COMPARISON OFEMA-SYNCHRONIZED AND STAND-ALONE SPEECH BASED ON SPEECH RECOGNITION

FANG Qiang

Abstract Synchronized acoustic-articulatory data is the basis of various applications, such as exploring the fundamental mechanisms of speech production, acoustic to articulatory inversion(AAI) , and articulatory to acoustic mapping(AAM) . Numerous studies have been conducted based on the synchronized ElectroMagnetic Articulograhy (EMA) data and acoustic data. Hence, it is necessary to make clear whether the EMA synchronized speech and stand-alone speech are different, and if so, how it affects the performance of the applications that are based on synchronized acoustic-articulatory data. In this study, we compare the difference between EMA-synchronized speech and stand-alone speech from the aspect of speech recognition based on the data of a male speaker. It is found that: i.) the general error rate of EMA-synchronized speech is much higher than that of stand-alone speech; ii.) apical vowels and apical/blade consonants are more significantly affected by the presence of EMA coils; iii) parts of vowel and consonant tokens are confused with the sounds who use the same articulator or the articulators nearby, such as confusion among apical vowels and confusion among apical and blade consonants; iv) the confusion of labial tokens demonstrates a diverse pattern.

Keywords EMA-synchronized speech, Stand-alone speech, Speech recognition, Confusion matrix

基于语音识别的 EMA 同步语音与独立语音的比较研究

方　强

摘　要　同步声学—发音信号同步记录了语音声学信号和对应的发音器官的位置和形状。因此,同步声学—发音信号是语音的生理发音机制、声学机制、发音逆推、基于发音数据的语音合成等领域研究的基石。时至今日,有多种不同的设备被用于采集同步的发音和声学数据,如超声、x-ray 微束、EMA、以及实时 MRI 等。在这些设备中,EMA 能同时记录多个发音器官的运动,具有高时间分辨率,采集的数据便于后期处理和分析,并且对被试无害,因而得到了广泛的应用。通常在使用 EMA 采集同步声学—发音信号时,我们将传感器粘贴在发音器官上(如上下唇、下颚、舌尖、舌叶、舌背等位置),而这些传感器通常占据一定的空间(2.2mm×2.4mm×1.8mm)并且通过线缆与外部采集模块相连。在发音过程中,传感器和线缆可能会影响发音时声道收紧部位的形状,和对发音的精准控制,导致产生的 EMA 同步语音与独立的语音在声学上产生差异,对语音的生理发音机制、声学机制、发音逆推、基于发音数据的语音合成等领域的研究产生不利影响。这一问题还没有得到广泛关注,目前只有少量的研究从感知的角度比较了 EMA 同步语音的发音正确率和可懂度,还没有研究考察感知正确的 EMA 同步语音与独立语音是否有差异,以及这些差异是否会影响其它任务。本文从语音识别的角度,通过对比分析 EMA 同步语音与独立语音的识别率和混淆矩阵我们发现:1)感知正确的 EMA 同步语音的正确率显著低于独立语音的正确率;2)舌尖元音和舌尖/舌叶辅音更容易受到传感器的影响,其它元音和辅音也受到不同程度的影响,这与人们通常的认知不同;3)元音和辅音容易与使用相同或相近发音器官的元音和辅音混淆;4)唇辅音的混淆模式有些发散,如/p/、/p^h/除部分被混淆成唇音外还被混淆成舌根擦音/x/,不少/f/被识别生成/k,k^h,x/。这些发现说明由于在发音器官上粘贴了传感器,除舌尖音外,其它发音的准确性和发音方式都受到一定程度的影响。

关键词　EMA 同步语音,独立语音,语音识别,混淆矩阵

1. INTRODUCTION

Synchronized acoustic-articulatory data is an important type of data for investigating the mechanism of speech production, AAI, and AAM. Nowadays, several types of equipment, such as ultrasound (Chiu *et al.* 2020), x – ray microbeam (Qin & Carreira – Perpinan 2010), EMA (Engwall 2003; Fang, Wei et al. 2013), real – time MRI (Lammert *et al.* 2010) have been deployed to collecting synchronized speech – articulatory data. Among them, EMA is widely used due to its harmless to subject, high time resolution, flexibility in monitoring multi speech organs, and relative less cost in processing the acquired data.

Typically, to record speech related articulatory motions, coils are attached to several articulators, namely upper lip (UL), lower lip (LL), lower incisor (LI), tongue tip (TT), tongue body (TB), tongue dorsum (TD) in the midsagittal plane (Richmond 2002).

The coil of EMA are about 2.2mm × 2.4mm × 1.8mm in size. The weight of the coil along with 1.2m coil cable is 2.47g. Moreover, the coating gun and glue covered the coil increase the effective size and weight of the coil. Therefore, the articulation might be affected by the following two factors: i.) the coils and wires may make subjects discomfort and influence their natural articulation; ii) the tiny coils may affect the formation of constriction in vocal tract due to their size. As a consequence, these factors might make EMA-synchronized speech different to stand-alone speech.

Hence, it is necessary to analyze the difference between EMA-synchronize speech and stand-alone speech, and their influence on the performance of AAI, AAM, and other applications. Unfortunately, most of the studies on AAI (Liu *et al.* 2015; Tobing *et al.* 2016; Udupa *et al.* 2022) and AAM (Aryal & Gutierrez-Osuna 2016; Liu *et al.* 2018) directly transformed EMA-synchronized speech signal into desired acoustic features and constructed mapping between acoustic feature and articulatory states with various statistic models. Few of them explored the potential difference between EMA-synchronized speech and stand-alone speech, let alone its influence on the performance of AAI and AAM.

Meenakshi et al. (Meenakshi *et al.* 2014) investigated the effect of the presence of EMA coils on speech spoken. In their experiment, a set of 19 VCV sequences were recorded in both coil present and absent conditions. 16 evaluators were involved in listening experiments to rate the recorded speech. They found that: i.) both the EMA-synchronized and stand-alone speech were not 100% correct; ii.) the human recognition score of EMA-synchronized speech was lower than that of stand-alone speech; iii.) the EMA-synchronized speech was perceptually different to the stand-alone speech; iv.) the spectral difference between EMA-synchronized speech and stand-alone speech was larger than within both EMA-synchronized speech and stand-alone speech. As far as we know, this is the only study that explores the difference of EMA-synchronized speech and stand-alone speech. And it reveales that the EMA synchronized speech and stand-alone speech are different both perceptually and objectively.

In this study, we would like to explore this based on utterance level speech. The considerationare two folds: i.) utterance level material is more natural for subjects to produce, and more similar to the circumstance of AAI and AAM; ii.) utterance level speech contains not only consonants but also vowels, which is more comprehensive than Meenakshi's study. Since the size of the utterance level speech is much larger than that in Meenakshi's study, it is not feasible to conduct listening experiments. Hence, we conduct speech recognitions on EMA-synchronized speech and stand-alone speech, respectively, and analyze the detail recognition results.

In this study, we concern the following 3 issues: i.) whether the performance of speech recognition is different for EMA-synchronized speech and stand-alone speech though both of them are perceptually correct; ii.) the collection of vowels and consonants that are prone to be affected by the presence of EMA coils; iii.) the sound change tendency of vowels and consonants affected by the presence of EMA coils.

2. DATASET

2.1 Subject and scripts

A male subject, about 35 years old, without known speech disorder, is recruited in the EMA experiment. And the subject lived in the region of Xiang Dialect for his first 18 years, and then lived in Beijing for about 20 years. In seldom circumstances, he mixes "z, c, s" up with "zh, ch, sh", and mixes nasal coda "n" up with "ng".

A set of 986 phonetically balanced sentences of standard Chinese arecollected for recording. It covers all the initials and finals of standard Chinese. The min, max and mean lengths of the utterance are 5 syllables, 30 syllables, 15.6 syllables, respectively. The standard deviation is 4.5 syllables. And most of the sentences contain 11 to 20 syllables.

2.2 Setup

The NDI Wave system is employed to record acoustic and articulatory signals simultaneously. To record the precise positions of articulators, several coils are glued to the surface of the articulators of interest. Fig 1 (a) presents the positions of coils roughly, and Fig 1 (b) presents the clouds of the coils registered to the MRI image which is overlayed with CBCT (Cone Beam Computer Tomograph) image of the same subjects in the midsagittal plane.

In our experiment, four coils are attached

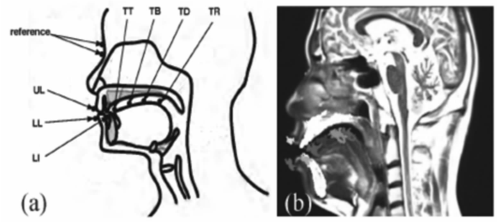

Fig. 1 (a) EMA sensors' placement and (b) the aligned MRI-CBCT-EMA volume.

to the tongue surface, namely Tongue Tip (TT), Tongue Blade (TB), Tongue Dorsum (TD), Tongue Rear (TR). Two coils are attached to the Lower Lip (LL) and Upper Lip (UL), respectively, and one coil is attached to the Lower Incisor (LI). Another two coils are attached to the ridge of nose to serve asthe references for removing artifacts introduced by head movement in recorded EMA data. Among them, TT is exactly attached to the surface of tongue tip to monitor the motion of tip that are import for apical consonants "t, d, n, l". TT, TB, TD, and TR are almost positioned along the tongue surface in the midsagittal plane with the interval about 1cm. Based on the EMA data registered to MRI image of the same subject (shown in Fig 1 (b)), we can assume that TB can be used to monitor the tongue motion for producing consonants "s, sh, z, zh, c, ch, j, q, x" and vowel "i, ix, iy", and TD and TR can be used to monitor the tongue motion for producing "g, k".

The EMA-synchronized speech is recorded simultaneously with the motion data. The sampling frequencies is 16,000Hz for acoustic signal and 100Hz for the articulatory signal, respectively. The stand-alone speech is recorded several days later in a separate sound proof room with the sampling frequency of 16,000Hz.

In the recording phase, an experimenter is asked to monitor the recoding process. The subject is asked to reproducethe sentence if a sentence is not correctly uttered. After recording, the subject himself and another experimenter are asked to check the recorded sentences to make sure that all the sentences are correct with reference to the recording prompt.

3. RECOGNITI EXPERIMENT

In this section we appliedthe speech recognition method to analyze the difference between EMA-synchronized speech and stand-alone speech.

3.1 Recognizer setup

In this experiment, a conformer model isemployed to serve as the recognizer (Guo *et al.* 2021). It consists of a conformer encoder and a transformer decoder, as shown in Fig 2. The encoder and decoder are both multi-blocked architectures. Each encoding block is stacked by a feedforward module, a multi-head self-attention module, a convolution module, and another feedforward module. And layer normalization is applied to the output of each encoding block in the end. For the details of each module in the encoding block, please refer to (Gulati *et al.* 2021). Each decoder block is stacked by a multi-head self-attention module, a multi-head cross-attention module, and a feedforward module. For the details of each module in the decoder block, please refer to (Vaswani *et al.* 2017). The detail configuration the conformer model and the setting for training the model can be found in the "train. yaml" file of the recipe of aishell in the ESPNET toolkit.

3.2 Peformance

A speech recognizer is trained by following the recipe of aishell in ESPNET. Because we want to compare the performance of recognizer on intials and finals, we take the intial-final streams rather than the word streams as the target. As shown in Table 1, the mean phone error rate is2.4%, and std is 1.3%.

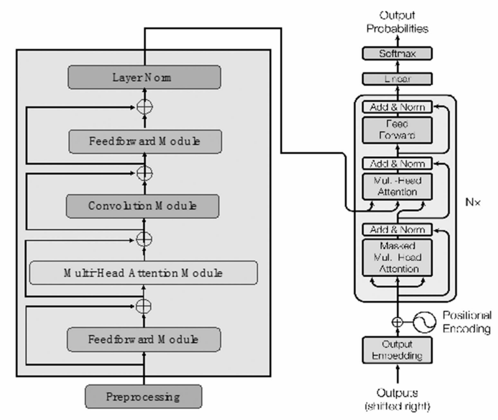

Fig. 2 Framework of the conformer recognizer.

Here phone means inital or final. This suggests that the trained recognizer is good enough for our purpose.

Table 1 Phone error rate on test set. (unit: %)

	Mean	Std	Median
Test set	2.4	1.3	2.1

4. RESUTLS

4.1 Overall phone error rates

The ASR system trained in Section 3 is applied to recognize the EMA-synchronized speech (denoted by Sync) and stand-alone speech (denoted by Stand-alone). As shown in Table 2, the mean phone error rate is 37.4% for EMA-synchronized speech and 7.8% for stand-alone speech. The phone error rate of EMA-synchronized speech is much higher than that of stand-alone speech. This indicates, the acoustic characteristics of EMA-synchronized speech have been significantly changed due to the presence of EMA coilsthough both of them are perceptually correct.

Table 2 Phone error rate on EMA-synchronized speech and stand-alone speech. (unit:%)

	Mean	Std	Median
Sync	37.4	—	—
Stand-alone	7.8	—	—

4.2 Phone-specific error rates

Moreover, we analyze difference of phone-specific error rates of consonants and vowels between EMA-synchronized speech and stand-alone speech to shed light on what is effect of the presence of EMA coil on detail consonants and vowels. The results are presented in Table 3 and Table 4. The first columns in Table 3 and Table 4 are the transcriptions of vowels and consonants in the recording prompt, and second columns are the corresponding IPAs.

To compare the degradation of the recognition performance on the consonants and vowels, we also compute the ratios with the following formular:

$$RT = \frac{err_{Sync}}{err_{stand\_alone}} \quad (1)$$

where err_{Sync} is the error rate of a consonant/vowel in EMA-synchronized speech, err_{stand_alone} is the error rate of the same consonant/vowel in stand-alone speech. The RT value is the ratio of the error rates, which denotes the relative increasement of ther error rates. Table 3 presents the error rates of consonants in EMA-synchronized speech and stand-alone speech. One can see that the error rates of all the consonants in EMA-synchronized speech are much larger than those in stand-alone speech.

As for apical consonants "d, t, l", dorsal stops "g, k", and the tongue blade consonants "j, q, x", the presence of EMA coils affects the recognition performance more significantly than most of the other consonants. For consonants "n, z, c, s, zh, ch, sh, r, h", the recognition performance also degrades significantly, but not as much as that in the sounds "d, t, l, g, k, j, q, x".

Table 3 Error rates of consonants in EMA-synchronized speech and stand-alone speech. (unit:%)

phone	IPA	Sync	Stand-alone	RT
b	/p/	30.90	1.132	27.30
p	/pʰ/	44.83	8.57	5.23
m	/m/	23.08	2.17	10.70
f	/f/	38.27	0.47	80.72
d	/t/	47.31	0.54	86.98
t	/tʰ/	76.82	3.13	24.58
n	/n/	30.93	6.17	5.04
l	/l/	21.51	2.12	10.14
j	/tɕ/	24.43	3.05	8.00
q	/tɕ/	38.78	5.11	7.58
x	/ɕ/	48.46	1.25	38.67
z	/ts/	38.81	8.11	4.79
c	/tsʰ/	68.75	16.94	4.06
s	/s/	74.38	23.29	3.197
zh	/tʂ/	32.08	11.67	2.75
ch	/tʂʰ/	40.63	20.70	1.96
sh	/ʂ/	13.21	4.00	3.307
r	/ɻ/	13.89	3.03	4.58
g	/k/	15.50	0.98	15.83
k	/kʰ/	41.20	4.19	9.82
h	/x/	12.90	2.45	5.28

Table 4 presents the error rates of vowels in EMA-synchronize speech and stand-alone speech. One can see that the error rates of all the vowels in EMA-synchronized speech are much larger than those in stand-alone speech.

As labial consonants "b, p, m, f", the EMA coils attached to the upper and lower lips do not affect the formation of constriction in the lip region, but the wires and coils may affect the subject's articulation. Hence, the acoustic characteristics of EMA-synchronized speech changed significantly.

The RTs of "i, ix, u" are significantly larger than those of "iy, v, e, o, a". The reason may be that the EMA coils happed to lie on the locations where the narrow portions are formed in the vocal tract for sound "i, ix, u". While for sound "iy, v, e, o, a", the reason may be that the coils are either just near the location of the narrow portions or the coil size is small compared with the size of narrow portion.

Table 4 Phone error rates of vowels in EMA-synchronized speech and stand-alone speech.

phone	IPA	Sync	Stand-alone	RT
i	/i/	28.10	2.33	12.05
v	/y/	30.20	6.45	4.68
ix	/ɿ/	49.77	4.46	11.15
iy	/ʅ/	19.62	3.40	5.77
u	/u/	24.03	1.86	12.91
o	/o/	45.26	13.54	3.34
e	/ɚ/	10.49	2.01	5.22
a	/a/	50.99	6.40	7.97

4.3 Confusion matrices of vowels and consonants

To analyze the details of recognition errors, we compute the confusion matrix of vowels and consonants for EMA-synchronized speech and stand-alone speech, respectively.

The confusion matrices of vowels are presented in Fig 3 and Fig 4 for EMA-synchronized speech and stand-alone speech, respectively. In these figures, the labels in vertical direction are the ground truth, and the labels in the horizontal direction are the recognition results. And the label 'else' in the figures means that the vowel is recognized as consonants or compound finals. As demonstrated by the color bar, the blue color stands for 0%, and the yellow color denotes 100%

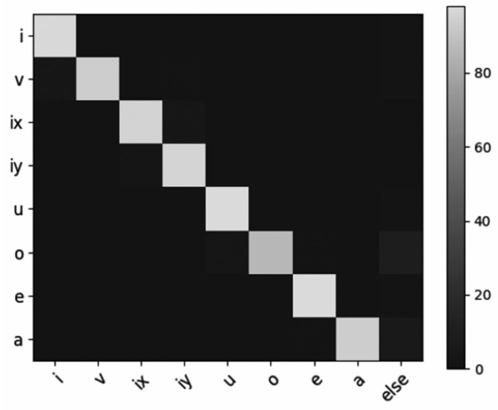

Fig. 3 Confusion matrix of vowels in stand-alone speech. (unit:%)

Fig 3 presents the confusion matrix of vowels in stand-alone speech. One can see that most of the vowel are recognized correctly. We also notice that a noticeable portion of "v" tokens are recognized as "i", a noticeable portion of "o" tokens are recognized as "u" and 'else', a noticeable portion of "a" tokens are recognized as 'else', and obvious confusion between "ix" and "iy".

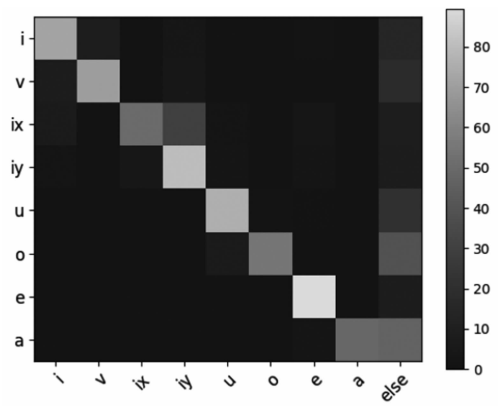

Fig. 4 Confusion matrix of vowels in EMA-synchronized speech. (unit:%)

Fig 4 presents the confusion matrix of

vowels in EMA-synchronized speech. One can see that Fig 4 is more colorful than Fig 3, which means that a larger portion of the tokens belong to each vowel are recognized as other vowels or ‘else’ . For each vowel, a noticeable portion of tokens are recognized to be ‘else’, especially vowel “o” and “a” . And more obvious confusion among vowels can be observed. The confusion between “i” and “v” is almost symmetrical. The portion of tokens of “i” recognized as “v” nearly equals to the portion of tokens of “v” recognized as “i” . However, the confusion between other vowel pairs is not symmetrical. For example, there are less “iy” tokens recognized as “ix” than “ix” recognized as “iy” . Generally, apical vowels tend to be recognized as vowels more posterior in the vowel chart.

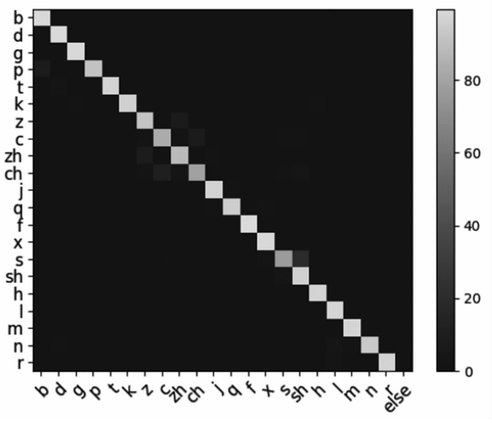

Fig. 5 Confusion matrix of consonants in stand-alone speech. (unit: %)

The confusion matrices of consonants are presented in Fig 5 and Fig 6 for EMA-synchronized speech and stand-alone speech, respectively. In these figures, the label ‘else’ means that the consonant is recognized as vowels or compound finals.

Fig 5 presents the confusion matrix of consonants in stand-alone speech. One can see that most of the consonants are recognized correctly. We also find that some of the “p”

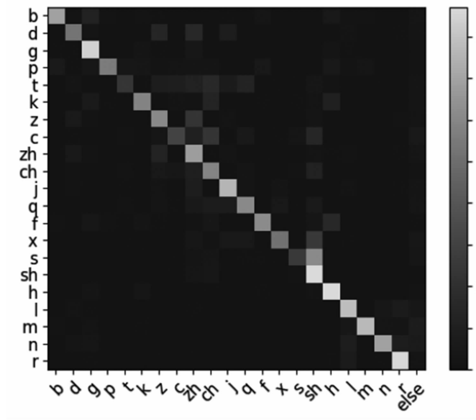

Fig. 6 Confusion matrix of consonants in EMA synchronized speech. (unit: %)

tokens are recognized as “b”, some of the “s” tokens are recognized as “sh”, and nearly symmetrical confusion between “c” and “ch”, and between “z” and “zh” .

Fig 6 presents the confusion matrix consonants in EMA-synchronized speech. We also find that Fig 6 is much more colorful than Fig 5. This indicates a larger portion of consonant tokens are recognized as other consonants or ‘else’ . Some detail results are also shown: i.) a noticeable portion of apical-alveolar stop tokens are recognized as apical-post alveolar consonants, such as “d” tokens are recognized as “z, zh, j”, and “t” tokens are recognized as “c, ch, j, q, z, zh”; ii.) obvious confusion can be found among voiceless blade-post alveolar consonants, for example “q” tokens are recognized as “zh, ch, j, x, sh, z”; iii.) salient confusion are observed among voiced consonants, for example “l” tokens are recognized as “n, r”; iv.) no obvious tendency are found for labial consonants, a noticeable portion of “f” tokens are recognized as “g, h, k”, a few of “b” tokens are mixed up with “p”, and a few of “m” tokens are mixed up with ‘else’; v.) few consonant tokens are recognized to be ‘else’ .

The above findings suggests that the presence of EMA coils affects the acoustic properties arise from two folds: i.) vowels and consonants are confused with the sounds who use the same articulator or nearby articulators, such as confusion among apical vowels and confusion among apical and blade consonants; ii.) some vowels and consonants are recognized as compound finals.

5. CONCLUSION

In this study, we record a corpus of parallel EMA-synchronized and stand-alone speech, and check both of them manually to make sure that the utterances are correct with reference to the recording prompts. The comparison between the EMA-synchronized speech and stand-alone speech is conduced based on the output of a conformer speech recognizer.

It is found that: i.) the general error rate of EMA-synchronized speech is much higher than that of stand-alone speech; ii.) apical vowels and apical/blade consonants are more significantly affected by the presence of EMA coils; iii) parts of vowels and consonants are confused with the sounds who use the same or nearly articulator articulator, such as confusion among apical vowels and confusion among apical and blade consonants; iv) the confusion of labial tokens demonstrates a diverse pattern.

This suggests that the presence of EMA coils affects both the precision and the natural pattern of articulation, hence, affects the characteristics of EMA-synchronized speech though these utterances are perceptually correct with reference to the prompts.

6. ACKOWNLEDGEMENTS

This workis supported by the National Natural Science Foundation of China (No. 61977049), Advanced Innovation Center for Language Resource and Intelligence (KYR17005), Innovation Program of Chinese Academy of Social Science.

REFERENCES

Aryal, S. , & Gutierrez-Osuna, R. (2016). "Data driven articulatory synthesis with deep neural networks." *Computer Speech and Language*, 36: pp. 260 – 273.

Chiu, C. , Wei, P. -C. , Noguch, M. et al. (2020) . is "Sibilant fricative merging in taiwan mandarin: An investigation of tongue postures using ultrasound imaging." *Language and speech,* 63(4): pp. 877 – 897.

Engwall, O. (2003). "Combining MRI, EMA and EPG measurements in a three-dimensional tongue. model." *Speech Communication,* 41: pp. 303 – 329.

Fang, Q. , Wei, J. , Hu, F. , et al. (2013) . Estimating the position of mistracked coil of EMA Data using GMM-based methods. *In proceeding of the* 2013 *Annual Summit and Conference of Asia Pacific Signal and Information Processing Association.*

Gulati, A. , Qin, J. , Chiu, C. , et al. (2021). Conformer: Convolution-augmented Transformer for Speech Recognition. *In proceeding of Interspeech* 2020.

Guo, P. , Boyer, F. , Chang, X. , et al. (2021). Recent Developments on ESPNET Tookit Boosted by Conformer. *In proceeding of* 2021 *IEEE International Conference on Acoustics, Speech and Signal Processing.*

Lammert, A. , Proctor, M. , & Narayanan, S. (2010). Data-driven analysis of realtime vocal tract MRI using correlated image regions. *In proceeding of Interspeech* 2010.

Liu, P. , Yu Q. , Wu, Z. , et al. (2015). A Deep Reccurent Approach for Acoustic-to-Articulatory Inversion. *In proceeding of* 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing.*

Liu, Z. , Ling, Z. , & Dai, L. (2018). Articulatory-to-Acoustic Conversion Using BLSTM-RNNs with Augmented Input Representation. *Speech Communication,* 99: pp. 161 – 172.

Meenakshi, N. , Yarra, C. , Yamini, B. K. , et al (2014). Comparison of speech quality with and without sensors in electromagnetic articulograph AG 501 recording. *In proceeding of*

Interspeech 2014.

Qin, C. , & Carreira-Perpinan, M. A. (2010). Reconstructing the full tongue contour from EMA/X-Ray microbeam. *In proceeding of* 2010 *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Richmond, K. (2002). *Estimating articulatory. parameters from the acoustic speech signal*. PhD diss. , University of Edinburgh.

Tobing, P. L. , Toda, T. , H. Kameoka, et al. (2016). Acoustic-to-Articulatory Inversion Mapping based on Latent Trajectory Gaussian Mixture Model. *In proceeding of Interspeech* 2016.

Udupa, S. , Illa, A. , & Ghosh, P. (2022). Streaming model for Acoustic to Articulatory Inversion with transformer networks. *In proceeding of Interspeech* 2022.

Vaswani, A. , Shazeer, N. , Parmar, N. , et al. (2017). Attention is All You Need. *In proceeding of NeurIPS*2017.

FANG Qiang

Ph. D, Associate Professor at Institute of Linguistics, Chinese Academy of Social Sciences. His research interest includes speech production, speech inversion, speech recognition, and speech syntheis.

E-mail：fangqiang@ cass. org. cn

[本文原载《中国语音学报》第20辑，2023年]