RAW SOURCE AND FILTER MODELLING FOR DYSARTHRIC SPEECH RECOGNITION

Zhengjun Yue\textsuperscript{1,2,†}, Erfan Loweimi\textsuperscript{1,3,†} and Zoran Cvetkovic\textsuperscript{1}

\textsuperscript{1} Department of Engineering, King’s College London, UK
\textsuperscript{2} Speech and Hearing Group (SPandH), University of Sheffield, UK
\textsuperscript{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.

\textsuperscript{†} 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)

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 10th 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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Severe</th>
<th>Moderate</th>
<th>Mild</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M01</td>
<td>M02</td>
<td>M04</td>
<td>M05</td>
</tr>
<tr>
<td>MFCC</td>
<td>56.1</td>
<td>86.6</td>
<td>56.3</td>
<td>80.4</td>
</tr>
<tr>
<td>FBank</td>
<td>69.5</td>
<td>93.4</td>
<td>67.8</td>
<td>80.7</td>
</tr>
<tr>
<td>Raw-wave</td>
<td>58.5</td>
<td>90.7</td>
<td>63.7</td>
<td>84.3</td>
</tr>
<tr>
<td>Mag</td>
<td>68.1</td>
<td>73.0</td>
<td>64.5</td>
<td>84.6</td>
</tr>
<tr>
<td>VT</td>
<td>59.9</td>
<td>68.5</td>
<td>59.5</td>
<td>80.6</td>
</tr>
<tr>
<td>Exc</td>
<td>98.4</td>
<td>102.6</td>
<td>103.5</td>
<td>97.2</td>
</tr>
<tr>
<td>VT+Exc</td>
<td>57.6</td>
<td>62.3</td>
<td>52.1</td>
<td>78.4</td>
</tr>
</tbody>
</table>

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.

5. CONCLUSION

In this paper, we developed an effective multi-stream acoustic model for ADSDR 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.
REFERENCES


