Multi-Stage Audio-Visual Fusion for Dysarthric Speech Recognition With Pre-Trained Models

Chongchong Yu, Xiaosu Su, and Zhaopeng Qian

***Abstract*— Dysarthric speech recognition helps speakers with dysarthria to enjoy better communication. However, collecting dysarthric speech is difficult. The machine learning models cannot be trained sufficiently using dysarthric speech. To further improve the accuracy of dysarthric speech recognition, we proposed a Multi-stage AV-HuBERT (MAV-HuBERT) framework by fusing the visual information and acoustic information of the dysarthric speech. During the first stage, we proposed to use convolutional neural networks model to encode the motor information by incorporating all facial speech function areas. This operation is different from the traditional approach solely based on the movement of lip in audio-visual fusion framework. During the second stage, we proposed to use the AV-HuBERT framework to pre-train the recognition architecture of fusing audio and visual information of the dysarthric speech. The knowledge gained by the pre-trained model is applied to address the overfitting problem of the model. The experiments based on UASpeech are designed to evaluate our proposed method. Compared with the results of the baseline method, the best word error rate (WER) of our proposed method was reduced by 13.5% on moderate dysarthric speech. In addition, for the mild dysarthric speech, our proposed method shows the best result that the WER of our proposed method arrives at 6.05%. Even for the extremely severe dysarthric speech, the WER of our proposed method achieves at 63.98%, which reduces by 2.72% and 4.02% compared with the WERs of wav2vec and HuBERT, respectively. The proposed method can effectively further reduce the WER of the dysarthric speech.**

***Index Terms*— Dysarthric speech recognition, pre-training and fine-tuning, multi-stage audio-visual fusion.**

I. Introduction

DYSARTHRIA is a neurological and muscular disorder which results in paralysis, weakened contractility and imprecise or uncoordinated movement of muscles. Speech subsystems including respiration, phonation, resonance, prosody and articulation are thus affected [1]. Dysarthria results in inaccurate pronunciation, slow speech and low voice and intelligibility. This disorder hampers dysarthric speakers' communication with other people, making it inefficient and inconvenient. Automatic speech recognition (ASR) makes communication of dysarthric speakers much more efficient [2]. Related studies thus abound. For better accuracy, researchers try to improve the acoustic model [3], acoustic feature extraction [4] or language/lexical model [5]. As collecting dysarthric speech is difficult, only limited resources are available. The machine learning models are insufficiently trained. With few speakers available for data collection, models are highly speaker-dependent.

To tackle the insufficient training of models, researchers have tried data enhancement. As shown in previous literatures [6], [7], [8], researchers have used normal speech to generate dysarthric speech. As the generated speech is similar to dysarthric speech both acoustically and perceptually, the approach makes up for the lack of dysarthric data. However, the approach cannot sufficiently improve models' generalization and as the rules are highly dependent on field knowledge, models cannot apply among multiple datasets. For the heavy reliance on individual speakers, LHUC (learning hidden unit contributions) [9], [10] proposed in 2014 and 2016 centers on speaker adaptation. One key breakthrough is for models to learn speaker-specific hidden unit contributions to improve the recognition of different speakers. Recently, audio-visual speech recognition (AVSR) has been proposed. According to the McGurk effect [11], people's speech perception is influenced by visual information. AVSR incorporates visual information and thus improves the accuracy of ASR [12], [13], [14]. The speech articulated is the coordinated result of vocal organs with most obvious contribution from the tongue, lip, teeth and nose [15]. The motor data of these organs have also been used in the automatic recognition of dysarthric speech [16], achieving satisfactory results. However, it is costly to use sensors to collect such data. We propose instead to collect facial signals with cameras and use the information collected for visual fusion.

Compared with traditional fusion approaches, our approach offers more and effective visual information. However, data scarcity still leads to low generalization of the models trained. In this aspect, transfer learning is the most applied and most effective solution [17], [18]. Hernandez et al. [19] proposed to use the pre-trained models by self-supervised learning to address the overfitting problem of the model due to the limited samples. The pre-trained audio-visual fusion models based on self-supervised learning put forward in recent years, for instance AV-HuBERT [20], substantially improve the accuracy

Manuscript received 22 October 2022; revised 28 January 2023 and 17 March 2023; accepted 21 March 2023. Date of publication 27 March 2023; date of current version 31 March 2023. This work was supported in part by the Humanity and Social Science Youth Foundation of Ministry of Education of China under Grant 21YJCZH117 and in part by the Humanities and Social Sciences Research Planning Fund of the Ministry of Education of China under Grant 21A10011003. *(Corresponding author: Zhaopeng Qian.)*

The authors are with the School of Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China (e-mail: chongzhy@vip.sina.com; mysxs@foxmail.com; qianzhaopeng@btbu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3262001

of lip reading and acoustic fusion. In this case, we propose a novel framework Multi-stage AV-HuBERT (MAV-HuBERT) to further improve the accuracy of dysarthric speech recognition.

In summary, our main contributions in this paper are as follows: 1) we propose the MAV-HuBERT framework to further improve the accuracy of dysarthric speech recognition; 2) in the MAV-HuBERT framework, two fusion stages including the visual information fusion and the audio-visual fusion are as follows: During the first stage the convolutional neural networks (CNN) is used to encode the facial speech function areas and during the second stage the AV-HuBERT is used to get the pre-trained model for fusing the acoustic and visual information; 3) the increment of data and ablation experiments are designed to evaluate the performance of our proposed framework.

II. Related Works

This section introduces the history of how to improve the accuracy of dysarthric speech recognition. We also introduce the audio-visual fusion models and self-supervised pre-training models.

A. ASR for Dysarthric Speech

Difficulties abound when ASR is applied to recognize dysarthric speech. Researchers have explored improvements among which the improvements based on acoustic feature extraction, the acoustic model or the language/lexical model are most common. For improvements based on acoustic feature extraction, Yalmaz et al. [4] proposed to use articulatory and bottleneck features to reduce the acoustic space difference due to speakers' different articulatory capabilities. Through convolutional restricted Boltzmann machine, Takashima et al. [21] tackled the local overfitting problem when a CNN with bottleneck is used get the pre-trained model. For improvement based on the acoustic model, Shahamiri et al. [3] proposed an artificial neural network of multi-view learning to reduce speech variation among dysarthric speakers. Bhat et al. [22] used multi-window spectral estimation and speaker self-adaptation to improve the accuracy of dysarthric speech recognition. Kim et al. used the Kullback-Leibler (KL) divergence between hidden Markov models [23] and convolutional recursive long short-term memory neural network [24] to recognize dysarthric speech. For improvements based on the language/lexical model, Takashima et al. [25] proposed an end-to-end speech recognition framework based on Listen, Attend and Spell [26]. The framework [25] includes an English and a Japanese model. Experiments show that using multiple databases for speech recognition as in this framework has its advantages [25]. Improvements based on language feature extraction, the acoustic model and the language/lexical model are not separated but complementary in making dysarthric speech recognition more accurate.

B. AVSR for Dysarthric Speech

Thanks to the bimodal nature of language perception and the successful application of AVSR in recognizing normal speech [27], [28], the researchers can take advantage of the visual information to improve the recognition of dysarthric speech. Also, dysarthric speakers themselves can have a better understanding of their articulatory difficulties and take targeted remedial trainings. Compared with traditional ASR with only audio information, AVSR has better robustness and accuracy. Liu [29] proposed an AVSR framework in dealing with the disorder speech recognition based on the Bayesian gated neural network which better fuses the audio and visual modalities than its baseline model (deep neural networks-based ASR). Salama et al. [30] explored using discrete cosine transform (DCT) to model the mouth area as the visual feature and applying the visual feature to recognize dysarthric speech. Miyamoto et al. [31] proposed a novel framework for dealing with the dysarthric speech. In this framework [31], the multiple acoustic frames are used as an acoustic feature to solve the problem that the degradation of speech recognition is caused by strain on speech-related muscles. In addition, an active appearance model is used to solve the problem for people with articulation disorders resulting from athetoid cerebral palsy. Insufficient audiovisual data is the bottleneck of applying AVSR to dysarthric speech. Researchers have thus developed an approach for cross-field generation of visual features [32]. LRS2, the lip-reading dataset, was used to build the audio-visual inversion system, generating visual features based on UASpeech's audio data. In this respect, Liu et al. [12] resorted to cross-field generation of visual features to effectively reduce the word error rate (WER).

C. Pre-Trained Models Based on Self-Supervised Learning

Normal speech recognition requires thousands of hours of audio data for the training. In this case, the scarcity of audio data for dysarthric speech leads to insufficient training of model. Self-supervised learning can effectively address the lack of annotated data. Through self-supervised learning, the network is pre-trained on large-scale unmarked corpus and then applied to downstream tasks. This scenario is especially successful in natural language processing [15], [33] and is also a dynamic research field of computer vision [34], [35]. The wav2vec 2.0 put forward by Baevski et al. [36] proves that speech recognition is feasible with limited annotated data. HuBERT proposed by Hsu et al. [37] is a self-supervised speech representation learning approach, which utilizes an offline clustering step to provide aligned target labels for a BERT-like prediction loss. In the task of dysarthric speech recognition, Hernandez et al. [19] tried to improve the accuracy through pre-training on dysarthric datasets. They used wav2vec, HuBERT and the cross-lingual data to train the acoustic model and designed the experiments based on UASpeech (English), PC-GITA (Spanish) [38] and EasyCall corpus (Italian) [39]. The experimental results [19] show that the pre-trained model based on large-scale unmarked data can effectively reduce the WER.

III. Methods

We designed a multi-stage fusion framework (MAV-HuBERT, shown in Fig. 1) to take advantage of the audio

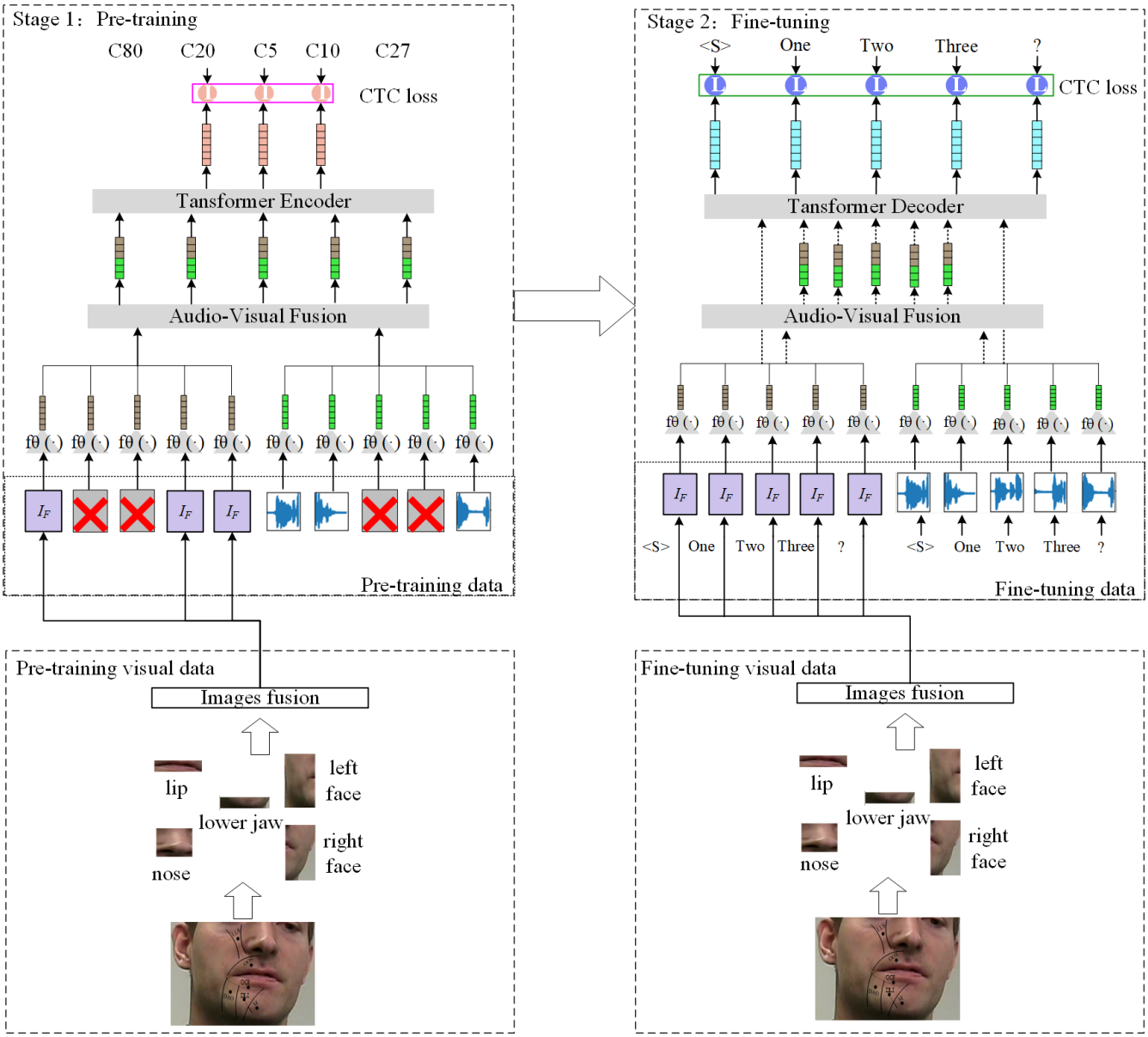


Fig. 1. Fusion Framework; *Cn*: Audio-visual clusters; X: mask.

and visual information of the dysarthric speech. During the first stage, the motor visual information of different facial speech function areas are fused; during the second stage, the pre-trained framework fusing the audio and visual information is applied to obtain the knowledge of audio-visual information of the speech.

In Fig. 1, the pre-training and fine-tuning stage are shown on the left and right respectively. During pre-training, the inputs are the audio and visual information and the visual information is obtained by the fusion of facial speech function areas. The acoustic and visual features are extracted separately. After that, the audio-visual fusion information is encoded by the fusion framework of MAV-HuBERT. Finally, the fusion features are decoded by Transformer. In addition, masking of input is used to improve the contextual presentation performance of the model. By masking the intermediate information, the model can capture the semantic association relationships between consecutive frames of actions during pre-training. During the fine-tuning, the inputs are the audio and visual information (fused by facial speech functional areas).

A. Fusion of Visual Articulatory Information

The fusion framework is shown in Fig. 2. The images are divided from videos at the frequency rate of 25 Hz per frame. And we undertake a frame-by-frame visual fusion. In this paper, we choose 5 areas of the lip (mouth), lower jaw, left and right cheeks and nose as the facial speech function areas for visual fusion, as shown in Fig. 3.

The face detector of Dlib toolkit is used to extract the images of facial speech function areas. Dlib incorporates a framework based on gradient boosting to learn learning an ensemble of regression trees. It extracts the regressed facial coordinates from input images, achieving super-real performance of high-quality prediction [41]. The CNN-based

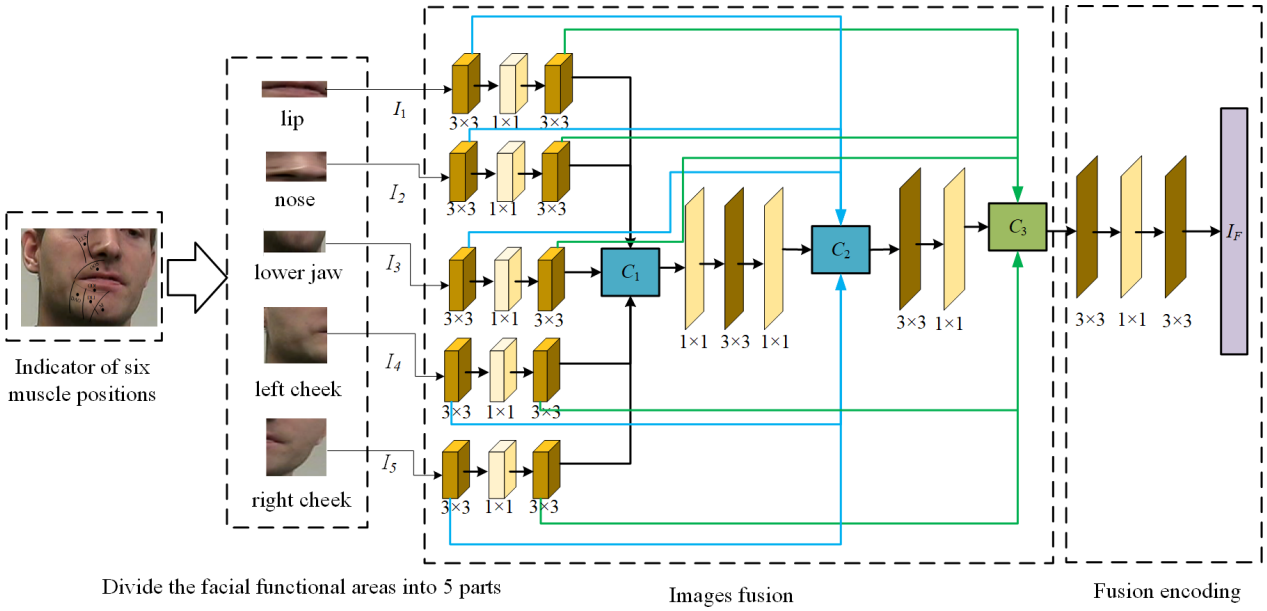
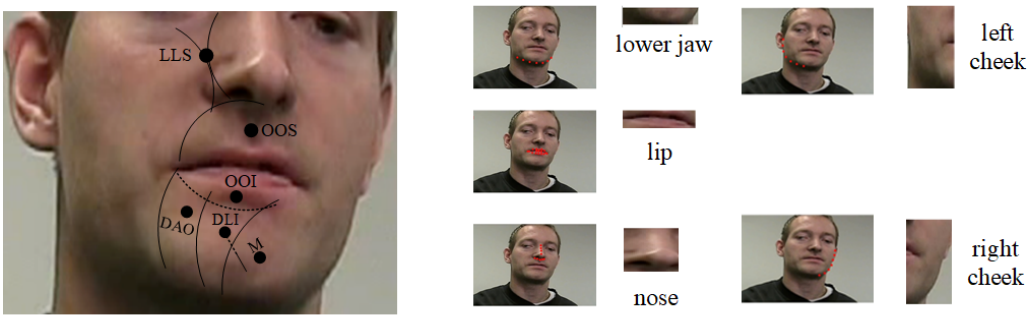


Fig. 2. Visual fusion framework of facial functional areas' movement. $I_1, \cdots, I_5$ stand for source images and I_F stands for fused images.



(a) Indicator of six muscle positions (b) Divide the facial functional areas into 5 parts

Fig. 3. Indicator of six muscle positions. LLS: levator labii superioris; OOS: orbicularis oris superior; OOI: orbicularis oris inferior; DLI: depressor labii inferioris; DAO: depressor anguli oris; M: mentalis.

architecture is used to fuse the different facial speech function areas, where the convolutional layers are defined as (1),

$$Y_i = F_k \odot F_{k-1} \cdots \odot F_2 \odot F_1 (I) = \odot_{i=1\ldots k} F_i (I) \quad (1)$$

In our visual fusion framework, the convolutional layer operates as $Y_i = F_i (X_i)$, where X_i and Y_i represent respectively the image features of the input facial speech function area and the output fusion image features of the i^{th} layer. $\odot$ denotes an element-wise multiplication operator. I is the source image. F_i represents the convolution operation of the i^{th} convolutional layer ($i = 1 \ldots k$). This CNN includes 3×3 and 1×1 convolutional kernels with the step length of 1. As the network does not use a fully-connected layer, the input images can be of any size. Apart from the last convolutional layer which is activated by tanh function, all other layers are activated by ReLU function. Our visual fusion framework constitutes of several modules including feature input, feature fusion and fusion encoding.

The feature input module includes 5 branches defined as $I_1, I_2, \cdots, I_5$. $O_j, j = 1, 2, \cdots, 5$ is the output of the convolutional layer calculated by (2). Each branch has 3 convolutional layers to deal with the different features of input images.

$$O_j = \odot_{i=1,2,3} \mathrm{Conv} \left(I_j\right), j = 1, \cdots, 5 \quad (2)$$

Conv means the convolutional layer of CNN; $\odot_{i=1,2,3}$ denotes the outputs of three convolutional layers are calculated by element-wise multiplication operator; and the outputs of 5 branches are cascaded to obtain the fusion feature c_1 as shown in (3).

$$c_1 = \textbf{cascade} \left(O_1, O_2, O_3, O_4, O_5\right) \quad (3)$$

Here **cascade** means the cascaded operation by cascaded structure. The fusion encoding operation here includes 8 convolutional layers as shown in Fig. 2, where the c_2 and c_3 are the intermediate features. The outputs of the first convolutional layer (used to calculate the c_1) and $\odot_{i=1,2,3} \mathrm{Conv} (c_1)$ are cascaded to calculate c_2 as shown in (4). The outputs O_j and $\odot_{i=4,5} \mathrm{Conv} (c_2)$ are cascaded to calculate c_3 as shown in (5). At last, we obtain the fused visual features I_F as shown

in (6).

$$c_2 = \textbf{cascade}\left(\odot_{i=1,2,3}\text{Conv}\left(c_1\right), \text{Conv}\left(I_1\right), \cdots, \text{Conv}\left(I_5\right)\right) \tag{4}$$

$$c_3 = \textbf{cascade}\left(\odot_{i=4,5}\text{Conv}\left(c_2\right), O_1, O_2, O_3, O_4, O_5\right) \tag{5}$$

$$I_F = \odot_{i=6,7,8}\text{Conv}\left(c_3\right) \tag{6}$$

Here, I_F serves as the fusion encoding vector.

B. Audio-Visual Fusion for Dysarthric Speech

In this paper, we aligned acoustic and visual features at the frame level. We extract 26-dimensional features of logarithm filter-bank energies (LFEs) as the acoustic features from the original waveform and extract images from videos. Four frames of consecutive acoustic features are integrated into one frame as the input.

To deal with the audio modality, a linear projection layer is used as the audio encoding module. Setting an audio sequence $A_{1:T}$, we undertake the operation $G(A_t), t \in \{1, 2, \ldots, V\}$ and cluster audio features into a discrete unit sequence $z^a_{1:T} = K-\text{Means}(G(A_t))$, where $G(A_t)$ is the LFEs. V represents the codebook size. For the video modality, after extracting facial images, the ResNet of hidden units are used to further deal with the fused visual features. Setting the fused visual feature sequence as $I_{1:T}$, we undertake the operation $G(I_t), t \in \{1, 2, \ldots, V\}$ and cluster visual features into a discrete unit sequence $z^i_{1:T} = K-\text{Means}(G(I_t))$, where G is the histogram of oriented gradients (HOG) of the fused visual features. The standard ResNet-18 is used as the encoder for dealing with the fused visual features.

The acoustic feature A and the visual fusion feature I are fused by a fusion layer based on the multi-layer Transformer-Encoder as shown in Fig. 1. The audio-visual fusion F is calculated as in (7).

$$F = \left\{f_1, f_2, \ldots, f_{n_t+n_i}\right\} = T_{E_M}\left(\textbf{concate}\left(A, I\right)\right) \tag{7}$$

TE_M means the Transformer-Encoder for fusing data; and **concate** denotes the channel-wise concatenation. The output of audio-visual fusion framework is the posterior probability of each input frame calculated by connectionist temporal classification (CTC) algorithm. The whole stack is pre-trained with CTC loss.

C. Pre-Training and Fine-Tuning

MAV-HuBERT is proposed to pre-train model in two major steps: audio-visual feature clustering and masked feature prediction. In the task of masked prediction, acoustic and image frames are used simultaneously to model and distill the correlation between the two modalities.

During the audio-visual clustering stage, the two modalities generating fusion cluster assignments are pre-trained, which serve as the target label for the next iteration of masked prediction. During the masking prediction stage by substitution, random segments of the same video are selected as the segments in the masked video stream.

We set the output probability as $p_{1:T}$ and target cluster assignment as $z_{1:T}$. During pre-training stage, the loss L of MAV-HuBERT is calculated by (8)

$$L = -\sum_{t\in M^a \cup M^v} \log p_t\left(z_t\right) - \alpha \sum_{t\notin M^a \cup M^v} \log p_t\left(z_t\right) \tag{8}$$

Here M^a and M^v represent the masked frames of audio and video streams, respectively; α is a hyper-parameter to weigh the contribution of unmasked regions in the overall objective.

During the fine-tuning stage, the pre-trained model needs to be fine-tuned in the task of dysarthric speech recognition. It is assumed that the feature sequence output of our pre-trained model is $e_{1:T}$, and the ground-truth transcription is $w = w_1, w_2, \ldots, w_s$. For CTC, a projection layer is used to map the input sequence onto the output probability p_t, which is calculated as in (9).

$$p_t = \text{Softmax}\left(\mathbf{W}^{ft} e_t + \mathbf{b}^{ft}\right) \tag{9}$$

$\mathbf{W}^{ft} \in \mathbb{R}^{d\times(U+1)}$ denotes weights matrix of this layer; $\mathbf{b}^{ft} \in \mathbb{R}^{U+1}$ denotes the bias vector of this layer; U is the output vocabulary size (+1 means plus one blank symbol). The model trained with CTC loss L_{ctc} is calculated as in (10).

$$L_{ctc} = -\log \sum_{\pi\in\mathcal{B}^{-1}(w)} p\left(\pi \mid e_{1:T}\right) \tag{10}$$

Here $\mathcal{B}$ maps an alignment sequence from π to w. It's assumed that the output probability per frame is y and $y^t_{\pi_t}$ represents the probability of the output π_t at the moment t. The output sequence probability is calculated as in (11).

$$p\left(\pi \mid e_{1:T}\right) = \prod_{t=1}^{T} y^t_{\pi_t} \tag{11}$$

Finally, we use a 4-gram language model, whose perplexity on the test set is 110.5, for the decoding. The text materials for training the language model are used from the LRS3 dataset. In particular, the beam width is tuned among {5, 10, 20, 50, 100, 150}, the language model weight, among {0, 1, 2, 4, 8} and word insertion penalty, among {±4, ±2, ±1, 0}.

IV. Experiment

A. Data Preparation

We draw our data from UASpeech [42], an English dataset supported by the University of Illinois. UASpeech included 102.7 hours of records from 29 speakers pronouncing individual words. Each speaker pronounces 765 words which are divided into 3 blocks.

The phonetic intelligibility score (ranging from 2% to 95%) of UASpeech is calculated based on the average performance in the listening test. Dysarthric speakers are thus divided into 4 intelligibility groups: 0-25%, 25-50%, 50-75%, and 75-100%. The 4 groups correspond to extremely severe, severe, moderate and mild dysarthria. Table I shows the intelligibility score of all available speakers in UASpeech. In addition, we use block 1 and block 3 as the training data and block 2 from UASpeech as the test data.

TABLE I
INTELLIGIBILITY SCORE AND DYSARTHRIA SEVERENESS OF UASPEECH SPEAKERS

Speaker Label	Audio	Video	Speech Intelligibility Score (%)	Dysarthric Severeness
M01	✓	×	15	Extremely Severe
M04	✓	×	2	Extremely Severe
M05	✓	✓	58	Moderate
M07	✓	✓	28	Severe
M08	✓	✓	93	Mild
M09	✓	✓	86	Mild
M10	✓	✓	93	Mild
M11	✓	✓	62	Moderate
M12	✓	✓	7.4	Extremely Severe
M14	✓	✓	90.4	Mild
M16	✓	✓	43	Severe
F02	✓	✓	29	Severe
F03	✓	×	6	Extremely Severe
F04	✓	✓	62	Moderate
F05	✓	✓	95	Mild

TABLE II
CLASSIFICATION OF VOWEL PHONES IN THE TEST SET

Classification	Phones
Vowels	AA-/ɑ/, AE-/æ/, AH-/ʌ/, AO-/ɔ/, AW-/aʊ/, AY-/aɪ/, EH-/ɛ/, ER-/ɝ/, EY-/eɪ/, IH-/ɪ/, IY-/i/, OW-/oʊ/, OY-/ɔɪ/, UH-/ʊ/, UW-/u/

TABLE III
CLASSIFICATION OF CONSONANT PHONES IN THE TEST SET

Classification		Phones
Consonants	Stops	P-/p/, B-/b/, T-/t/, D-/d/, K-/k/, G-/g/
	Affricates	CH-/tʃ/, JH-/dʒ/
	Fricatives	F-/f/, V-/v/, S-/s/, Z-/z/, SH-/ʃ/, ZH-/ʒ/, TH-/θ/, DH-/ð/, HH-/h/, R-/ɹ/
	Nasals	M-/m/, N-/n/, NG-/ŋ/
	Laterals	L-/l/
	Zero Initials	W-/w/, Y-/j/

B. Experiment Setup

We use word error rate (WER) to assess the experiment of dysarthric speech recognition. WER is an important metric for the performance of speech recognition and reveals the error rate of predicted texts as compared with the transcripts. Therefore, lower WER means better performance. WER is based on Levenshtein distance and calculated on the level of words instead of phonemes. WER is calculated as in (12). In addition, the Kaldi toolkit is used in our experiments.

$$\mathrm{WER} = \frac{\mathrm{S}+\mathrm{D}+\mathrm{I}}{\mathrm{N}} = \frac{\mathrm{S}+\mathrm{D}+\mathrm{I}}{\mathrm{S}+\mathrm{D}+\mathrm{C}} \tag{12}$$

S stands for the number of substitutions, D for the number of deletions, I for the number of insertions, C for the number of correct words and N for the number of words in the reference.

C. Vowels and Consonants

The test set includes 39 phones (15 vowels and 24 consonants). Table II illustrates the vowel phones and Table III, the consonant phones.

V. RESULTS

We designed the comparative, data increment, multi-stage fusion and speaker dependency experiment to evaluate the performance of our proposed method. In particular, the data increment experiment is used to explore the performance of our proposed method under limited amount of training data; the speaker dependency experiment is used to evaluate whether our proposed method can overcome speaker restriction. During our testing stage, the mild, moderate, severe, and extremely severe dysarthric speech are used, respectively.

A. Experiment of AVSR With Pre-Trained Models for Dysarthric Speech

For dysarthric speech recognition, we use MAV-HuBERT to pre-train models on the UASpeech dataset. The results are shown as in Table IV. Wav2vec 2.0 and HuBERT serve as the baseline systems. We use data mixed from LRS3 [43] and UASpeech in the pre-training of MAV-HuBERT with a mixing ratio of 8:1. Then we use UASpeech data for fine-tuning. The audio-visual speech recognition method refers to the audio-visual fusion part in MAV-HuBERT.

Results of Table IV show that the MAV-HuBERT pre-trained by the mixed training data from the LRS3 and UASpeech achieves the best results of dysarthric speech recognition. The best WERs of our proposed method are 6.05% (mild dysarthric speech), 22.8% (moderate dysarthric speech), 30.77% (severe dysarthric speech) and 63.98% (extremely dysarthric speech), respectively. Compared with the results of only using audio modality, the WERs of our proposed method for mild, moderate, severe and extremely severe dysarthric speech all exhibited a reduction. For instance, compared with the results of wav2vec (audio-only), the WERs of our proposed method were reduced by 0.65%, 13.5%, 2.53% and 2.72%. Similarly, compared with the results of HuBERT, the WERs of our proposed method reduced by 0.15%, 0.6%, 5.23% and 4.02%.

B. Experiment With Increment Data

To evaluate how the performance of our proposed method is influenced by the data amount, we used the training data with different amount (30 hours of LRS3, 433 hours of LRS3, mixed data from LRS3 and UASpeech (10:1), mixed data from LRS3 and UASpeech (8:1)) in our experiment. Then we use UASpeech data for fine-tuning. Our experiment is designed based on the audio, visual and audio-visual modalities.

Wav2Vec 2.0 and HuBERT only use audio modality of dysarthric speech for pre-training. When using normal speech for pre-training, AV-HuBERT has a worse performance than the baseline. When adding dysarthric speech to the pre-training, AV-HuBERT sees its WER drop substantially. AV-HuBERT has a better performance using data from LRS3: UASpeech=8:1 than from LRS3: UASpeech=10:1. In the audio-visual fusion condition, compared to the results using a data rate of 10:1, the WERs on mild, moderate, severe, and extremely severe dysarthric speech using a data rate of 8:1 were reduced by 1.09%, 2.24%, 1.75%, and 2.67%, respectively.

TABLE IV
EXPERIMENT RESULTS FOR AVSR OF DYSARTHRIC SPEECH

Systems	Datasets	Modality	WER% / Standard Deviations			
			Mild	Moderate	Severe	Extremely Severe
wav2vec 2.0 [36]	UASpeech	audio-only	6.7	36.3	33.3	66.7
HuBERT [37]	UASpeech	audio-only	6.2	23.4	36.0	68.0
MAV-HuBERT	Mixed Data (LRS3: UASpeech = 8: 1)	audio-only	8.24/1.7	26.10/1.9	32.03/3.2	68.15/5.7
		video-only	7.58/1.7	24.54/1.9	36.25/4.0	72.84/6.3
		audio-video	**6.05/1.7**	**22.80/1.9**	**30.77/3.2**	**63.98/5.7**
	UASpeech	audio-only	17.98/1.7	35.20/1.7	36.73/2.8	70.22/4.5
		video-only	15.78/1.7	38.24/1.7	38.24/2.8	75.15/4.5
		audio-video	11.07/1.7	30.05/1.7	34.03/2.8	66.78/4.5

TABLE V
RESULTS OF THE DATA INCREMENT EXPERIMENT

Systems	Datasets	Modality	WER% / Standard Deviations			
			Mild	Moderate	Severe	Extremely Severe
wav2vec 2.0 [36]	UASpeech	audio-only	6.7	36.3	33.3	66.7
HuBERT [37]	UASpeech	audio-only	6.2	23.4	36.0	68.0
MAV-HuBERT	LRS3(30h)	audio-only	26.34	33.10	46.34	89.65
		video-only	33.82	34.86	47.63	93.54
		audio-video	24.55/1.5	33.00/2.5	42.74/4.7	89.32/10.5
	LRS3(433h)	audio-only	23.13	33.60	49.96	91.39
		video-only	31.47	34.86	48.98	93.32
		audio-video	23.21/0.8	33.00/2.8	47.32/5.2	90.63/13.5
	LRS3：UASpeech=10:1	audio-only	8.79	29.40	33.64	69.73
		video-only	7.93	28.51	38.24	72.24
		audio-video	7.14/1.7	25.04/2.2	32.52/3.8	66.65/7.5
	LRS3：UASpeech=8:1	audio-only	8.24/1.7	26.10/1.9	32.03/3.2	68.15/5.7
		video-only	7.58/1.7	24.54/1.9	36.25/4.0	72.84/6.3
		audio-video	**6.05/1.7**	**22.80/1.9**	**30.77/3.2**	**63.98/5.7**
		audio-only	40.21/15.4			
		video-only	44.92/18.0			
		audio-video	41.23/15.4			

TABLE VI
EXPERIMENT RESULTS OF FUSING FACIAL FUNCTIONAL AREAS

visual inputs	Modality	WER% / Standard Deviations			
		Mild	Moderate	Severe	Extremely Severe
visual fusion	video-only	**7.58/1.7**	**24.54/1.9**	**36.25/4.0**	72.84/6.3
	audio-video	**6.05/1.7**	**22.80/1.9**	**30.77/3.2**	**63.98/5.7**
lip	video-only	13.42/1.9	26.80/1.9	42.71/4.5	**71.34/7.0**
	audio-video	8.08/1.9	26.00/1.9	30.82/3.2	64.76/5.7

Results of Table V show that the increment amount of training data can effectively improve the performance of our proposed method. When the amount of UASpeech increased, the WER of our proposed method decreased significantly. However, if our proposed method is pre-trained only using LRS3 dataset, the WER of our proposed method would increase when the amount of training data increases.

C. Experiment on Multi-Stage Fusion Model

The experimental results of motor visual information of facial speech function areas are shown in Table VI. The visual inputs are the visual fusion and lip movement, respectively. Both inputs are processed in pre-training and fine -tuning based on MAV-HuBERT. The pre-training data is the mixed data from LRS3 and UASpeech (8:1).

By comparing the methods using visual fusion information and lip information, we find that fusing facial speech function areas effectively reduces WER. In the video-only modality, the WER of using visual information is 5.84%, 1.74% and 6.46% lower than that of using lip information for mild, moderate and severe dysarthric speech. However, the WER of using visual fusion information is 0.78% higher than that of using lip information for extremely severe dysarthric speech. Speakers with extremely severe dysarthria often suffer major diseases and they move their heads excessively when speaking. All this makes it extremely difficult to capture their facial speech function areas. With audio-visual fusion, the WER of using visual fusion information is 2.03%, 3.2%, 0.05% and 0.78% lower than that of using lip information for mild, moderate, severe and extremely severe dysarthric speech, respectively. Results of Table VI show that the visual fusion operation can effectively improve the performance of AVSR in dealing with the dysarthric speech.

D. Experiment on Speaker Dependency for Dysarthric Speech

The experiment on speaker dependency is to evaluate whether our proposed method can get rid of the restriction of speaker. During the testing stage, one of the speakers (F05) with mild dysarthria was chosen. Speech of speaker F05 was not used to pre-train (or train) the MAV-HuBERT. We only used speech of speaker F05 to test our proposed method. The results can be found in Table VII.

TABLE VII
EXPERIMENT ON SPEAKER DEPENDENCY FOR DYSARTHRIC SPEECH RECOGNITION

Systems	Datasets	Modality	WER%	
			F05	Mild
MAV-HuBERT	LRS3: UASpeech=8:1	audio-only	**8.00**	8.24
		video-only	8.00	**7.58**
		audio-video	**6.00**	6.05

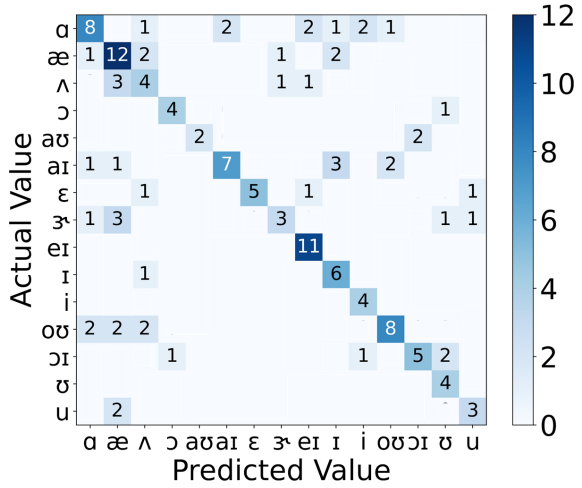


Fig. 4. Vowel confusion matrix for mild dysarthric speech recognition.

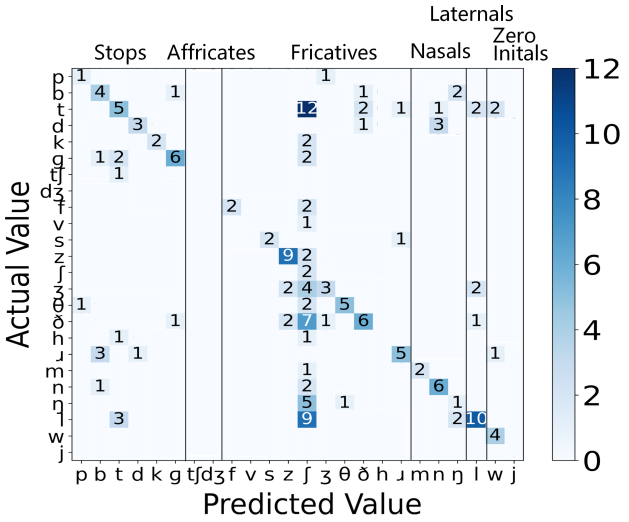


Fig. 5. Consonant confusion matrix for mild dysarthric speech recognition.

Results of Table VII show that the WERs between F05's speech and mild dysarthric speech have no significant difference. The experimental results illustrate that our proposed method enjoys the speaker independent characteristic for the mildly dysarthric speech.

E. Confusion Matrix Analysis of Pronunciations for Dysarthric Speech

We also analyzed the recognition results of confused pronunciations. The results are shown in Fig. 4 and Fig. 5.

Fig. 4 shows the vowel confusion matrix. The recognition of vowels is better than that of consonants, with that of AE-/æ/, EY-/eI/, AA-/a/ and OW-/oʊ/ being the best. Fig. 5 shows the consonant confusion matrix. The lateral L-/l/ and fricative Z-/z/ are recognized with high accuracy, while stops and fricatives are more easily confused, particularly the fricative SH-/ʃ/, which has the highest confusion rate. When the fricative /ʃ/ is pronounced, the tip of the tongue is upturned and is close to the front of the hard palate. The complicated actions of pronunciation are very hard for the speakers with dysarthria. Therefore, the fricative /ʃ/ has the lowest accuracy among all pronunciations.

VI. DISCUSSION

To further improve the accuracy of dysarthric speech recognition, we designed a novel framework MAV-HuBERT. MAV-HuBERT is designed based on self-supervised pre-training technology. Our proposed method can recognize the text content from audio-visual files in real time. The experimental results show that our proposed method significantly reduced WER compared to the baseline methods.

In the comparative experiment, In the comparative experiment, our proposed method achieved the best WER reduction of 13.5% for moderate dysarthric speech compared with the baseline method. Additionally, our proposed method achieved the best result of 6.05% WER for mild dysarthric speech. Even for the extremely severe dysarthric speech, the WER of our proposed arrives at 63.98%, which reduces by 2.72% and 4.02% compared with the WERs of wav2vec and HuBERT, respectively. Furthermore, we note that the results of baseline methods (wav2vec and HuBERT) only using audio modality are not stable. For example, the WERs of HuBERT were lower than that of wav2vec for mild and moderate dysarthric speech, while for severe and extremely severe dysarthric speech, the opposite was true. Moreover, the experimental results show that the multistage fusion performs better than the fusion of only lip visual information and acoustic information.

The results of data increment show that when the pre-training data is from LRS3, the accuracy is low. LRS3 is a dataset of normal speech, which is significantly different from the dysarthric speech, such as that from UASpeech. The difference results in the extremely large variation of data space. Therefore, the pre-trained model only using normal speech can hardly be effectively dealing with the dysarthric speech. However, adding dysarthric data during pre-training stage can improve accuracy substantially. The more dysarthric data are used during pre-training stage, the better performance of model would be.

The results of visual fusion experiment show that compared with only using lip information, using visual information of facial speech function areas can effectively improve the accuracy of dysarthric speech recognition. Although for extremely severe dysarthric speech, only using lip information has better accuracy than using the visual fusion information based on the single visual modality, the problem has been addressed by our proposed method based on audio-visual modalities. This is because the speakers with severe dysarthria (suffering from severe diseases like Parkinson disease or cerebral palsy) have excessive movements of the faces and heads when speaking. In this case, the extracted images cannot accurately reflect the

TABLE VIII
GLOSSARY OF TERMS

Acronyms	Description
wav2Vec 2.0	A Framework for Self-Supervised Learning of Speech Representations
HuBERT	A neural network framework based on deep learning to process the long-dependency time series data.
AV-HuBERT	The framework is used to pre-train the audio-visual fusion model based on the HuBERT.
Transformer	Transformer is the neural networks architecture designed based only on attention mechanism. Transformer can be used to capture the long-distance dependencies of sequential data.
tanh function	The hyperbolic tangent function is abbreviated as tanh function, which is commonly used as the activation function in deep learning models.
ReLU function	The linear rectification function is abbreviated as ReLU function, which is also commonly used as the activation function in deep learning models.
ResNet	The residual network is abbreviated as the ResNet. The internal residual block uses the skip connection architectures, which alleviate the gradient disappearance caused by the addition of depth in the deep neural network.

movement of facial muscles, leading to worse quality of visual data. However, audio-visual fusion operation can effectively address this problem.

The results of speaker-dependency experiment show that our proposed method can get rid of the restriction of speaker in dealing with the mildly dysarthric speech. However, if the speakers suffer from severe dysarthria, the generalization ability of our proposed method would be at discount. This is because the differences among the speakers with severe dysarthria are very huge. The acoustic and visual spaces would be too large to be modeled.

After analyzing the confusion matrix for consonants, the results indicate that stops and fricatives exhibit a higher confusion rate compared to other consonants. In particular, SH-/ʃ/ has the highest confusion rate. This is likely due to the precise control of tongue, teeth, and lip movements required for the pronunciation of stops and fricatives, as well as the impact of airflow from the lungs on the vocal tract. However, individuals with dysarthria cannot precisely control their tongue, lip, and teeth movements, resulting in greater confusion for stops and fricatives compared to other consonants. In addition, the experiments of our research in this paper are carried out by graphics processing unit (GPU) servers. Pre-training our proposed model requires more than 24 GB of GPU memory. In the future, we will continue to investigate methods to reduce resource consumption and enable offline deployment of the model. Our goal is to ensure that individuals with dysarthria are able to conveniently utilize the model on their mobile devices.

VII. CONCLUSION

To make up for the insufficient training due to scarce dysarthric data, we designed a MAV-HuBERT fusion structure. During the first stage, we fused the motor visual information of facial speech function areas. During the second stage, we proposed to use the pre-training framework to further fuse audio and visual information. We explored the effectiveness of MAV-HuBERT in task of dysarthric speech recognition. The experimental results show that our proposed method can effectively improve the accuracy of dysarthric speech recognition. Our proposed method has a significance of effectively incorporating the audio and visual information to recognize the dysarthric speech.

APPENDIX

See Table VIII.

ACKNOWLEDGMENT

The authors wish to extend their gratitude to Prof. Heejin Kim from the University of Illinois, USA, for graciously granting permission to access the UASpeech database utilized in this study.

REFERENCES

[1] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *J. Speech Hearing Res.*, vol. 12, no. 2, pp. 246–269, 1969.

[2] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assist. Technol.*, vol. 22, no. 2, pp. 99–112, 2010.

[3] S. R. Shahamiri and S. S. B. Salim, "A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 5, pp. 1053–1063, Sep. 2014.

[4] E. Yılmaz, V. Mitra, G. Sivaraman, and H. Franco, "Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech," *Comput. Speech Lang.*, vol. 58, pp. 319–334, Nov. 2019.

[5] M. Kim, Y. Kim, J. Yoo, J. Wang, and H. Kim, "Regularized speaker adaptation of KL-HMM for dysarthric speech recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1581–1591, Sep. 2017, doi: 10.1109/TNSRE.2017.2681691.

[6] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6009–6013.

[7] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Interspeech*, 2018, pp. 471–475.

[8] L. Zhu, Y. Nie, C. Chang, J. H. Gao, and Z. Niu, "Different patterns and development characteristics of processing written logographic characters and alphabetic words: An ALE meta-analysis," *Hum. Brain Mapping*, vol. 35, no. 6, pp. 2607–2618, Jun. 2014, doi: 10.1002/hbm.22354.

[9] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2014, pp. 171–176.

[10] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 8, pp. 1450–1463, Aug. 2016.

[11] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[12] S. Liu et al., "Recent progress in the CUHK dysarthric speech recognition system," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2267–2281, 2021, doi: 10.1109/TASLP.2021.3091805.

[13] R. Su, X. Liu, L. Wang, and J. Yang, "Cross-domain deep visual feature generation for Mandarin audio–visual speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 185–197, 2020.

[14] S. R. Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 852–861, 2021, doi: 10.1109/TNSRE.2021.3076778.

[15] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *J. Phonetics*, vol. 30, no. 3, pp. 533–553, Jul. 2002.

[16] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 947–960, May 2011.

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[18] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2010.

[19] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," 2022, *arXiv:2204.01670*.

[20] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," 2022, *arXiv:2201.02184*.

[21] Y. Takashima, T. Nakashika, T. Takiguchi, and Y. Ariki, "Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 1411–1415.

[22] C. Bhat, B. Vachhani, and S. Kopparapu, "Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation," in *Proc. Interspeech*, Sep. 2016, pp. 228–232.

[23] M. Kim, J. Wang, and H. Kim, "Dysarthric speech recognition using Kullback–Leibler divergence-based hidden Markov model," in *Proc. Interspeech*, Sep. 2016, pp. 2671–2675.

[24] M. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional LSTM neural network," in *Proc. Interspeech*, Sep. 2018, pp. 2948–2952.

[25] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6395–6399.

[26] 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 Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960–4964.

[27] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[28] J. Yu et al., "Audio-visual multi-channel recognition of overlapped speech," 2020, *arXiv:2005.08571*.

[29] S. Liu et al., "Exploiting visual features using Bayesian gated neural networks for disordered speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 4120–4124.

[30] E. S. Salama, R. A. El-Khoribi, and M. E. Shoman, "Audio-visual speech recognition for people with speech disorders," *Int. J. Comput. Appl.*, vol. 96, no. 2, pp. 51–56, Jun. 2014.

[31] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2010, pp. 517–520.

[32] S. Liu et al., "Exploiting cross-domain visual feature generation for disordered speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 711–715.

[33] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.

[34] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[35] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.

[36] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[37] W.-N. Hsu, B. Bolte, Y.-H.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[38] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. González-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 342–347.

[39] R. Turrisi et al., "EasyCall corpus: A dysarthric speech dataset," 2021, *arXiv:2104.02542*.

[40] R. Leanderson, A. Persson, and S. Öhman, "Electromyographic studies of facial muscle activity in speech," *Acta Oto-Laryngol.*, vol. 72, nos. 1–6, pp. 361–369, 1971.

[41] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[42] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, and K. Watkin, "Dysarthric speech database for universal access research," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, Brisbane, QLD, Australia, Sep. 2008, pp. 1741–1744.

[43] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," 2018, *arXiv:1809.00496*.