INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

End-to-End Articulatory Attribute Modeling for Low-resource Multilingual Speech Recognition

Sheng Li[1]*, *Chenchen Ding*[1]*, *Xugang Lu*[1], *Peng Shen*[1], *Tatsuya Kawahara*[1,2]*, *Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Kyoto, Japan
[2]Kyoto University, Kyoto, Japan

sheng.li@nict.go.jp

Abstract

The end-to-end (E2E) model allows for training of automatic speech recognition (ASR) systems without the hand-designed language-specific pronunciation lexicons. However, constructing the multilingual low-resource E2E ASR system is still challenging due to the vast number of symbols (e.g., words and characters). In this paper, we investigate an efficient method of encoding multilingual transcriptions for training E2E ASR systems. We directly encode the symbols of multilingual writing systems to universal articulatory representations, which is much more compact than characters and words. Compared with traditional multilingual modeling methods, we directly build a single acoustic-articulatory within recent transformer-based E2E framework for ASR tasks. The speech recognition results of our proposed method significantly outperform the conventional word-based and character-based E2E models.

Index Terms: Speech recognition, acoustic model, End-to-End multilingual model, universal articulatory attributes

1. Introduction

Nowadays, people using different languages require a reliable speech-to-speech (S2S) translation system for communications in international affairs. As one of the most critical components in the S2S systems, multilingual speech recognition has been investigated for many years [1, 2, 3, 4, 5] and achieved encouraging results. For traditional context-dependent deep neural network hidden Markov models (CD-DNN-HMM) [6], hand-designed language-specific pronunciation lexicons must be employed. This is not a big problem for the widely used languages (e.g., English), because their modeling techniques have been fully studied for decades. However, this severely limits their application to low-resource languages. Recent End-to-End (E2E) attention-based models remove this dependency on the pronunciation lexicon [7]. Instead of alphabets, many languages using other writing systems (such as Abugidas) typically have more symbols than alphabets. As a result, the output nodes of the softmax layer are too large and make constructing multilingual E2E ASR system very challenging.

The articulatory attribute modeling, also known as "acoustic-to-articulatory(-attribute) modeling", is widely used to describe the attributes of different articulators during human speech production. Articulation means the movement of the tongue, lips, and other organs to make speech sounds, which share in common universally by all human beings. Articulatory information has been demonstrated useful in many related areas, such as speech comprehension improvement [8], speech therapy [9, 10, 11], pronunciation perceptual training [12, 13], robust speech recognition [14, 15] and large vocabulary continuous speech recognition (LVCSR) [16]. We can also encode many languages with articulatory sequences using very compact universal representations. Although the existing languages in the world are different in grammar and syntax, their pronunciations can be decomposed into a set of "atom" units of the articulation.

In previous researches on automatic speech attribute transcription [17, 18, 19, 20, 16], many detectors are trained to generate a bank of speech attributes. These attributes are either concatenated with speech features or used separately to detect a specific pronunciation. These methods are based on DNN-HMM models requiring context-dependent frame-level articulatory labels.

In this study, we investigate an efficient method of encoding multilingual transcriptions with universal articulatory representations. Using the articulatory representations as labels, we directly build a single articulatory attribute model based on recent transformer-based E2E framework for multilingual ASR tasks.

The remainder of this paper is organized as follows. Section 2 describes our proposed method. Section 3 provides experimental evaluations with different tasks. Conclusions and future works are given in Section 4.

2. Proposed Method

The proposed method of this paper (shown in Figure 1) is introduced in following subsections.

2.1. E2E Multiligual Articulatory Attribute Model

As we introduced in Section 1, multilingual speech recognition has been investigated for many years. For GMM-based systems, global Phones or IPA-like symbols [5] was commonly used for multilingual ASR. For DNN-based systems, multitask learning integrated with Global Phones achieved encouraging results [1, 2, 3, 4]. In these methods, the hidden layers are shared across multiple languages while the softmax layers are language dependent. They are optimized jointly by specifying the primary task and secondary task in the objective function. Multilingual articulatory attribute modeling [21, 22] relies on the multitask learning method [1, 2, 3, 4].

We follow the state-of-the-art multilingual transformer-based model [23] to train a single E2E model for multilingual articulatory attributes. The most significant difference between the transformer and commonly used E2E models [24, 25] is that the transformer entirely relies on attention and feedforward components [26]. In our proposed model, the log-Mel filterbank features of an input sequence are mapped to an output sequence of articulatory attributes. The detailed settings for training are described in Section 3.2.

* Corresponding authors and main contributors.

http://dx.doi.org/10.21437/Interspeech.2019-2092

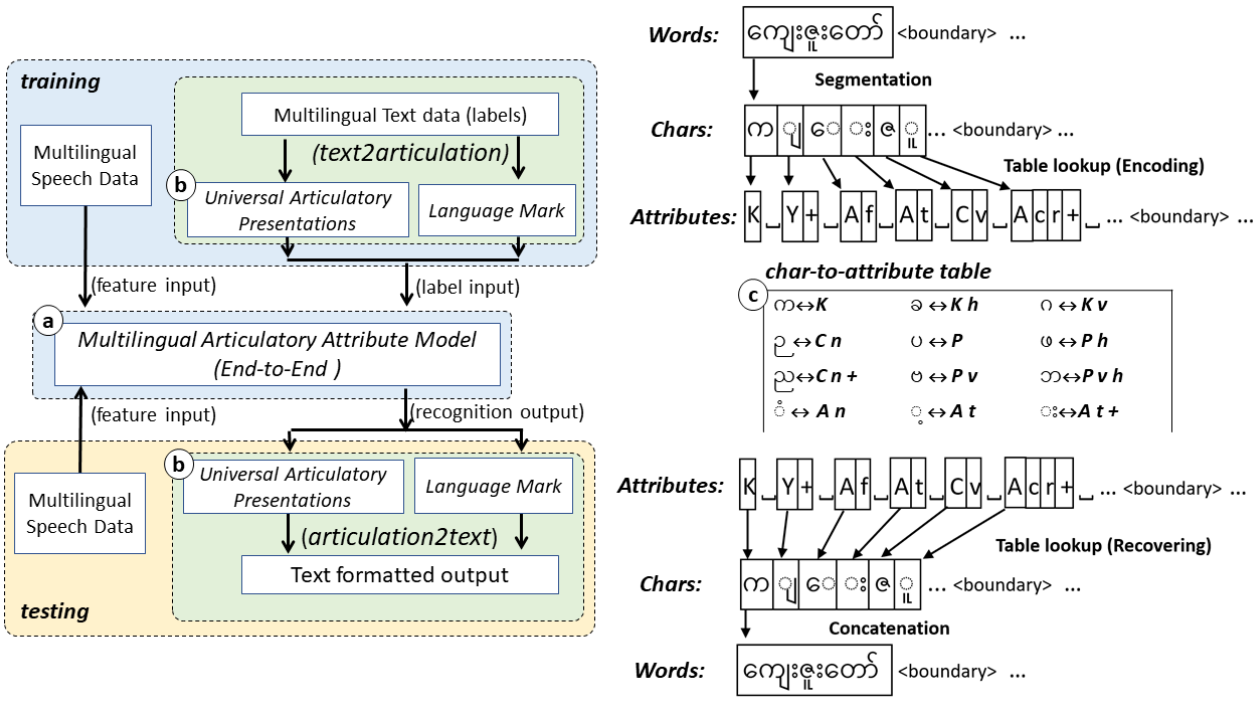

Figure 1: *The flowchart of the proposed method (taking Myanmar language as example).*

2.2. Universal Articulatory Representations

Table 1: *Universal Articulatory Representations*

Catergories	Attributes
Consonants (placement)	Velar (**K**) Palatal (**C**) Coronal (**T**) Labial (**P**) Glottal (**Q**)
Consonants (manner)	Aspired (**h**) Voiced (**v**) Nasal (**n**) Trill (**R**) Lateral approximant (**L**) Labial/Labio-velar approximant (**W**) Palatal approximant (**Y**) Sibilant fricative (**S**) Non-sibilant fricative (**H**)
Vowel (**A**)	Round (**r**) Front (**f**) Close (**c**) Tonal (**t**) Visarga (**h**) Anunasika (**n**)
Special Marks	Repeat Removal (**+**)

Visarga is an allophone of /r/ and /s/ in pausa (at the end of an utterance). Anunasika (anunasika) is a form of vowel nasalization.

As it is pointed out in Section 1, pronunciations from all of the existing languages can be decomposed into a set of "atom" units of the articulation. To retrieve the articulatory dynamics from speech signal is known as the acoustic-to-articulatory-attribute modeling. Taking advantage of modern devices such as EMA [27], X-ray [28], ultrasound [29] and MRI [30], it has also been used to search the acoustic correlates with the speech pronunciation variations.

In this paper, we are not using the recorded data to train the articulatory model, because the data collection may result in unnatural pronunciation and there are also many other problems [31], e.g., the synchronization among data streams, speaker normalization, data calibration, and data smoothing.

With the help of linguists of low-resource languages addressed in this study, each consonant is represented as placement attributes (**K**, **C**, **T**, **P**, **Q**) and/or manner attributes (**h**, **v**, **n**, **R**, **L**, **W**, **Y**, **S**, **H**), while vowels are described by other several attributes (**A**, **r**, **f**, **c**, **t**, **h**, **n**) as shown in Table 1. We define all articulatory attributes in a context-independent way. Inspired by [21, 22, 32, 33], we use a much easier method mapping the text to articulatory attributes using the table-lookup method as shown in Figure 1 and Figure 2.

2.3. Character-to-attribute Mapping

We take following steps to encode texts to universal articulatory representations (**Text2Articulation** in Figure 1).

Firstly, the texts are segmented into characters. The spaces between neighboring words are preserved using a word boundary mark ("<**boundary**>" in Figure 1).

To make the char-to-attribute table (as shown in Figure 2) with one-to-one mapping within the same language, each character has a unique representation by adding single or multiple "**+**" to the end of the duplicated representations. We design shorter articulatory representations according to the unigram frequencies. Moreover, the other un-pronounced tokens such as punctuation marks are not considered in this processing.

For recovering from the outputs of articulatory representations to original texts (**Articulation2Text** in Figure 1) when

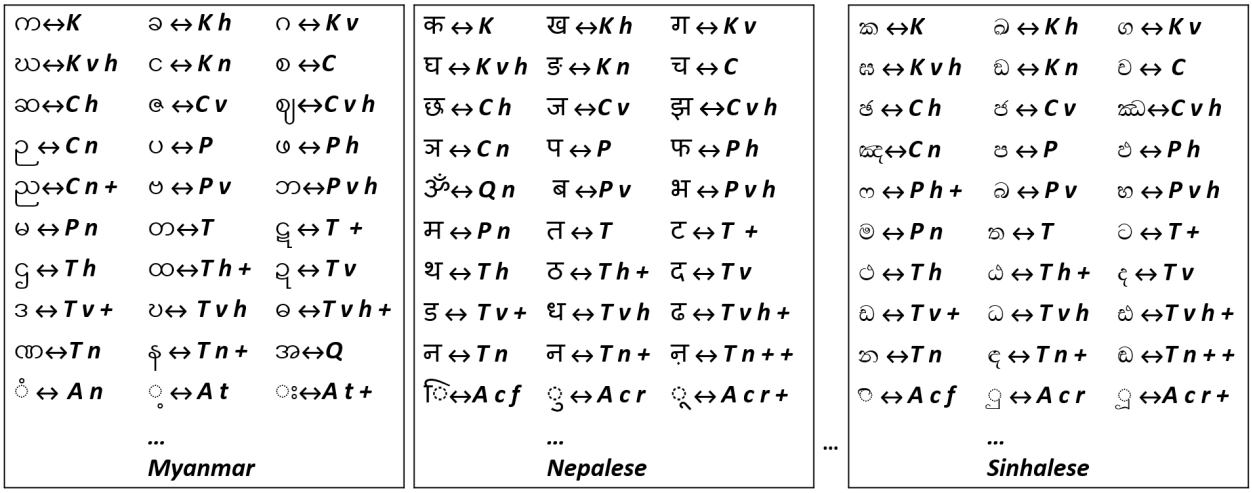

Figure 2: *Examples of converting multilingual characters to articulatory representations (assumed one-to-one mapping within the same language).*

testing, we refer to the char-to-attribute mapping table (as shown in Figure 2), and the word boundary marks detected from the recognition output.

3. Experiment Evaluations

3.1. Data Description

In this paper, we focus on four low-resourced South/Southeast Asian languages (Myanmar, Khmer, Sinhalese and Nepalese) as shown in Table 2. Unlike the widely used languages (e.g., English), these languages are not thoroughly studied for speech recognition. The Myanmar and Khmer data sets are the same with our previous work [34, 35]. The data sets of the other two languages are selected from Google's opensource databases [1]. These four datasets are all from smartphone input in tourist scenarios. The quality of the Khmer speech data is the best.

For each language, we select speech data from two hours to five hours as the test set and the rest is used for training.

Table 2: *Multilingual Datasets*

Language	Dataset	Hours
Myanmar (**MY**)	Training	54.4
	Testing	2.3
Khmer (**KH**)	Training	102.2
	Testing	5.5
Sinhalese (**SI**)	Training	27.9
	Testing	2.7
Nepalese (**NE**)	Training	38.7
	Testing	2.7

We used 120-dim filterbank features (40-dim static +Δ +$\Delta\Delta$), which were mean and variance normalized per speaker, and four frames were spliced (four left, one current and zero right). Speed-perturbation [36] was not used to save training time.

[1] http://www.openslr.org/52/ and http://www.openslr.org/54/

3.2. Model Training

We used the implementation of the Transformer-based neural machine translation (NMT-Transformer) [26] in tensor2tensor [2] for all our experiments. The training and testing settings listed in Table 3 are similar to those in [37].

Table 3: *Major Experimental Settings*

Model structure			
Attention-heads	8	Decoder-blocks	6
Hidden-units	512	Residual-drop	0.3
Encoder-blocks	6	Attention-drop	0.0
Training settings			
Max-length	5000	GPUs (K40m)	4
Tokens/batch	10000	Warmup-steps	12000
Epochs	30	Steps	300000
Label-smooth	0.1	Optimizer	Adam
Testing settings			
Ave. chkpoints	last 20	Batch-size	100
Length-penalty	0.6	Beam-size	13
Max-length	200	GPUs (K40m)	4

When training single-language models for reference, we find using the well-trained model as seed model to initialize the low-resourced speech data is very useful. However, this initialization strategy does not work when training the multilingual model. In this paper, we use a Mandarin transformer-based model (eight head-attention, six encoder-blocks and six decoder-blocks with 512 nodes) trained from 178 hours of speech data selected from AIShell dataset [38] with the CER of 9.0% to initialize the single-language models.

When training a multilingual model, we add the particular words <Language Mark> (e.g., <MY>, <KH>, <SI> and <NE>) to the beginning of the labels for every utterance. The training labels are organized as "<S> <Language Mark> labels </S>".

[2] https://github.com/tensorflow/tensor2tensor

3.3. Multilingual Speech Recognition Evaluation

To evaluate our proposed method, we trained a set of E2E speech recognition systems with word-based (**w**), character-based (**c**), globalphone (**p**) and (articulatory) attribute-based labels (**a**). Both single-language models (**separate-w, c, p, a**) and multilingal models (**multilingual-w, c, p, a**) are compared.

Table 4: *ASR performance (CER%) of acoustic models with different settings ("**w**" means word-based model, "**c**" means character-based model, "**p**" means globalphone-based model and "**a**" means articulatory-attribute-based model)*

	Multilingual Evaluation Sets			
	MY	KH	SI	NE
separate models (**w**) MY	23.9	/	/	/
KH	/	1.3	/	/
SI	/	/	23.7	/
NE	/	/	/	26.1
separate models (**c**) MY	49.2	/	/	/
KH	/	5.0	/	/
SI	/	/	22.1	/
NE	/	/	/	13.8
separate models (**p**) MY	23.4	86.4	90.6	93.4
KH	85.4	2.2	90.5	87.3
SI	83.6	85.7	21.0	71.1
NE	83.7	78.7	64.5	14.2
separate models (**a**) MY	23.2	86.0	89.7	92.1
KH	86.3	2.2	95.4	95.3
SI	84.0	85.0	22.8	70.8
NE	83.9	78.6	64.1	16.3
multi-lingual model (**w**)	23.7	1.5	34.2	36.7
multi-lingual model (**c**)	26.6	**0.7**	14.5	**10.8**
multi-lingual model (**p**)	28.1	2.0	**13.3**	**10.3**
multi-lingual model (**a**)	**21.6**	2.4	**13.6**	**10.7**

The results compared to the lowest result without statistical significance (from two-tailed t-test at significant level of p-value < 0.05) are shown in bold fonts.

From the results in Table 4, we observe that single-language models trained with globalphone and attribute-based labels (**separate-p, a**) are not good for recognizing speech of other languages (shown in small font size). We also find that training with the attribute-based labels (**separate-a**) for the single-language model does not provide the better performance, compared to character-based (**separate-c**), word-based (**separate-w**) and globalphone-based models (**separate-p**).

However, the attribute-based model with multilingual training (**multilingual-a**) can significantly outperform other models. The multilingual training can make use of more data from other languages, especially when there is not enough data for a specific language. This is very similar to the transfer learning mechanism. For the globalphone-based multilingual training (**multilingual-p**), since some phones are missing in a specific language, the knowledge of these phones cannot be shared between languages. For example, some phones of Myanmar do not exist in Khmer, Sinhalese, Nepalese, the globalphone-based multilingual training (**multilingual-p**) does not work well for Myanmar. Our proposed universal articulatory representation can share knowledge between different languages more effectively.

The other advantage of our proposed method can be found in Table 5. The universal articulatory attributes provide more compact representations compared to characters, words, and globalphones. No matter how many languages are dealt with, the number of classes is always very limited.

Table 5: *Number of Different Modeling Units ("**w**" means word, "**c**" means character, "**p**" means globalphone, "**a**" means articulatory attribute)*

Language	**#w** → **#c** → **#p** → **#a**
Myanmar (**MY**)	17,951 → 67 → 182 → 23
Khmer (**KH**)	3,129 → 79 → 182 → 23
Sinhalese (**SI**)	24,803 → 153 → 182 → 23
Nepalese (**NE**)	25,929 → 100 → 182 → 23
Multilingual (**MY+KH+SI+NE**)	71,812 → 365 → 182 → 23

For an objective evaluation, we also notice in Table 4 that Khmer is the only language that has slight performance degradation by using the proposed method. It has high-quality data over 100 hours and a well-tuned baseline. This result indicates that it is more appropriate to use the proposed multilingual training with universal attributes for the low-resource languages.

The proposed method can also be easily extended to other widely used languages (e.g., English and Chinese) by using pronuncication dictionaries. We first convert the words to phoneme sequences with dis-ambiguous marks to distinguish the homonyms (e.g., write → /r ai t @1/ and right → /r ai t @2/), and then map the phonemes to articulatory attribute representations using the phone-to-attribute table proposed in this paper.

4. Conclusions and Future Work

In this paper, we investigate an efficient method of encoding multilingual transcriptions for training E2E ASR systems. Compared with traditional multilingual modeling methods, we directly build a single acoustic-articulatory within the transformer-based E2E framework for multilingual ASR. The speech recognition results of our proposed model significantly outperform the traditional word-based and character-based E2E models. Moreover, universal articulatory attributes provide more compact representations than characters and words. In the future, we will test our method on more languages from all of the world.

5. References

[1] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. IEEE-ICASSP*, 2013, pp. 7304–7308.

[2] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proc. IEEE-ICASSP*, 2014, pp. 7639–7643.

[3] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," in *Proc. IEEE-ICASSP*, 2015, pp. 4994–4998.

[4] R. Sahraeian and D. V. Compernolle, "A study of rank-constrained multilingual dnns for low-resource ASR," in *Proc. IEEE-ICASSP*, 2016, pp. 5420–5424.

[5] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[6] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[7] T. N. Sainath, R. Prabhavalkar, S. Kumar, S. Lee, A. Kannan, D. Rybach, V. Schogol, P. Nguyen, B. Li, and Y. W. et al., "No need for a lexicon? evaluating the value of the pronunciation lexica in End-to-End models," in *CoRR abs/1712.01864*, 2017.

[8] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you read tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication*, vol. 52, pp. 493–503, 2010.

[9] S. Fagel and K. Madany, "A 3D virtual head as a tool for speech therapy for children," in *Proc. INTERSPEECH*, 2008.

[10] V. Mitra and E. Shriberg, "Effects of feature type, learning algorithm and speaking style for depression detection from speech," in *Proc. IEEE-ICASSP*, 2015.

[11] E. Yilmaz, V. Mitra, C. Bartels, and H. Franco, "Articulatory features for ASR of pathological speech," in *CoRR abs/1807.10948*, 2018.

[12] A. Rathinavelu, H. Thiagarajan, and A. Rajkumar, "Three dimensional articulator model for speech acquisition by children with hearing loss," in *Proc. the 4th International Conference on Universal Access in Human Computer Interaction*, vol. 4554, 2007, pp. 786–794.

[13] L. Wang, H. Chen, S. Li, and H. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication*, vol. 54, no. 7, pp. 845—856, 2012.

[14] K. Kirchhoff, "Robust speech recognition using articulatory information," *PhD thesis, Univ. of Bielefeld, Germany*, 1999.

[15] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, and M. Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Communication*, vol. 89, pp. 103–112, 2017.

[16] V. Mitra, W. Wang, C. Bartels, H. Franco, and D. Vergyri, "Articulatory information and multiview features for large vocabulary continuous speech recognition," in *CoRR abs/1802.05853*, 2018.

[17] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. INTERSPEECH*, 2004.

[18] C.-H. Lee, M. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L. Rabiner, "An overview on automatic speech attribute transcription (asat)," in *Proc. INTERSPEECH*, 2007, pp. 1825—1828.

[19] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," in *Proc. INTERSPEECH*, 2006, pp. 1287—1291.

[20] C.-Y. Lin and H.-C. Wang, "Attribute-based mandarin speech recognition using conditional random fields," in *Proc. INTERSPEECH*, 2007, pp. 1833—1836.

[21] W. Li, S. Siniscalchi, N. Chen, and C. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *Proc. IEEE-ICASSP*, 2016, pp. 6135–6139.

[22] R.Duan, T.Kawahara, M.Dantsuji, and J.Zhang, "Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data," in *Proc. IEEE-ICASSP*, 2017, pp. 5815–5819.

[23] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," in *CoRR abs/1806.05059*, 2018.

[24] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015.

[25] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE-ICASSP*, 2016.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *CoRR abs/1706.03762*, 2017.

[27] A. Wrench, "Multi-channel/multi-speaker articulatory database for continuous speech recognition research," *Phonus*, vol. 5, pp. 1–13, 2000.

[28] J. Westbury, *X-ray Microbeam Speech Production Database Users Handbook, version 1.0 edition*. Waisman Center on Mental Retardation and Human Development. University of Wisconsin, Madison, WI, USA, 1994.

[29] M. Grimaldi, B. Gili, F. Sigona, M. Tavella, P. Fitzpatrick, L. Craighero, L. Fadiga, G. Sandini, and G. Metta, "New technologies for simultaneous acquisition of speech articulatory data: 3D articulograph, ultrasound and electroglottograph," in *Proc. LangTech*, 2008.

[30] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1771–1776, 2004.

[31] H. Li, J. Tao, M. Yang, and B. Liu, "Estimate articulatory MRI series from acoustic signal using deep architecture," in *Proc. IEEE-ICASSP*, 2015.

[32] B. Abraham and S. Umesh, "An automated technique to generate phone-to-articulatory label mapping," *Speech Communication*, vol. 86, pp. 107–120, 2017.

[33] H. Zheng, Z. Yang, L. Qiao, J. Li, and W. Liu, "Attribute knowledge integration for speech recognition based on multi-task learning neural networks," in *Proc. INTERSPEECH*, 2015.

[34] H. Naing, A. Hlaing, W. Pa, X. Hu, Y. Thu, C. Hori, and H. Kawai, "A myanmar large vocabulary continuous speech recognition system," in *Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015.

[35] K. Soky, X. Lu, P. Shen, H. Kato, H. Kawai, C. Vanna, and V. Chea, "Building wfst based grapheme to phoneme conversion for khmer," in *Proc. Khmer Natural Language Processing (KNLP)*, 2016.

[36] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015.

[37] S.Zhou, L.Dong, S.Xu, and B.Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," in *Proc. INTERSPEECH*, 2018.

[38] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. Oriental COCOSDA*, 2017.