MULTI-MODAL ACOUSTIC-ARTICULATORY FEATURE FUSION FOR DYSARTHRIC SPEECH RECOGNITION

Zhengjun Yue[1], *Erfan Loweimi*[2], *Zoran Cvetkovic*[2], *Heidi Christensen*[1] *and Jon Barker*[1]

[1] Speech and Hearing Research Group (SPandH), University of Sheffield, UK
[2] Department of Engineering, King's College London, UK

ABSTRACT

Building automatic speech recognition (ASR) systems for speakers with dysarthria is a very challenging task. Although multi-modal ASR has received increasing attention recently, incorporating real articulatory data with acoustic features has not been widely explored in the dysarthric speech community. This paper investigates the effectiveness of multi-modal acoustic modelling for dysarthric speech recognition using acoustic features along with articulatory information. The proposed multi-stream architectures consist of convolutional, recurrent and fully-connected layers allowing for bespoke per-stream pre-processing, fusion at the optimal level of abstraction and post-processing. We study the optimal fusion level/scheme as well as training dynamics in terms of cross-entropy and WER using the popular TORGO dysarthric speech database. Experimental results show that fusing the acoustic and articulatory features at the empirically found optimal level of abstraction achieves a remarkable performance gain, leading to up to 4.6% absolute (9.6% relative) WER reduction for speakers with dysarthria.

Index Terms— Multi-modal dysarthric speech recognition, multi-stream acoustic modelling, feature fusion

1. INTRODUCTION

Dysarthria is a speech disorder caused by a neuro-motor interface disruption [1]. People with dysarthria often lack motor-control of their speech articulators resulting in abnormal speech. Due to the data sparsity, the automatic dysarthric speech recognition (ADSR) task is still challenging, performance-wise lagging far behind the mainstream ASR systems for typical speech.

Incorporating features from other modalities which contain information correlated with the audio modality, such as visual and articulatory data, has shown promise for ADSR [2–4]. Articulatory measurements captures the movements of speakers' articulators, e.g., lips and tongue, and directly model the signal in the speech production domain. Compared with acoustic representations, articulatory information has been shown to be less speaker-variant [5] and more suitable to model the coarticulation [6]. As such it can complement the acoustic information and help towards better handling the dysarthric speech.

Recent studies [4, 7] have employed pseudo dysarthric articulatory data obtained via learnt acoustic-to-articulatory mappings for ADSR. A potential drawback is that the synthetic articulations can be an inaccurate representation of the real dysarthric articulatory space as the mapping is normally learned using typical speech. Therefore, using real dysarthric articulatory data is a more reliable choice. TORGO [8] is a dysarthric speech database, containing aligned articulatory and acoustic data. The authors have conducted several studies showing the usefulness of applying both articulatory and acoustic features for the GMM-HMM and simple DNN-based acoustic modelling [3,9]. We extend this work to using recent state-of-the-art acoustic modelling architectures.

Besides, few studies have studied the optimal information fusion scheme for combining these two representations. It should be noted that concatenation of the acoustic and articulatory features in the input level, although simple, is suboptimal. That is, the acoustic and articulatory representations encode different information, in different formats and with different importance to the task. Consequently, the optimal set of filters to process each stream will be different. This necessitates pre-processing each stream individually, and then fusing the processed streams at a higher level. Direct concatenation at the input level gives rise to pre-mature information fusion and does not permit such per stream pre-processing.

In this paper, we build multi-modal acoustic-articulatory ADSR systems using convolutional, recurrent and fully-connected layers along with normalisation techniques applied by state-of-the-art ASR systems. The proposed architectures allow various fusion schemes to be investigated. As well as empirically finding the optimal fusion level, we analyse the training dynamics of the models with various fusion schemes in terms of cross-entropy loss and WER, for both typical and dysarthric speakers. Experimental results show up to 4.6% absolute (9.6% relative) performance gain for dysarthric speech, when optimal fusion scheme is employed.

ZY, JB and HC are supported by the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Grant Agreement No. 766287). EL and ZC are supported by EPSRC Project EP/R012180/1 (SpeechWave).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

2. RELATED WORK

Multi-modal ASR has received increasing attention over recent years owing to deep learning which provides an effective framework for fusing different modalities. Visual and articulatory data are among the most commonly used data modalities. Previous studies have demonstrated the benefit of employing these two features in audio-visual [10] and acoustic-articulatory ASR systems for typical speech [11, 12].

However, most of the research on dysarthric speech recognition has solely focused on applying the acoustic representations. Research on multi-modal ADSR has been limited by the lack of parallel multi-modal data in the dysarthric domain. Visual features were first introduced for dysarthric speech in the UASpeech [13] dysarthric corpus in [2] and shown to be beneficial towards elevating the recognition performance.

Utilising the acoustic and articulatory features jointly, has also led to achieving higher performance in the ADSR task [4, 7]. Due to the limited amount of dysarthric articulatory data, synthetic (often referred to as pseudo) articulatory data has been commonly used to support acoustic features for improving acoustic modelling of dysarthric speech [4, 7]. By learning the mapping from acoustic to articulatory features, the synthesiser estimates the articulatory data from the acoustic representations. *Gnuspeech* [14] and TADA [15] are the two popular articulatory data synthesisers.

The potential drawback of synthetic articulatory features is that they might not represent the real dysarthric articulatory space effectively. That is, the synthesisers are normally trained on typical speech and then applied to generate dysarthric speech. Given the substantial differences between the typical and dysarthric speech signals, there is a notable mismatch between the acoustic-articulatory mappings for typical and dysarthric speech. Therefore, using acoustic-articulatory mappings learned for typical speech is suboptimal and accompanied with a significant error. However, using the real dysarthric articulatory data is devoid of such error and is a more reliable speech representation.

TORGO is a widely used dysarthric speech dataset [8], providing aligned acoustic and articulatory data for both typical and dysarthric speech. There are several TORGO-based studies which employ its articulatory data, e.g., [3, 9, 16] which illustrate that deploying articulatory information alongside acoustic features results in significantly higher performance. However, most of this work is based on GMM-HMM or simple DNN acoustic models. Whether the articulatory information is still beneficial in state-of-the-art acoustic models and optimal fusion scheme are remained to be explored.

3. THE PROPOSED SYSTEM

Fig. 1 depicts the structure of the proposed acoustic-articulatory multi-modal speech recognition system. The left side (pink) illustrates the articulatory data pre-processing. The right side (blue) is the proposed multi-stream acoustic model where each stream is first pre-processed and then, the fused streams are post-processed before reaching the output layer.

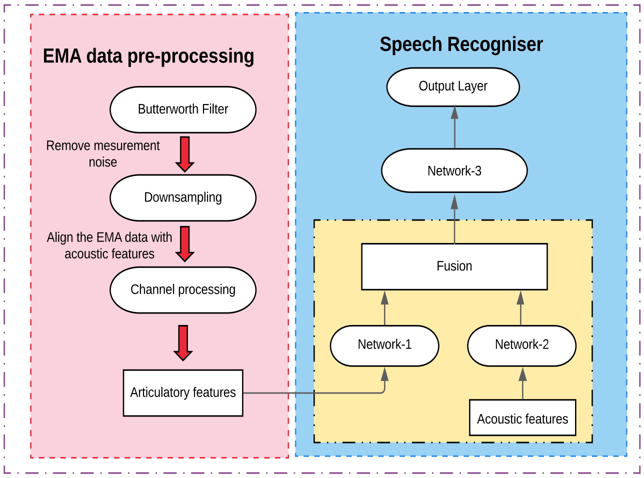

Fig. 1: The architecture of the proposed model. Net-1 and Net-2 perform stream-wise per-processing while Net-3 carries out post-processing after fusion and before output layer.

3.1. EMA data pre-processing

The articulatory measurements in TORGO were collected through a 3D AG500 electromagnetic articulography (EMA) system. As shown in Fig. 1, the EMA data is pre-processed by low-pass filtering, downsampling and channel processing. A Butterworth low-pass filter was first applied to reduce the measurement noise existing in most of the EMA channels. The high cutoff frequency of the filter was set to 10 Hz, and the order is 5. The acoustic features (e.g., MFCCs) and the articulatory features have different frame rates, namely 100 Hz and 200 Hz, respectively. To align and synchronise them, the EMA data was downsampled to 100 Hz. Finally, the task-beneficial EMA channels which could complement the acoustic features towards achieving the highest recognition result were selected as the articulatory features.

3.2. Acoustic-articulatory features fusion

Most of the previous works have integrated articulatory features by concatenating them with the acoustic features in a single vector on a frame-by-frame basis and at the network's input level [4]. However, a direct concatenation of these features is suboptimal. This is owing to the fact that these two information streams carry different types of information, encoded in different forms with different importance to the given task. This necessitates applying a bespoke per-stream pre-processing before fusion at optimal level of abstraction.

In [17], multi-stream acoustic modelling and fusion at low, medium and high levels were studied. In this work, the individually processed streams are concatenated at three stages to explore the optimal fusion level. Inspired by such a framework, we study the following fusion levels: at the input level (**concat-1**), at the medium level after the last convolutional layer (**concat-2**), and at the high level after the last recurrent layer and before the output layer (**concat-3**). Fig. 2 illustrates various concatenation levels and fusion schemes.

3.3. Acoustic model

Our acoustic models are cascades of convolutional neural networks (CNN), fully-connected multi-layer perceptrons (MLP) and Light Gated recurrent units (LiGRU) [18]. Monophone regularisation is also employed as an auxiliary task. For more details please refer to our previous work [19].

4. EXPERIMENTAL RESULTS

4.1. Data description

TORGO [8] is one of the few available and widely used dysarthric speech databases. It has 15 speakers. Eight of the speakers (5 males, 3 females) have dysarthria ranging from mild to severe degrees, while the other seven are typical speakers (4 males, 3 females).

What makes TORGO attractive is that it contains aligned acoustic and articulatory recordings. Besides 16363 audio recordings, TORGO consists of 7177 articulatory measurement recordings collected from the 3D AG500 EMA system. Since the acoustic data is recorded by both head-mounted and array microphones simultaneously, one set of EMA data is normally associated with two sets of acoustic data. Therefore, 13127 utterances (which have both EMA and audio data) are used in the following experiments.

EMA data samples are measured by 12 sensors capturing articulatory movements in 3D, each returning sensor positions in Cartesian coordinates (x, y, z) along with the spatial orientation angles. The sensors are attached to the tongue back, tongue middle, tongue tip, forehead, bridge of the nose, upper lip (UL), lower lip (LL), lower incisor, left and right mouth, left and right ear.

4.2. Experimental setup

The 39D MFCCs and 3D EMA features are used as inputs, with splicing of ±5 contextual frames. The training data is augmented using speed perturbation (using factors 0.9, 1.0 and 1.1). The CNNs are a cascade of three 1D convolutional layers. The LiGRU layers are designed based on [18], consisting of one fully-connected layer, a stack of five bidirectional LSTM layers [20] followed by another fully-connected layer and two softmax classifiers (estimating the context-independent and context-dependent states). The dropout (0.15) [21], layer normalisation [22] and batch normalisation [23] are also applied along with RMSProp optimisation [24]. Learning-rate annealing is deployed with a factor of 0.5. The 5-fold cross-training TORGO setup proposed in [19] is applied. An independent 200k vocabulary size LibriSpeech trigram LM, as proposed in [25], was employed for decoding.

4.3. Results and discussion

Results are reported in Table 1. The first row displays the baseline system using only MFCC acoustic features. Then, we integrated the lip articulatory data with MFCC (MFCC+Lip). To explore the most task-beneficial articulatory measurements, firstly, we employed the Cartesian coordinates (x,y,z) of the articulators as the articulatory features, similar to most of the previous work [3, 16]. However, this does not provide consistent improvement across all speakers. We then used the Euclidean distance (ED) between the articulators positions and the origin. Similar to the previous approach, consistent improvement across all speakers was not achieved. Next, we used the pair-wise EDs between the articulators in the lip region[1] as articulatory features. This representation outperformed the baseline across all speakers. We hypothesise it removes the influence of the head movement and implicitly normalises the articulatory features.

The results of the MFCCs concatenated with the proposed lip features are reported in the second row. As seen, direct feature fusion at the input level outperforms the baseline (MFCC alone). On average it reduces WER by 1.9% and 0.5% (absolute) for dysarthric and typical speech, respectively.

The 3rd to 5th rows in Table 1 show the results of introducing the multi-stream CNN feature fusion schemes. Comparing the *concat-1* and *concat-2* with *concat-0* shows 0.5% and 2.7% absolute WER reductions for dysarthric speech, respectively. The *concat-2* system appears to be the best fusion scheme while *concat-3* leads to the poorest performance.

We also counted the number of trainable parameters (*#params*) for models in Table 2. As seen, fusion at higher levels notably increases *#params*. For example, *#params* of the *concat-3* system is 1.5 times higher than that of the *concat-2* systems. This makes the model more liable to overfitting, especially in low-resource data scenarios. Furthermore, concatenating the streams close to the output layer could give rise to insufficient post-processing (after fusion). Our experimental results for dysarthric speech verify the conclusion in [17] regarding the optimal fusion level: it should be high enough to effectively pre-process each information stream for the given task and low enough to leave sufficient capacity after fusion for post-processing the fused streams.

Fig. 3 compares the evolution of the cross-entropy (CE) loss of various proposed fusion schemes during training. It illustrates that the *concat-3* system converges faster than other models and *concat-2* tends to have the lowest training and

[1] The ED between the UL (UL_x,UL_y,UL_z) and LL (LL_x,LL_y,LL_z).

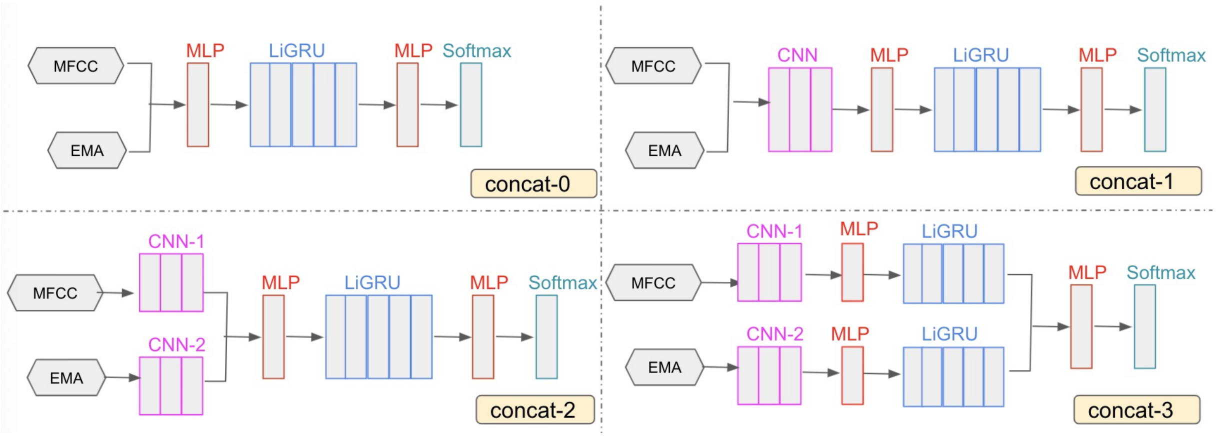

Fig. 2: The proposed architectures, fusing the acoustic and articulatory features at low/medium/high levels.

Table 1: ASR performance (WER) for different features per (F)emale and (M)ale speakers with different dysarthria severity, along with the averaged results for all speakers. 'M/S' indicates speakers with Moderate to Severe levels of dysarthria.

		Severe			**M/S**	**Moderate**	**Mild**		**Average**	
Input features	**Systems**	M01	M02	M04	M05	F03	F04	M03	Dys	Typ
MFCC	**baseline**	74.4	73.5	83.8	57.9	48.0	19.5	18.4	47.8	16.4
MFCC+Lip	**concat-0**	74.4	71.8	82.0	57.2	44.7	18.3	15.5	45.9	15.9
MFCC+Lip	**concat-1**	71.6	70.3	79.2	60.2	45.3	**15.8**	16.3	45.4	15.0
MFCC+Lip	**concat-2**	**67.3**	**66.5**	**75.9**	**56.3**	**38.4**	17.2	**10.1**	**43.2**	**12.9**
MFCC+Lip	**concat-3**	98.0	103.2	107.8	91.6	63.6	34.6	26.51	60.4	35.2

Table 2: *#params* (in millions) for different fusion schemes.

	Baseline	Concat-0	Concat-1	Concat-2	Concat-3
#params	11.1	11.3	15.1	15.0	24.9

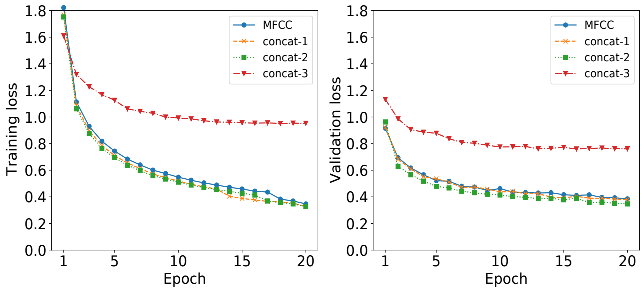

Fig. 3: Cross-entropy loss for different fusion schemes.

validation loss which is due to the fact that it provides the best trade-off in terms of pre- and post-processing.

We also explored the performance evolution of the *concat-2* system in terms of WER across different epochs. The results for speakers with dysarthric and typical speech are plotted in Fig. 4. As seen, the WER improvement for speakers with severe dysarthria is notably limited and does not continuously improve during training, contrary to the typical or mild conditions. Moreover, the performance reaches a plateau after 10 epochs for dysarthric speech whilst for typical speech the performance keeps significantly improving up to 15 epochs.

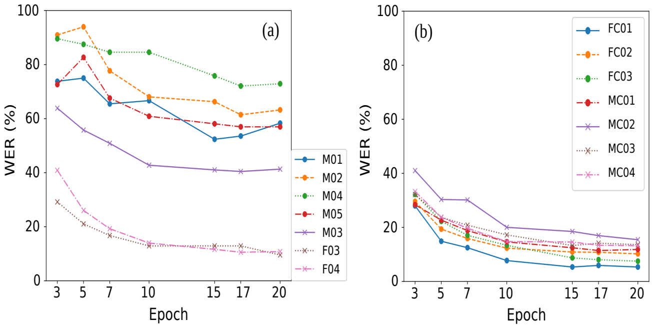

Fig. 4: WER vs epoch in the *concat-2* system for various speakers. (a) Dysarthric speech, (b) Typical speech.

5. CONCLUSION

Using TORGO dysarthric speech database, we constructed effective multi-modal speech recognition systems by combining the acoustic and articulatory features. The proposed multi-stream acoustic models consist of convolutional, recurrent and fully-connected layers allowing the multi-modal features to be fused via various schemes and at different levels of abstraction. Optimal fusion level and training dynamics of the models were studied. The optimal fusion scheme resulted in up to 4.6% (absolute) WER reduction, with best improvement for speakers with severe dysarthria. Future work includes exploring usefulness of other modalities and applying more advanced architectures and more sophisticated fusion schemes.

6. REFERENCES

[1] W.R. Gowers, "Clinical speech syndromes of the motor systems," *Neurology for the Speech-Language Pathologist. Fifth edition. Philadelphia: Butter worth_ Heinemenn*, pp. 196–203, 2001.

[2] E. Salama, R. El-Khoribi, and M. Shoman, "Audio-visual speech recognition for people with speech disorders," *International Journal of Computer Applications*, vol. 96, no. 2, 2014.

[3] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *NIPS*, 2010, pp. 70–78.

[4] F. Xiong, J. Barker, and H. Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018.

[5] O. Fujimura, "Relative invariance of articulatory movements, in invariance and variability in speech processes," *Lawrence Erlbaum*, pp. 226–242, 1986.

[6] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *ICSLP*, 2000.

[7] E. Yılmaz, V. Mitra, C. Bartels, and H. Franco, "Articulatory features for asr of pathological speech," *arXiv preprint arXiv:1807.10948*, 2018.

[8] F. Rudzicz, A. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[9] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2010.

[10] T. Afouras, J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[11] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "Integrating articulatory data in deep neural network-based acoustic modeling," *Computer Speech & Language*, vol. 36, pp. 173–195, 2016.

[12] V. Mitra, G. Sivaraman, et al., "Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks," in *ICASSP*. IEEE, 2017, pp. 5205–5209.

[13] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008.

[14] D. Hill, C. Taube-Schock, and L. Manzara, "Low-level articulatory synthesis: A working text-to-speech solution and a linguistic tool," *The Canadian Journal of Linguistics/La revue canadienne de linguistique*, vol. 62, no. 3, pp. 371–410, 2017.

[15] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "Tada: An enhanced, portable task dynamics model in matlab," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[16] F. Rudzicz, G. Hirst, and P. van Lieshout, "Vocal tract representation in the recognition of cerebral palsied speech," *Journal of Speech, Language, and Hearing Research*, 2012.

[17] E. Loweimi, P. Bell, and S. Renals, "Raw sign and magnitude spectra for multi-head acoustic modelling.," in *INTERSPEECH*, 2020, pp. 1644–1648.

[18] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[19] Z. Yue, H. Christensen, and J. Barker, "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," in *INTERSPEECH*, 2020.

[20] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *ASRU*. IEEE, 2013, pp. 273–278.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] Kiros J Ba, J and G Hinton, "Layer normalization," *Deep Learning Symposium, NIPS*, 2016.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[24] T. Tieleman and G. Hinton, "Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," *COURSERA Neural Networks Mach. Learn*, 2012.

[25] Z. Yue, F. Xiong, H. Christensen, and J. Barker, "Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition," in *ICASSP*. IEEE, 2020.