Multi-stream Acoustic Modelling using Raw Real and Imaginary Parts of the Fourier Transform

Multi-stream Acoustic Modelling using Raw Real and Imaginary Parts of the Fourier Transform

Erfan Loweimi *(Member, IEEE)*, Zhengjun Yue *(Member, IEEE)*, Peter Bell *(Member, IEEE)*, Steve Renals *(Fellow, IEEE)*, Zoran Cvetkovic *(Senior Member, IEEE)*

***Abstract*—In this paper, we investigate multi-stream acoustic modelling using the raw real and imaginary parts of the Fourier transform of speech signals. Using the raw magnitude spectrum, or features derived from it, as a proxy for the real and imaginary parts leads to irreversible information loss and suboptimal information fusion. We discuss and quantify the importance of such information in terms of speech quality and intelligibility. In the proposed framework, the real and imaginary parts are treated as two streams of information, pre-processed via separate convolutional networks, and then combined at an optimal level of abstraction, followed by further post-processing via recurrent and fully-connected layers. The optimal level of information fusion in various architectures, training dynamics in terms of cross-entropy loss, frame classification accuracy and WER as well as the shape and properties of the filters learned in the first convolutional layer of single- and multi-stream models are analysed. We investigated the effectiveness of the proposed systems in various tasks: TIMIT/NTIMIT (phone recognition), Aurora-4 (noise robustness), WSJ (read speech), AMI (meeting) and TORGO (dysarthric speech). Across all tasks we achieved competitive performance: in Aurora-4, down to 4.6% WER on average, in WSJ down to 4.6% and 6.2% WERs for Eval-92 and Eval-93, for Dev/Eval sets of the AMI-IHM down to 23.3%/23.8% WERs and in the AMI-SDM down to 43.7%/47.6% WERs have been achieved. In TORGO, for dysarthric and typical speech we achieved down to 31.7% and 10.2% WERs, respectively.**

***Index Terms*—Raw signal representation, Fourier transform, automatic speech recognition, multi-stream acoustic modelling**

I. Introduction

HANDCRAFTED magnitude spectrum based features such as MFCC [1], PLP [2] and filterbank energies (FBank) are widely employed in various automatic speech recognition (ASR) and speech classification tasks. The corresponding pipelines of such front-ends are engineered based on properties of the human speech production and perception systems, along with considerations related to requirements of the back-end (*e.g.*, an acoustic model or a classifier). However, they are task-blind and do not directly take into account its specific requirements. That is, it is likely to irreversibly discard some task-useful information along the pipeline whilst passing through task-irrelevant information. If there is no redundancy, there is no chance to recover the lost information: essentially, the back-end can only process the given information; based on the *data processing inequality* [3] a chain of sequential transformations cannot generate new information. On the other hand, passing irrelevant information to the back-end, complicates the back-end's function as it should filter out noisy information and carry out the decision making. A well-trained sufficiently-deep structure is capable of filtering out the nuisance information and representing the data properly. However, it cannot compensate for the useful information lost along the front-end pipeline. This can be costly performance-wise, regardless of the back-end's capabilities.

Direct employment of the raw signal representations with minimal or no information loss, along with powerful learning architectures can effectively tackle the aforementioned issue. This has led to successful single-stream acoustic modelling using raw waveforms [4]–[12], raw magnitude [13], deep scattering spectrum [14] and raw phase spectrum [15] with better or comparable performance to the classic features across various tasks, even for databases as small as TIMIT [16]–[18].

Along with benefits stemming from preserving all signal information, acoustic modelling can also take advantage of multi-stream processing [19] where the model is presented with multiple information streams at the *input* level[1]. Examples of recent multi-stream acoustic models are the raw magnitude spectrum along with the *sign spectrum* [22] and the raw source and filter components [15], [23], [24]. Compared with single-stream raw waveform modelling, the multi-stream approach can take advantage of some prior knowledge about the input (*e.g.*, source-filter separation [25] for the speech signal) to decompose the input into multiple complementary information streams. It also allows for learning a bespoke pre-processing per stream, weighing each stream based on its importance to the given task and fusing the streams in an optimal level of abstraction. For example, using the raw waveform or raw magnitude spectrum implicitly means fusing the source and filter components at the input level whilst in the multi-stream approach the vocal tract and excitation streams are pre-processed individually based on their usefulness to the task. When fused at an optimal level of abstraction, the multi-stream approach provides a more effective framework for information processing than its single-stream counterparts.

In this paper, we study multi-stream acoustic modelling using the raw real and imaginary parts of the Fourier transform

Manuscript received dd mmmm yyyy; revised dd mmmm yyyy; revised dd mmmm yyyy; accepted dd mmmm yyyy. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rohit Prabhavalkar.

E. Loweimi (Corresponding author), Z. Yue and Z. Cvetkovic are with King's College London (KCL); e-mail: {erfan.loweimi, zhengjun.yue, zoran.cvetkovic}@kcl.ac.uk. E. Loweimi, P. Bell and S. Renals are with The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK; e-mail: {e.loweimi, peter.bell, s.renals}@ed.ac.uk.

Supported by EPSRC Project EP/R012180/1 (SpeechWave).

[1]The multi-streaming can also occur in the *medium* levels by creating multiple branches from a single stream, *e.g.*, as in [20] or using multiple heads in a Transformer [21] block. These are outside the scope of this paper.

(FT). That is, the FT is used to convert the single-stream raw waveform into two orthogonal streams which together uniquely characterise the signal, without any information loss. Combining and representing them in the form of the magnitude spectrum leads to irreversible information loss. In the signal reconstruction literature (*e.g.*, [22], [26]), it is well-established that, in general, a signal cannot be uniquely characterised or equivalently perfectly reconstructed using only its magnitude spectrum. In particular, mixed-phase signals like speech [27], [28] are not uniquely specifiable by their magnitude spectrum. As such once the real and imaginary parts are fused as the magnitude spectrum via summing their squares, the phase spectrum information is irreversibly lost. In the proposed multi-stream framework, not only is such information loss avoided, but also these two parts are pre-processed and fused in a task-specific way, leveraging the data-driven learning paradigm rather than a task-blind rule-based combination.

Compared with source-filter separation, factorising the signal into the FT's real and imaginary parts predicates on a much weaker prior knowledge. That is, it is no longer limited to speech and is applicable to all sequences with FT such as audio, image and biomedical signals. In comparison with a multi-stream system fed with the magnitude and phase spectra, the real and imaginary parts are easier to compute as the former may involve phase unwrapping which is challenging. In fact, the phase and magnitude spectra are lossy non-linear representations of these two parts. The model will learn appropriate non-linear representations when trained effectively.

The proposed framework is also comparable with *complex-valued neural networks* (CVNNs) [29], that have been employed in ASR [30], speech enhancement [31], [32] and audio [33] applications. These networks involve redesigning some atomic components such as activation function, differentiation, convolution, batch normalisation and even initialisation [34] to handle complex operations while the proposed approach is entirely based on the real-valued well-established operations. Further, in CVNNs the real and imaginary parts are effectively combined after the first layer, whilst the proposed framework is more flexible, allowing for pre-processing the streams individually by two distinct chains of non-linear transformations and fusing them at an arbitrarily higher level of abstraction. Complex linear projection (CLP) [30] layer is an example of CVNNs where the real and imaginary parts are combined by a complex fully-connected layer. It is equivalent to convolution and average pooling in the time domain [30]. In contrast, we pre-process the real and imaginary parts with two separate branches of real-valued CNNs along with max-pooling and the fusion can occur in an arbitrarily higher level.

Having reviewed the information content of the magnitude spectrum relative to the real and imaginary parts and characterised the information gain in Section II, we put forward multiple architectures to efficiently pre-process and fuse these two information streams in Section III. Section IV includes experimental results along with a discussion of various observations and modelling issues. In Section V, the learned filters in the first convolutional layer of single- and multi-stream models are analysed. Conclusions and directions for future work are presented in Section VI.

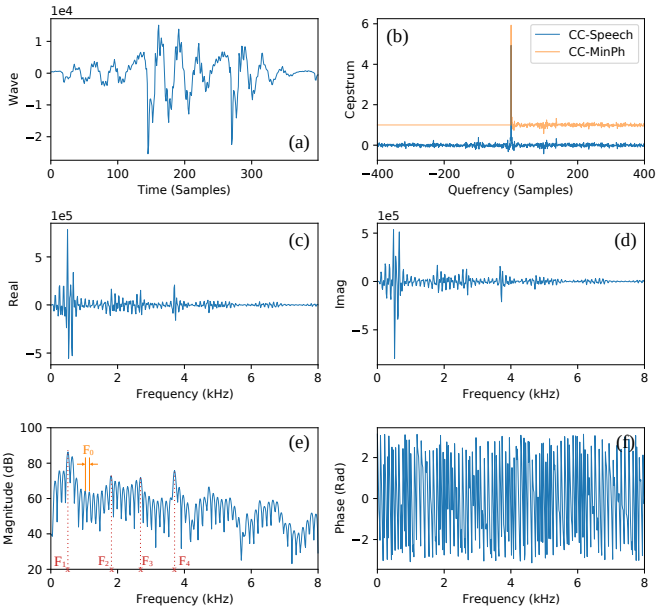

Fig. 1. Various signal representations for a speech frame. (a) waveform, (b) complex cepstrum (CC) of the frame along with the complex cepstrum of its Minimum-Phase counterpart (CC-MinPh), (c) Real part, (d) Imaginary part, (e) Magnitude spectrum, (f) Principle (wrapped) phase spectrum. Formants (F_1, F_2, F_3 and F_4) and fundamental frequency (F_0) are highlighted in the magnitude spectrum. To better illustrate the causality of the complex cepstrum of the MinPh signal (CC-MinPh), we added a small DC offset to it in (b).

II. Information Content of the FT's Components

The Fourier transform returns a sequence of complex numbers which can be represented in the Cartesian or polar coordinates via the real and imaginary parts or magnitude and phase spectra, respectively. The magnitude spectrum is the most widely used spectral representation: it is shift-invariant and relatively easy to model, process and enhance [35]. For each frequency bin, it is a non-negative quantity, proportional to the signal energy in the corresponding bin. Fig. 1 shows a speech frame along with its various representations. As seen, some characteristics of the speech signal such as fundamental frequency (F_0) and formants (F_1, F_2, ...) are clearly highlighted in the magnitude spectrum and can be easily extracted.

However, information-wise, the magnitude spectrum is not the most informative component of the FT. Consider a single speech frame, $x[n]$; it can be reconstructed up to a scale error from its phase spectrum [26]. Also, assuming the sequence is *causal* (equals zero up to a certain time instant), it can be exactly reconstructed from the real part using the Hilbert transform [36]. Utilising the imaginary part, causal signals can be reconstructed up to an additive error [36] (proportional to the value of the signal at zero, $x[0]$).

On the other hand, the magnitude spectrum can only uniquely characterise minimum-phase (MinPh) signals [36]. For these signals all the zeros and poles are located inside the unit circle or equivalently, their complex cepstrum (CC) is causal, *i.e.*, equals zero in the negative quefrencies [36]. As seen in Fig. 1 (b), speech is not a minimum phase signal because its CC is not causal. Therefore, this mixed-phase signal is not uniquely specifiable from its magnitude spectrum

and even if the entire magnitude spectrum is employed, some information is already irreversibly discarded.

The lost information is uniquely captured by the phase spectrum and shown to be important from the perceptual standpoint [37]. That is, the quality/intelligibility of the magnitude-only reconstructed signal is significantly less than the original one. This point has been verified by both subjective [38], [39] and objective [28], [40]–[43] tests. Such detrimental information loss which is important from both theoretical and perceptual perspectives, potentially harms the performance of the raw magnitude-based systems.

To be more precise, the magnitude spectrum is partially blind to the phase spectrum information. Since speech is a mixed-phase signal, it can be factorised using the *minimum-phase/all-pass decomposition* [36]

$$X(\omega) = X_{Re}(\omega) + jX_{Im}(\omega) = X_{MinPh}(\omega)X_{AllP}(\omega) \tag{1}$$

$$|X(\omega)| = |X_{MinPh}(\omega)| \tag{2}$$

$$\arg\{X(\omega)\} = \arg\{X_{MinPh}(\omega)\} + \arg\{X_{AllP}(\omega)\} \tag{3}$$

where X, ω, $|.|$, arg, Re, Im, $AllP$ and $MinPh$ denote the (short-time) Fourier transform, angular frequency, magnitude, unwrapped phase, real part, imaginary part, all-pass and minimum-phase components, respectively.

For the minimum-phase signals, the phase spectrum is the Hilbert transform ($\mathcal{H}$) of the $\log|X(\omega)|$

$$\arg\{X_{MinPh}(\omega)\} = \mathcal{H}\{\log|X(\omega)|\} \tag{4}$$

$$\Rightarrow X(\omega) = |X(\omega)|\ e^{j\mathcal{H}\{\log|X(\omega)|\}}\ e^{\phi_{AllP}(\omega)}. \tag{5}$$

This implies that the information encoded in the MinPh component of the phase spectrum is already captured by the magnitude spectrum owing to this one-to-one relationship. However, the magnitude spectrum is blind to the all-pass element. Therefore, when using the real and imaginary parts, the information gain relative to applying the raw magnitude spectrum (or its representations) will be as much as the information encoded in the (phase of the) all-pass component.

To quantify the importance of the all-pass component, we compare the quality and intelligibility of the speech reconstructed from magnitude spectrum with the original one. Information-wise, the information of a mixed-phase signal is the union of the information encoded in the minimum-phase and all-pass components (Eq. 1). The former is entirely captured by the magnitude spectrum (Eqs. 2 and 4). Therefore, the distance between the quality/intelligibility of the original signal with those of the magnitude-only reconstructed speech can serve as a proxy for the all-pass phase information content.

For signal reconstruction from magnitude spectrum, we use the well-established Griffin-Lim method [44]. This iterative analysis-modification-synthesis-based algorithm aims at recovering the phase information by exploiting the contextual/overlapping frames. To this end, we used Hamming windows with 25 ms frame length and 75% overlap and the number of iterations was set to 100. The quality and intelligibility were measured during the process via PESQ [45] and STOI [46] metrics, respectively. We used random (uniformly distributed in $[-\pi, \pi]$), zero and phase of the minimum-phase component for initialising the phase spectrum. NOIZEUS [47] database was used in this set of experiments and the error bars at Fig. 2 show the mean and standard deviations over 30 signals.

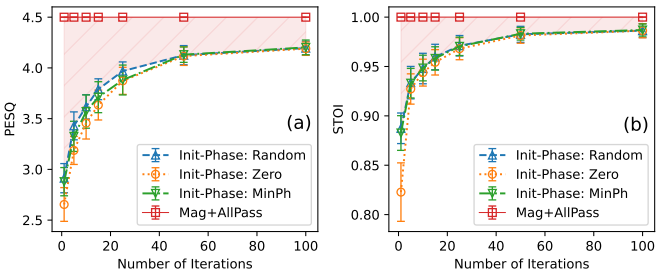

Fig. 2. Iterative magnitude-only signal reconstruction via Griffin-Lim method along with various phase initialisation (init-phase). The red zone demonstrates the difference between the original and the magnitude-only reconstructed signals and is a proxy for the importance of the all-pass component in terms of the speech quality (PESQ) and intelligibility (STOI).

The red zones in Fig. 2 (a) and (b) show the contribution of the all-pass component in terms of the speech quality (PESQ) and intelligibility (STOI), respectively. The quality and intelligibility of the reconstructed signals improve along the iterative reconstruction process, but it never reaches the perfect level. That is, using the contextual information and magnitude spectrum, the Griffin-Lim algorithm can only partially recover the all-pass information. As seen, the all-pass component has a noteworthy information content, contributing towards improving both quality and intelligibility of the speech.

The all-pass component is also instrumental in beamforming and microphone array processing applications. For example, in the well-established delay-and-sum method [48], the differences between delays with which sensors receive signals play a key role in steering the beam towards the direction of interest. These delays do not affect the magnitude spectrum and consequently the phase of the minimum-phase component (Eq. 4), and are uniquely captured by the all-pass components of the phase spectra of sensor signals. As such making the model aware of the all-pass component can also be helpful towards enhancing the directivity, particularly when learning the beamforming jointly with the end goal. For example, in LIMABEAM [49] and neural network adaptive beamforming (NAB) [50]–[52] the beamforming is jointly learned/done with acoustic modelling.

Despite being blind to the all-pass part, the magnitude spectrum is widely employed in speech processing owing to its relatively clear behaviour which greatly facilitates engineering a pipeline to process, normalise noise and represent the speech signal. However, in a data-driven framework, this advantage is no longer a necessity because the representation is learned rather than being handcrafted based on some prior expert knowledge and/or rule-based transformations.

Therefore, although the real and imaginary parts' behaviour might not be as favourable as the magnitude spectrum from a feature engineering perspective, in learning-based frameworks this issue no longer poses an obstacle. Besides, information-wise they include all the information encoded in the signal and inherit some desirable properties of spectral representations.

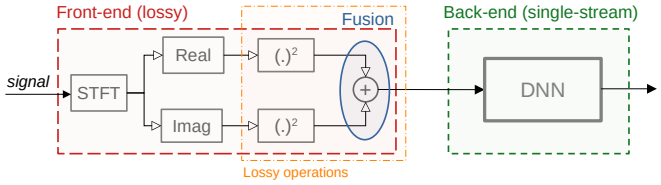

Fig. 3. The baseline single-stream raw magnitude-based model. The red zone is a rule-based front-end, the green part is a data-driven back-end and the orange zone indicates lossy operations irreversibly discard information.

III. Multi-Stream Acoustic Modelling Using Real and Imaginary Parts

As discussed, the real and imaginary parts together uniquely characterise the signal, while when using the magnitude spectrum or its representations, some perceptually important information is irreversibly discarded. Therefore, there is a potential perceptually significant information gain when employing the raw real and imaginary parts instead of the magnitude spectrum. This is, however, a theoretical benefit and we need a framework to practically realise such potential. To this end, we first enumerate the shortcomings of a baseline single-stream raw magnitude-based system and considering these, propound some multi-stream architectures which allow for addressing those limitations and leveraging the extra information.

A. The Baseline System

Fig. 3 illustrates a single-stream acoustic model fed with the raw magnitude spectrum. The workflow starts with the short-time Fourier transform (STFT), transforming the signal using a set of complex exponential basis functions. Taking the real and imaginary parts returns two information streams which together uniquely specify the signal. In the next stage, these two streams pass through a non-linear quadratic function which is a lossy operation as the sign information determining the quadrants is lost. Fusing the streams via summing them, outputs the raw magnitude spectrum and results in losing the phase information. Finally, the raw magnitude spectrum is passed to a single-stream acoustic model.

The front-end consists of a series of heuristic operations which give rise to the following issues:

- First, as elaborated in Section II, it is a lossy workflow leading to irreversible information loss. The discarded information, namely the phase spectrum, is perceptually significant and contributes towards improving the quality and/or intelligibility of the speech signal.
- Second, for all of the streams, identical pre-processing[2] (a quadratic function) is used, which is not necessarily optimal. Each information stream encodes a different set of information because all streams are needed for a unique signal characterisation. As such their information content and consequently the per-stream pre-processing transformations should be different and stream-specific.
- Third, it is very challenging, if not impossible, to design such stream-wise pre-processing stages using hand-crafted rule-based techniques. In fact, only a data-driven framework (with adequate data, architecture and training regime) can learn such a complicated series of linear/non-linear task- and stream-specific chain of transformations.
- Fourth, fusion happens at a very low level after merely squaring each stream. Fusion at higher levels paves the way for conducting more effective stream-wise information filtering and representation learning. It will partially filter out nuisance variabilities and represent each information stream in an optimal form before fusion.
- Fifth, summing the streams whilst giving the same weight to both is suboptimal. It neglects the relative importance and/or contribution of different streams to the task. Some weighting or gating mechanisms should be built into the fusion apparatus to account for the significance and relevance of each stream to the given task.

[2]Here, the pre- and post- processing are defined relative to the fusion point.

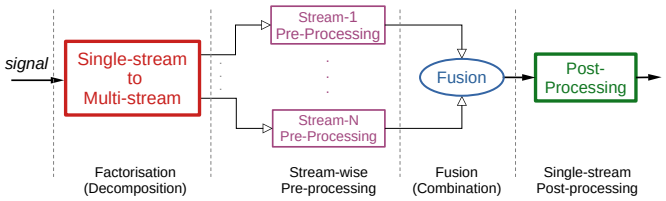

Fig. 4. Canonical structure of a multi-stream information processing system consisting of a front-end extracts multiple information streams from a single-stream input, per-stream pre-processing, fusion and post-processing stages.

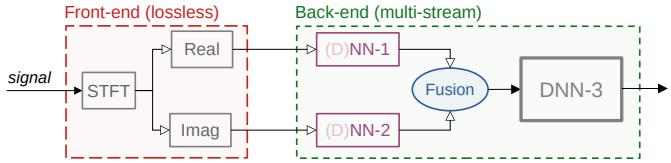

Fig. 5. The proposed multi-stream system fed with the real and imaginary parts of the FT. Each stream is pre-processed by (D)NN-1 and (D)NN-2 which can be (deep or) shallow. DNN-3 post-processes the fused streams.

B. Canonical Structure of a Multi-Stream System

To address and tackle these shortcomings of the single-stream raw magnitude spectrum based system, there is a need for a multi-stream framework that involves

- replacement of heuristic operations causing task-blind information loss with a data-driven stream-wise pre-processing leading to task-dependent information loss;
- decomposition of the single-stream input into multiple information streams;
- appropriate pre-processing for each information stream;
- a fusion mechanism at an optimal level of abstraction;
- adequate single-stream post-processing.

Fig. 4 shows the canonical structure for a multi-stream system.

As seen in Fig. 5, in the proposed multi-stream framework, the Sine and Cosine basis functions of the Fourier transform factorise the single-stream raw waveform while collectively preserving all signal information. For per-stream pre-processing before fusion, a sub-network – such as a series of convolutional layers – can be employed to effectively carry

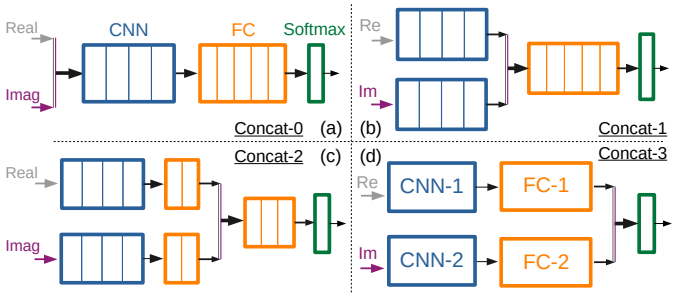

Fig. 6. Multi-stream acoustic models consisting of only convolutional and fully-connected (FC) layers. Fusion can happen at the (a) input (Concat-0), (b) medium (Concat-1), (c) high (Concat-2) and (d) very high (Concat-3) levels.

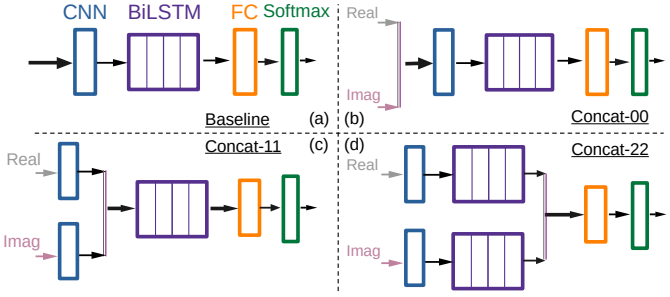

Fig. 7. Multi-stream acoustic models consisting of convolutional, recurrent (BiLSTM) and fully-connected (FC) layers. (a) Single-stream baseline, (b)-(d) multi-stream with fusion at different abstraction levels.

out the representation learning. For fusion, we concatenate the pre-processed streams and linearly combine them as follows

$$Fusion(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N) = W \underbrace{\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N\}}_{Concatenate}, \quad (6)$$

where $\{\mathbf{s}_i, \mathbf{s}_j\}$ indicates concatenating the i^{th} and j^{th} pre-processed streams and W denotes a learnable weight matrix that linearly combines the streams. Finally, the output of the fusion block is post-processed via a single-stream model.

C. Implementation of Multi-Stream Systems

For multi-stream acoustic modelling using the raw real and imaginary parts, we employ two sets of architectures consisting of a cascade of convolutional and fully-connected (FC) layers, without and with intermediate recurrent layers.

1) CNN+FC: Fig. 6 shows the first series of the multi-stream models which comprise only convolutional and fully-connected layers. As seen, the following fusion/concatenation (Concat) levels are plausible:

- **Concat-0**: fusion at the input (low) level (Fig. 6 (a)).
- **Concat-1**: fusion at a medium level after a series of convolutional layers operated on each stream and before FC layers (Fig. 6 (b)). In this scenario, the pre-processing sub-networks are purely convolutional and the post-processing one is solely fully-connected.
- **Concat-2**: fusion at a high level. The pre-processing block (per-stream) consists of a cascade of convolutional and fully-connected layers whilst post-processing is done via fully-connected layers (Fig. 6 (c)).
- **Concat-3**: fusion at a very high level just before the output layer. The pre-processing is carried out via convolutional and FC layers. The post-processing is a linear combination of the pre-processed streams (Fig. 6 (d)).

2) CNN+BiLSTM+FC: In the second series of multi-stream models (Fig. 7), we investigate a cascade of convolutional, recurrent and fully-connected layers. This is similar to the *CLDNN* architecture proposed in [8] for the single-stream acoustic modelling from raw waveform. Here, we apply the CLDNNs in a multi-stream mode (MS-CLDNN) along with investigating three fusion schemes at different levels.

When there is no recurrent layer (*e.g.*, CNN+FC), the temporal modelling is carried out implicitly, by augmenting each frame with $\pm m$ contextual frames while here the internal memories of the bi-directional LSTMs (BiLSTMs) [53] handle the forward and backward sequential modelling. Therefore, there is no need for appending with the neighbouring frames. Further, the model potentially benefits from a longer context.

Fig. 7 shows various fusion schemes: in the low (**Concat-00**), medium (**Concat-11**) and high (**Concat-22**) levels. We noticed that compared with other schemes, the training complexity (time and memory footprint) for Concat-22 is remarkably higher. This is owing to including two recurrent sub-networks which dramatically enlarge the computational graph.

Overall, assuming a fixed budget in terms of the number of layers, the higher the fusion point, the richer the pre-processing and the lower the remaining capacity for post-processing. The optimal fusion scheme depends on the information content of each stream, data and architecture's depth/width/type; the best trade-off remains to be found empirically. Although the number of model parameters ($\#Params$) for various fusion schemes is different, in all models each individual stream passes through the same number of convolutional, recurrent and fully-connected layers, with the identical width and depth.

IV. Experimental Results

A. Setup

The CNN+FC architectures consist of a cascade of four 1D convolutional layers along with five fully-connected layers. The convolutional layers consist of 128/60/60/60 filters, of length 129/5/5/3 samples along with max-pooling size of 3 and stride size of 1. The FC sub-network consists of five layers including 1024 units and batch normalisation [54]. Dropout (0.15) [55] and ReLU activation [56] were used in all layers.

The single and multi-stream CLDNNs consist of one (or two or three)[3] 1D convolutional layer with 128 (and 80, 60) filters of length 129 (and 7, 5) samples and max-pooling size of 3 (and 3, 2) samples. The length of kernel size of the first convolution layer for MFCC and FBank features could not be more than feature size (39 and 80, respectively). For consistency with raw spectral feature, it was set to 19 and 41 (half of the feature length). The BiLSTM sub-network consists of four layers containing 550 units in each direction followed by batch normalisation and dropout (0.2). Batch size was set to

[3]We will investigate multiple configurations in various experiments.

128 for CNN+FC and 8 for CLDNN architectures. DNNs were trained by PyTorch-Kaldi [57]–[59] on single GPUs (GeForce GTX-1080 and RTX-2080) with RMSprop [60] optimiser.

Experiments conducted on TIMIT, NTIMIT [61], Aurora-4 [62] (multi-style), TORGO [63], WSJ [64] and AMI's IHM[4] and SDM[5] [65] tasks. Aurora-4 includes four test sets: A (clean), B (additive noise), C (channel mismatch) and D (additive and channel) and average is calculated as $\frac{A+6B+C+6D}{14}$.

Sizes of the MFCC and raw spectral features are 39 and 257, respectively. FBank features' default size is 80 unless mentioned otherwise. The raw waveform features' size in CNN+FC and CLDNN systems is 3200 (200 ms) and 400 samples (25 ms), respectively. Except for raw waveform, all features were augmented with ± 5 contextual frames in the CNN+FC systems. Mag and Mag$^{0.1}$ denote the raw magnitude spectrum and its 10^{th} root, respectively.

The ASR systems are hybrid and trained by the cross-entropy loss. For each dataset, the language models (pruned tri-gram) and alignments were taken from the respective Kaldi's [59] standard recipes. We have not employed any data augmentation (except for TORGO), speaker-related embeddings or lattice re-scoring with RNN language models. For decoding, beam size and acoustic model weight [59] were set to 13 and 0.2 in TIMIT/NTIMIT and, 18 and 0.1 in other tasks, respectively.

We also compared the proposed framework with the CLP [30] network which is a cascade of a CLP layer, a recurrent block and a fully-connected sub-network. The CLP layer fuses the real and imaginary parts of the Fourier transform through a complex-valued linear transformation, $W = W_{Re} + jW_{Im}$, that operates on the short-time FFT, X, and returns Y

$$\begin{aligned} Y = W\ X &= (W_{Re} + jW_{Im})\ (X_{Re} + jX_{Im}) \\ &= (W_{Re}X_{Re} - W_{Im}X_{Im}) + j(W_{Re}X_{Im} + W_{Im}X_{Re}) \\ &= Y_{Re} + Y_{Im}. \end{aligned} \tag{7}$$

Y is further processed by $f(|Y|)$ where $|Y| = \sqrt{Y_{Re}^2 + Y_{Im}^2}$ and f is the $\log$ function in the original formulation [30].

The W_{Re} and W_{Im} matrices have the same dimensions: $\mathbb{R}^{P \text{x} N}$ where P is the projection size and N is $\frac{FFT_{Size}}{2} + 1$. Here, FFT_{Size} is 512; hence, N is 257. In [30], 128 and 1280 were applied for P. We will examine both values and investigate the effect of replacing the logarithm (CLP-$\log$) with the 10^{th} root (CLP-0.1) and *identity* (CLP-w/o $\log$) functions. We will also study the usefulness of putting one convolutional layer between the CLP and recurrent layers (CLP-$\log$-Conv.).

As Eq. 7 shows, CLP fuses the real and imaginary parts by taking into account their special relationship as two elements of a complex quantity while the proposed framework treats them as two independent information streams. Note that CLP is mostly comparable with the Concat-11 architecture with a difference that in Concat-11 the real and imaginary parts are pre-processed by (arbitrarily deep) CNNs and then fused via a real-valued linear transformation (Eq. 6) while in CLP the fusion is carried out immediately by a complex-valued linear transformation (Eq. 7). Other than the pre-fusion stage, the configuration of the other blocks (recurrent and FC layers) in the CLP and Concat-11 networks is identical.

[4]Individual Headset Microphone; close-talking scenario.

[5]Single Distant Microphone; far-field scenario.

TABLE I
WER of different front-ends on Aurora-4 (CNN+FC). $\#Params$ is in millions.

Feature	A	B	C	D	Avg	$\#Params$
MFCC	3.5	6.8	7.1	16.5	10.7	9.4
FBank-80	2.9	5.9	4.5	14.5	9.2	9.5
FBank-128	2.6	5.6	4.4	14.0	8.9	9.5
FBank-256	2.6	5.4	4.5	14.2	8.9	9.6
Mag	2.7	5.5	4.7	14.3	9.0	9.6
Mag$^{0.1}$	2.6	5.3	4.3	14.1	8.8	9.6
Raw-wave	3.1	5.7	7.5	16.5	10.3	10.1
FBank-80-Concat-2	2.6	5.5	4.5	14.0	8.9	14.1
Mag$^{0.1}$-Concat-2	**2.4**	**5.2**	**4.2**	13.6	8.5	15.1
Real	2.8	6.1	5.1	14.5	9.4	9.6
Imag	2.7	6.1	5.0	14.7	9.5	9.6
Concat-0	**2.4**	5.8	4.7	14.5	9.2	13.1
Concat-1	2.5	5.5	4.6	13.7	8.7	13.0
Concat-2	2.6	**5.2**	4.8	**13.5**	**8.5**	15.1
Concat-3	2.5	5.6	4.7	14.1	9.0	19.3

B. Results and Discussion

Now we investigate the performance of the proposed systems and on each task, discuss a number of modelling issues.

1) Aurora-4: Table I reports the performance of different features on Aurora-4 for CNN+FC architecture. As seen, while the real and imaginary parts individually lead to slightly poorer results than other raw spectral features, fusing them based on the proposed schemes yields a consistent gain over all test sets. With a narrow margin, Concat-1 and Concat-2 outperform the best baseline system, namely Mag$^{0.1}$. Compared with the widely-used FBank features (e.g., FBank-80, FBank-128), Mag$^{0.1}$ can be thought of as a filterbank with higher spectral resolution and narrower non-overlapping rectangular filters, uniformly distributed across the spectrum.

Using FBank features with higher spectral resolutions such as FBank-256 has been shown to be useful in the joint neural beamforming and acoustic modelling [51]. This feature has almost the same length (256 vs 257) as the raw magnitude features, with (50%) overlapping triangular filters (mimicking spectral masking) and a denser sampling at low frequencies – which are considered to be perceptually more important. As seen, despite embedding such prior knowledge in the FBank feature extraction pipeline, FBank-256's performance is on par with the raw magnitude spectrum which is extracted without taking advantage of any domain-specific expert knowledge.

Also note that multi-streaming by replicating the same input, e.g., Mag$^{0.1}$-Concat-2, where both input streams of the Concat-2 architecture (Fig. 6) are fed with the Mag$^{0.1}$ features is helpful. Similar observation was made for the FBank feature (FBank-80-Concat-2), too.

Compressing the dynamic range of the magnitude spectrum by taking 10^{th} root (Mag$^{0.1}$), leads to a significant and consistent performance gain relative to Mag ($\equiv$ Mag$^{1.0}$). This poses two questions: why such dynamic range compression is helpful and whether it will offer a similar advantage for the real and imaginary parts, if their dynamic range is compressed.

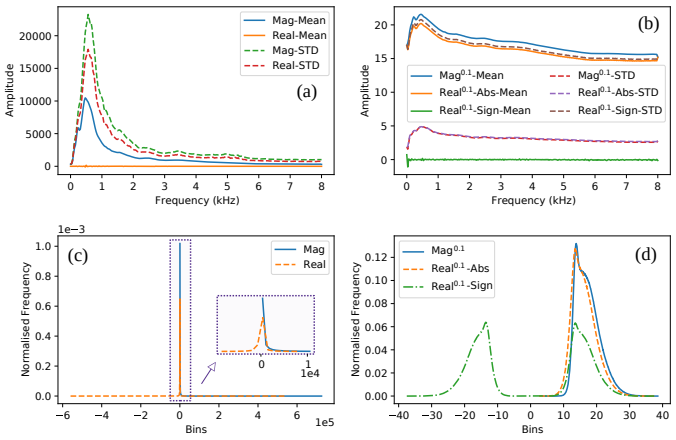

Fig. 8. Statistics of the magnitude spectrum and real part, estimated using 500 signals ($\sim$ 170k frames). Mean and standard deviation (STD) per spectral bin: (a) without, (b) with dynamic range compression (10^{th} root). Distribution of the magnitude spectrum and real part: (c) without, (d) with compression.

Contrary to the magnitude spectrum, the X_{Re} and X_{Im} can be negative; hence, the root compression or log are not directly applicable. Nonetheless, we can do either of the followings

$$\tilde{Z}_{Sign} = sign(Z)\ |Z|^{\alpha} \tag{8}$$

$$\tilde{Z}_{Abs} = |Z|^{\alpha} \tag{9}$$

where $sign$ is the Signum function, α is the power (0.1 here) and Z could be either the real or imaginary part. The $\tilde{Z}_{Sign}$ preserves the sign information while $\tilde{Z}_{Abs}$ discards it.

Fig. 8 presents the mean, standard deviation (STD) and distribution of the real[6] and the magnitude spectrum. Fig. 8 (a) and (b) illustrate that if the sign is preserved the mean for the real (and imaginary) part at all frequencies will be (almost) zero. The variance peaks at around 1 kHz and becomes very small in frequencies larger than 2 kHz. Furthermore, dynamic range compression dramatically decreases the spikiness of the distribution and lightens its tail (Fig. 8 (c) and (d)).

DNNs do not make any statistical assumption about their input. Nevertheless, the mean-variance normalisation (MVN) of the input is widely applied in practice and is considerably helpful [66]. When the dynamic range is compressed, the distribution gets closer to Gaussian. While being quasi-Gaussian on its own is not important, the expressive power of the mean and variance in representing the underlying distribution and consequently the effectiveness of the MVN increases. We believe this is why the dynamic range compression is helpful. It should be noted that contrary to $\tilde{Z}_{Abs}$, the mean of $\tilde{Z}_{Sign}$ is always (almost) zero (Fig. 8 (a) and (b)). This will nullify the mean normalisation contribution towards mismatch reduction.

While keeping the sign leads to a bi-modal and (almost) symmetric distribution, discarding the sign information results in a uni-modal one (Fig. 8 (d)). The symmetricity of the distribution implies that the amount of information in the positive and negatives sides is (approximately) identical. Therefore, discarding the sign information induces a minor information loss. Overall, we expect $\tilde{Z}_{Abs}$ to be more effectual because the lost information is approximately redundant (owing to symmetricity) and normalising its mean is more influential.

[6]We noticed the statistical properties of the real and imaginary parts are very similar; hence, to avoid clutter only display the statistics of the real part.

TABLE II
WER on Aurora-4 after dynamic range compression (CNN+FC). $\#Params$ is in millions.

Feature	A	B	C	D	Avg	$\#Params$
Concat-0-0.1-Sign	2.4	5.8	4.8	14.9	9.4	13.1
Concat-1-0.1-Sign	2.4	5.8	4.7	14.9	9.4	13.0
Concat-2-0.1-Sign	2.3	5.6	4.3	14.7	9.2	15.1
Concat-3-0.1-Sign	2.4	5.8	4.7	15.0	9.4	19.3
Concat-0-0.1-Abs	2.5	5.2	4.0	13.4	8.4	13.1
Concat-1-0.1-Abs	2.5	5.2	4.0	**13.2**	8.3	13.0
Concat-2-0.1-Abs	**2.3**	**4.9**	**3.8**	13.3	**8.2**	15.1
Concat-3-0.1-Abs	2.6	5.5	4.2	13.7	8.7	19.3

TABLE III
WER on Aurora-4 after using clean alignment (CNN+FC).

Feature	A	B	C	D	Avg
MFCC	3.4	5.8	4.5	7.9	6.4
FBank-80	2.8	5.1	3.2	6.3	5.3
FBank-128	2.7	4.8	3.3	6.0	5.1
FBank-256	2.6	4.6	3.2	5.8	4.9
Mag$^{0.1}$	2.7	4.7	3.3	5.8	4.9
Raw-wave	2.7	**4.4**	4.0	6.4	5.1
FBank-80-Concat-2	2.8	4.7	3.1	6.1	5.1
Mag$^{0.1}$-Concat-2	2.6	**4.4**	3.1	5.7	4.7
Concat-0-0.1-Abs	2.4	4.6	2.8	5.9	4.8
Concat-1-0.1-Abs	2.4	4.5	2.9	5.7	4.7
Concat-2-0.1-Abs	**2.3**	4.5	**2.5**	**5.6**	**4.6**
Concat-3-0.1-Abs	2.5	4.8	3.0	6.2	5.1

Table II shows the effect of dynamic range compression on the performance. ReIm-1.0, ReIm-0.1-Sign and ReIm-0.1-Abs refer to using the original real and imaginary parts, the compressed version via Eq. (8) and Eq. (9), respectively. As seen, while ReIm-0.1-Sign has a poorer performance than ReIm-1.0, ReIm-0.1-Abs leads to a notable gain. We made a similar and consistent observation on other tasks, too, corroborating the merit of $\tilde{Z}_{Abs}$ relative to $\tilde{Z}_{Sign}$.

We also studied the training dynamics of the CNN+FC models in terms of cross-entropy (CE) loss, frame classification accuracy and WER, vs epoch. Fig. 9 (a) shows the CE loss for the Dev data. While the FBank feature has the highest CE loss, the raw waveform model returns the lowest loss, with a notable margin relative to the ReIm-0.1-Abs (Concat-1) which has the second lowest CE loss. Compressing the dynamic range of the real and imaginary parts by keeping/discarding the sign leads to increasing/decreasing the CE loss. The knee point for all features is around 15 epochs; after that the CE loss remains almost constant. Similar trends are observed for the frame classification accuracy in Fig. 9 (b).

Fig. 10 illustrates the WER evolution of various features vs epoch for Aurora-4's four test sets. Fig. 10 (a) demonstrates the WERs for the test set A which includes only clean signals. As seen, ReIm systems outperform others and dynamic range compression has a marginal effect on the performance. Compared with Mag$^{0.1}$, the ReIm-based models require more epochs: for up to 15 epochs their performance is similar; after that, the ReIm systems outperform Mag$^{0.1}$. This observation holds for other test sets, too.

The advantage of dynamic range compression becomes

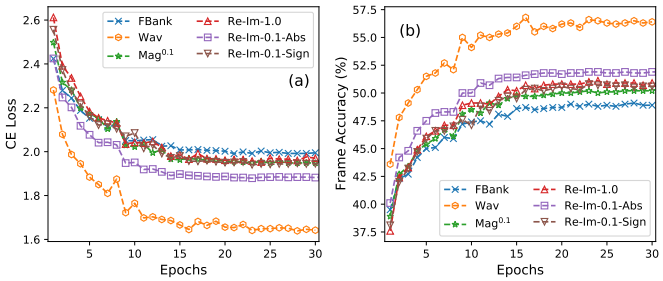

Fig. 9. Performance vs epoch for various features on Aurora-4's Dev set (architecture: CNN+FC). (a) CE loss, (b) frame classification accuracy.

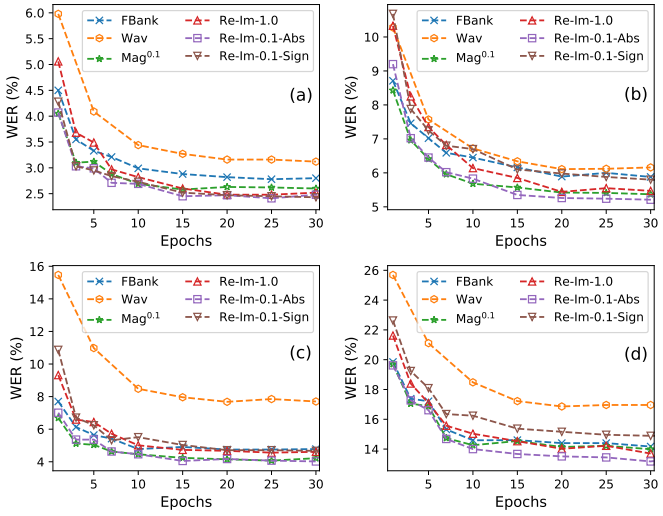

Fig. 10. WER vs epoch for Aurora-4's test sets (TS): (a) TS A (clean), (b) TS B (additive noise), (c) TS C (channel noise), (d) TS D (additive and channel noise). Re-Im indicates fusing the real and imaginary parts via Concat-1.

significant in noisy conditions. In the presence of additive noise (Fig. 10 (b)), while ReIm-0.1-Sign leads to a poorer performance than ReIm-1.0, ReIm-0.1-Abs significantly outperforms it. When there is a channel mismatch, ReIm-0.1-Abs and Mag$^{0.1}$ render the best performance whereas ReIm-0.1-Sign performance is on par with ReIm-1.0. The suboptimality of ReIm-0.1-Sign gets highlighted in the test set D (Fig. 10 (d)) where it returns the second worst results after the raw waveform model. While ReIm-1.0, FBank and Mag$^{0.1}$ have almost similar performance on test set D, ReIm-0.1-Abs outperforms them when the model is trained with at least 15 epochs.

Another noteworthy observation is the weakness of the raw waveform model in handling the channel mismatch (Fig. 10 (c) and (d)). Furthermore, despite having the lowest CE loss and the highest frame accuracy, it returns the poorest WER.

Since in Aurora-4 the noise signal is added synthetically, the alignments can also be taken from a model trained with only clean data (*clean-align*), supplying a higher quality alignment. As shown in Table II, using the clean-align alignment leads to achieving up to 4.6% average WER which is among the best reported results for Aurora-4. Similar to previous experiments, Concat-2 renders the best performance.

We also tried the Re-Im features in the MS-CLDNN architecture (with two convolutional layers). Comparing Table IV

TABLE IV
WER on Aurora-4 for single- and multi-stream CLDNNs.

Feature	A	B	C	D	Avg	$\#Params$
FBank	3.3	6.1	5.1	13.4	9.0	18.7
Mag$^{0.1}$	3.2	6.2	4.9	14.4	9.4	27.5
Raw-wave	3.2	6.1	6.6	14.1	9.3	40.7
Concat-00-0.1-Abs	3.2	6.0	4.8	**13.3**	**8.8**	51.5
Concat-11-0.1-Abs	3.1	**5.9**	**4.6**	13.4	**8.8**	39.6
CLP-log-P:128	2.8	6.2	5.2	14.1	9.3	15.7
CLP-log-P:1280	3.0	6.4	5.4	14.7	9.6	18.9
CLP-0.1-P:1280	3.0	6.3	5.4	14.9	9.7	18.9
CLP-w/o log-P:1280	2.9	6.1	5.1	14.3	9.3	18.9
CLP-log-P:128-Conv.	**2.7**	5.9	5.5	14.2	9.2	23.6

TABLE V
WER on Aurora-4 for CLDNN models after using clean alignment.

Feature	A	B	C	D	Avg
FBank	3.2	5.1	4.1	6.7	5.6
Mag$^{0.1}$	3.1	5.5	4.1	6.9	5.8
Raw-wave	3.0	**5.0**	4.6	6.8	5.6
Concat-00-0.1-Abs	2.9	5.1	4.0	6.4	5.4
Concat-11-0.1-Abs	3.0	**5.0**	3.8	**6.3**	**5.3**
CLP-log-P:128	2.7	5.4	3.8	6.7	5.7
CLP-log-P:1280	2.7	5.3	4.0	7.1	5.8
CLP-0.1-P:1280	2.8	5.5	3.9	7.2	5.9
CLP-w/o log-P:1280	2.9	5.3	3.8	6.8	5.7
CLP-log-P:128-Conv.	**2.6**	5.3	**3.7**	6.7	5.6

with I shows the CNN+FC outperforms the CLDNN models. As we will see later, this observation is limited to Aurora-4 and might be owing to noise and amount of the training data. Table V shows the performance of CLDNNs after applying clean-align. It notably improves the performance, although still the performance lags behind the CNN+FC models (Table III).

Tables IV and V also illustrate the performance of various variants of the CLP models with multi-condition and clean alignments, respectively. As seen, while using 10^{th} root instead of log does not reduce the WER, replacing log with the identity function noticeably improves the performance. Comparing WERs of the CLP networks with projection sizes of 128 and 1280 shows both return similar results. Further, inserting one convolutional layer between the CLP and recurrent blocks slightly improves the performance. On average, the proposed system outperforms CLP and leads to more than 8% relative WER reduction.

In these set of experiments, the Concat-11 CNNs include 56k parameters while the projection layer in CLP-P:128 and CLP-P:1280 contains 68k and 680k parameters, respectively. However, despite using exactly the same post-fusion (recurrent, fully-connected and softmax) layers, #Params of the Concat-11 model is twice as many as those of the CLP models (Table IV). This is owing to flattening the CNN outputs which results in a layer with about 11k parameters right before the BLSTM block while the CLP output size is 128 or 1280.

Finally, we studied the effect of applying the clean-align on the training dynamics of the Concat-1 and Concat-11 architectures. As shown in Fig. 11, CLDNN has a notably smaller loss than CNN+FC while its WER (in this task) is slightly larger. In spite of having a remarkable impact on the WER, using clean-align does not have a major effect on the loss value and the convergence behaviour of the models.

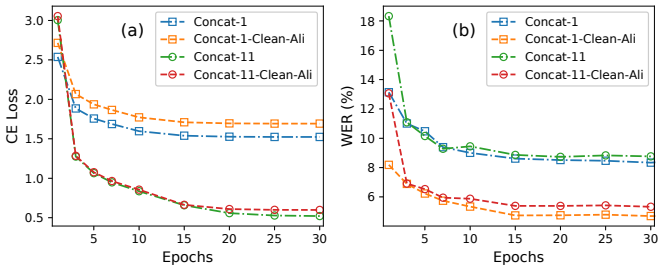

Fig. 11. Dynamics of CNN+FC (Concat-1) vs CLDNN (Concat-11) on Aurora-4 in terms of (a) CE and (b) WER (Avg), without/with clean-align.

TABLE VI
*PER on TIMIT and NTIMIT for single- and multi-stream CLDNNs. Number of * denotes number of convolutional layers.*

	TIMIT		NTIMIT		$\#Params$
	Dev	Test	Dev	Test	(in Millions)
MFCC*	14.4	16.3	21.7	23.7	17.0
FBank*	13.1	14.6	**19.0**	20.6	18.9
Raw-wave*	15.2	16.7	22.4	24.0	40.7
Concat-00*-0.1-Abs	14.4	15.7	22.9	23.9	51.4
Concat-11*-0.1-Abs	13.6	15.2	22.1	22.6	39.5
Concat-22*-0.1-Abs	14.0	15.9	22.7	23.2	52.8
CLP-log-P:128	15.1	16.9	22.0	23.4	15.6
CLP-w/o log-P:128	15.0	17.0	22.5	23.6	15.6
CLP-log-P:1280	16.1	17.3	24.8	26.2	18.8
CLP-w/o log-P:1280	15.9	17.3	23.9	24.9	18.8
Raw-wave**	14.9	16.5	22.1	23.3	20.3
Mag**	13.3	15.5	20.0	21.2	17.5
$\text{Mag}^{0.1}$**	13.2	15.1	19.6	21.0	17.5
Concat-00**-0.1-Abs	13.2	**14.5**	20.2	21.5	20.8
Concat-11**-0.1-Abs	**13.0**	14.8	19.9	**20.5**	19.7
Concat-22**-0.1-Abs	13.8	15.2	20.8	21.6	32.1
Raw-wave***	15.8	17.3	20.9	22.3	17.0
Concat-00***-0.1-Abs	13.8	15.6	20.8	21.9	17.8
Concat-11***-0.1-Abs	13.4	15.1	20.2	20.5	16.6
Concat-22***-0.1-Abs	14.1	15.6	21.4	22.1	29.9

2) TIMIT and NTIMIT: TIMIT and its noisy version NTIMIT (TIMIT transmitted over some telephone networks), include just 3.14 hours of training data and this is not favourable for acoustic modelling using raw signal representations. However, it is still insightful to evaluate the performance of the proposed model in such small tasks.

Table VI reports the phone error rate (PER) for ReIm-0.1-Abs using MS-CLDNNs along with various fusion schemes. The results are compared with the MFCC, FBank, single-stream raw waveform and raw magnitude spectrum models. On average, the Concat-11 fusion scheme returns the highest performance. On TIMIT's Dev/Eval data, it results in 13.0%/14.8% PER and on NTIMIT, it leads to 19.9%/20.5%, a competitive performance. Concat-00 outperforms Concat-11 only on TIMIT/Test, resulting in 14.5% PER. The Concat-22 model returns significantly poorer performance and training-wise it is notably slower than other models.

Table VII shows the PER for the CNN+FC. In comparison with CLDNN, it returns notably poorer PERs and Concat-1 is the most optimal fusion scheme for this architecture.

Comparing the amount of (N)TIMIT's training data ($\sim$ 1.13 M frames) and $\#Params$ shows that all models achieve a reasonable to competitive performance despite being highly over-parameterised. This verifies that $\#Params$ is not a reliable proxy for model's complexity and/or capacity [67].

TABLE VII
PER on TIMIT and NTIMIT for ReIm-0.1-Abs (CNN+FC).

	TIMIT		NTIMIT		$\#Params$
	Dev	Test	Dev	Test	(in Millions)
Concat-0	16.0	17.6	24.4	25.3	13.1
Concat-1	**15.6**	**17.4**	**24.0**	**25.0**	12.9
Concat-2	16.2	18.2	25.0	25.9	15.0
Concat-3	16.4	18.4	25.5	26.3	19.1

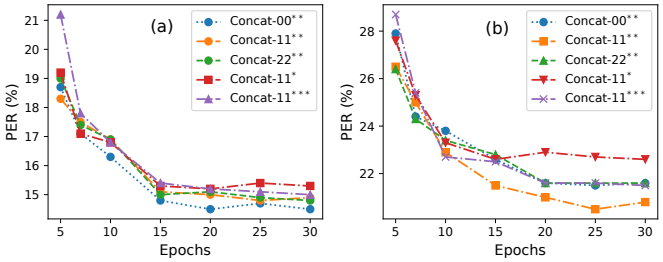

Fig. 12. PER vs epoch for various fusion schemes (ReIm-0.1-Abs, CLDNN). (a) TIMIT, (b) NTIMIT. Number of * denotes number of convolutional layers.

Fig. 12 illustrates the training dynamics (PER vs epoch) of different models. The differences are more pronounced in case of NTIMIT which is a more challenging task. For TIMIT, the models converge after almost 15 epochs and different systems show relatively similar dynamics. For NTIMIT the convergence rate is slower which is expected considering the fact that learning from data with similar lingual content in the presence of noise is more complicated and slower [11].

Fig. 13 compares the CE loss and PER of the best performing systems, namely Concat-11 (CLDNN) and Concat-1 (CNN+FC). The CLDNN models return smaller loss, even when the loss of Concat-11-NTIMIT is compared with the loss of Concat-1-TIMIT, despite the fact that the PER of Concat-1-TIMIT is remarkably lower than Concat-11-NTIMIT. The lower loss of CLDNNs stems from having a higher modelling capacity (beyond depth/width/#Params) owing to presence of the recurrent layers, which make the model more flexible and capable of learning from sequences. Such a lower loss, however, does not guarantee a lower PER (or WER).

3) WSJ: Tables VIII and IX report WER for the WSJ task using the CNN+FC and CLDNN networks. As seen, although for the Eval-92, the performance of both architectures is on par (4.75%), for Eval-93, the CLDNN leads to absolute/relative 1.2%/16.2% lower WER, reaching 6.2% while the CNN+FC results in 7.4% WER. Similar to other tasks, compression of the dynamic range is helpful and using ReIm-0.1-Abs leads to the highest performance on average.

In Table X, we compare various combinations of raw spectral representations, including systems with similar input streams, as well as combinations of the magnitude with the wrapped (principle) phase ($\text{Mag}^{0.1}$+wrapped-phase) and cosine of the phase, ($\text{Mag}^{0.1}$+cos(phase)). As can be observed, the combination of the real and imaginary parts achieves the best performance. Note that using cosine of phase, although solves the wrapping issue, leads to information loss. That is, information distinguishing the first and fourth quadrants ($\cos(\theta) = \cos(-\theta)$) or second and third quadrants ($\cos(\pi -$

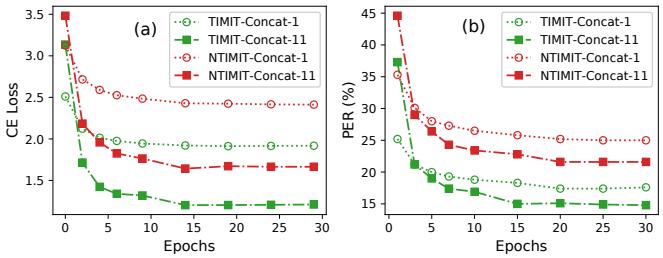

Fig. 13. Performance evolution of the Concat-1 and Concat-11 architectures on TIMIT/NTIMIT. (a) CE loss vs epoch, (b) PER vs epoch.

TABLE VIII
WER of different front-ends on WSJ (CNN+FC). $\#Params$ is in millions.

	Dev	Eval-92	Eval-93	$\#Params$
Real	8.5	5.0	7.8	11.1
Imag	8.4	5.0	7.4	11.1
Concat-0	8.4	5.0	7.6	14.6
Concat-1	8.4	4.8	7.4	14.4
Concat-2	8.2	4.9	7.5	16.5
Concat-3	8.2	4.7	7.7	22.1
Concat-1-0.1-Sign	8.2	4.8	7.7	14.4
Concat-1-0.1-Abs	**8.2**	**4.6**	**7.4**	14.4

$\theta) = \cos(\pi + \theta))$ is discarded. On the other hand, using the wrapped phase leads to a sequence with a highly complex data structure and unwrapping the phase leads to inconsistent results, with a negative effect on the performance [15].

Fig. 14 shows the dynamics of the best fusion scheme per architecture (Concat-1 for CNN+FC and Concat-11 for CLDNN) in terms of the CE loss and WER vs epoch for the WSJ's dev and test sets. As seen, the CNN+FC model shows a faster convergence: while CNN+FC's performance in terms of CE loss and WER gets barely improved after 10 epochs, the performance of the multi-stream CLDNN model keeps elevating for at least five extra epochs.

Despite having a noticeably smaller loss than CNN+FC, the MS-CLDNN's performance on the Eval-92 is comparable and on the Eval-93 is notably better (16.2% relative). This is another evidence for the non-perfect correlation between the CE loss as a general-purpose training criterion and WER as a task-specific performance metric. Similar observations have been reported in the literature and some solutions proposed, *e.g.*, training with the *expected WER* [68] or with the *entropy regularised log loss* [69] criteria.

Finally, comparing Table IX with VI and Table VIII with VII show that $\#Params$ for the same models over different databases is slightly different. This stems from having a different number of nodes in the output layer (Aurora-4:2016, TIMIT:1936, WSJ:3400, AMI:3992 and TORGO:440). For example, assuming the layer before the output has 1024 nodes, the same model on WSJ will have 1.5 M ((3400-1936) $\times$ 1024) more parameters than its TIMIT counterpart.

4) AMI: Table XI shows the WER for AMI's IHM and SDM tasks. In these experiments we explore the effect of the batch size (BS), dynamic range compression and the number of convolutional layers. The dynamic range compression even when the training data is as large as 100 hours, is still useful. Also increasing the batch size from 4 to 8 results in up to 4% relative WER reduction. Such gain warrants exploring larger batch sizes. Unfortunately we could not successfully train the models with a larger batch size (*e.g.*, 16) owing to the memory limitation (11 GB) of our computing infrastructure.

TABLE IX
*WER of different front-ends on WSJ (CLDNN). Number of * denotes number of convolutional layers.*

	Dev	Eval-92	Eval-93	$\#Params$
Raw-wave*	7.6	**4.7**	6.7	52.9
FBank-80*	7.4	4.9	6.8	20.5
FBank-128*	7.3	4.8	6.6	22.7
FBank-256*	7.3	5.0	6.3	28.6
Mag*	7.7	5.0	6.9	28.9
$\text{Mag}^{0.1}$*	7.3	4.9	6.6	28.9
$\text{Mag}^{0.1}$**	7.3	5.1	6.5	18.6
Real*	7.5	5.1	6.7	28.9
Imag*	7.7	4.9	6.9	28.9
CLP-log-P:128	7.7	5.1	6.6	16.3
CLP-log-P:1280	7.5	5.2	6.5	19.4
CLP-w/o log-P:1280	7.6	5.1	6.5	19.4
Concat-00*	7.6	5.0	6.4	52.9
Concat-11*	7.5	**4.8**	**6.2**	41.1
Concat-22*	7.5	5.1	6.8	54.3
Concat-11*-0.1-Sign	7.6	5.0	6.7	41.1
Concat-11*-0.1-Abs	**7.2**	4.9	6.5	41.1
Concat-11**-0.1-Abs	7.3	4.8	6.4	21.2

TABLE X
WER of different feature combinations on WSJ. Architecture: Concat-11 (one convolutional layer per stream). Note that $|Real|^{0.1}+|Imag|^{0.1}$ is equivalent to Concat-11*-0.1-Abs in Table IX.*

	Dev	Eval-92	Eval-93	$\#Params$
$\|\text{Real}\|^{0.1}+\|\text{Imag}\|^{0.1}$	**7.2**	**4.9**	**6.5**	41.1
$\|\text{Real}\|^{0.1}+\|\text{Real}\|^{0.1}$	7.3	5.0	6.9	41.1
$\text{Mag}^{0.1}+\text{Mag}^{0.1}$	7.4	5.5	6.6	41.1
$\text{Mag}^{0.1}$+wrapped-phase	7.4	5.5	**6.5**	41.1
$\text{Mag}^{0.1}$+cos(phase)	7.3	5.4	6.6	41.1

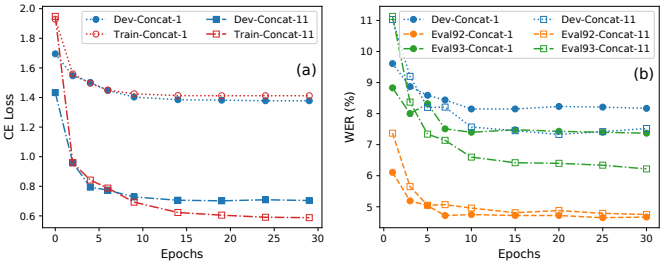

Fig. 14. Loss and WER for WSJ's dev and test (Eval-92 and Eval-93) sets.

In terms of the optimal number of convolutional layers, we noticed that in this task one convolutional layer is sufficient. Using more convolutional layers makes the model deeper and decreases the model parameters (Table VI). Contrary to the TIMIT task, such parameter reduction reduces the performance. The best WER for the proposed multi-stream system in the IHM and SDM scenarios on the Dev/Eval sets are 23.3/23.8 and 43.8/47.7, respectively. This is a competitive performance achieved with no speaker adaptation, data augmentation and re-scoring via advanced RNN language models.

We also compared the proposed models with other alternative systems including the end-to-end (E2E) stochastic attention head removal (SAHR) [70], multi-stream E2E [71] and the hybrid multi-scale octave CNNs [72], parametric (Parznet) 2-

TABLE XI
*WER on AMI-IHM and AMI-SDM (CLDNN). Number of * denotes number of convolutional layers. BS: batch size.*

	IHM		SDM	
	Dev	Eval	Dev	Eval
Raw-wave* (BS:8)	24.1	24.5	47.3	50.8
FBank* (BS:8)	23.8	24.4	44.2	48.1
$\text{Mag}^{0.1}$* (BS:8)	23.4	24.3	43.8	47.8
Concat-11* (BS:4)	24.4	25.7	45.9	50.7
Concat-11* (BS:8)	24.0	25.1	45.2	49.7
Concat-11*-0.1-Abs (BS:4)	24.1	24.8	45.2	49.1
Concat-00*-0.1-Abs (BS:8)	23.9	24.3	**43.5**	47.6
Concat-11*-0.1-Abs (BS:8)	**23.3**	**23.8**	43.8	47.7
Concat-11**-0.1-Abs (BS:8)	23.4	24.2	43.7	**47.6**
Concat-11***-0.1-Abs (BS:8)	23.7	24.4	44.3	48.6
SAHR-Transformer (E2E) [70]	24.2	24.6	-	-
SAHR-Conformer (E2E) [70]	24.1	24.2	-	-
Multi-stream (E2E) [71]	-	-	-	54.9
Multi-scale Octave CNN (Hybrid) [72]	32.2	37.2	48.2	53.3
Parznet 2D-CNN (Hybrid) [73]	24.9	26.0	-	-
Parznet 2D-CNN+VI (Hybrid) [12]	24.7	25.7	-	-

TABLE XII
*WER on TORGO (Dys: Dysarthric, Typ: Typical), with and without applying speed perturbation (sp). Number of * denotes number of convolutional layers (default setting [75] includes three convolutional layers).*

Setup	without sp		with sp	
Feature	**Dys**	**Typ**	**Dys**	**Typ**
MFCC	47.5±3.5	15.8±2.7	39.2±3.2	12.1±1.2
FBank-80	45.3±2.1	15.0±1.6	36.5±1.4	11.3±0.6
FBank-128	-	-	35.0±1.8	11.9±1.1
Raw-wave	57.4±3.4	23.5±2.3	38.8±2.0	13.8±0.8
Mag	48.4±4.9	17.3±3.0	39.6±3.8	12.1±1.5
$\text{Mag}^{0.1}$	51.8±5.8	20.4±3.9	35.7±3.4	11.1±1.2
Concat-00	47.3±3.8	18.2±3.0	37.3±2.7	12.4±0.8
Concat-11	53.1±2.9	21.7±2.1	42.2±2.3	14.6±1.1
Concat-00-0.1-Sign	44.2±2.8	15.5±1.3	36.1±2.9	12.4±1.1
Concat-11-0.1-Sign	46.9±2.7	17.7±1.2	38.6±2.9	13.5±1.7
Concat-00-0.1-Abs	44.5±1.7	15.0±1.3	33.3±1.7	10.6±0.6
Concat-11-0.1-Abs	44.2±3.0	14.8±1.5	34.7±2.9	11.0±1.0
Concat-00*-0.1-Abs	-	-	33.5±2.4	10.9±0.8
Concat-11*-0.1-Abs	-	-	**31.7**±2.3	**10.2**±0.6

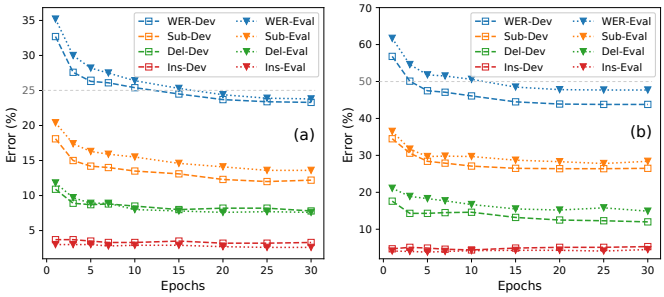

Fig. 15. ASR error components (WER, Sub, Del and Ins) vs epoch on AMI for Concat-22*-0.1-Abs (batch size:8) configuration. (a) IHM, (b) SDM.

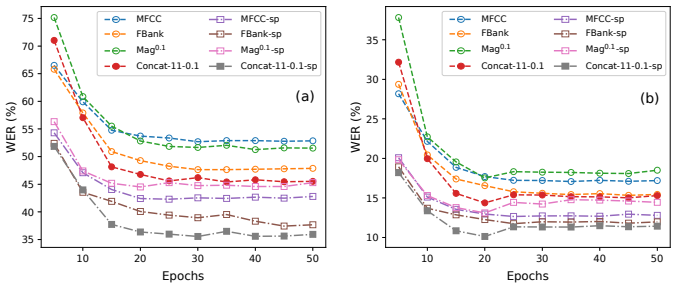

Fig. 16. WER vs epoch for TORGO (fold 1). (a) Dysarthric, (b) Typical.

D CNNs without [73] and with variational inference (VI) [12]. As seen in Table XI, SAHR systems in which the Transformer [21] or Conformer [74] heads are randomly dropped out, return the highest performance, although still lagging behind the proposed Concat-11*-0.1-Abs (BS:8) system.

Fig. 15 shows the training dynamics of the proposed model in terms of WER, substitution (Sub), deletion (Del) and insertion (Ins) for both IHM and SDM conditions. As seen, the main source of error is the substitution and its dynamics closely resembles the dynamics of the WER. The deletions are the second important source of error and insertions are the smallest component of the WER. System's performance in terms of insertion is almost constant in both IHM and SDM scenarios during training. In terms of deletions, there is barely any significant improvement after 10^{th} epoch.

5) Dysarthric Speech Recognition: We also investigated the efficacy of the proposed real-imaginary based MS-CLDNN systems on the TORGO [63] dysarthric speech dataset. In this series of experiments two setups were explored: without and with data augmentation via speed perturbation (*sp*) with the following speed change factors: 0.9 (slower), 1.0 (original) and 1.1 (faster). We used the 5-fold cross-training setup proposed in [75] and report the mean±STD WER in Table XII. To make our systems comparable with the baseline models developed in [75], we used three convolutional and five recurrent layers.

Among the classic features, FBank returns the highest performance, with and without data augmentation, on both dysarthric (Dys) and typical (Typ) speech. Among the ReIm systems, we can see similar trends to other datasets, namely dynamic range compression helps and ReIm-0.1-Abs outperforms ReIm-0.1-Sign. Comparing the ReIm-0.1-Abs and FBank systems when speed perturbation *is not*/is used shows up to 1.1%/3.2% and 0.2%/0.7% lower WER (absolute) for the dysarthric and typical speech, respectively.

We also studied the training dynamics of the models (WER vs epoch) for different features and for both dysarthric and typical speech. As seen in Fig. 16, when speed perturbation is applied, the model benefits from further training epochs and convergence gets slower. This is due to the fact that speed perturbation, without keeping F0 fixed, simulates many speakers and increases the data three fold. It makes the training data richer and helps the model to learn to normalise the speaker. Comparing dysarthric with typical speech also shows that in contrast to dysarthric speech, for typical speech the models barely benefit from more than 20 epochs. For some systems, including ReIm ones, the performance on typical speech even gets worse by over-training.

One challenge in dysarthric ASR is high inter- and intra-speaker variability. As mentioned earlier, the speed perturbation is helpful to cope with this issue to some extent. In Section V, we will show that multi-streaming also can help towards speaker normalisation. This will partially explain why the proposed system outperforms other front-ends in this task.

V. Analysis of the Learned Filters

In this section, we analyse the filters learned in the first convolutional layer (Conv-L1) and compare their functionality in the single- and multi-stream models.

To investigate the collective behaviour of the Conv-L1's filters, we study their statistics per sample. Fig. 17 (a) and (b) show the mean and STD for the single-stream systems fed with the real and imaginary parts while Fig. 17 (c) and (d) depict the mean and STD of the multi-stream Concat-1 model. Comparing the filters' mean and STD in Fig. 17 (a) and (b) shows that the Conv-L1's filters in the single-stream systems fed with the real or imaginary parts have a similar behaviour. Such similarity is not surprising considering the similarity of the real and imaginary parts in encoding speech information (Fig. 1 (c) and (d)). The shaded area in these figures demonstrates the evolution of the mean and STD of the filters during training and shows a similar trend for both models, too. In addition, in the single-stream systems the STD and the mean are considerably large in the last 30 bins. This illustrates that in the single-stream mode, the filters mostly model some short-range dependencies within this range.

On the other hand, the behaviour of the filters fed with the real and imaginary parts in the multi-stream system is significantly and interestingly different. Although (on average) the filters operating on the real part remain similar, the filters fed with the imaginary part capture a different and complementary aspect of the input. In particular, while the former still models the short-range dependencies, the latter captures the medium-range relationships which are highlighted by the green and red zones in Fig. 17 (c) and (d), respectively.

To further investigate such short- and medium-range dependencies, we plotted the mean and STD of the magnitude of the FT of the filters in Fig. 18. Note that since the filters operate directly on the compressed real and imaginary parts (Eq. 9), the domain after taking the FT would be comparable with the *generalised cepstrum* [76] (because of using 10^{th} root instead of log). We refer to this domain as *cepstrum**. Similarly to the cepstral domain, the low quefrency components are associated with the envelope of the compressed real and imaginary parts. Such envelope is highly correlated with the envelope of the compressed magnitude spectrum ($\text{Mag}^{0.1}$) and consequently the vocal tract. Therefore, in ASR applications one might expect the FT of the filters to further attend the low quefrency components to capture the vocal tract element.

As seen in Fig. 18, for both real and imaginary base single-stream systems, the filters on average are more focused on the low quefrency components. However, in the multi-stream systems, while the filters operating on the real part behave similarly and capture the vocal tract, the filters fed with the imaginary part focus on the medium-range cepstral coefficients which are mostly correlated with the speaker. It is beneficial for the modelling to normalise the speaker-related attributes whilst the vocal tract information is captured by the other stream. This is highly desirable in acoustic modelling for ASR and improves the robustness of the model in handling and normalising the nuisance factors.

Finally, we look over the shape of the learned filters in the Conv-L1. Although, in general, filters have an ambiguous shape difficult to understand, we noticed some filters bear a resemblance to some well-known parametric functions and wavelets. For example, as seen in Fig. 19 (a), indices 8-10 look like the first and second-order derivatives of the Gaussian wavelet, similar to the so-called MRASTA [77] filters. Further, filters shown in Fig. 19 (b) resembles filters of the parametric CNNs such as SincNet [16], Sinc^{2}Net and GaussNet [17]. This encourages exploring the usefulness of the parametric CNNs in this context. We also depicted the |FFT| of these filters in Fig. 19 (c) and (d) which shows they collectively act as a bank of overlapping band-pass lifters with different bandwidths.

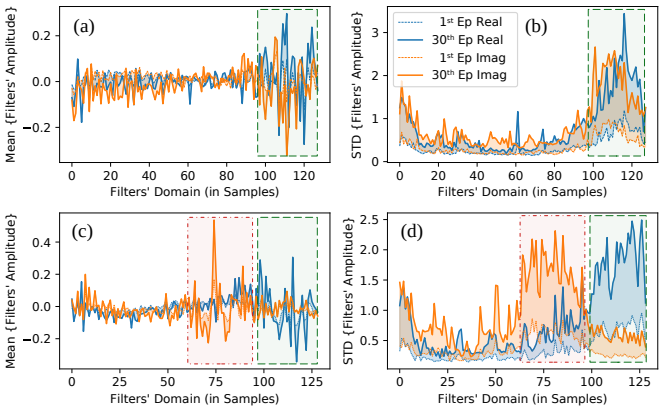

Fig. 17. Mean and STD of the Conv-L1's filters for the CNN+FC architecture (WSJ task). The shaded area is delimited by the 1^{st} and 30^{th} epochs (Ep). The green and red zones indicate the short- and medium-range dependencies modelling. (a) and (b) single-stream, (c) and (d) multi-stream (Concat-1).

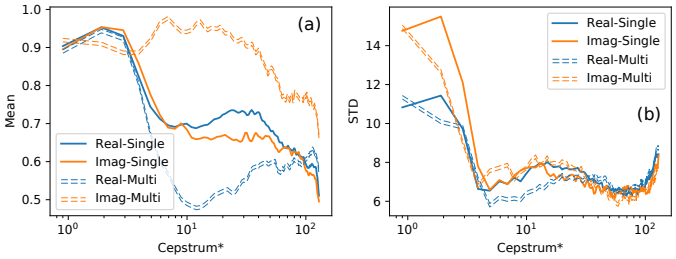

Fig. 18. Mean and STD of the Conv-L1's filters after taking |FFT| from the filters in the single-stream (Single) and multi-stream (Multi) systems (CNN+FC,WSJ) fed with the real and/or imaginary parts. (a) Mean, (b) STD.

VI. Conclusions and Scopes for Future Work

In this paper, we investigated the usefulness of the multi-stream acoustic modelling using the raw real and imaginary parts of the Fourier transform. Applying the magnitude spectrum as their proxy, leads to irreversible information loss of the all-pass component and pre-mature information fusion. In the proposed multi-stream framework, the real and imaginary parts were fused after being pre-processed via convolutional layers. Then, they were post-processed through recurrent and fully-connected layers. We investigated various modelling issues including different architectures and fusion levels as well as the training dynamics of models in terms of cross-entropy loss, frame classification accuracy and WER. We also analysed

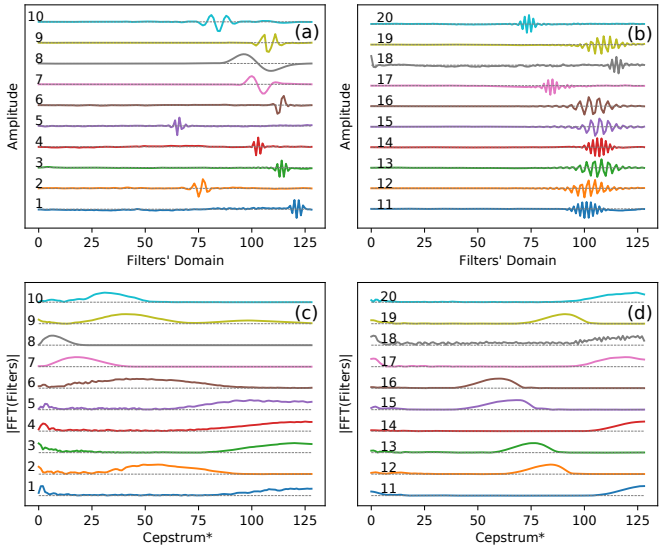

Fig. 19. The learned filters in the Conv-L1 which resemble (a) wavelets and (b) parametric functions (*e.g.*, sinc), along with their |FFT| (c) and (d).

the filters learned in the first convolutional layer of single and multi-stream models and illustrated that in a multi-stream architecture, filters operating on each stream play complementary roles. The effectiveness of the proposed systems was successfully demonstrated by achieving consistent gains across various tasks including TIMIT/NTIMIT (phone recognition), Aurora-4 (noise robustness), WSJ (read), TORGO (dysarthric) and AMI's (meeting) IHM (close-talking) and SDM (far-field) scenarios. The proposed framework is generic and can be employed in recognition and/or classification tasks for all sequences with the Fourier transform such as audio, image and biomedical signals, opening a broad avenue for future work.

REFERENCES

[1] 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*, pp. 357–366, 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Wiley-Interscience, 2006.

[4] J. Yousafzai, P. Sollich, Z. Cvetkovic, and B. Yu, "Combined features and kernel design for noise robust phoneme classification using support vector machines," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 19, pp. 1396–1407, 2011.

[5] M. Ager, Z. Cvetković, and P. Sollich, "Combined waveform-cepstral representation for robust speech recognition," in *2011 IEEE International Symposium on Information Theory Proceedings*, 2011, pp. 864–868.

[6] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *INTERSPEECH*, 2014.

[7] D. Palaz, M. Magimai-Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *ICASSP*, 2015.

[8] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, 2015, pp. 1–5.

[9] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *INTERSPEECH*, 2016.

[10] P. Agrawal and S. Ganapathy, "Interpretable representation learning for speech and audio signals based on relevance weighting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2823–2836, 2020.

[11] E. Loweimi, P. Bell, and S. Renals, "On the Robustness and Training Dynamics of Raw Waveform Models," in *INTERSPEECH*, 2020, pp. 1001–1005.

[12] D. Oglic, Z. Cvetkovic, and P. Sollich, "Learning waveform-based acoustic models using deep variational convolutional neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 2850–2863, 2021.

[13] E. Loweimi, P. Bell, and S. Renals, "Raw Sign and Magnitude Spectra for Multi-Head Acoustic Modelling," in *INTERSPEECH*, 2020.

[14] N. M. Joy, D. Oglic, Z. Cvetkovic, P. Bell, and S. Renals, "Deep scattering power spectrum features for robust speech recognition," in *INTERSPEECH*, 2020.

[15] E. Loweimi, Z. Cvetkovic, P. Bell, and S. Renals, "Speech acoustic modelling from raw phase spectrum," in *ICASSP*, 2021.

[16] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[17] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," NIST, Tech. Rep., 1993.

[19] H. Hermansky, "Multistream recognition of speech: Dealing with unknown unknowns," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1076–1088, 2013.

[20] K. J. Han, J. Pan, V. K. N. Tadala, T. Ma, and D. Povey, "Multistream cnn for robust acoustic modeling," *ICASSP*, pp. 6873–6877, 2021.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, vol. 30, 2017.

[22] P. Van Hove, M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from signed fourier transform magnitude," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 31, no. 5, pp. 1286–1293, 1983.

[23] E. Loweimi, Z. Cvetkovic, P. Bell, and S. Renals, "Speech Acoustic Modelling Using Raw Source and Filter Components," in *INTERSPEECH*, 2021.

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

[25] G. Fant, *Acoustic Theory of Speech Production*. Berlin, Boston: De Gruyter Mouton, 1971.

[26] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 6, pp. 672–680, 1980.

[27] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[28] E. Loweimi, S. M. Ahadi, and H. Sheikhzadeh, "Phase-only speech reconstruction using very short frames," in *Proc. Interspeech 2011*, 2011, pp. 2501–2504.

[29] A. Hirose, *Complex-Valued Neural Networks: Advances and Applications*, ser. IEEE Press Series on Computational Intelligence. Wiley, 2013.

[30] E. Variani, T. N. Sainath, I. Shafran, and M. Bacchiani, "Complex linear projection (CLP): A discriminative approach to joint feature extraction and acoustic modeling," in *INTERSPEECH*, 2016.

[31] L. Drude, B. Raj, and R. Haeb-Umbach, "On the appropriateness of complex-valued neural networks for speech enhancement," in *INTERSPEECH*, 2016, pp. 1745–1749.

[32] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *ICASSP*, 2019, pp. 6885–6889.

[33] A. M. Sarroff, "Complex neural networks for audio," Ph.D. dissertation, Dartmouth College, 2018.

[34] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *ICLR*, 2018, pp. 1 – 19.

[35] P. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, Inc., 2013.

[36] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Prentice Hall, 2009.

[37] E. Loweimi, "Robust phase-based speech signal processing; from source-filter separation to model-based robust asr," Ph.D. dissertation, University of Sheffield, 2018. [Online]. Available: http://etheses.whiterose.ac.uk/19409/

[38] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[39] K. Paliwal and L. Alsteris, "On the usefulness of stft phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[40] E. Loveimi and S. Ahadi, "Objective evaluation of magnitude and phase only spectrum-based reconstruction of the speech signal," in *ISCCSP*, 2010.

[41] ——, "Objective evaluation of phase and magnitude only reconstructed speech: New considerations," in *IEEE ISSPA*, May 2010, pp. 117–120.

[42] E. Loweimi, S. Ahadi, and S. Loveymi, "On the importance of phase and magnitude spectra in speech enhancement," in *ICEE*, 2011.

[43] E. Loweimi, J. Barker, and T. Hain, "Statistical normalisation of phase-based feature representation for robust speech recognition," in *ICASSP*, 2017, pp. 5310–5314.

[44] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 2, pp. 236–243, 1984.

[45] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

[46] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*, 2010, pp. 4214–4217.

[47] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[48] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[49] M. Seltzer, B. Raj, and R. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, 2004.

[50] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *INTERSPEECH*, 2016.

[51] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *ASRU*, 2015, pp. 30–36.

[52] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[53] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *ICANN*, 2005.

[54] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

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

[57] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *ICASSP*, 2019.

[58] 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, pp. 1–4.

[59] 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.

[60] 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.

[61] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *ICASSP*, vol. 1, no. 109–112, 1990.

[62] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal and Information Process, Mississippi State University, Tech. Rep., 2002.

[63] 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.

[64] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *ICASSP*, 1992, pp. 899–902.

[65] I. McCowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," in *MLMI*, 2005.

[66] Y. LeCun, L. Bottou, G. B. Orr, and K. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade - Second Edition*, ser. Lecture Notes in Computer Science, G. Montavon, G. B. Orr, and K. Müller, Eds. Springer, 2012, vol. 7700, pp. 9–48.

[67] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *ICLR*, 2017, pp. 107–115.

[68] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," *ICASSP*, 2018.

[69] A. May, A. B. Garakani, Z. Lu, D. Guo, K. Liu, A. Bellet, L. Fan, M. Collins, D. Hsu, B. Kingsbury, M. Picheny, and F. Sha, "Kernel approximation methods for speech recognition," *J. Mach. Learn. Res.*, vol. 20, pp. 59:1–59:36, 2019.

[70] S. Zhang, E. Loweimi, P. Bell, and S. Renals, "Stochastic attention head removal: A simple and effective method for improving transformer based asr models," in *INTERSPEECH*, 2021, pp. 2541–2545.

[71] R. Li, X. Wang, S. H. Mallidi, S. Watanabe, T. Hori, and H. Hermansky, "Multi-stream end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 646–655, 2020.

[72] J. Rownicka, P. Bell, and S. Renals, "Multi-scale octave convolutions for robust speech recognition," in *ICASSP*, 2020.

[73] D. Oglic, Z. Cvetkovic, P. Bell, and S. Renals, "A deep 2d convolutional network for waveform-based speech recognition," in *INTERSPEECH*, 2020.

[74] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020.

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

[76] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 5, pp. 1087–1089, 1984.

[77] H. Hermansky and P. Fousek, "Multi-resolution rasta filtering for tandem-based asr," in *INTERSPEECH*, 2005, pp. 361–364.

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.

Zhengjun Yue is a research associate in King's College London (KCL). She received her B.Sc. degree from Shanghai University, in 2017, and the M.Sc. degree in Artificial Intelligence from the University of Edinburgh. She is a Ph.D. candidate in the Speech and Hearing Research Group (SPandH), the University of Sheffield, since 2018. Her research interests include acoustic modelling and acoustic-articulatory multi-modal speech recognition for dysarthric speech, and end-to-end ASR.

Peter Bell received the B.A. degree in mathematics in 2002 and the M.Phil. degree in computer speech, text and Internet technology in 2005 from the University of Cambridge, and the Ph.D. degree in automatic speech recognition from the University of Edinburgh, in 2010. He is a reader in speech technology with the School of Informatics, University of Edinburgh. His research interests include domain adaptation, regularization, and low-resource methods for acoustic modeling.

Steve Renals (M'91 — SM'11 – F'14) received the B.Sc. degree in chemistry from the University of Sheffield, in 1986 and the M.Sc. degree in artificial intelligence in 1987 and the Ph.D. degree in neural networks and speech recognition from the University of Edinburgh, in 1991. He is Professor of speech technology with the School of Informatics, University of Edinburgh, having previously held positions at ICSI Berkeley, the University of Cambridge, and the University of Sheffield. His research interests include ASR, spoken language processing, and machine learning. Dr Renals is a fellow of ISCA (2016) and a Senior Area Editor of the IEEE OPEN JOURNAL OF SIGNAL PROCESSING.

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.