrating Articulatory Information in Deep Learning-based Text-to-Speech Synthesis.* 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), Vols 1-6, 2017: p. 254-258.
6. Bocquelet, F., et al., *Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces.* Plos Computational Biology, 2016. **12**(11).
7. Mitra, V., et al., *Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition.* Speech Communication, 2017. **89**: p. 103-112.
8. Hahm, S. and J. Wang, *Parkinson's Condition Estimation using Speech Acoustic and Inversely Mapped Articulatory Data.* 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Vols 1-5, 2015: p. 513-517.
9. Liu, P., et al., *A Deep Recurrent Approach for Acoustic-to-Articulatory Inversion.* 2015 Ieee International Conference on Acoustics, Speech, and Signal Processing (Icassp), 2015: p. 4450-4454.
10. Rebernik, T., et al., *JOURNAL A review of data collection practices using electromagnetic articulography.* Laboratory Phonology, 2021. **12**(1).
11. Richmond, K., Z. Ling, and J. Yamagishi, *The use of articulatory movement data in speech synthesis applications: An overview — Application of articulatory movements using machine learning algorithms &#8212.* Acoustical Science and Technology, 2015. **36**(6): p. 467-477.
12. Badin, P., et al., *Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding.* Speech Communication, 2010. **52**(6): p. 493-503.
13. Narayanan, S., et al., *Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC).* Journal of the

Acoustical Society of America, 2014. **136**(3): p. 1307-1311.
14. Wrench, A., *MOCHA: multichannel articulatory database.* 1999.
15. Rudzicz, F., A.K. Namasivayam, and T. Wolff, *The TORGO database of acoustic and articulatory speech from speakers with dysarthria.* Language Resources and Evaluation, 2012. **46**(4): p. 523-541.
16. Ji, A., J.J. Berry, and M.T. Johnson, *The Electromagnetic Articulography Mandarin Accented English (Ema-Mae) Corpus of Acoustic and 3d Articulatory Kinematic Data.* 2014 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp), 2014.
17. Richmond, K., P. Hoole, and S. King, *Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus.* 12th Annual Conference of the International Speech Communication Association 2011 (Interspeech 2011), Vols 1-5, 2011: p. 1516-+.
18. Tiede, M., et al., *Quantifying kinematic aspects of reduction in a contrasting rate production task.* The Journal of the Acoustical Society of America, 2017. **141**(5): p. 3580-3580.
19. Canevari, C., L. Badino, and L. Fadiga, *A new Italian dataset of parallel acoustic and articulatory data.* 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Vols 1-5, 2015: p. 2152-2156.
20. Cai, Z.X., et al., *The DKU-JNU-EMA Electromagnetic Articulography Database on Mandarin and Chinese Dialects with Tandem Feature based Acoustic-to-Articulatory Inversion.* 2018 11th International Symposium on Chinese Spoken Language Processing (Iscslp), 2018: p. 235-239.
21. Lobsang, G., et al., *Tibetan Vowel Analysis with a Multi-Modal Mandarin-Tibetan Speech Corpus.* 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (Apsipa), 2016.
22. Moen, I., H. Gram Simonsen, and A.M. Lindstad, *An electronic database of Norwegian speech sounds: clinical aspects.* Journal of Multilingual Communication Disorders, 2004. **2**(1): p. 43-49.
23. Berry, J.J., *Accuracy of the NDI Wave Speech Research System.* Journal of Speech Language and Hearing Research, 2011. **54**(5): p. 1295-1301.
24. Kurematsu, A., et al., *Atr Japanese Speech Database as a Tool of Speech Recognition and Synthesis.* Speech Communication, 1990. **9**(4): p. 357-363.
25. *The MathWorks, Inc. (2022) MATLAB version: 9.13.0 (R2022b).*
26. Lee, A. and T. Kawahara, *Julius v4.5.* 2019.
27. Muda, L., M. Begam, and I. Elamvazuthi, *Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques.* Journal of Computing, 2010. **Volume 2**(Issue 3).
28. Uria, B., et al., *Deep Architectures for Articulatory Inversion.* 13th Annual Conference of the International Speech Communication Association 2012 (Interspeech 2012), Vols 1-3, 2012: p. 866-869.
29. Parrot, M., J. Millet, and E. Dunbar, *Independent and automatic evaluation of speaker-independent acoustic-to-articulatory reconstruction*, in *Interspeech 2020-21st Annual Conference of the International Speech Communication Association*. 2020.