This is a repository copy of *Acoustic modelling from raw source and filter components for dysarthric speech recognition*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/192463/

Version: Accepted Version

Article:

Yue, Z., Loweimi, E., Christensen, H. orcid.org/0000-0003-3028-5062 et al. (2 more authors) (2022) Acoustic modelling from raw source and filter components for dysarthric speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30. pp. 2968-2980. ISSN 2329-9290

https://doi.org/10.1109/taslp.2022.3205766

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works. Reproduced in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Acoustic Modelling from Raw Source and Filter Components for Dysarthric Speech Recognition

Zhengjun Yue[†] *(Member, IEEE)*, Erfan Loweimi[†] *(Member, IEEE)*, Heidi Christensen *(Member, IEEE)*,
Jon Barker *(Member, IEEE)*, Zoran Cvetkovic *(Senior Member, IEEE)*

***Abstract*—Acoustic modelling for automatic dysarthric speech recognition (ADSR) is a challenging task. Data deficiency is a major problem and substantial differences between typical and dysarthric speech complicate the transfer learning. In this paper, we aim at building acoustic models using the raw magnitude spectra of the source and filter components for ADSR. The proposed multi-stream models consist of convolutional, recurrent and fully-connected layers allowing for pre-processing various information streams and fusing them at an optimal level of abstraction. We demonstrate that such a multi-stream processing leverages information encoded in the vocal tract and excitation components and leads to normalising nuisance factors such as speaker attributes and speaking style. This leads to a better handling of dysarthric speech that exhibits large inter- and intra-speaker variabilities and results in a notable performance gain. Furthermore, we analyse the learned convolutional filters and visualise the outputs of different layers after dimensionality reduction to demonstrate how the speaker-related attributes are normalised along the pipeline. We also compare the proposed multi-stream model with various systems based on MFCC, FBank, raw waveform and i-vector, and, study the training dynamics as well as usefulness of the feature normalisation and data augmentation via speed perturbation. On the widely used TORGO and UASpeech dysarthric speech corpora, the proposed approach leads to a competitive performance of up to 35.3% and 30.3% WERs for dysarthric speech, respectively.**

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

I. Introduction

DYSARTHRIA is a common speech disorder stemming from damage to the central or peripheral nervous system [1] that causes the muscles involved in the speech production process to get weak or uncoordinated. The produced speech is often characterised by heavily slurred articulation, slower speaking rate, abnormal pauses, false starts and repetitions [2]. As a result, the intelligibility of dysarthric speech is low. People with dysarthria can be difficult to understand and those with severe dysarthria are often only intelligible to their close friends and family. This can negatively affect a person's social interaction, employment, education and other aspects of life.

People with dysarthria also typically have other physical disabilities, manifested by constrained or involuntary body movements such as stroke survivors or people with cerebral palsy. This limits their ability to interact with physical devices such as switches, keyboards and touch screens. As such, developing a speech-driven assistive technology to facilitate a reliable human-machine interaction is highly desirable. Automatic speech recognition (ASR) plays a key role in implementing such interfaces because other pipeline modules such as natural language understanding reside in its downstream. Therefore, developing ASR systems with satisfactory performance on dysarthric speech is a crucial step towards enabling a reliable human-machine interaction. Such technology has the potential to greatly improve the quality of life and well-being of people with dysarthria by helping them to communicate effectively with others and live more independently.

Commercial ASR systems, although capable of achieving a high performance for typical speech, perform poorly on dysarthric speech. The large acoustic mismatch between dysarthric and typical speech, reduced speaking rate [3], less distinctive phone classes and boundary position shifts [4], all hinder achieving good performance by the mainstream ASR systems trained with typical speech [5], [6]. High inter- and intra-speaker variabilities are also inherent to the dysarthric speech [7] and pose further challenge for the speech recogniser. In [8], [9] speaker adaptive training (SAT) has been used to address this issue.

To handle the mismatch and variability, a large amount of training data is required to learn an adequate model. However, there are only a few publicly available dysarthric datasets [10]–[13] and each has only a limited amount of data. Such data scarcity restricts the performance of the mainstream ASR systems which heavily rely on the data-demanding feed-forward deep neural networks (DNNs) [14], convolutional neural networks (CNNs) [15] and recurrent neural networks (RNNs) [16]. This highlights the need for developing particular approaches functional in the domain of dysarthria.

To address the data scarcity issue, data augmentation techniques have been extensively deployed in developing automatic dysarthric speech recognition (ADSR) systems. Motivated by the spectral and temporal differences between the dysarthric and typical speech, recent studies have been mainly focused on tempo adjustments [9], speed perturbation [17], [18] and vocal tract length perturbation [19]. In [20], [21], adversarial training along with voice conversion was applied to transform typical speech towards dysarthric speech. The out-of-domain typical speech data has also been exploited to address data scarcity [22]–[24].

Most of the studies in ADSR have been focused on building

[†] Zhengjun Yue and Erfan Loweimi made equal contribution. Correspondence to: {zhengjun.yue,erfan.loweimi}@kcl.ac.uk

Zhengjun Yue, Erfan Loweimi and Zoran Cvetkovic are with King's College London (KCL), and are supported by EPSRC Project EP/R012180/1 SpeechWave; e-mail: {zhengjun.yue,erfan.loweimi,zoran.cvetkovic}@kcl.ac.uk.

Jon Barker and Heidi Christensen are with Speech and Hearing Research Group (SPandH), University of Sheffield and are supported by the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Grant Agreement No. 766287); e-mail: {j.p.barker,heidi.christensen}@sheffield.ac.uk

acoustic models using handcrafted features such as MFCCs [25] and filterbank energies (FBank). Previous studies have also demonstrated the benefit of utilising other representations such as articulatory [23], [26]–[28] and bottleneck features [29] to improve the acoustic modelling for dysarthric speech. These magnitude spectrum based handcrafted features are, however, lossy representations and discard the signal information without considering the downstream task.

The raw magnitude spectrum is a richer representation and includes both vocal tract (VT) and excitation (Exc) components. Acoustic modelling using the raw magnitude [30] and raw phase [31] spectra of the source and filter components in ASR has been recently explored for typical speech, leading to significant performance gains. However, the usefulness of such an approach in the context of ADSR is under-explored.

In this paper, we build on [30] and our previous work [32] and construct multi-stream acoustic models from the raw magnitude spectra of the source and filter components for the ADSR task. In this framework, the raw magnitude spectra of the VT and Exc components are first pre-processed individually via convolutional layers. After fusion, they are further post-processed through recurrent and fully-connected (FC) layers. This approach offers several 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 these information streams to be fused at an optimal level of abstraction.

Although this framework is generic and applicable in any speech recognition/classification task, it offers some special advantages in the context of ADSR. When the model is fed by inputs characterising the lingual content and speaker attributes, it learns to normalise the speaker-associated variabilities captured by the source component while extracting the lingual content from the filter part. Such implicit speaker normalisation is highly desirable when recognising dysarthric speech given that it has high inter- and intra-speaker variabilities.

The main contributions of this paper are summarised below:

- Acoustic modelling using the raw magnitude spectra of the signal, source and filter components for ADSR;
- Effect of information fusion at various abstraction levels in the proposed multi-stream model is investigated;
- An analysis of the learned filters in the first convolutional layer is presented;
- The output of different layers is analysed to demonstrate how speaker attributes are normalised along the pipeline.

The rest of this paper is organised as follows. In Section II, the source-filter modelling and separation are briefly reviewed and some spectral properties of the dysarthric and typical speech are compared. The source and filter components are recombined via the proposed multi-stream architectures in Section III and the potential advantages of such separation and recombination are discussed. The experimental setup, results and discussion are presented in Section IV. Section V is dedicated to understanding and visualising some aspects of the learned models. Finally, Section VI concludes the paper.

II. Source-filter Modelling

A. Source-filter separation

Cepstral liftering [33] offers a straightforward framework to dissociate the vocal tract and excitation elements. These two speech components are assumed to be convolved in the time domain; therefore, they are multiplied in the frequency domain. After taking the log from the magnitude spectrum, they become additive and remain additive in the cepstral domain. The log of the magnitude spectrum as a curve can be interpreted as a superposition of two components with different changing rates. That is, it is composed of a rapidly oscillating element modulated by a slowly varying envelope. It can be shown that (approximately) the former is associated with the excitation component and the latter reflects the vocal tract. Applying a low-pass lifter in the cepstral domain returns the VT component. Taking advantage of the additivity of the source and filter in the cepstral domain, the Exc component can be extracted via subtraction.

One particular parameter of the low-pass lifter which should be adjusted is the high cut-off quefrency (L_0). It is related to the fundamental frequency (F_0) and this requires tracking F_0 per frame. In [30], it was argued that tracking F_0 can be bypassed, 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 extracted filter component is devoid of any source information, although it leads to some erroneous VT leakage to the Exc component. It highly simplifies the workflow whilst the VT residues in the Exc component, extracted in this way, were shown to be insignificant [30]. We use the same configuration and set L_0 to 50, equivalent to F_0^{Max} of 320 Hz when the sampling rate is 16 kHz.

Figs. 1 and 2 illustrate the separated source and filter components using cepstral liftering for a severely dysarthric speech sample and a typical speech sample, respectively, from the TORGO database. As seen, both the fine and coarse structures of the magnitude spectrum, which are respectively associated with the Exc and VT components, are notably distorted in the dysarthric speech.

B. Spectral differences between dysarthric and typical speech

To study the spectral differences of the typical and dysarthric speech, we computed the mean and standard deviation (STD) of the magnitude spectra of a randomly selected sample of one hundred typical and one hundred dysarthric speech signals from the TORGO database [13]. Fig. 3 illustrates the average raw magnitude spectra of the signals, filter and source components along with the corresponding STD.

As seen, on average, the dysarthric speech has lower energy and higher STD at high-frequency components. The higher STD implies larger variability which makes the learning process more challenging. Another interesting observation is that contrary to the average VT, the average Exc components for the dysarthric and typical speech are very similar. This could imply the source component is less affected than the filter part in the dysarthric speech. This point, however, warrants further investigation before drawing a firm conclusion.

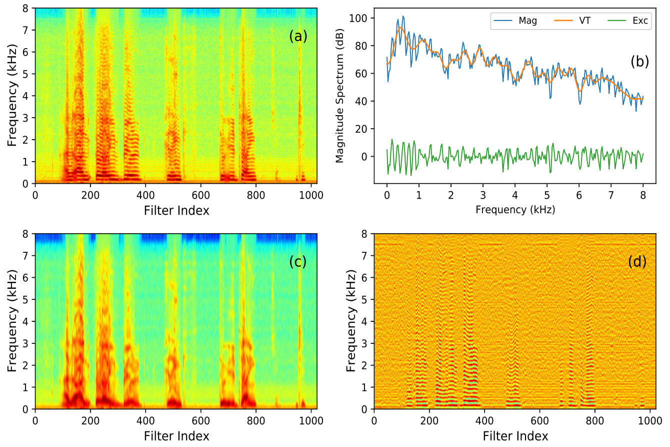

Fig. 1: Source-filter separation via cepstral liftering for a *dysarthric* speech. (a) Spectrogram, (b) Raw magnitude spectrum, VT and Exc components for a frame, (c) VT, (d) Exc.

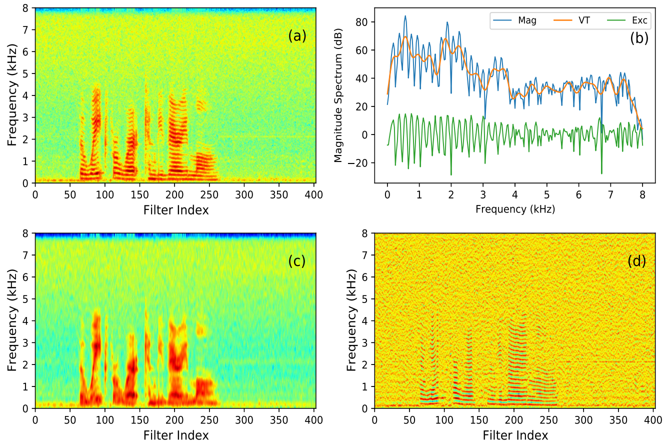

Fig. 2: Source-filter separation via cepstral liftering for a *typical* speech. (a) Spectrogram, (b) Raw magnitude spectrum, VT and Exc components for a frame, (c) VT, (d) Exc.

III. Multi-stream acoustic modelling

We aim to construct multi-stream acoustic models using the raw magnitude spectra of the source and filter components. Unlike the MFCC and filterbank features, which are lossy representations for the VT component, the raw magnitude spectrum provides the model with a more informative representation, preserving both VT and Exc elements. It is also devoid of potentially suboptimal information loss that inadvertently occurs along typical engineered front-end pipelines.

However, one shortcoming of feeding the model with the raw magnitude spectrum is that it implicitly fuses the source and filter components at the input level, whereas these two orthogonal components carry complementary information, encode this information differently, and are not equally important to a given task. As such, the optimal pre-processing pipeline for each stream is different. To address this issue, one needs a multi-stream system which allows for bespoke individual pre-processing, fusion at an optimal level of abstraction, and finally post-processing of the recombined streams.

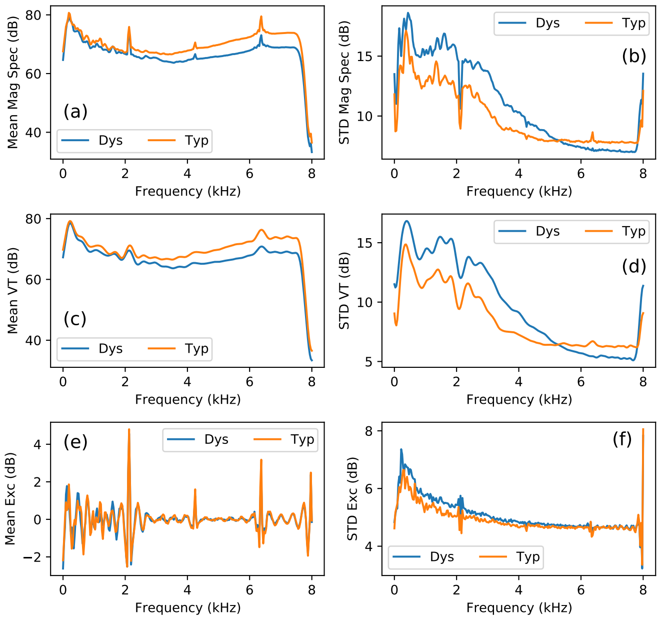

Fig. 3: Spectral mean and standard deviation (STD) of the dysarthric and typical speech extracted using randomly selected 100 signals for each speech type (The dysarthric samples are equally distributed over the severity levels: 33 severe, 33 moderate, and 34 mild samples). Mean of the raw magnitude spectra of (a) signals, (c) VT, (e) Exc. STD of the raw magnitude spectra of (b) signals, (d) VT, (f) Exc.

An optimal information processing system should only allow the task-relevant information to pass through while filtering out that which is irrelevant. In the context of ASR, applying speaker-invariant representations is desirable. However, using seemingly task-irrelevant information such as the Exc part, which essentially captures speech's speaker-correlated characteristics, can also be helpful. That is, this component informs the model about the speaker and thus allows it to normalise out the speaker-related attributes. Such normalisation is instrumental in recognising dysarthric speech with its large inter- and intra-speaker variabilities and, potentially contributes towards enhancing the robustness of the model.

One issue in the multi-stream processing is ensuring fusion occurs at an optimal level of abstraction. The fusion level should be high enough to allocate enough layers to the per-stream pre-processing while being sufficiently low to leave enough layers for post-processing the recombined streams. It should be also noted that an excessively deep structure, although allows for allocating many layers to both pre- and post-processing, will have too many parameters and consequently will be difficult to train, particularly in scenarios with limited training data. According to findings in [30]–[32], [34], [35], fusion after pre-processing each stream via multiple convolutional layers appears to be a reasonable design choice.

Fig. 4 illustrates the proposed multi-stream acoustic models along with a baseline single-stream system. In the proposed systems, the source and filter components are first pre-processed via three convolutional layers. Then, they are fused via a multi-layer perceptron (MLP-1) and post-processed by a

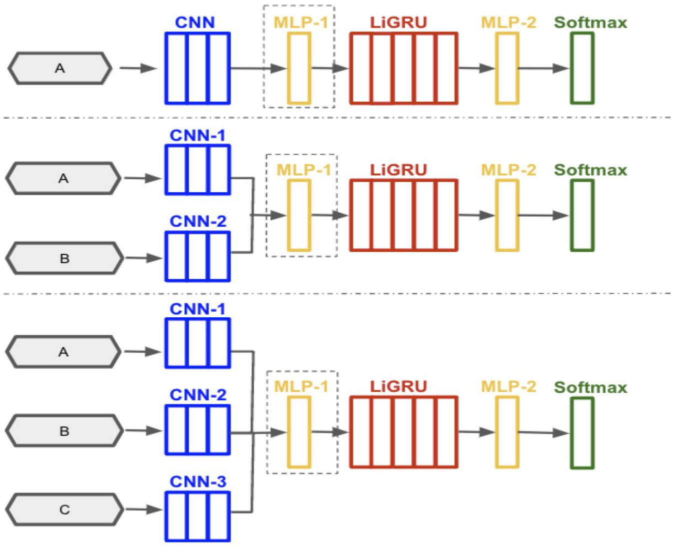

Fig. 4: The proposed multi-stream models vs. single-stream baseline, consisting of convolutional, MLP and recurrent layers. A, B and C represent various input features.

TABLE I: Duration (hours) of the training and test data per fold in the employed 5-fold cross-training setup.

Subset	fold 1	fold 2	fold 3	fold 4	fold 5	Mean±STD
Train+Dev	10.71	10.69	10.71	10.83	10.57	10.70±0.09
Test	2.71	2.73	2.72	2.59	2.67	2.68±0.06

stack of five layers of bidirectional Light-gated recurrent units (LiGRU) [36] and a fully-connected network (MLP-2). These components play complementary roles: the convolutional layers represent the input via learned feature maps, MLP-1 fuses the pre-processed streams, recurrent layers conduct sequential modelling and MLP-2 provides further post-processing to make the data linearly separable just before the softmax layer.

IV. Experimental Setup, Results and Discussion

A. Dataset, architecture and front-end configuration

Acoustic models are built using the TORGO [13] dysarthric speech dataset. 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 in degree from mild to severe, while others are non-dysarthric typical speakers. The acoustic data is simultaneously recorded by a head-mounted microphone and a single directional microphone. TORGO consists of both word and sentence prompts: 615 unique words as well as 354 unique sentences with a total vocabulary size of 1573 tokens.

An N*-fold cross-training* setup, as proposed in [37], is applied for training and evaluating the systems. The total dataset is divided into five folds[1] with allocating 70% of data to training, 10% to validation (dev) and 20% for testing. This maximises the use of the available training and test data while maintaining the need for disjoint training and test sets. Table I summarises the amount of train/test data per fold.

As shown in Fig. 4, the pre-fusion CNNs are cascades of three 1D convolutional layers with 128, 60 and 60 feature maps of kernel sizes of 129, 5 and 5 samples along with max pooling of length 3 for all layers. In all of the convolutional layers ReLU activation [38], layer normalisation [39] and dropout [40] (0.15) were applied. The post-fusion sub-network consists of one fully-connected layer (MLP-1) with 1024 nodes and a stack of five bidirectional [41] LiGRU [36] layers with 550 units per direction, followed by another fully-connected layer with 1024 nodes and a softmax classifier. Dropout (0.15) and batch normalisation [42] were used in all the post-processing layers (except for the softmax layer). The batch size was set to 8 and the network was trained using an RMSProp optimiser [43]. For training and decoding we used PyTorch-Kaldi [44]–[46].

The dimensions of MFCC, FBank and raw waveform and raw spectral features (per frame) are 39 (including delta and delta-delta), 83 (80 log-FBank + 3 pitch-related [47]), 400 and 257, respectively. Mag$^{0.1}$, VT and Exc correspond to the 10^{th} root of the raw magnitude spectra of the signal, vocal tract and excitation components, respectively. All features are extracted with 25 ms frame length and 10 ms shift size. Unless mentioned otherwise, mean-variance normalisation (MVN) at the speaker level is applied to all features. An independent 200k vocabulary size Librispeech [48] trigram language model, as proposed in [49], was employed for decoding.

We also investigate the usefulness of the data augmentation via speed perturbation. The training data is augmented by changing the articulation speed with the following factors: 0.9 (slower), 1.0 (original) and 1.1 (faster). This expands the training data by a factor of three. In addition, we examine the effect of speed perturbation with and without keeping the fundamental frequency (F0) fixed which are referred to as *sp-FixedF0* and *sp*, respectively.

B. Performance of various input features

Table II reports the recognition results for various features. Beginning with the handcrafted 39D MFCCs and 83D FBanks features, MFCC outperforms FBank by a noticeable margin in this task, for both dysarthric and typical scenarios.

The second part of this table is dedicated to the single-stream raw magnitude and raw waveform acoustic models. Although these representations are more informative than MFCC, their performance is significantly poorer. Note that more informative sequences can potentially lead to higher performance, but only when an adequate architecture is employed and sufficient training data is available.

The third part of Table II displays the WERs when the source and filter components are applied individually and jointly (VT+Exc). The raw VT representation outperforms Mag$^{0.1}$ by 1.6% (absolute) for dysarthric speech and 1.6% for typical speech. Using only the Exc component leads to very poor performance for both dysarthric and typical speech. However, compared with the single-stream VT system, better performance is achieved when this seemingly task-irrelevant

[1] The pre-defined training and test partition sets are available at `https://github.com/zhengjunyue/bntg`.

TABLE II: Mean±STD of WER in the 5-fold training setup for various models (MVN at speaker level). '+' indicates concatenation at the multi-stream processing.

	Average	
Feature	**Dysarthric**	**Typical**
MFCC	49.0±3.5	16.3±2.7
FBank	51.3±2.1	19.6±1.6
Raw-wave	57.2±3.3	23.6±2.3
$\text{Mag}^{0.1}$	54.4±5.9	21.3±3.9
VT	52.8±4.6	19.7±3.7
Exc	96.8±3.3	94.8±2.5
VT+Exc	**47.4**±1.9	**15.7**±1.1
$\text{Mag}^{0.1}$+VT	48.5±3.8	16.4±1.6
$\text{Mag}^{0.1}$+Exc	48.8±4.5	16.8±2.0
$\text{Mag}^{0.1}$+VT+Exc	48.0±2.4	16.0±0.7

TABLE III: WER after MVN at the utterance level (gains(+)/losses(-) relative to the speaker level MVN, Table II). * refers to applying only utterance level mean normalisation.

	Average	
Feature	**Dysarthric**	**Typical**
MFCC	47.0 (+4.1%)	15.9 (+2.2%)
FBank	50.4 (+1.8%)	17.7 (+9.7%)
Raw-wave	62.7 (-9.6%)	27.5 (-16.8%)
$\text{Mag}^{0.1}$	65.0 (-19.5%)	35.9 (-68.7%)
VT	64.3 (**-21.9%**)	38.1 (**-93.5%**)
Exc	103.5 (-6.9%)	96.0 (-1.3%)
VT+Exc	56.2 (-18.6%)	22.6 (-43.9%)
FBank*	48.3 (+5.8%)	16.3 (+16.8%)
VT+Exc*	47.6 (-0.4%)	15.8 (-0.6%)

representation is employed jointly with the VT within the proposed multi-stream architecture.

This interesting observation illustrates that although Exc does not carry information directly applicable to ASR, it informs the model about the speaker, allowing it to better normalise out the speaker-related attributes, returning the best performance for both typical and dysarthric speech. Factoring the raw magnitude spectrum into the VT and Exc components outperforms MFCC by 1.6% and 0.6% (absolute) in terms of WER for dysarthric and typical speech, respectively, in spite of the data scarcity problem.

The last part of this table displays the WERs when the Exc and VT components are applied jointly with the magnitude spectrum. As seen, the best performance is still achieved by the VT+Exc system. It is also observed that the combination of the magnitude spectrum and VT or Exc or both improves the performance (relative to $\text{Mag}^{0.1}$). However, adding $\text{Mag}^{0.1}$ to the VT+Exc system worsens the performance.

To investigate the statistical significance of the WER improvements, we applied the Matched-Pair Sentence-Segment Word Error (MAPSSWE) [50] which compares the recognition errors made by the two systems. Computing the p-value of the test between WERs of the FBank and VT+Exc systems returns 0.045, which shows the gain is statistically significant.

Table III reports the results after utterance-level mean-variance normalisation along with the relative gain w.r.t. the speaker-level normalisation results shown in Table II. It is observed that the utterance level mean-variance normalisation leads to higher performance for the handcrafted features (MFCC and FBank) whilst decreasing the performance on the raw waveform and raw magnitude-based representations. Quantitatively, maximum gains occur when using the FBank features with improvements of up to 10% (relative) while the biggest performance drop occurs for the VT feature, with a WER increase of more than 93% relative.

The last two rows in Table III show the effect of applying only utterance-level mean normalisation. As seen, skipping the variance normalisation is notably beneficial for both FBank and VT+Exc features. This is owing to the fact that some of the utterances in TORGO are very short and this hinders reliably estimating the variance owing to data deficiency. In case of the speaker-level normalisation, since the data is pooled across all utterances belonging to each speaker, the variances are estimated using more data and therefore, are more reliable.

C. Data augmentation via speed perturbation

We augment the data via a speed perturbation (sp) scheme with 0.9 (slower), 1.0 (original) and 1.1 (faster) speed change factors, increasing the training data by three times. The speed perturbation can be carried out with and without keeping the fundamental frequency (F0) fixed, referred to as sp-F0Fixed and sp, respectively. The difference between these two approaches is substantial: the former keeps the pitch (F0) of the speaker fixed and only simulates various articulation speeds while the latter simulates speakers of different identities (different F0s) speaking with various speeds which renders richer training data [51]. The results of the two speed perturbation methods are reported in Table IV along with Fig. 5 which demonstrates the relative gain for each feature.

As seen, perturbing the articulation speed while keeping F0 fixed, barely provides any performance gain (relative to Table II) despite increasing the training data by three folds. However, allowing F0 to vary offers a notable performance gain for both handcrafted and raw feature representations. As mentioned earlier, the sp scheme allows for simulating many speakers during training which in turn, helps the model to normalise the speaker attributes while sp-F0Fixed only helps the model to normalise the articulation speed. When the sp data augmentation scheme is used, VT+Exc results in the best performance with a WER of 36.1% (23.8% relative gain with respect to Table II) for the dysarthric speech. The best WER for the typical speech is 11.2%, achieved by $\text{Mag}^{0.1}$+VT and $\text{Mag}^{0.1}$+VT+Exc. It slightly outperforms VT+Exc with 11.3% WER (28% relative gain with respect to Table II).

Another important observation is the performance gap between the VT and $\text{Mag}^{0.1}$ features. Although without data augmentation VT outperforms $\text{Mag}^{0.1}$, after applying the sp scheme, $\text{Mag}^{0.1}$ results in a lower WER than VT with a notable margin. Their performance is comparable in case the sp-FixedF0 approach is applied. When the model is fed with

TABLE IV: Mean±STD of WER for dysarthric (Dys) and typical (Typ) speech after data augmentation via speed perturbation without (sp) and with (sp-F0Fixed) keeping F0 fixed.

Setup	sp		sp-FixedF0	
Feature	**Dys**	**Typ**	**Dys**	**Typ**
MFCC	42.3±3.2	12.3±1.2	**50.6**	**18.1**
FBank	47.6±1.4	14.2±0.6	56.9	21.7
Raw-wave	46.0±3.4	17.5±1.5	56.5	25.7
Mag$^{0.1}$	40.6±3.7	11.8±2.6	53.2	18.8
VT	42.8±2.7	12.7±1.1	52.9	20.0
VT+Exc	**36.1**±2.2	11.3±1.2	53.7	20.9
Mag$^{0.1}$+VT	37.1±2.2	**11.2**±1.1	51.0	20.2
Mag$^{0.1}$+Exc	39.2±3.1	12.2±1.5	52.3	19.3
Mag$^{0.1}$+VT+Exc	36.4±2.0	**11.2**±1.7	50.6	19.0

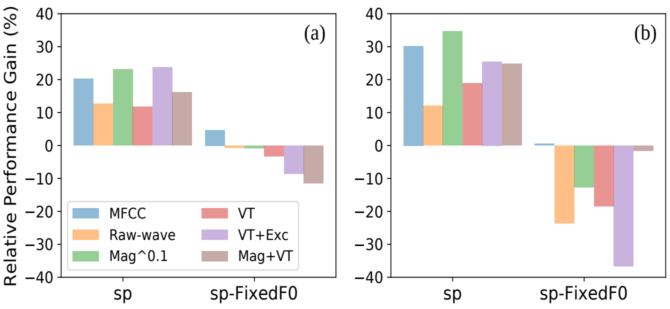

Fig. 5: The relative performance gain(+)/loss(-) after data augmentation via speed perturbation. (a) Dysarthric, (b) Typical.

VT, the speaker-related attributes encapsulated in the Exc are already discarded. This, to a great extent, nullifies the variability induced by the sp scheme in terms of speaker identities and subsequently limits the model's capability in learning the speakers' attributes for normalisation purposes. That is why the VT model benefits the least from the speech perturbation, as depicted in Fig. 5.

D. Training dynamics

Now, we explore the training dynamics of the single-stream Mag$^{0.1}$ and VT acoustic models along with the multi-stream VT+Exc system. By training dynamics we mean the evolution of some performance metrics such as cross entropy (CE) loss or WER vs. training epochs.

Fig. 6 depicts the training dynamics of various models in terms of the CE loss on the validation set in three training conditions: without data augmentation (original training data) and with data augmentation via the sp and sp-FixedF0 schemes. Comparing the results shows that using sp-FixedF0 leads to the highest CE loss. This is in agreement with the WER results reported in Table IV where it led to the highest WERs.

Furthermore, the convergence rate for this scheme is the fastest compared with the models using the original data and sp data augmentation. Except for Mag$^{0.1}$, in all other features the training converges (reaches a plateau) by 10 epochs. This could be explained by considering the fact that the sp-FixedF0 training set simulates various speaking rates for the same set of speakers. This helps the model to only normalise the data variability along the speaking rate dimension which is not helpful in this task. Therefore, there is not much to learn from such extra information, leading to relatively fast convergence.

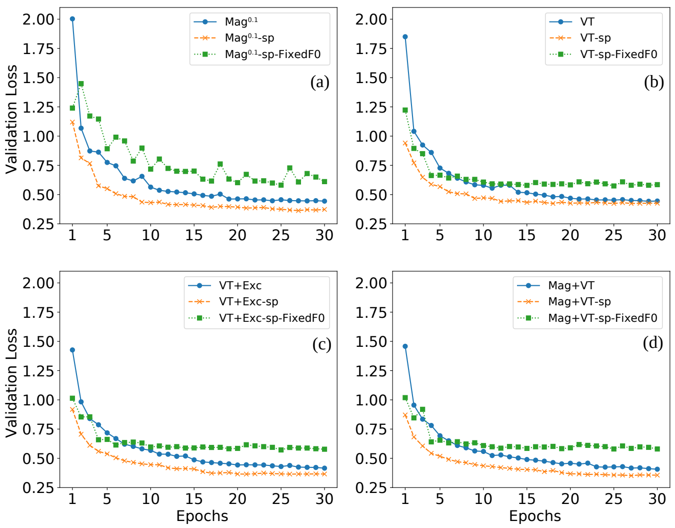

Fig. 6: Training dynamics in terms of CE loss vs. epoch on the validation set for four different inputs in three training conditions. (a) Mag$^{0.1}$, (b) VT, (c) VT+Exc, (d) Mag$^{0.1}$+VT.

On the other hand, the sp scheme leads to the lowest loss value at all epochs compared with training with the original training data and, of course, sp-FixedF0. The only exception to this is the VT feature where the loss after 30 epochs is equal for both the original and sp training conditions (Fig. 6 (b)). As explained earlier, the VT feature benefits the least from the sp scheme as the excitation part is already removed. Another observation is that the sp scheme results in a persistent loss reduction even at high epochs. As seen, even after 30 epochs the loss is still slowly decreasing, implying that more epochs can further improve the performance. Compared with the single-stream systems (Fig. 6 (a) and (b)), this slope is larger for the multi-stream systems (Fig. 6 (c) and (d)), demonstrating that the latter require more epochs to be well trained.

The CE loss is a general criterion used in training the DNNs. It, however, may not strongly correlate with the task-specific performance metrics such as WER. To better study the training dynamics, we computed the WER evolution during training, too. Fig. 7 shows the WER vs. epoch on the test set for the dysarthric and typical speech, with and without data augmentation via the sp scheme. Since the dynamic range of the WER during training is large, we have zoomed in on the results for the last epochs to better highlight the differences.

As seen, data augmentation significantly improves the WER for both dysarthric and typical speech and in all epochs. Comparing the training dynamics of the dysarthric and typical speech also shows that the dysarthric speech requires more epochs and has a slower rate of convergence. This is explainable considering the fact that the dysarthric speech is more complicated than the typical speech (involves more variability); hence, the learning process is slower and demands more training epochs.

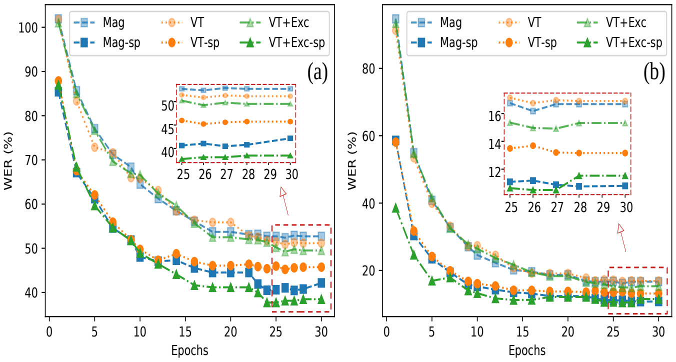

Fig. 7: Training dynamics in terms of WER vs. epoch on the test set for various features. (a) Dysarthric, (b) Typical.

For both dysarthric and typical speech, the overall trend of the WER evolution vs. epoch for the models fed with the raw Mag, VT and VT+Exc features is similar. However, VT+Exc further benefits from extra training epochs and its advantages relative to the single-stream models become more pronounced in the later epochs. Also note that in the Mag-sp and VT+Exc-sp models, around epoch 25, performance jumps out of a local optimum reached around epoch 17 and falls into a better local optimum – in terms of having a lower WER.

E. Speech recognition error analysis

Table V presents the insertion (Ins), deletion (Del) and substitution (Sub) errors along with the corresponding WER for each system. Fig. 8 compares the normalised[2] Ins, Del and Sub errors for the Mag, VT and VT+Exc systems.

It is observed that for the dysarthric speech and without speed perturbation, the Mag system has the fewest Ins errors whilst the VT+Exc system results in the fewest Del and Sub errors. The speed perturbation technique (without keeping F0 fixed) although increasing the number of Ins errors for all three systems, leads to a substantial reduction in the number of deletion and substitution errors. The recognition performance gain achieved by using the speed perturbation technique primarily comes from the reduction in the substitution error. Compared with the sp data augmentation scheme, the sp-FixedF0 method increases all three error types.

Fig. 9 illustrates the evolution of the Ins, Del and Sub errors vs. epoch for the dysarthric and typical speech. It is observed that the Del error converges faster than other errors for both dysarthric and typical speech. The Ins error continuously decreases for the typical speech during training while it increases for the dysarthric speech after 20 epochs. More training is mostly beneficial in decreasing the Sub errors which is the primary source of error. Comparing three systems, VT+Exc-sp outperforms others on the Sub error by a significant margin, while it has a slightly higher Ins error.

TABLE V: Speech recognition errors (Ins, Del, Sub, WER) for the first fold

Feature	Dysarthric Ins	Del	Sub	WER	Typical Ins	Del	Sub	WER
Baseline								
MFCC	30	27	135	53.1	17	25	132	18.2
$Mag^{0.1}$	25	26	137	52.4	13	19	121	16.2
VT	28	22	136	50.9	18	20	119	16.8
VT+Exc	27	20	130	49.2	13	20	108	15.0
$Mag^{0.1}$+VT	28	19	115	**45.8**	13	22	105	**14.8**
$Mag^{0.1}$+VT+Exc	31	24	134	52.3	13	27	138	19.0
Speed perturbation via sp scheme								
MFCC	28	17	106	42.3	14	15	92	12.7
$Mag^{0.1}$	28	15	100	40.5	13	20	109	**10.9**
VT	28	16	118	45.2	14	16	100	13.7
VT+Exc	28	13	94	**37.7**	7	12	59	11.4
$Mag^{0.1}$+VT	25	16	96	38.4	9	17	82	11.2
$Mag^{0.1}$+Exc	24	18	96	38.4	9	20	88	12.4
$Mag^{0.1}$+VT+Exc	24	19	96	38.6	8	19	83	11.7
Speed perturbation via sp-FixedF0 scheme								
$Mag^{0.1}$	33	19	136	52.2	18	18	121	16.5
VT	27	23	138	52.0	17	22	138	18.8
VT+Exc	30	22	139	52.8	20	23	136	19.0
$Mag^{0.1}$+VT	26	24	134	51.0	14	33	143	20.2
$Mag^{0.1}$+Exc	29	24	136	52.3	15	29	138	19.3
$Mag^{0.1}$+VT+Exc	29	21	128	**49.4**	10	26	118	**16.4**

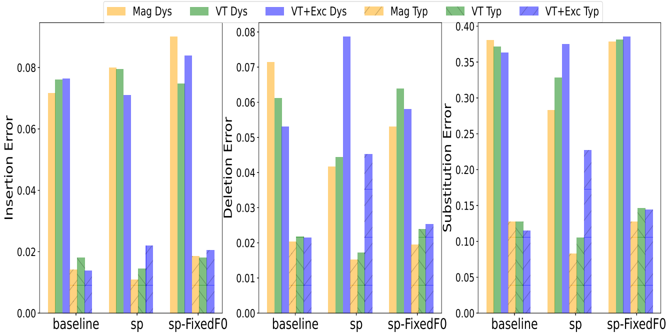

Fig. 8: The normalised insertion, deletion and substitution errors of different models for the dysarthric and typical speech.

We also studied the optimal value for the language model weight (LMWT) that obtains the best WER vs. epoch for speakers in different severity levels in the proposed VT+Exc-sp system. As seen in Fig. 10, the optimal LMWTs is almost fixed across various epochs and is strongly influenced by the severity level. Dysarthric speech requires higher LMWTs than typical speech and the higher the severity level, the higher the optimal LMWTs. The optimal LMWT for the mild dysarthric speech is similar to that for typical speech.

The LMWT is a hyperparameter used to balance the dynamic range of the acoustic model and the language model: a higher LMWT elevates the contribution of the language model in the decoding stage. For severely dysarthric speech, the output of the acoustic model is less reliable and therefore, noisier; hence, allocating a higher weight to the language model and consequently reducing the contribution of the acoustic model improves the WER during the decoding process.

F. Importance of MLP-1

As mentioned earlier, the role of the MLP-1 is primarily fusing the pre-processed information streams. To this end, a linear or non-linear FC layer may be used. We also examine the effect of removing this layer. Table VI shows the WERs for

[2] Normalised by the number of tokens for the dysarthric and typical speech.

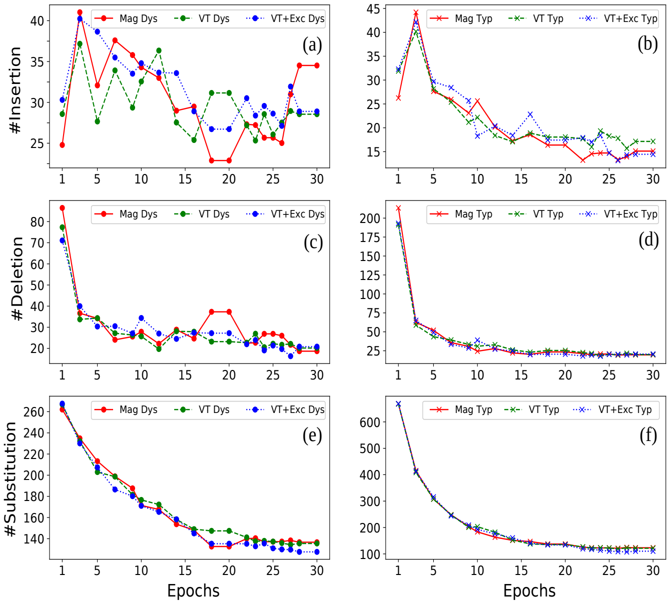

Fig. 9: Training dynamics of the insertion, deletion and substitution errors vs. epoch for various systems.

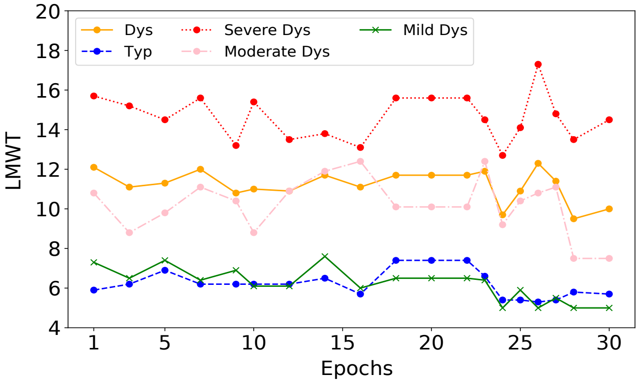

Fig. 10: Optimal LMWTs vs. epoch for various severity levels. Dys denotes the average of the severe, moderate and mild.

TABLE VI: WER of different systems for dysarthric and typical speech. NL: non-linear, L: linear.

Arch Feature	NL-MLP1 Dys	 Typ	L-MLP1 Dys	 Typ	No-MLP1 Dys	 Typ
MFCC	49.0	16.8	47.0	15.5	47.5	15.8
FBank	54.9	20.6	48.7	15.9	47.7	15.3
Mag	55.8	22.1	56.1	20.3	49.8	17.5
VT+Exc	47.4	15.7	47.9	15.7	45.3	14.1
(VT+Exc)-sp	36.1	**11.3**	37.3	11.9	**35.9**	11.7

various features when a non-linear (NL-MLP1) or linear (L-MLP1) fully-connected layer is used as MLP-1 and also when MLP-1 is removed (No-MLP1). Fig. 11 shows the relative (to the default setting, namely NL-MLP1) gain for various configurations. As seen, although the results do not reflect a clear trend, for the dysarthric and typical speech, No-MLP1 and NL-MLP1 lead to the highest performance, respectively.

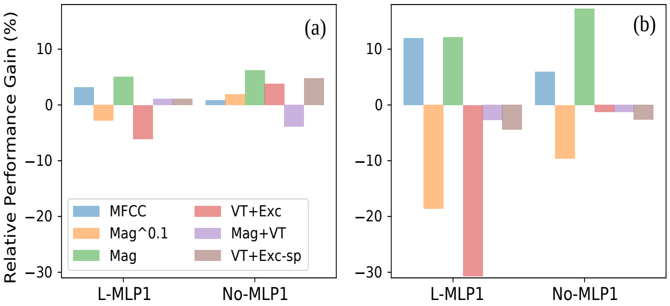

Fig. 11: The relative performance gain/loss of different architectures. (a) Dysarthric, (b) Typical.

TABLE VII: The best WER along with the corresponding epoch (Best-Ep) for various systems (architecture: No-MLP1). * refers to concatenation at the input level.

Feature	Best-Ep	Average Dysarthric	 Typical
MFCC	35	47.5	15.8
FBank	45	45.3	15.0
Mag	30	49.8	17.5
Mag$^{0.1}$	40	52.0	20.5
VT+Exc*	35	44.0	14.0
VT+Exc	45	**43.6**	**13.5**
FBank-sp	45	36.5	11.3
Mag-sp	30	40.0	12.5
VT+Exc*-sp	40	35.5	11.1
VT+Exc-sp	45	**35.3**	**11.0**

We trained the No-MLP1 models which returns the best results for the dysarthric speech, for an extra 20 epochs (50 epochs in total), considering the training dynamics depicted in Fig. 6 where the loss value was still improving with a mild slope around 30 epochs. The best WERs and the epoch at which they were achieved (Best-Ep) are reported in Table VII. In this set of experiments, we added an $VT + Exc^*$ system where * indicates directly concatenating the VT and Exc components at the input level. Compared with $VT + Exc^*$, VT+Exc achieves higher recognition performance by concatenating the two information streams after pre-processing each stream using CNN layers (as shown in Fig. 4). This observation is in line with what was found in [35] that the optimal fusion level should be high enough to effectively pre-process each information stream and low enough to leave sufficient capacity after fusion for post-processing the fused streams. Compared with other features, both VT+Exc and VT+Exc-sp systems further benefit from the extra training epochs. The proposed multi-stream VT+Exc-sp system achieves competitive performance with WERs of 35.3% and 11.0% for the dysarthric and typical speech, respectively.

G. UASpeech database

To explore the generalisation of the proposed method, we applied our best system (the proposed architecture with No-MLP1) to another widely used dysarthric speech corpus,

TABLE VIII: WERs of the best systems for UASpeech. Dys and Both denote using only dysarthric speech and both dysarthric and typical speech for training, respectively.

Training data	**Feature**	**Dysarthric**
Dys	FBank	43.1
Dys	VT+Exc	42.0
Dys	[17]	48.5
Both	FBank	42.9
Both	VT+Exc	42.2
Both-sp	FBank	31.7
Both-sp	VT+Exc	30.3
Both-sp	[24]	32.4
Both-sp	[52]	30.5

namely UASpeech [16]. It contains 102.7 hours of speech recorded from 29 speakers (16 speakers with dysarthria and 13 typical speakers). The dataset only includes isolated word utterances based on single word utterances of digits, computer commands, radio alphabet letters, common and uncommon words. It contains two training sets: only dysarthric and a combination of both dysarthric and typical speech, which we refer to them as *Dys* and *Both* , respectively. The speed perturbation expands the data threefold.

Table VIII shows the performance of the proposed VT+Exc and FBank systems on UASpeech along with results from some recent studies. When the system is trained with only dysarthric speech (Dys) and without speech perturbation, both FBank and VT+Exc systems outperform the result obtained in [17]. The absolute WER reduction of the VT+Exc system over the FBank baseline is 1.1%.

Adding typical speech to the training data (Both) have no significant and consistent positive effect on WER in recognising dysarthric speech. With speed perturbation, the VT+Exc model outperforms FBank by 1.4% absolute WER reduction. Additionally, the proposed VT+Exc system outperform both [24] and the highly competitive system in [52] which utilises QuartzNet [53], CTC [54], meta-leaning [55], [56] and SAT [5]. This indicates that the proposed VT+Exc based system effectively generalise over other dysarthric datasets.

H. Usefulness of speaker adaptation by i-vector

To further highlight the capability of the proposed VT+Exc system in normalising the speaker variability, we compared it with the speaker adaptation via i-vectors [57], [58]. The results on TORGO are shown in Table IX. As seen, adding the i-vectors consistently improves the performance on typical speech by 0.2% to 0.3% absolute WER reduction. For dysarthric speech, the FBank and VT systems benefit from the i-vectors while the VT+Exc system performs slightly worse when i-vector is added. This can be explained by considering the fact that both i-vector and Exc represent the speaker related attributes and having both is rather redundant.

TABLE IX: WERs on TORGO for different systems along with applying i-vector for speaker adaptation.

System	Average	
Feature	**Dysarthric**	**Typical**
FBank	36.5	11.3
VT	36.6	11.2
VT+Exc	35.3	11.0
FBank + i-vector	36.3	11.0
VT + i-vector	36.0	11.0
VT+Exc + i-vector	35.4	10.8

Comparing VT, VT+i-vector and VT+Exc on typical speech shows utilising the i-vector and Exc component are equally useful, leading to identical performance. However, on the dysarthric speech adding Exc to the VT feature results in 1.3% absolute WER reduction while adding i-vector to VT improves the WER by 0.6%. Also note that extracting i-vector involves learning a universal GMM and developing a separate system to extract such embeddings while extracting the Exc component is as easy and as fast as computing FFT and a liftering process.

V. Towards Understanding the Learned Models

In this section, we take some steps towards understanding the models' behaviour. In particular, we analyse the filters in the first convolutional layer as well as how the model normalises the nuisance variabilities such as speech style (dysarthric vs. typical) or the speaker identities.

A. Interpretation of the learned filters

Fig. 12 illustrates the average of the first convolutional layer (Conv-L1) for models shown in Fig. 4 along with the magnitudes of their FFT. Note that these filters are fed with the compressed magnitude spectrum (of the signal, VT or Exc), not the time domain signal; hence, after taking the Fourier transform (FT), the domain will be analogous to the cepstral domain. We noticed that the filters' shapes in the original domain are not readily interpretable. However, taking the FT sheds further light on the behaviour of the learned filters, providing some informative insights into their functionality.

As seen in Fig. 12 (a), on average, the filters fed with the $\text{Mag}^{0.1}$ act like a low-pass lifter. The low cepstral components are correlated with the vocal tract and speech's lingual content. The acoustic model trained for ASR learns about such variability and more heavily focuses on these components. When the model is fed with VT, similar behaviour is exhibited, with the difference that the model does not pay attention to cepstral components beyond 50 samples. This is an interesting observation, too, because the information encoded in the VT part is concentrated within the first 50 samples and essentially the VT component has no information beyond 50 samples (recall how the VT component was computed in Section 2). On the other hand, the Exc part has no information in low quefrencies; the model learns about this property, too, and on average disregards the first 50 samples.

Fig. 12 (b), (c) and (d) show the average of the magnitude of filters' FFT for filters in the first convolutional layer in various single and multi-stream systems. It is observed that in the multi-stream systems, the role which each stream can

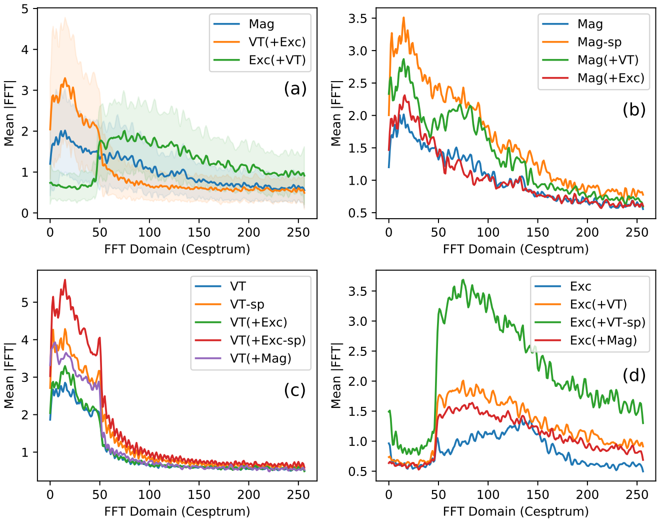

Fig. 12: Average of the learned filters in the first convolutional layer. A(+B) indicates the average of the filters of the head fed with A in a multi-stream system fed with A and B features.

play is enhanced and the streams complement each other. Let A(+B) in this figure mean the average of the filters in the first layer of a sub-network (head) takes A input, in a multi-stream system fed with both A and B features. As seen in Fig. 12 (b), comparing the single-stream Mag system with the multi-stream Mag(+VT) shows that the latter pays extra attention to the quefrencies beyond 50 which are missed by the VT and are uniquely captured by the Mag input. On the other hand, comparing the single-stream Mag system with the multi-stream Mag(+Exc) illustrates that the Mag head in Mag(+Exc) gives relatively less attention to components beyond 50 as the Exc head will handle that part better. Similar complementary behaviour is seen in Fig. 12 (c) and (d).

B. Speaker normalisation across layers

Fig. 13 depicts the scatter plot for the VT+Exc system after applying dimensionality reduction on the outputs of different layers for the dysarthric and typical speech. Dimensionality reduction to 2D was carried out in two stages after initially standardising the activations (mean-variance normalisation) per dimension. First, for each layer the activations vector was reduced to 50D using the principal component analysis (PCA) [59]; second, the dimensionality was further reduced to 2D through t-distributed stochastic neighbouring entities (t-SNE) [60]. The activations were calculated by the forward propagation of a randomly selected portion of the dev set.

We computed the centroids for the typical and dysarthric speech at different layers and calculated the Euclidean distance between them. Ideally, the distance between the centroids representing each speech type should decrease for layers closer to the output as this implies the model is normalising the dysarthric and typical speech types to some canonical representation. Aligned with this insight, we noticed that this is indeed the case. As seen in Fig. 13, the distance between the centroids after the last recurrent layer, MLP-2 and the softmax output layer is 13.4, 11.6 and 0.7, respectively.

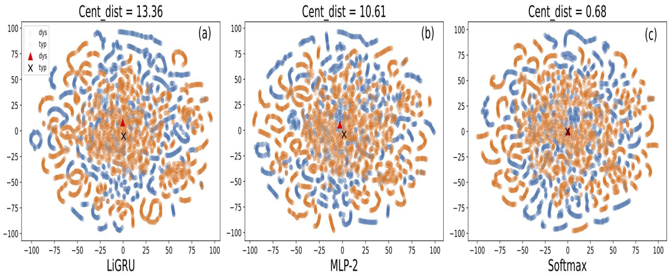

Fig. 13: Scatter plot after applying dimensionality reduction to 2D on activations of the (a) LiGRU block, (b) MLP-2 layer and (c) Softmax layer. The red triangle and black cross indicate the centroids for dysarthric and typical speech, respectively.

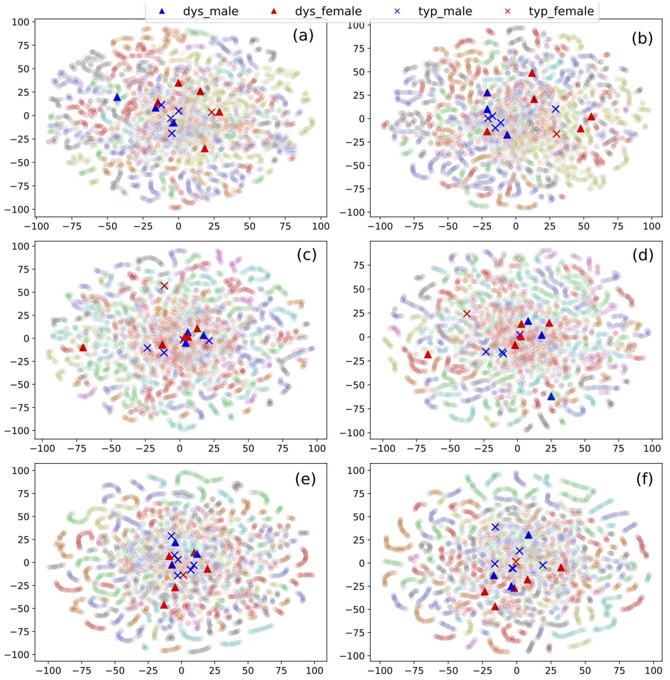

Fig. 14: Visualisation of the softmax layer activations after dimensionality reduction to 2D for different models. Triangles and crosses denote the centroids for the dysarthric and typical speech types whilst the red and blue markers denoted male and female speakers, respectively. (a) Mag$^{0.1}$ epoch 10, (b) VT+Exc epoch 10, (c) Mag$^{0.1}$ epoch 30, (d) VT+Exc epoch 30, (e) Mag$^{0.1}$-sp epoch 30, (f) VT+Exc-sp epoch 30.

We conducted a similar study, but this time the centroids were computed considering the speaker identities (IDs). Based on the aforementioned insight, the model should learn to normalise the nuisance and irrelevant variabilities such as speaker IDs. As such we expect that during training, the corresponding clusters get closer to each other. Comparing the 10^{th} and 30^{th} epochs in Fig. 14 (a)-(d) shows that for both single-stream Mag$^{0.1}$ and multi-stream VT+Exc systems, the centroids get closer to each other by epoch 30.

Furthermore, although the male and female speakers are still separable by epoch 10, by epoch 30 they are further mixed and harder to separate. This is another indicator of speaker normalisation by the model during training process. As seen in Fig. 14 (e) and (f), data augmentation also contributes towards further restricting the spread of the speaker centroids and consequently a better speaker normalisation.

Finally, comparing the centroids representing the dysarthric speakers (triangles) with the typical ones (crosses) demonstrates that the dysarthric speakers are distributed in a larger subspace. This implies a higher speaker variability than the typical speech which makes ADSR a more challenging task.

VI. Conclusion and scope for future work

In this paper, we developed a 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 a multi-stream architecture. Having pre-processed each stream by a CNN, they were fused via a fully-connected layer and post-processed by a cascade of recurrent and fully-connected layers. We studied the effects of data augmentation via speed perturbation and feature normalisation and, scrutinised the training dynamics of the models in terms of the CE loss and WER for both dysarthric and typical speech. In addition, the role of the excitation component for speaker normalisation were investigated and compared with speaker adaptation via i-vector. We also analysed the learned filters in the first convolutional layer for various input features and showed that the proposed model normalises the speech type and speakers attributes. The proposed system achieved a competitive performance with up to 35.3% and 11.0% WERs for the dysarthric and typical speech on the TORGO and 30.3% on the UASpeech databases. Future work includes learning domain-invariant features via adversarial training, leveraging unlabelled data via semi-supervised learning and transfer learning by employing the self-supervised pre-trained models.

References

[1] J. Duffy, *Motor Speech disorders-E-Book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[2] 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.

[3] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[4] J. Wilson, B., "Acoustic variability in dysarthria and computer speech recognition," *Clinical Linguistics & Phonetics*, vol. 14, no. 4, pp. 307–327, 2000.

[5] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *INTERSPEECH*, 2012, pp. 1776–1779.

[6] N. Joy and S. Umesh, "Improving acoustic models in torgo dysarthric speech database," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 637–645, 2018.

[7] B. Blaney and J. Wilson, "Acoustic variability in dysarthria and computer speech recognition," *Clinical Linguistics & Phonetics*, vol. 14, no. 4, pp. 307–327, 2000.

[8] K. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *ICASSP*. IEEE, 2011, pp. 4924–4927.

[9] 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.

[10] J. Deller, M. Liu, L. Ferrier, and P. Robichaud, "The Whitaker database of dysarthric (cerebral palsy) speech," *The Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.

[11] X. Menendez-Pidal, J. Polikoff, S. Peters, J. Leonzio, and H. Bunnell, "The Nemours database of dysarthric speech," in *International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, Oct. 1996, pp. 1962–1965.

[12] 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, pp. 1741–1744.

[13] 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.

[14] E. Hermann *et al.*, "Dysarthric speech recognition with lattice-free mmi," in *ICASSP*. IEEE, 2020.

[15] C. Espana-Bonet and J. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 97–107.

[16] M. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional lstm neural network," *INTERSPEECH*, pp. 2948–2952, 2018.

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

[18] 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.

[19] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, "Investigation of data augmentation techniques for disordered speech recognition." in *INTERSPEECH*, 2020, pp. 696–700.

[20] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *ICASSP*. IEEE, 2018, pp. 6009–6013.

[21] Z. Jin, M. Geng, X. Xie, J. Yu, S. Liu, X. Liu, and H. Meng, "Adversarial data augmentation for disordered speech recognition," *INTERSPEECH*, 2021.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

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

[31] ——, "Speech acoustic modelling from raw phase spectrum," in *ICASSP*. IEEE, 2021, pp. 6738–6742.

[32] Z. Yue, E. Loweimi, and Z. Cvetkovic, "Raw source and filter modelling for dysarthric speech recognition," in *ICASSP*. IEEE, 2022.

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

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

[35] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition," in *ICASSP*. IEEE, 2022.

[36] 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.

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

[38] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines." in *ICML*, 2010, pp. 807–814.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[45] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Workshop on Autodiff*, 2017.

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

[47] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," *ICASSP*, pp. 2494–2498, 2014.

[48] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[49] 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.

[50] D. Pallet, W. Fisher, and J. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *ICASSP*. IEEE, 1990, pp. 97–100.

[51] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *ICASSP*. IEEE, 1993, pp. 554–557.

[52] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu, and H. Meng, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[53] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP*, 2020, pp. 6124–6128.

[54] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *ICML*, 2006.

[55] A. Nichol, J. Achiam, and J. Schulman, "On First-Order Meta-Learning Algorithms," *arXiv e-prints*, 2018.

[56] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," ser. ICML, 2017, p. 1126–1135.

[57] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[58] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 55–59.

[59] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, no. 1-3, pp. 37–52, 1987.

[60] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Zhengjun Yue is a research associate in King's College London (KCL). She received her B.Sc. degree in telecommunication engineering from Shanghai University, China, in 2017, and the M.Sc. degree in Artificial Intelligence from the University of Edinburgh, United Kingdom, and is doing Ph.D. in the University of Sheffield, United Kingdom since 2018. She is now the Ph.D. candidate in Speech and Hearing Group from the University of Sheffield, UK. Her research interests lie in acoustic modelling and acoustic-articulatory multi-modal speech recognition for dysarthric speech, end-to-end and robust ASR, and deep learning.

Erfan Loweimi (S'10 — M'18) is a research associate with King's College London (KCL) and a visiting researcher in the Centre for Speech Technology Research (CSTR) in the University of Edinburgh where he was a post-doc in 2018-2021. He received the B.Sc. (2007), M.Sc. (2011) and Ph.D. (2018) degrees from the Shahid Chamran University of Ahvaz, Amirkabir University of Technology (Tehran Polytechnic) and University of Sheffield, respectively. His research interests lie in the area of acoustic modelling from raw signal representations, end-to-end ASR, robust model-based ASR and phase-based speech signal processing.

Heidi Christensen received the M.Sc. and Ph.D. degrees from Aalborg University, Denmark, in 1996 and 2002, respectively. She is a Professor in Spoken Language Technologies in the Computer Science department at the University of Sheffield. Before that she has held post-doc positions at the University of Sheffield, IDIAP, Switzerland and Aalborg University, Denmark. Her main research interests are in the areas of recognition of disordered speech, automatic processing of conversations, and the automatic detection and tracking of paralinguistic information such as emotions and general interactional behaviours.

Jon Barker received the B.A. degree in electrical and information sciences from the University of Cambridge, UK, in 1991 and the Ph.D. degree in computer science from the University of Sheffield, UK, in 1998. He has worked as a Researcher at GIPSA-lab, Grenoble, France, studying audiovisual speech perception and has spent time as a Visiting Research Scientist at IDIAP, ICSI (Berkeley, CA) and the Columbia University. He is currently a Professor in Computer Science at the University of Sheffield. His research interests include human speech processing, robust automatic speech recognition and machine listening.

Zoran Cvetkovic (Senior Member, IEEE) received the Dipl.Ing. and Mag. degrees from the University of Belgrade, the M.Phil. degree from Columbia University, and the Ph.D. degree in electrical engineering from the University of California, Berkeley. He is currently a Professor of Signal Processing with King's College London. He held research positions with EPFL (1996), and with Harvard University (2002–2004). Between 1997 and 2002, he was a member of the technical staff of AT&T Shannon Laboratory. His research interests are in the broad area of signal processing, ranging from theoretical aspects of signal analysis to applications in audio and speech technology. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.