Dysarthric Speech Recognition From Raw Waveform with Parametric CNNs
Zhengjun Yue\textsuperscript{1,2,\dag}, Erfan Loweimi\textsuperscript{1,3,\dag}, Heidi Christensen\textsuperscript{2}, Jon Barker\textsuperscript{2}, Zoran Cvetkovic\textsuperscript{1}
\textsuperscript{1} Department of Engineering, King’s College London, UK \\
\textsuperscript{2} Speech and Hearing Research Group (SPandH), University of Sheffield, UK \\
\textsuperscript{3} Centre for Speech Technology Research (CSTR), University of Edinburgh, UK \\
{zhengjun.yue,erfan.loweimi,zoran.cvetkovic}@kcl.ac.uk, \\
{heidi.christensen,j.p.barker}@sheffield.ac.uk

Abstract
Raw waveform acoustic modelling has recently received increasing attention. Compared with the task-blind hand-crafted features which may discard useful information, representations directly learned from the raw waveform are task-specific and potentially include all task-relevant information. In the context of automatic dysarthric speech recognition (ADSR), raw waveform acoustic modelling is under-explored owing to data scarcity. Parametric convolutional neural networks (CNNs) can compensate for this problem due to having notably fewer parameters and requiring less training data in comparison with conventional non-parametric CNNs. In this paper, we explore the usefulness of raw waveform acoustic modelling using various parametric CNNs for ADSR. We investigate the properties of the learned filters and monitor the training dynamics of various models. Furthermore, we study the effectiveness of data augmentation and multi-stream acoustic modelling through combining the non-parametric and parametric CNNs fed by hand-crafted and raw waveform features. Experimental results on the TORG0 dysarthric database show that the parametric CNNs significantly outperform the non-parametric CNNs, reaching up to 36.2\% and 12.6\% WERs (up to 3.4\% and 1.1\% absolute error reduction) for dysarthric and typical speech, respectively. Multi-stream acoustic modelling further improves the performance resulting in up to 33.2\% and 10.3\% WERs for dysarthric and typical speech, respectively.

Index Terms: dysarthric automatic speech recognition, raw waveform acoustic modelling, parametric CNNs

1. Introduction
Dysarthria is a neurological speech disorder that affects the intelligibility of human speech due to the weakening or incoordination of muscles and articulators involved in the articulation process. People with dysarthria also often have physical motor disabilities, leading to limited or involuntary body movements. As a result, they not only face problems in human-human communication but also have difficulty in human-machine interaction through smart devices. A reliable automatic dysarthric speech recognition (ADSR) technology can make a difference in their lives by enabling voice-driven interfaces which facilitates the communication with humans and machines. However, the currently available commercial ASR systems perform poorly on dysarthric speech, especially for severe dysarthria. This is owing to the scarcity of annotated dysarthric speech data and high inter- and intra-speaker variabilities. Moreover, the mismatch between dysarthric and typical speech hinders leveraging typical speech data in developing ADSR systems.

Acoustic modelling for ADSR is typically carried out using hand-crafted features such as MFCC [1] and Mel-filter bank (FBank) [2–4] or by utilising data from other modalities (e.g., visual [5] or articulatory [6, 7]). A potential drawback of hand-crafted features is that they are lossy and task-blind, giving rise to losing task-useful information at the outset of the process. Such irreversible information loss can in turn affect the performance of ASR systems. In raw waveform acoustic modelling, the representation learning process (front-end) is conducted jointly with learning the back-end, guided by the optimiser, objective function and labels (task). Further, raw waveform modelling allows incorporation of the phase spectrum information neglected in the Fourier transform magnitude-based features such as MFCC and FBank. The phase spectrum has been shown to be useful in ASR [8, 9], ADSR [10] and dysarthric speech detection [11]. Despite being rich information-wise, acoustic modelling from raw waveform is greatly data demanding. This poses a considerable challenge in the context of ADSR because the amount of available transcribed dysarthric speech data is too small to train an effective model. As a result, raw waveform modelling is under-explored in the dysarthric speech domain.

One solution to tackle this challenge is employing parametric CNNs (PCNNs). SincNet [12], Sinc$^2$Net [13], GaussNet [13], GammaNet [13] and ParzNet [14] are examples of PCNNs and have been successfully applied in ASR [13–15]. Compared with conventional non-parametric CNNs, the parametric filters are characterised by many fewer parameters, leading to faster learning and requiring fewer data. This is particularly interesting for ADSR as a low-resource task. In addition, PCNNs are more amenable to human interpretation and provide a framework for embedding some perceptual prior knowledge into the architecture. For example, GammaNet [13] is built on the Gammatone filters [16] proposed for modelling the cochlea [17] and provides a framework for learning the model parameters.

In this paper, as well as single-stream raw waveform modelling, we construct multi-stream acoustic models using both raw waveform and hand-crafted features. This is motivated by [9, 18, 19] and [20, 21] which have demonstrated the advantages of multi-stream acoustic modelling in ASR and ADSR tasks. We also explore the usefulness of data augmentation and information fusion level along with analysing the learned filters and monitoring the training dynamics of various models.

Having reviewed PCNNs in Section 2, in Section 3 we present the single- and multi-stream architectures used for acoustic modelling. In Section 4 experimental setup and results are presented and discussed. The learned filters are analysed and visualised in Section 5 and Section 6 concludes the paper.
2. Parametric CNNs

A parametric convolutional layer is a bank of bandpass filters where each one is characterised by an impulse response \( h(t) \) equal to the product of a baseband kernel \( K \) and a carrier \[21\]

\[
h^{(i)}(t) \theta_K^{(i)} f_c^{(i)} = K(t) \theta_K^{(i)} \text{carrier}(t) f_c^{(i)},
\]

where \( \theta_K^{(i)} \) is the parameter set of the kernel and \( f_c^{(i)} \) is the centre frequency of the \( i \)-th filter. The kernel determines the shape of the baseband filter and the carrier controls its location in the frequency spectrum. While the carrier is a sinusoid with one parameter, namely \( f_c^{(i)} \), the kernel could take different forms with one or more learnable (and interpretable) parameters.

Table 1 lists some of the parametric kernel-based models along with the corresponding parameters. Most filters characterised by \( A^{(i)} \) and \( B^{(i)} \) which denote \( i \)-th filter’s maximum amplitude (gain) and bandwidth, respectively. Filters like Gamma-tone include extra parameters, namely \( N^{(i)} \), called order. Typically, \( A^{(i)} \) is normalised to one, therefore, is not learnable.

<table>
<thead>
<tr>
<th>PCNNs</th>
<th>Kernel ( \theta^{(i)} )</th>
<th>Carrier ( f_c^{(i)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SincNet [12]</td>
<td>( A^{(i)} 2f_c^{(i)} \sin(2\pi f_c^{(i)} t) )</td>
<td>( \cos(2f_c^{(i)} t) )</td>
</tr>
<tr>
<td>Sinc2Net [13]</td>
<td>( A^{(i)} \sin^2(B^{(i)} t) )</td>
<td>( \cos(2f_c^{(i)} t) )</td>
</tr>
<tr>
<td>GaussNet [13]</td>
<td>( A^{(i)} \exp(-B^{(i)} t^2) )</td>
<td>( \cos(2f_c^{(i)} t) )</td>
</tr>
<tr>
<td>GammaNet [13]</td>
<td>( A^{(i)} \exp(-B^{(i)} t^2) )</td>
<td>( \cos(2f_c^{(i)} t) )</td>
</tr>
<tr>
<td>ParNet [14]</td>
<td>( A^{(i)} \max(0, 1 - \gamma^{(i)} t^2) )</td>
<td>( \cos(2f_c^{(i)} t) )</td>
</tr>
</tbody>
</table>

For SincNet, the kernel is a cardinal sine (sinc) function in the time domain rendering rectangular filters in the frequency domain. Sinc2Net leads to triangular filters approximately \[1\] resemble Mel filterbank. The first convolutional layer of the GammaNet is a learnable version of the Gammatone filters. GaussNet filters in both time and frequency domains are bell-shaped.

These parametric kernels are biologically plausible: triangular and tapered filters (along with their smoother versions like Gaussian filters) mimic frequency masking and GammaNet filters simulated filtering takes place in the Cochlea. As such, raw waveform modelling using PCNNs provides a framework for incorporating some prior perceptual knowledge in the architecture along with learning the involved parameters.

Let \( L \) denote a convolutional filter length in samples. Training a non-parametric conventional CNN involves learning \( L \) parameters per filter. However, as can be inferred from Eq. 1 and Table 1, training a parametric filter of length \( L \) usually involves learning many fewer parameters (e.g., only centre frequency and bandwidth). This implies the degrees of freedom of parametric models are one or two orders of magnitude less than their non-parametric counterparts. Fewer trainable parameters lead to requiring less data for effective training, faster learning, and better generalisation. This is particularly useful in low-resource domains such as dysarthric speech recognition.

Note that considering a parametric form for the filters imposes a strong prior on the model and notably reduces its flexibility and modelling capacity. However, assuming the role of the first layer is learning a quasi filterbank, not necessarily extracting a highly abstract representation (which is left to the higher layers), any collection of the bandpass filters seems to be sufficient from modelling standpoint.

3. Architectures for Acoustic Modelling

Fig. 1 (a) shows the single-stream architecture employed for raw waveform acoustic modelling. It is similar to CLDNNs proposed in [22] consisting of a cascade of convolutional, recurrent and fully-connected layers which perform feature learning, sequential modelling and further abstraction extraction (and linear separability enhancement), respectively. Our architecture is different from CLDNN in two ways: first, the conventional non-parametric CNNs are replaced with parametric CNNs. Second, as well as single-stream modelling, we study the usefulness of multi-stream acoustic modelling where both raw waveform and hand-crafted features are employed for acoustic modelling.

Figs. 1 (b) and (c) show multi-stream architectures. We use both raw waveform and hand-crafted features as input streams to take advantage of both representations. For fusing the information streams, we apply two schemes denoted by concat-1 and concat-2 in which the input streams are concatenated at low (input) and medium levels, respectively. In concat-2, streams are fused after being pre-processed via CNNs. Such stream-specific pre-fusion processing facilitates more effective information processing and significantly improves the performance.

Note that the raw waveform acoustic models with parametric and non-parametric CNNs are only different in the first convolutional layer. In addition, although the concat-1 and concat-2 fusion schemes have different numbers of parameters, they are still comparable with each other and also with the single-stream system: in all of them each individual stream is passed through the same number and sequence of layers. In the multi-stream setup, the hand-crafted features are always processed by non-parametric CNNs while the raw waveform models are processed by both types of CNNs.

4. Experimental Results

4.1. Experiment Setup

Acoustic models are built using the TORGO [23] dysarthric speech dataset including 21 h (7.3h dysarthric and 13.7h typical) speech data. The architectures shown in Fig. 1 are a cascade of convolutional, Light Gated recurrent units (LiGRU) [24] and fully-connected multi-layer perceptrons (MLP) layers, similar to the setup in [20, 21]. The only difference is that we removed the MLP layer between the CNN and LiGRU lay-

![Figure 1](image-url)
ers since we found that the current architecture yields better performance. The CNN block is a stack of three 1-D convolutional layers. The LiGRU block includes five bidirectional LiGRU layers with 550 units in each direction, followed by an MLP layer with 1024 units and a softmax classifier. The dropout [25] layer, with probability 0.15, is used after each LiGRU layer and the batch normalisation [26] layer is employed along with RMSProp optimiser [28].

In all layers, ReLU [29] activation function was used. The 5-fold cross-training TORGO setup proposed in [30] is applied. An independent 200k vocabulary size Librispeech [31] trigram language model [32], is employed for decoding. For training and decoding we use PyTorch-Kaldi toolkit [33–35]. The size of the hand-crafted MFCC and FBank features is 39 and 83-D (80 FBank + 3 pitch [36]), respectively, while the size of the raw waveform features is 400 samples (sampling rate: 16kHz, frame length: 25 ms). Hand-crafted features are normalised at the speaker level. We also investigate the usefulness of data augmentation via speed perturbation (sp) using following speed change factors: 0.9 (slower), 1.0 (original) and 1.1 (faster). This increases the amount of training data by three folds.

4.2. Results and Discussion

Table 2 reports the ASR performance for various systems in terms of word error rate (WER) for dysarthric and typical speech. Comparing the hand-crafted features shows that FBank outperforms MFCC by a significant margin. Raw waveform acoustic modelling using conventional non-parametric CNNs yields poor results despite the fact that the raw waveform feature is more informative than MFCC or FBank features. Such lower performance is greatly owing to data scarcity which hinders realising the potentials of raw waveform features. That is, for effective learning in high dimensional space, further training data is needed. Employing data augmentation via speed perturbation can notably decrease the gap between the performance of the hand-crafted and raw waveform models by conventional CNNs. The relative performance gain after applying speed perturbation is the largest for raw waveform models constructed by non-parametric CNNs (30.7% and 41.9% for dysarthric and typical speech).

Applying various PCNNs with different kernels markedly improves the performance relative to conventional CNNs. Compared with FBank and without applying data augmentation, parametric raw waveform models return comparable performance on typical speech and noticeably outperform FBank on dysarthric speech. When speed perturbation is utilised, although PCNNs still outperform the conventional CNNs, the gap is reduced notably. Furthermore, after data augmentation the performance of FBank and PCNN models becomes comparable on dysarthric speech while FBank outperforms them on typical speech, contrary to what happened for systems built without data augmentation. These two observations imply the FBank and raw waveform models respond differently to data augmentation on dysarthric and typical speech and encourage combining them as complementary representations in a multi-stream setup to benefit from both of them. We address this at the end of this section. Among the parametric raw waveform models, GammaNet returns the best performance, with a narrow margin.

Fig. 2 demonstrates the training dynamics of various models in terms of the evolution of the cross-entropy (CE) loss vs. epoch. Without data augmentation, the hand-crafted features show a relatively lower loss especially in the first 20 epochs. After the 20th epoch, the loss of MFCC does not decrease while the loss for FBank keeps decreasing. The faster convergence and reaching a plateau for the MFCC model stems from its information content which is the lowest compared with FBank and raw waveform. Although the CE loss for the parametric raw waveform models in early epochs is larger than FBank, after the 20th epoch they achieve lower or comparable loss. This implies that these models have a slower convergence rate and require further training epochs than hand-crafted features. Also note that the raw waveform model with non-parametric CNN has the largest loss and using PCNNs notably reduces the CE loss. Applying speed perturbation slightly decreases the loss value while has a notable effect on WER. In this case, for up to 20 epochs the FBank model has the lowest loss but after the 25th epoch, Sinc2Net and ParzNet lead to lower CE loss.

Fig. 3 illustrates the training dynamics of various systems in terms of WER vs. epoch for the dysarthric and typical speech, with and without speed perturbation. As seen, modifying the first convolutional layer by replacing the non-parametric filters with their parametric counterparts has a major effect on the dynamics of the raw waveform models. While data augmentation has a significant effect on the WER, it marginally affects the convergence rate, i.e., still around 30 epochs are required for effective training. Further, comparing the FBank and Sinc2Net models – where both have triangular filters – shows that the FBank model learns faster. That is, Sinc2Net should learn the \( f_{s2}(\cdot) \) and \( B(\cdot) \) (Table 1) as two trainable parameters while in the FBank they are engineered using the perceptually-motivated Mel scale and the 50% overlap between adjacent filters.

Comparing the training dynamics in terms of CE loss vs. epoch and WER vs. epoch, shows that the latter is a more revealing and expressive metric for studying the performance of different ASR systems and should be prioritised in any analysis. This is not surprising as the CE loss is a general-purpose train-

<table>
<thead>
<tr>
<th>Feature</th>
<th>without sp</th>
<th>with sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>47.5</td>
<td>39.2</td>
</tr>
<tr>
<td>FBank</td>
<td>45.3</td>
<td>36.5</td>
</tr>
<tr>
<td>Raw CNN</td>
<td>57.2</td>
<td>39.6</td>
</tr>
<tr>
<td>SincNet</td>
<td>43.6</td>
<td>36.5</td>
</tr>
<tr>
<td>Sinc2Net</td>
<td>43.7</td>
<td>36.9</td>
</tr>
<tr>
<td>GaussNet</td>
<td>43.0</td>
<td>36.4</td>
</tr>
<tr>
<td>GammaNet</td>
<td>44.6</td>
<td>36.7</td>
</tr>
<tr>
<td>ParzNet</td>
<td>43.3</td>
<td>37.1</td>
</tr>
</tbody>
</table>

Table 2: WER on TORGO, without and with speed perturbation.
Table 3: WER for PCNNs when filter gain ($A^{(i)}$) is learned.

<table>
<thead>
<tr>
<th>Feature</th>
<th>without sp</th>
<th>with sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinc2Net</td>
<td>48.8</td>
<td>17.5</td>
</tr>
<tr>
<td>GaussNet</td>
<td>50.2</td>
<td>18.8</td>
</tr>
<tr>
<td>ParzNet</td>
<td>50.9</td>
<td>20.0</td>
</tr>
</tbody>
</table>

Table 4: WER for multi-stream systems (Raw: raw waveform).

<table>
<thead>
<tr>
<th>Fusion level Systems</th>
<th>Concat-1</th>
<th>Concat-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dys Typ</td>
<td>Dys Typ</td>
</tr>
<tr>
<td>FBank(CNN)+Raw(CNN)</td>
<td>35.8</td>
<td>11.8</td>
</tr>
<tr>
<td>FBank(CNN)+Raw(Sinc2)</td>
<td>35.1</td>
<td>11.5</td>
</tr>
<tr>
<td>FBank(CNN)+Raw(Parz)</td>
<td>36.1</td>
<td>12.3</td>
</tr>
<tr>
<td>FBank(CNN)+Raw(Gauss)</td>
<td>36.9</td>
<td>12.2</td>
</tr>
</tbody>
</table>

We also studied the usefulness of multi-stream acoustic modelling by combining the hand-crafted and raw waveform features. Table 4 presents the performance of the multi-stream systems in Fig. 1 (b) and (c). As seen, combining the handcrafted features processed by non-parametric CNNs and the raw waveforms processed by non-parametric or parametric CNNs leads to consistent and significant gain. Furthermore, fusion level has a noteworthy impact on the performance of the multi-stream systems. Comparing the concat-1 and concat-2 fusion levels shows that the latter offers a more powerful framework and renders higher performance. This is owing to the bespoke pre-processing learned and applied on each stream by a cascade of convolutional layers before fusion.

The best multi-stream system (Table 4) in comparison with the best single-stream raw waveform model (Table 2) has 3.2% (33.2 vs 36.4) and 2.2% (10.3 vs 12.5) absolute lower WERs for dysarthric and typical speech, respectively. Such a significant gain corroborates the claim we put forward earlier that the hand-crafted representations and raw waveform features can play complementary roles in acoustic modelling.

5. Interpretation and Visualisation

To investigate the collective behaviour of the filters in the first convolutional layer, we use the average frequency response (AFR) [13, 37]: mean of the magnitude spectra of the impulse responses of the filters. The higher the AFR, the higher the overall attention of the first layer to the corresponding frequency bins. Fig. 4 illustrates AFR for various parametric and non-parametric raw waveform models. As seen, the parametric models mostly focus on frequencies below 2kHz. We also plot the mean and standard deviation (STD) of the short-time magnitude spectra of 200 signals from TORGO. Fig. 4 (b) shows that the STD is significantly large below 2kHz, implying this spectral subband is more informative. Furthermore, the mean in Fig. 4 (a) helps us to justify the local peaks around 6kHz in AFR of non-parametric and some of the parametric models which is owing to recording conditions of the TORGO data.

Finally, the data augmentation has a marginal effect on AFR of the parametric CNN models. Note that the parameterisation of the filters is equivalent to imposing a strong prior on a non-parametric model with a heavy regulatory effect. This minimises the effect of data augmentation on the first layer’s AFR as the model is already greatly constrained. Of course higher layers benefit from data augmentation, leading to a remarkable performance gain, as shown in Table 2.

6. Conclusions

In this paper, we explored and demonstrated the effectiveness of raw waveform acoustic modelling for automatic dysarthric speech recognition (ADSR) using various parametric CNNs. We built single-stream raw waveform models as well as multi-stream acoustic models fusing raw waveform and hand-crafted features. The optimal fusion level, training dynamics of the models and the filters learned in the first convolutional layer were studied and analysed. We showed that using parametric models along with data augmentation can effectively address the data scarcity problem in low-resource scenarios like ADSR. The best performance was achieved by the multi-stream model fusing FBank and raw waveform features, leading to 33.2% and 10.3% WERs for dysarthric and typical speech on the TORGO database. Future work includes transfer learning from high resource typical speech to low-resource dysarthric speech by adapting the filters’ parameters, e.g., similar to [38].
7. References


