ACOUSTIC-TO-ARTICULATORY INVERSION BASED ON SPEECH DECOMPOSITION AND AUXILIARY FEATURE

Jianrong Wang[1] *Jinyu Liu*[2] *Longxuan Zhao*[2] *Shanyu Wang*[2] *Ruiguo Yu*[1] *Li Liu*[3⋆]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Tianjin International Engineering Institute, Tianjin University, Tianjin, China
[3] Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Acoustic-to-articulatory inversion (AAI) is to obtain the movement of articulators from speech signals. Until now, achieving a speaker-independent AAI remains a challenge given the limited data. Besides, most current works only use audio speech as input, causing an inevitable performance bottleneck. To solve these problems, firstly, we pre-train a speech decomposition network to decompose audio speech into speaker embedding and content embedding as the new personalized speech features to adapt to the speaker-independent case. Secondly, to further improve the AAI, we propose a novel auxiliary feature network to estimate the lip auxiliary features from the above personalized speech features. Experimental results on three public datasets show that, compared with the state-of-the-art only using the audio speech feature, the proposed method reduces the average RMSE by 0.25 and increases the average correlation coefficient by 2.0% in the speaker-dependent case. More importantly, the average RMSE decreases by 0.29 and the average correlation coefficient increases by 5.0% in the speaker-independent case.

***Index Terms*—** Acoustic-to-articulatory inversion, Speech decomposition, Personalized speech feature, Auxiliary feature, Speaker-independent

1. INTRODUCTION

The conversion from acoustic speech to articulatory movement is called the acoustic-to-articulatory inversion (AAI) [1], which plays a significant role in many applications (*e.g.*, pronunciation guidance [2], helping patients with vocal or hearing impairments [3] and speech recognition [4]).

Early in [5], the acoustic speech was mapped to articulatory movement with the codebook. However, the results of their inversion were highly dependent on the quality of the codebook. Then, with the publishing of corpora containing parallel acoustic and articulatory data, data-driven inversion frameworks based on machine learning were proposed. And the Mel-scale frequency cepstral coefficients (MFCC) of the speech signals were first accepted as inputs and then mapped to articulatory movements. Later, other methods like hidden markov model [6], mixture density network [7], and deep belief network [8] were proposed. Recently, with the rise of deep neural networks (DNNs), the deep bidirectional long short-term memory (DBLSTM) was used by [9] in the AAI. Since then, most of AAI related works (*e.g.*, [10, 11, 12, 13]) used DBLSTM to deal with various applications but the inputs of the models were only audio speech features. As for the speaker-independent AAI, the vocal tract length normalization [11] was proposed to transform the acoustic spaces of different speakers to a target one. [12] proposed the idea of pre-train and fine-tune to improve the generalization performance on their own dataset. Especially, [14] used one-dimensional convolution of different sizes to extract the audio feature. It improved the performance in speaker-independent case by adding extra phoneme information, achieving the state-of-the-art (SOTA) result on the public Haskins Production Rate Comparison (HPRC) [15] dataset.

Until now, there are two main challenges leading to performance bottlenecks in AAI. Firstly, most of existing works only used the audio speech feature to predict the articulatory movement without exploiting any additional features (*i.e.*, lip feature, speaker identity feature or the content feature). Secondly, some works devoted to improve the generalization performance for the speaker-independent AAI, but their methods either lost the personalized information [11], needed large amounts of data [12] or required additional phoneme information [14].

To address the above two challenges, we propose a novel SAF network composed of **S**peech Decomposition Network (SDN), **A**uxiliary Feature Network (AFN) and **F**eature Transformation Network (FTN), which we call SAFN in brief. Firstly, in order to adapt to the speaker-independent case, we explore a SDN inspired by the idea of speech synthesis [16] to obtain the personalized speech features. Then, to further improve the performance of AAI, we propose a novel AFN to estimate the lip auxiliary features as the prior information from the personalized speech features. Then, we design a FTN to generate feature pairs by transforming the personalized speech features and the lip auxiliary features. Last but not least, though AAI is not an one-to-one maping task

⋆Corresponding author.

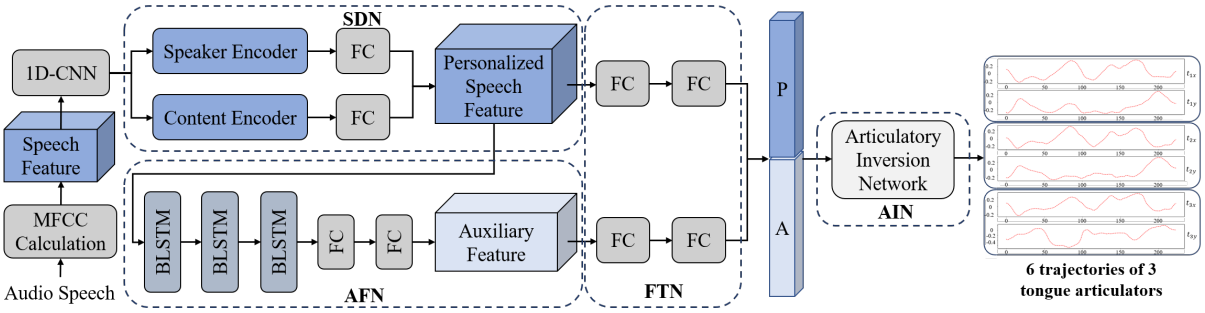

Fig. 1. The overview of our proposed SAFN. The P is the personalized speech features and the A is the auxiliary features.

(*i.e.*, different articulatory movements may correspond to the same speech), this problem can be alleviated by ensuring the smoothness of the articulatory movement on consecutive frames. An overview of our SAFN is shown in Fig. 1.

In summary, this work has following three contributions: 1) To adapt to the speaker-independent case, a SDN is proposed to obtain the speaker embedding and the content embedding as personalized speech features. 2) To further improve the performance of AAI, a novel AFN is proposed to estimate the lip auxiliary features as prior knowledge. To the best of our knowledge, this is the first work that introduces an auxiliary feature instead of directly using the speech feature as input. 3) Speaker-independent AAI experimental results show the superior performance of the SAFN. On the public HPRC dataset, the SAFN outperforms the SOTA by a large margin (*i.e.*, the average RMSE decreases by 0.29 and the average correlation coefficient increases by 5.0%).

2. PROPOSED APPROACH

2.1. Overall Framework

In Fig. 1, the core of SAFN are SDN and AFN. The SDN is pretrained to obtain the speaker embedding and the content embedding, which are served as the personalized acoustic features. The AFN is trained to estimate the corresponding lip auxiliary features as prior knowledge to help the prediction of tongue organ. To get such a network, the MFCC of the speech audio are extracted as speech features, and then the features are further encoded by multi-scale one-dimensional convolution. Besides, the speech features are sent to the pretrained SDN to obtain the corresponding two embeddings, which are used as personalized speech features and sent to the AFN to obtain the lip auxiliary features. Later, the personalized speech features and the lip auxiliary features are feature fused [17] as multi-features. These multi-features are sent to the articulatory inversion network (AIN) to predict the movements of the tongue organs. The objective function for training is to minimize the combination of the reconstruction L2 loss of AFN and reconstruction L2 loss of AIN, which is given as:

$$L = \alpha \times \sum_{i=0}^{m} \left(y_l^i - \hat{y}_l^i\right)^2 + \beta \times \sum_{i=0}^{m} \left(y_t^i - \hat{y}_t^i\right)^2, \quad (1)$$

where α and β are set as 0.5 and 0.5 experimentally. y_l^i and y_t^i refer to the estimated lip auxiliary features by AFN and tongue movements by AIN, respectively. $\hat{y}_l^i$ and $\hat{y}_t^i$ refer to the corresponding real labels.

2.2. Speech Decomposition Network

It was shown in [16] that speech signals inherently carry both non-linguistic information and linguistic information. The non-linguistic part refer to speaker identity, which is time-independent. While the linguistic part refer to content, which changes dramatically every several frames. On this basis, we pre-train the SDN (see Fig. 2) to obtain the representation of speaker and content. Then the two representations obtained by the SDN, are fed as prior knowledge to adapt to the multi-speaker case, further improving the speaker generalization ability of SAFN.

The SDN is a self-supervised model, composed of a speaker encoder, a content encoder and a decoder. The speaker encoder is trained to encode the non-linguistic information into the speaker representation. The content encoder is trained to encode the linguistic information into the content representation. And then the decoder is aimed to synthesize the speech feature by combining these two representations. However, we are just concerned about the part of speaker representation and content representation, so we pretrain the SDN beforehand to obtain the speaker encoder and content encoder. Moreover, the SDN is trained on the whole dataset without using the articulatory labels. Thus, the proposed method is speaker-independent. We consider that the SDN has learned all the speakers acoustic information including identify information and content information.

The core of SDN is that, by normalizing the channel statistics which control the global information, the instance normalization (IN) [18] enforces the content encoder to focus on the linguistic part and remove the global information (*i.e*, speaker information), while the averge-pool enforces the speaker encoder to focus on the non-linguistic part and learn the global information. Besides, the convolutional layer is used to capture long-term information. The dense layer [19] is used to enhance feature reuse and network training. And the adaptive instance normalization (AdaIN) [20] is utilized in decoder to bring the global information to the predicted

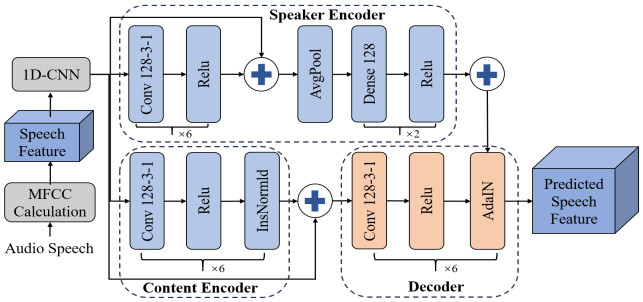

Fig. 2. Structure of the speech decomposition network.

speech feature though the corresponding parameters provided by speaker encoder. By doing this, the global information needed in the decoder is controlled by the speaker encoder. Thus, the SDN is encouraged to learn factorized representations. The loss of SDN is the reconstruction L1 loss between the input speech feature and the predicted speech feature. IN is expressed as:

$$M_c^{'} = \frac{M_c[w] - u_c}{\sigma_c}, \tag{2}$$

where M_c represents the c-th channel with dimension w. $M_c[w]$ is the w-th element in M_c. To obtain IN, we first compute the mean $u_c = \frac{1}{W}\sum_{w=1}^{W} M_c[w]$, the standard variation $\sigma_c = \sqrt{\frac{1}{W}\sum_{w=1}^{W}(M_c[w] - u_c)^2 + \epsilon}$, where ϵ is a small value to avoid numerical instability. Each element in the array M_c is normalized into $M_c^{'}$.

2.3. Auxiliary Feature Network

A new auxiliary feature is defined based on the EMA dataset [21], where parallel acoustic and articulatory data are collected. Multiple sensors are attached to the pre-specified positions in EMA recordings. There are totally six sensors placed on the six articulators, namely tongue tip (T1), tongue blade (T2), tongue rear (T3), upper lip (UL), lower lip (LL), and lower incisors (LI). In particular, we divide these positions into two categories: outside visible positions (UL, LL, and LI) called lip auxiliary features in this work, and inner invisible positions (T1, T2, and T3) called the movements of tongue organ.

Then the above lip auxiliary features are estimated from the audio speech by AFN, instead of using the true label directly detected by the sensors. The reason is that we want to keep the same input (only acoustic speech) as the previous works. Besides, we do not freeze the parameters of the AFN during the training of the whole network.

The AFN is to estimate the lip auxiliary features from the personalized speech features, which is composed of three BLSTM layers to extract the contextual information, and two FC layers are followed by the BLSTM layers to generate trajectories of lip auxiliary features. The core of the AFN is expressed as:

$$O_t^l = \sigma(W_{io}^l x_t + b_{io}^l + W_{ho}^l h_{t-1} + b_{ho}^l),$$
$$O_t^r = \sigma(W_{io}^r x_t + b_{io}^r + W_{ho}^r h_{t+1} + b_{ho}^r),$$
$$O_t = \frac{1}{2} \times (O_t^l + O_t^r), \tag{3}$$

where O_t^l is the lip auxiliary features estimated at the frame t and the frame $t-1$, O_t^r is the lip auxiliary features estimated at the frame t and the frame $t+1$. x_t is the input personalized speech features at frame t, h_t is the temporary state at frame t and W_{io}^l, b_{io}^l are the corresponding transformation matrix and bias from i to o. By processing the forward and backward iteration, we obtain the O_t, which is the lip auxiliary features estimated at contextual information.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets The public MOCHA-TIMIT [22], MNGU0 [23], and HPRC [15] speech corpora include six reading locations set on T1, T2, T3, UL, LL and LI. In this work, we use the three tongue locations of X and Z directions (*i.e.*, T1, T2, T3) as our experimental predicted target. The MOCHA-TIMIT dataset consists of 460 utterances and EMA data recorded for one male and one female speaker, who speak British English. The MNGU0 dataset consists of 1263 utterances and EMA data recorded for one male speaking British English. The HPRC dataset consists of 720 utterances and EMA data recorded for eight native American English speakers.

Performance Metrics The performance is evaluated by two classical metrics, *i.e.*, root mean square error (RMSE) and correlation coefficient (CC) [9].

Implementation Details In addition to the modules described in section 2, the AIN contains three BLSTM layers with 100 units in each layer, followed by 2 FC layers. We train the proposed SAFN for 28800 iterations by Adam optimizer with a 1e-4 learning rate and the batch size is set as 5. Besides, the SOTA in [14] is reproduced as the baseline of our experiments. The SDN (shown in Fig. 2) is pre-trained beforehand by Adam optimizer with a 5e-4 learning rate and the batch size is set as 25. Datasets are divided into the training set, the validation set, and the test set according to the proportion 8:1:1, respectively.

3.2. Comparisons with the SOTA

To verify the generalization ability of SAFN, we conduct experiments by comparing the proposed SAFN with the SOTA [14] in two directions (*i.e.*, single speaker and multiple speakers) and four scenarios according to Table 1. S1 represents the experiment on single speaker. S2 represents the experiment on multi-speakers. S3 represents the speaker adaption experiment. More precisely, we first pool the training data from the whole dataset except the target speaker data to train a generic model. Then we fine tune the generic model weights using the target speaker data. S4 represents the speaker-independent

experiment. The RMSE and CC in four scenarios are shown in Table 2.

Table 1. Experimental setup for 4 different scenarios. S1 means single speaker, S2 means multi-speaker, S3 means speaker adaptation and S4 means speaker-independent. * means taking the corresponding proportion of data from each speaker. G, M and H represent dataset MNGU0, MOCHA and HPRC, respectively. --- means no action.

Scenarios	Dataset	#Speaker	Train	Validation	Fine-tune	Test
S1	G, M, H	1	80%	10%	---	10%
S2	H	N	80%*	10%	---	10%
S3	H	N-1 1	80%* ---	20%* ---	--- 80%	--- 20%
S4	H	N-1 1	80%* ---	20%* ---	--- ---	--- 100%

Table 2. RMSE and CC for SOTA and SAFN in four scenarios.

Scenarios	Model	t_{1x}	t_{1z}	t_{2x}	t_{2z}	t_{3x}	t_{3z}	RMSE	CC
S1(G)	SOTA	0.886	0.792	1.061	0.707	1.106	0.911	1.014	0.922
	SAFN	0.789	0.738	0.990	0.619	1.051	0.796	**0.830**	**0.941**
S1(M)	SOTA	1.520	1.868	1.869	1.568	1.535	1.822	1.697	0.906
	SAFN	1.289	1.497	1.403	1.442	1.571	1.551	**1.459**	**0.924**
S1(H)	SOTA	1.725	1.760	1.871	1.684	2.060	2.170	1.881	0.901
	SAFN	1.419	1.530	1.601	1.552	1.559	1.443	**1.517**	**0.922**
S2	SOTA	1.730	1.840	1.901	1.721	2.100	2.210	1.917	0.890
	SAFN	1.488	1.857	1.701	1.631	1.709	1.589	**1.662**	**0.903**
S3	SOTA	1.730	1.759	1.830	1.651	2.030	1.850	1.808	0.911
	SAFN	1.411	1.509	1.551	1.563	1.534	1.535	**1.507**	**0.925**
S4	SOTA	2.675	3.803	3.384	2.102	2.878	3.227	3.009	0.701
	SAFN	2.184	3.077	2.938	2.621	2.412	3.096	**2.721**	**0.751**

Table 2 shows RMSE and CC value in four scenarios among three public datasets. Basically, we can observe that the proposed SAFN outperforms SOTA by almost 0.18mm $\sim$ 0.36mm on RMSE. Besides, CC scores show a similar trend (increase by 1.4% $\sim$ 5.0%). It is obvious to see that the improvement of CC in the speaker-independent case (S4) is 5.0%, which is much higher than that in the speaker-dependent cases (S1, S2 and S3). It indicates that the prior speaker identity features obtained by SDN can effectively alleviate the mismatch between the acoustic space of the speakers in the training set and those in the test set, further to adapt to the speaker-independent case.

3.3. Ablation Study

To verify the effectiveness of the proposed modules in Section 2, we carry out the ablation experiment according to Table 3, and the results are shown in Fig. 3. Obviously, in the speaker-dependent cases (S1, S2 and S3), those models with AFN (*i.e.*, SAFN, SAFN-S-A and SAFN-A) outperform those without AFN (*i.e.*, SOTA and SAFN-S). However, in the speaker-independent case (S4), those models with SDN (*i.e.*, SAFN, SAFN-S-A and SAFN-S) outperform those without SDN (*i.e.*, SOTA, SAFN-A). Based on the above results, it is demonstrated that in the speaker-dependent case, the AFN improves the performance by adding lip auxiliary features as prior knowledge. In the speaker-independent case, the prior

Table 3. Ablation experiment verifies the performance of each module. SAFN-S, SAFN-A and SAFN-S-A represents the AIN comparing with SDN, AFN and both of the above two parts, respectively.

	SDN	AFN	FTN
SOTA	×	×	×
SAFN-S	✓	×	×
SAFN-A	×	✓	×
SAFN-S-A	✓	✓	×
SAFN	✓	✓	✓

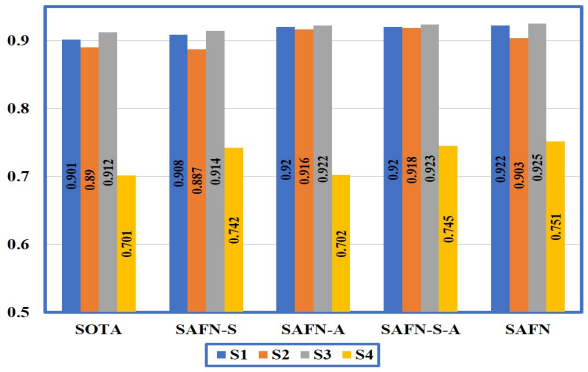

Fig. 3. CC of the different neural network in four scenarios.

personalized speaker information obtained by SDN brings a large gain. Besides, comparing SAFN with SAFN-S-A, we hypothesis that FTN improves the performance of SAFN by enhancing the correlations between the personalized speech features and the lip auxiliary features.

4. CONCLUSION

In this work, we propose a novel network SAFN to promote the generalization ability of speaker-independent AAI and further improve AAI performance. Firstly, to improve the generalization ability of the proposed SAFN, a SDN is presented to obtain the speaker embedding and content embedding as the personalized speech features. Besides, to further improve the performance of AAI, a new AFN is proposed to obtain the lip auxiliary features as prior knowledge to help the prediction of the tongue organ. Experimental results on three public datasets demonstrate that both in speaker-dependent and speaker-independent scenarios, the SAFN outperforms SOTA by a large margin. For the future work, the self-supervised method based on Meta Learning will be applied to the speaker-independent AAI task.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China(No. 61977049), National Natural Science Foundation of China (No. 62101351) and the Tianjin Key Laboratory of Advanced Networking.

6. REFERENCES

[1] Korin Richmond, "Estimating articulatory parameters from the acoustic speech signal," *Annexe Thesis Digitisation Project*, 2002.

[2] Li Liu, Gang Feng, and Denis Beautemps, "Inner lips feature extraction based on clnf with hybrid dynamic template for cued speech," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–15, 2017.

[3] Junhong Zhao, Hua Yuan, and WaiKim Leung, "Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training," *in Proc. ICASSP*, pp. 8218–8222, 2013.

[4] Li Liu, Gang Feng, Denis Beautemps, and Xiao-Ping Zhang, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 292–305, 2020.

[5] Sadao Hiroya and Masaaki Honda, "Determination of articulatory movements from speech acoustics using an hmm-based speech production model," *in Proc. ICASSP*, vol. 1, pp. 433–437, 2002.

[6] Masaaki Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *in Proc. TSAP*, vol. 12, no. 2, pp. 175–185, 2004.

[7] Korin Richmond, Simon King, and Paul Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 153–172, 2003.

[8] Leonardo Badino, Claudia Canevari, and Lucian Fadiga, "Deep-level acoustic-to-articulatory mapping for dbn-hmm based phone recognition," *in Proc. SLT*, pp. 370–375, 2012.

[9] Peng Liu, Quanjie Yu, and Zhiyong Wu, "A deep recurrent approach for acoustic-to-articulatory inversion," *in Proc. ICASSP*, pp. 4450–4454, 2015.

[10] Aravind Illa and Prasanta Kumar Ghosh, "The impact of speaking rate on acoustic-to-articulatory inversion," *Computer Speech and Language*, vol. 59, pp. 75–90, 2020.

[11] Ganesh Sivaraman, Vikramjit Mitra, and Hosung Nam, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion," *in Proc. Interspeech*, pp. 455–459, 2016.

[12] Aravind Illa and Prasanta Kumar Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," *in Proc. Interspeech*, pp. 3122–3126, 2018.

[13] Maud Parrot, Juliette Millet, and Ewan Dunbar, "Independent and automatic evaluation of speaker-independent acoustic-to-articulatory reconstruction," in *Proc. Interspeech*, 2020.

[14] Abdolreza Sabzi Shahrebabaki, Sabato Marco Siniscalchi, and Giampiero Salvi, "Sequence-to-sequence articulatory inversion through time convolution of sub band frequency signals," *in Proc. Interspeech*, pp. 2882–2886, 2020.

[15] Mark Tiede, Carol Y Espy-Wilson, and Dolly Goldenberg, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[16] Juchieh Chou and HungYi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *in Proc. Interspeech*, pp. 664–668, 2019.

[17] Mohammad Haghighat, Mohamed Abdel-Mottaleb, and Wadee Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition," *in Proc. TIFS*, vol. 11, no. 9, pp. 1984–1996, 2016.

[18] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 4700–4708.

[20] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[21] N. Meenakshi, C. Yarra, and P. K. Ghosh, "Comparison of speech quality with and without sensors in electromagnetic articulograph 501 recording," *in Proc. Interspeech*, 2014.

[22] Alan Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research.," *Phonus.*, 2000.

[23] R. Korin, H. Phil, and K. Simon, "Announcing the electromagnetic articulography subset of the mngu0 articulatory corpus," *in Proc. Interspeech*, 2011.