Proceedings Paper:
Yue, Z., Loweimi, E. and Cvetkovic, Z. (2022) Raw source and filter modelling for dysarthric speech recognition. In: Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23-27 May 2022, Singapore, Singapore. IEEE, pp. 7377-7381. ISSN: 1520-6149. EISSN: 2379-190X.

https://doi.org/10.1109/icassp43922.2022.9746553

RAW SOURCE AND FILTER MODELLING FOR DYSARTHRIC SPEECH RECOGNITION

Zhengjun Yue[1,2,†], *Erfan Loweimi*[1,3,†] *and Zoran Cvetkovic*[1]

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

ABSTRACT

Acoustic modelling for automatic dysarthric speech recognition (ADSR) is a challenging task. Data deficiency is a major problem and substantial differences between the typical and dysarthric speech complicates transfer learning. In this paper, we build acoustic models using the raw magnitude spectra of the source and filter components. The proposed multi-stream model consists of convolutional and recurrent layers. It allows for fusing the vocal tract and excitation components at different levels of abstraction and after per-stream pre-processing. We show that such a multi-stream processing leverages these two information streams and helps the model towards normalising the speaker attributes and speaking style. This potentially leads to better handling of the dysarthric speech with a large inter-speaker and intra-speaker variability. We compare the proposed system with various features, study the training dynamics, explore usefulness of the data augmentation and provide interpretation for the learned convolutional filters. On the widely used TORGO dysarthric speech corpus, the proposed approach results in up to 1.7% absolute WER reduction for dysarthric speech compared with the MFCC baseline. Our best model reaches up to 40.6% and 11.8% WER for dysarthric and typical speech, respectively.

Index Terms— Dysarthric speech recognition, source-filter separation and fusion, multi-stream acoustic modelling

1. INTRODUCTION

People with dysarthria often have impaired motor-control over their speech articulation. The reduced articulation control often leads to heavily slurred speech, slower speaking rate, abnormal pauses, false starts and repetitions [1]. As a result, the dysarthric speech, depending on severity, can sound very different from the typical speech and is less intelligible. This makes building ASR systems for dysarthric speech very challenging. In particular, state-of-the-art ASR systems built for typical speech do not work well for dysarthric speech, whilst dysarthric speech data paucity limits the efficacy of the automatic dysarthric speech recognition (ADSR) systems.

Data augmentation techniques such as speed perturbation [2, 3] and tempo adjustment [4] have been shown to be useful and offer some limited improvement. Out-of-domain data has been also exploited to address data sparsity [5–7]. However, the substantial differences between typical and dysarthric speech limits the usefulness of transfer learning from typical to dysarthric speech. Previous studies also have demonstrated the benefit of employing effective speech representations such as articulatory [6, 8] and bottleneck features [9] to improve acoustic modeling of dysarthric speech.

In this paper, we build on the recent work on multi-stream acoustic modelling from the raw source and filter components [10, 11]. In this framework, the vocal tract (VT) and excitation (Exc) components are pre-processed individually, and post-processed after fusion. This approach offers a number of advantages: the source and filter components are pre-processed based on their contribution to the task and by considering how they encode information. Moreover, it allows for fusing the streams at an optimal level of abstraction.

Although this framework is generic and applicable in any speech recognition/classification task, it can offer a special advantage in the context of ADSR. When the model takes two inputs characterising the lingual content (vocal tract) and speaker attributes (excitation), among others, it learns to normalise the speaker-associated properties reflected in the source component whilst extracting the lingual content of the speech from the filter component. Such implicit speaker normalisation is highly desirable in recognising dysarthric speech with a high inter and intra-speaker variability.

Based on this rationale, we build acoustic models for ADSR from the raw magnitude spectra of the VT and Exc components. We first separate the source and filter elements via cepstral processing. Having pre-processed each stream by a convolutional neural network (CNN), we recombine them and pass them through a stack of recurrent layers. We also study the effect of data augmentation via speed perturbation, analyse the training dynamics in terms of cross entropy (CE) loss and provide some interpretation for the learned filters. We achieved up to 40.6% and 11.8% WER on TORGO for dysarthric and typical speech, respectively.

[†] Equal contribution.

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

2. PROPOSED SYSTEMS

In this section, we briefly review how the source and filter components are separated and recombined.

2.1. Source-filter Separation

Cepstral low-pass liftering (CLPL) [12] is a straightforward method to extract the source and filter components. The underlying premise of CLPL is that the log of the magnitude spectrum can be interpreted as a superposition of the two components: a rapidly oscillating component modulated by a slowly varying envelop. The former is associated with the excitation component and the latter reflects the vocal tract. Applying a low-pass lifter returns the VT component and by taking advantage of the additivity of the source and filter in the cepstral domain, the Exc component is extracted.

The high cut-off quefrency of the low-pass lifter (L_0) should be adjusted based on the fundamental frequency (F_0) which necessitates tracking F_0 per frame. In [11], it was argued that tracking F_0 can be avoided, if L_0 is chosen based on the minimum possible fundamental periodicity T_0^{Min} (or equivalently F_0^{Max}). This design choice ensures the filter component is devoid of any source information and highly simplifies the setup. It, however, results in some error (VT residues exist in the Exc component) which was shown to be insignificant [11]. We use the same configuration and set L_0 to 50, equivalent to F_0^{Max} of 320 Hz.

Fig. 1 illustrates the separated source and filter components using CLPL for a severely dysarthric speech utterance. Compared with typical speech signals, the spectral energy of the dysarthric speech is more limited to the low frequencies and has weak high frequency components. Besides, as the VT spectral component demonstrates, the formants structure is highly distorted; the spectral energy above 1.5 kHz is notably low, making higher order formants particularly affected.

2.2. Source-filter Fusion (Recombination)

We wish to construct a multi-stream acoustic model using the raw magnitude spectra of the source and filter components. Unlike the MFCC [13] and FBank features, which are lossy representations, raw magnitude spectrum provides the model with a more informative signal representation, preserving both vocal tract and excitation components. It is also devoid of suboptimal information loss inadvertently occurs along the handcrafted parametrisation pipelines.

However, one shortcoming of feeding the model with the raw magnitude spectrum is that it implicitly fuses the VT and Exc components at the input level, whereas these two components carry complementary information, encode it differently, and are not equally important. As such optimal preprocessing for each stream is different. To address this issue, one needs a multi-stream system which allows for individual

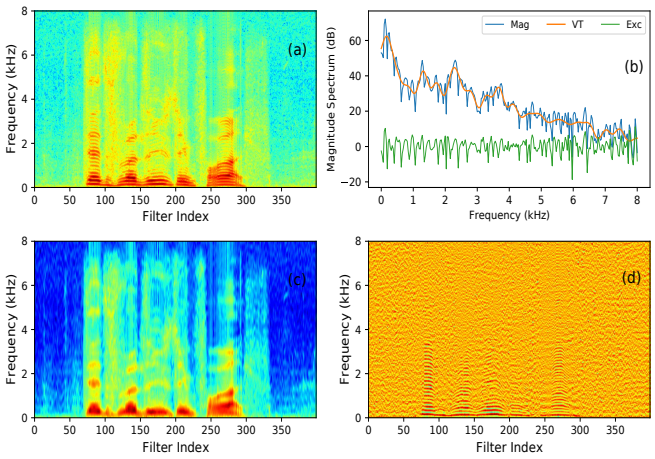

Fig. 1: Source-filter separation for a severe dysarthric speech. (a) Spectrogram, (b) Magnitude spectrum along with the VT and Exc components, (c) VT component, (d) Exc component.

pre-processing, fusion at optimal level of abstraction, and finally post-processing the recombined streams.

One issue in the multi-stream processing is fusion at an *optimal* level of abstraction. Assuming a fixed budget in terms of number of layers, the fusion level should be high enough to allocate enough layers to per-stream pre-processing while being sufficiently low to leave enough layers for post-processing the recombined streams. According to findings in [10, 11, 14], fusion after pre-processing each stream via multiple convolutional layers appears to be a reasonable design choice.

Ideally, an optimal information processing system should only pass through the task-relevant information whilst filtering out the irrelevant one. In the context of ASR, using speaker-invariant representations like VT component is desirable. However, using task-irrelevant information such as Exc component, which essentially captures speaker-correlated characteristics of speech, could be helpful, too. That is, this component can inform the model about the speaker and helps it to normalise the speaker-related attributes and style. Such normalisation, among others, is potentially very instrumental in recognising the dysarthric speech with a large inter and intra-speaker variability and, can contribute towards enhancing the generalisation and robustness of the acoustic model.

Fig. 2 illustrates the proposed multi-stream acoustic model along with the single-stream baseline system. In the proposed system the source and filter components are first pre-processed via three convolutional layers. Then, they are fused via a fully-connected multi-layer perception (MLP) and post-processed by a stack of five layers of LiGRU [15].

3. INTERPRETATION OF THE LEARNED FILTERS

Fig. 3 illustrates the average of the 128 learned filters, 129 samples long, along with the magnitudes of their FFT, for the first convolutional layer (ConvL-1) for models shown in

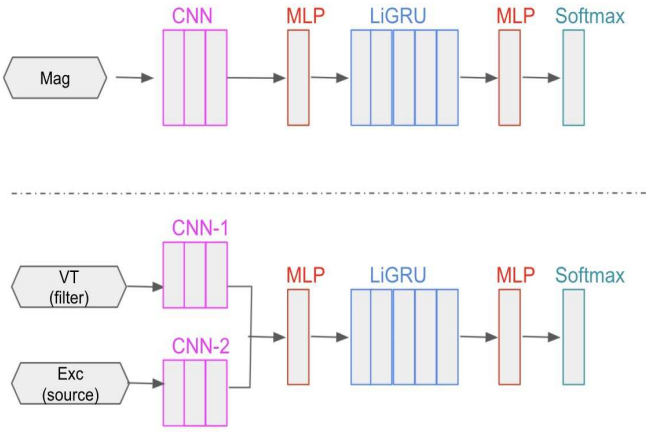

Fig. 2: Proposed multi-stream model vs single-stream baseline, consisting of convolutional, MLP and recurrent layers.

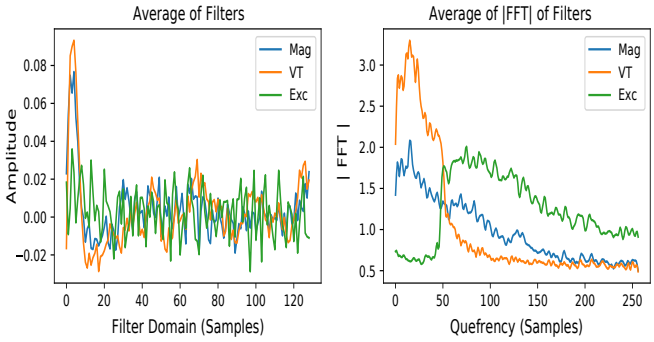

Fig. 3: Average of learnt filters in the first convolutional layer. (left) learnt filters, (right) magnitude of FFT of the filters.

Fig. 2. Note that these filters operate on the magnitude spectrum, not the time-domain. Hence, after taking Fourier transform (FT), the domain is analogous to the cepstral domain.

As seen, the average responses in the original domain are not very insightful. However, taking Fourier transform shows that the Mag filters act like low-pass lifters. This demonstrates the model implicitly learns to pay more attention to the low cepstral components which are associated with the VT and discards the Exc part. The VT filters behave similarly to Mag while filters operating on the Exc component pay no attention to the low cepstral components. This is an interesting observation: the corresponding input (Exc component) has no component in low quefrencies; the model learns this and disregards such components.

4. EXPERIMENTAL RESULTS

4.1. Data Description

Acoustic models are built using TORGO [16] dysarthric speech datasets. It contains 21 hours (7.3 hours for dysarthric and 13.7 hours for typical speech) of acoustic recordings collected from 15 speakers. Eight of the speakers have dysarthria ranging from mild to severe, while others are non-dysarthric typical speakers. The total vocabulary size is 1573.

4.2. Experimental Setup

The networks are trained using PyTorch-Kaldi [17]. To build acoustic models for raw signal representations, raw waveform model configuration was used. As shown in Fig. 2, the pre-fusion CNNs are cascades of three 1D convolutional layers. The post-fusion sub-network consists of one fully-connected layer, a stack of five bidirectional [18] LiGRU [15] layers, followed by another fully-connected layer and a softmax classifier. The dropout [19] (0.15), layer normalisation [20] and batch normalisation [21] are also employed along with RMSProp optimiser [22]. Learning rate annealing with a factor of 0.5 was applied. The dimensions of MFCC, FBank and raw spectral features (per frame) are 39 (including delta and delta-delta), 83 (80 FBank + 3 pitch) and 257, respectively. The Mag, VT and Exc are all 10^{th} root of the corresponding raw magnitude spectra. The 5-fold cross-training TORGO setup proposed in [23] is applied. An independent 200k vocabulary size Librispeech trigram language model, as proposed in [24], was employed for decoding.

4.3. Results and Discussion

Table 1 reports the results for various features. The first two rows show the performance of the handcrafted 39-D MFCCs and 83-D FBank features. As seen, MFCC outperforms FBank by a significant margin in this task. The third and fourth rows illustrate the performance of the raw waveform and raw magnitude spectrum. These two signal representations are more informative than MFCC and FBank, however, return poorer results. This is primarily due to the data scarcity problem in dysarthric speech and the fact that the amount of training data is more critical for high dimensional features.

The last three rows display the WERs when the source (Exc) and filter (VT) components are applied individually and jointly (VT+Exc). The raw VT representation outperforms the raw magnitude spectrum by 3.9% and 3.0% (absolute) for dysarthric and typical speech, respectively. Although VT outperforms FBank feature, it is still behind MFCC by a noticeable margin for both dysarthric and typical speech.

Using only the Exc component leads to a very poor performance for both dysarthric and typical speech. However, when this seemingly task-irrelevant representation is employed jointly with the VT within the proposed architecture, the best performance is achieved. This is a very interesting observation showing that although Exc does not carry information directly applicable to ASR, it helps the VT component to normalise the speaker related attributes, returning the best performance for both typical and dysarthric speech. It outperforms MFCC by 1.6% and 0.6% (absolute) in terms of WER for dysarthric and typical speech, respectively.

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	
Feature	F01	M01	M02	M04	M05	F03	F04	M03	**Dysarthric**	**Typical**
MFCC	**56.1**	86.6	56.3	80.4	**62.6**	**40.1**	26.6	13.1	49.0	16.3
FBank	69.5	93.4	67.8	80.7	61.5	47.9	29.7	22.1	54.9	20.6
Raw-wave	58.5	90.7	63.7	84.3	76.7	40.4	28.0	16.9	57.2	23.6
Mag	68.1	73.0	64.5	84.6	68.6	53.2	28.4	22.3	56.7	22.7
VT	59.9	68.5	59.5	80.6	67.6	50.2	24.7	17.6	52.8	19.7
Exc	98.4	102.6	103.5	97.2	96.7	91.4	95.3	93.1	96.8	94.8
VT+Exc	57.6	**62.3**	**52.1**	**78.4**	**62.6**	42.2	**20.2**	**12.9**	**47.4**	**15.7**

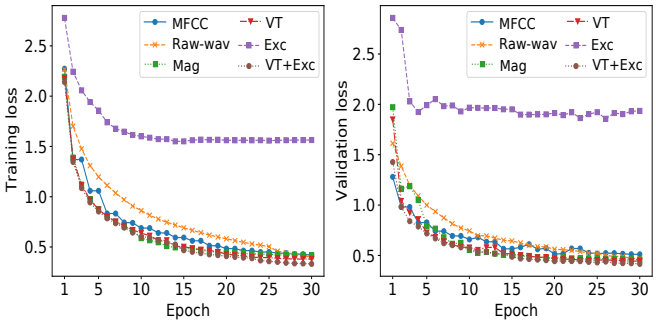

Fig. 4: Training dynamics (CE vs Epoch) for various models.

Fig. 4 illustrates the evolution of the cross-entropy (CE) loss for various features during training. Except for Exc, other features have similar convergence pattern and on average training converges after 15 epochs. As the WER shows, Exc has a very poor performance even though the CE dynamics shows a fast convergence. Such a fast convergence is owing to the fact that the model finds out that there is not much to learn from this stream when used individually. On the other hand, the CE for *VT+Exc* is similar to others whilst in terms of WER it outperforms them with a significant margin.

Finally, we augment the data via speed perturbation, using 0.9, 1.0 and 1.1 factors and *without* keeping the pitch fixed. The results are reported in Table 2 along with the relative gain for each feature. With more data for training, the raw magnitude spectrum achieves the highest performance with relative gain of 28% and 48% for dysarthric and typical speech, respectively. The *VT* system, although benefiting relatively more than the handcrafted features, returns poorer results than the *Mag* system. The results obtained by *VT+Exc* system are behind the raw magnitude spectrum on both typical and dysarthric speech and despite benefiting from data augmentation, its relative performance gain is less than other features.

Why is the relative gain after data augmentation the most for Mag and low for VT+Exc? The raw magnitude spectrum, individually, is the most informative spectral representation but cannot handle the speaker variability when data is limited. By perturbing the speed without keeping the fundamental frequency fixed, this data augmentation scheme implicitly simulates many speakers and consequently helps the model to learn to normalise the speaker variability. Additionally, it undermines the speaker normalising role of the Exc component.

Table 2: Data augmented using speed perturbation for various features (along with the corresponding relative gain).

	Average	
Feature	**Dysarthric**	**Typical**
MFCC	42.3 (13.7%)	12.3 (24.5%)
FBank	47.6 (13.3%)	14.2 (31.1%)
Raw-wave	46.0 (19.6%)	17.5 (25.8%)
Mag	**40.6 (28.4%)**	**11.8 (48.0%)**
VT	42.8 (18.9%)	12.7 (35.5%)
VT+Exc	40.8 (14.8%)	12.7 (19.1%)

5. CONCLUSION

In this paper, we developed an effective multi-stream acoustic model for ADSR using raw magnitude spectra of the source and filter components. We separated the excitation and vocal tract elements via cepstral processing and recombined them by multi-stream CNNs. Having pre-processed each stream with CNNs, the streams are fused through fully-connected layers and post-processed via LiGRU layers. Training dynamics of the model as well as the learned filters in the first convolutional layer were studied and up to 1.6% absolute WER reduction for dysarthric speech was achieved. We also employed data augmentation by speed perturbation which further improved the performance, reaching state-of-the-art results compared with the previous TORGO-based work. Future work includes disentangled representation learning and employing pre-trained models using out-of-domain data.

6. REFERENCES

[1] F. Darley, A. Aronson, and J. Brown, "Clusters of deviant speech dimensions in the dysarthrias," *Journal of speech and hearing research*, vol. 12, no. 3, pp. 462–496, 1969.

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

[3] S. Liu, S. Hu, X. Xie, and H. Meng, "Recent progress in the cuhk dysarthric speech recognition system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[4] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP*. IEEE, 2019, pp. 5836–5840.

[5] H. Christensen, M. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech.," in *INTERSPEECH*, 2013, pp. 3642–3645.

[6] V. Yılmaz, E.and Mitra, G. Sivaraman, and H. Franco, "Articulatory and bottleneck features for speaker-independent asr of dysarthric speech," *Computer Speech and Language*, vol. 58, pp. 319–334, 2019.

[7] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *ICASSP*. IEEE, 2020.

[8] F. Xiong and H. Barker, J.and Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[9] Y. Takashima, T. Nakashika, T. Takiguchi, and Y. Ariki, "Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 1411–1415.

[10] E. Loweimi, Z. Cvetkovic, P. Bell, and S. Renals, "Speech acoustic modelling from raw phase spectrum," in *ICASSP*. IEEE, 2021, pp. 6738–6742.

[11] E. Loweimi, Z. Cvetkovic, P. Bell, and S. Renals, "Speech acoustic modelling using raw source and filter components," *INTERSPEECH*, pp. 276–280, 2021.

[12] L. Rabiner and R. Schafer, "Digital processing of speech signals prentice hall," *New Jersey*, pp. 121–123, 1978.

[13] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

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

[15] 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.

[16] 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.

[17] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *ICASSP*. IEEE, 2019, pp. 6465–6469.

[18] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.