End-to-end speech recognition from raw speech: Multi time-frequency resolution CNN architecture for efficient representation learning

Dhanya Eledath
International Institute of Information Technology - Bangalore (IIIT-B)
Bangalore, India

P. Inbarajan
Samsung R&D Institute, Bangalore (SRIB)
Bangalore, India

Anurag Biradar
Samsung R&D Institute, Bangalore (SRIB)
Bangalore, India

Sathwick Mahadeva
Samsung R&D Institute, Bangalore (SRIB)
Bangalore, India

V. Ramasubramanian
IIIT-Bangalore
Bangalore, India

***Abstract*—We propose a multi time-frequency (t-f) resolution CNN architecture for end-to-end speech recognition from raw speech waveform. We address issues related to the nature of front-end convolutional kernels and the kind of multi t-f spectrographic feature maps formed and the back-end convolutional processing of the feature maps within two tasks, namely, frame-wise phoneme classification and an encoder-decoder (with CTC-Attention loss) based continuous phoneme decoding. Our multi t-f resolution CNN (MT-CNN) architecture works with unconstrained learnt kernels and back-end 2D-convolutional layers to process the multi t-f spectrographic feature maps for these tasks. We contrast this architecture with two other variants in recent work - a multi-scale feature based system and the SincNet (which uses parameterized convolutional kernels constrained in the form of Sinc functions with learnable bandwidths). We show a consistent performance gain of the proposed multi t-f architecture over these two variants - a 3-8% accuracy (absolute gain) in the frame-wise classification task and 3% PER (absolute gain) in the continuous phoneme decoding task. These two performance gains together establish the effectiveness of the proposed architecture in using the multi t-f unconstrained variable length 1-D convolutional kernels, 2-D multi t-f spectrographic feature maps and the back-end 2-D convolution layers.**

***Index Terms*—multi time-frequency resolution, CNN, end-to-end, speech recognition**

I. Introduction

We address the problem of fully end-to-end (E2E) automatic speech recognition (ASR) using a multi time-frequency (t-f) resolution CNN architecture, with emphasis on the ability of this new architecture to perform enhanced representation learning from 1-dimensional signals such as speech waveform. This enhanced representation learning comes from the architecture's ability to perform a multi time-frequency analysis on the input waveform using variable-sized kernels in its first convolution layer and thereby create 2-dimensional t-f feature maps that correspond to multiple spectrographs, each equivalent to a filter-bank analysis with variable kernel (convolving filter) sizes.

Starting from the early introduction of the convolutional neural-network (CNN) by Le Cun [1] for successful recognition of handwritten digit images, CNNs have come to be a well established framework for end-to-end approaches (i.e. from raw input), combining a powerful representational learning mechanism [2] in its lower convolution layers and discriminative fully-connected higher layers for multi-class classification tasks such as from raw images [3], speech spectrographic images [4], speech-waveform [5], [6], audio-waveform [7], [8] and music-waveform [9], [10].

In this paper, we focus on a specific aspect of CNNs, namely, the kernel sizes used in the convolutional kernels, and point out that for applying CNNs on raw 1-dimensional signals such as speech-, audio- and music-waveforms, it becomes important to 'provide' for a variable kernel size, to exploit and resolve the well known time-frequency trade-off inherent in such 1-dimensional convolution operation. While this applies to 2-dimensional images also, this issue of having to address the time-frequency trade-off in the application of a filter-bank kind of operation (what a set of kernels in a CNN layer do) has been more or less overlooked in the image-CNN community. The closest treatment in the image-CNN literature to this notion of using variable kernel sizes is in the now well known Inception network (or the GoogleNet) [11], where multiple image kernels of sizes 1×1, 3×3 and 5×5 have been used in the early CNN layers. However, the motivation for providing for these variable sized kernels has been more or less very different from the fundamental time-frequency (spatial intensity variation vs spatial frequency in the case of images) trade-off, and as a consequence, the advent of Inception did not really see the emergence of a strong line of enquiry into such architectures with variable kernel sizes in the 1-d signal community in order to address the time-frequency trade-off using multi-temporal convolutional analysis.

II. Relation to prior work

In processing 1-d speech/audio waveform for various tasks such as audio scene classification, speaker-recognition, speech recognition etc. the problem of time-frequency trade-off aris-

ing in end-to-end frameworks using convolutional layers has indeed received some attention in the recent years, and we note all of such work in the following.

Earliest among such work is that of [12], [13] who address this issue for the first time, and propose a multi-temporal architecture for audio-scene classification (ASC), taking into account the need for a variable time-frequency representational analysis of the 1-d signal such as audio-signal for the ASC task. We extended and generalized the multi-temporal architecture of [12], [13] to a highly scaled number of multi-temporal branches (e.g. 12) for the ASC task, allowing for creating multiple spectrographic feature maps with a wide range of time-frequency resolution trade-offs [14]. By this, we showed a very significant performance gain (11-15% absolute) by the multi t-f architecture (with 12 branches) over a conventional single-branch CNN operating at any of the kernel sizes that is part of the multi t-f architecture.

With regard to ASR, the only relevant work in this direction is by Zhu et al. [15] which proposes a multi-scale feature learning framework in an end-to-end architecture. This work uses multiple (specifically 3) branches of convolutional layers operating on a frame of input waveform samples to yield an aggressively pooled feature vector per input frame and which is further processed as a sequence of feature vectors by a Bi-LSTM encoder and CTC decoder system for continuous speech recognition. The crux of this work is in essentially addressing the time-frequency tradeoff required to be handled in efficient representation learning from raw speech waveform. However, we point that, importantly, by performing an aggressive pooling on the feature map produced by the 1-d convolutional kernels, this work loses on the rich t-f spectrographic information available (and much needed) to represent the multi t-f resolution signal such as speech in the input. We propose to address this specific deficiency of this architecture and propose to use 2-D convolutional layers to process the 2-D multi t-f spectrographic feature map stack to exploit the t-f correlations and acoustic-phonetic information inherent in such a multi t-f analysis of the input raw speech waveform. By this, we show significant performance gain (e.g. 3% absolute in PER) over Zhu et al.'s [15] system.

Secondly, we note the other relevant work in this direction for ASR comes from the E2E-SincNet [16] architecture which exploits the recently proposed SincNet convolutional layer based speech processing [17], [18] proposed originally for speaker recognition, frame-wise phoneme recognition and with DNN-HMM systems. While this framework does use variable length kernels arising from the learnable band-widths of the Sinc kernel's rectangular frequency response, this framework does not address the need for handling the time-frequency tradeoff inherent in 1-d signal analysis by means of the convolutional network in learning short-time and running spectral representations. In other words, this work uses a 'single' branch of Sinc-constrained kernels - which, while being capable of learning variable bandwidths and center frequencies and thereby approximate a mel-scale filter-bank - is not adequate to represent the time-frequency trade-offs that can be handled only by a multi-branch convolutional system. We address this and propose to have multiple branches of Sinc-constrained kernels to provide for such a CNN architecture to learn multi t-f resolution representation from all the branches in addition to the individual branch's variable kernel / bandwidth representation by virtue of the rectangular band-pass filters of the Sinc kernels.

The above two constitute the context of our present work here, and we now set out to outline the overall framework to position our present contribution with respect to the above two recent work [15] and [16].

III. Overall framework

Fig. 1 provides the overview of the framework-pipeline and architecture within which we propose our multi time-frequency resolution CNN based ASR, particularly with respect to the other two works referred above, namely, Zhu et al's [15] multi-scale feature based ASR and the E2E-SincNet [16]. These three approaches outlined in these 2 figures are termed i) Proposed, ii) Zhu's multi-scale feature and iii) SincNet for further discussion. The 3 approaches are grouped into two pipelines - shown in Panel A and Panel B - both with the following broad components:

1) **Input:** Raw waveform (a single frame of 25 ms duration or 400 samples for frame-wise phoneme classification) or a sequence of such single frames for continuous phoneme decoding.
2) **Representation learning by 1-D Conv Branches:** 1-D conv layers convolving on the input raw waveform to yield multiple t-f spectrographic feature maps.
3) **Back-end:** A back-end processing of the spectographic feature maps in the form of 2-D conv layers (in the case of the proposed system, following a flattening of the reduced 2-D feature map) or aggressive pooling (in the case of the Zhu's multi-scale feature and SincNet) to yield a flattened feature vector
4) **Frame-wise phoneme classification:** The feature vector being fed to fully connected layers with soft-max and cross-entropy loss training and corresponding maximum a posterior classification in inference for a frame-wise phoneme classification on which %Acc is measured.
5) **Encoder-Decoder based continuous phoneme decoding:** The feature vector sequence (corresponding to a sequence of input raw waveform frames) being fed to an Encoder-Decoder architecture for a sequence-to-sequence learning with joint CTC-Attention loss and corresponding CTC decoding to yield a continuous phoneme label sequence on which PER is measured.

The main differences between Panel A and Panel B are in terms of the essential differences between the Proposed approach and those of Zhu's and SincNet as outlined below:

1) **Nature of 1-D Conv Kernels:** In the proposed approach in Panel A, the 3 1-D conv layers work on the input speech waveform as parallel branches, termed Br-1, Br-2 and Br-3, each with 32 kernels of duration

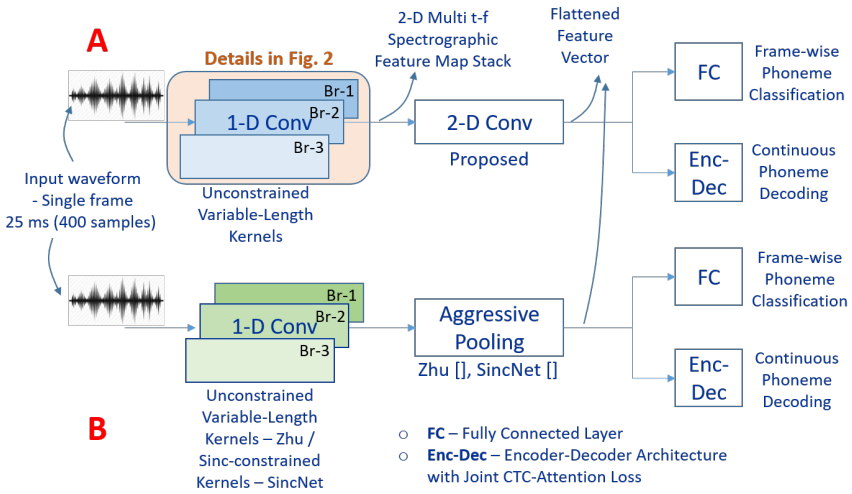

Fig. 1. *Multi time-frequency resolution CNN architecture - Overview of A) proposed and B) Zhu [15] and SincNet [16]*

1ms, 5ms and 10ms respectively. This corresponds to a 'unconstrained variable-length kernels' scenario which offers an unrestricted learning of the kernel weights (i.e., filter impulse responses of the convolving filter and corresponding band-pass characteristics of each such kernel) for a given task. In Panel B, the 3 1-D conv layers correspond to those of Zhu's system (also 'unconstrained variable-length kernels') and to those of SincNet where the kernels are 'fixed-length constrained kernels' in the sense of being constrained to be Sinc-functions - each parameterized to different band-pass cut-off frequencies of a rectangular band-pass filter - within a fixed duration kernel size for a specific branch. Note that this is a multi-branch generalization of the original E2E-SincNet architecture [16] (which worked with only one branch).

2) **Nature of back-end processing:** Panel A shows the proposed approach to use a 2-D Conv Layers to work on the multi t-f spectrographic feature map 'stack' viewed as a composite image - which has a rich t-f representation of the input signal in a complementary manner across the 3 maps (narrowband to wideband) - and from which it is important to extract further features across t-f cells of the acoustic-phonetic manifestation of different phone-classes for further classification (via fully connected layers) / decoding (via an Encoder-Decoder and CTC/Attention). Panel B, on the contrary, shows this back-end to be an 'Aggressive Pooling' as performed in these respective works (Zhu and E2E-SincNet) which fails to retain or exploit the rich 2-D t-f map 'stack' - but reduces it directly into a flattened feature vector (of dim 480) - which is further fed to a fully-connected layer or encoder-decoder for frame-wise phoneme classification or continuous phoneme decoding.

We argue that Panel A (the proposed pipeline) can offer superior performance to the pipeline in Panel B (representing the Zhu's system or the E2E-SincNet) in differentiating itself with respect to the above two central aspects of the multi t-f resolution CNN representation learning, namely, i) the front-end 1-D Conv 'unconstrained variable-length kernels' facilitating the very formation of the multi t-f spectrographic feature map 'stack' by multiple branches and ii) the back-end 2-D Conv layers to provide for back-end processing of such a multi t-f 'stack' for more efficient 2-dimensional t-f feature extraction for subsequent down-stream classification and decoding.

IV. Multi Time-Frequency Resolution CNN

We now elaborate on the '1-D Conv' block in Panel A of Fig. 1 (marked 'Details in Fig. 2'), as this forms the central part of the multi t-f CNN architecture being proposed and studied here. This '1-D Conv' block is shown in Fig. 2 in an expanded form highlighting the formation of the multi time-frequency spectrographic feature maps.

The 1-D Conv layers corresponding to a multi-branch CNN architecture is capable of processing the raw 1-d signal input (speech waveform for ASR) to create multiple spectrographic feature maps with a wide range of time-frequency resolution trade-offs. It can be seen that the input raw signal (shown as 25 ms sec duration here, made of 400 samples corresponding to a sampling rate of 16 kHz), is fed to 3 branches, each with a set of 32 kernels, with each branch having a fixed kernel size (e.g. branch 1 has kernel size of 1 ms or 16 samples, branch 2 has kernel size of 5 ms of 80 samples aand branch 3 has kernel size of 10 ms or 160 samples).

To provide a reference, a conventional CNN has only one branch (with multiple kernels, e.g. 32 here), with some fixed kernel size, e.g. 5 ms (in the 2nd branch). In such a conventional CNN branch, each kernel convolves with the 1-d signal input and yields an output that is a linearly filtered version of the signal through each of the 32 kernels in that branch. As the CNN learns to map the input to the classes in the fully connected layer in the output (for e.g. as in the frame-wise phoneme classification pipeline in Panel A or B in Fig. 1), the kernels (the filter coefficients) are optimized to learn to extract an appropriate feature from the input signal, and create a 'feature map' which is one spectrogram-like

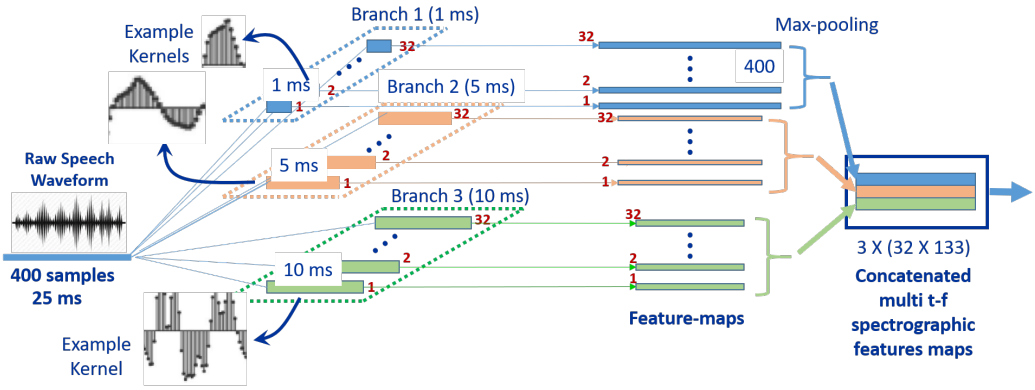

Fig. 2. *Multi time-frequency resolution CNN conv layer - formation of multi time-frequency spectrographic feature maps*

output made of 32 channels each with its time varying filter outputs, which is further down sampled (e.g. by max pooling or average pooling). This 'single' spectrogram is governed by the time-frequency trade-off inherent and defined by the kernel size (of the single branch, taken as 5 ms or 80 samples in this discussion), i.e., the kernel as a filter defines an impulse response of length 80. Its corresponding frequency response has a typical band-pass characteristic with the band-pass bandwidth determined by the kernel length (80 here); the actual frequency response is itself determined by the kernel values which in turn are determined by the CNN's weight learning for the given task, typically with different kernels tuning-in to different parts of the spectral range.

The resultant spectrogram-like feature map can be viewed as a narrow-band or wide-band spectrogram depending on the kernel size, as is well known for instance in speech signal processing [19], i.e., small kernels yield high temporal resolution and poor frequency resolution resulting in a wide-band spectrogram and long kernels yield poor temporal resolution and very good frequency resolution resulting in a narrow-band spectrogram. This can also be viewed as equivalent to a filter-bank analysis of the input signal with the filter-banks' filter's spectral characteristics (mainly the band-pass bandwidths) determined by the kernel size. Thus a single branch CNN performs the equivalent of a filter bank analysis, with each filter in the filter bank having a fixed kernel size, and possibly a fixed stride (corresponding to the hops of the filters), followed by a down sampling of the filter-bank outputs via max- or average-pooling in CNN terminology.

It is clear that such a 'single' branch and the corresponding spectrogram with a time-frequency trade-off specific to the kernel size of that branch is highly restricted in the kind of time-frequency analysis it can perform on the input 1-d signal. For instance, in a wide class of 1-d signal classification problems such as speech recognition, audio-classification or music-genre classification problems, the signal is highly non-stationary with the spectral dynamics changing at varying rates in time, and with various spectral events localized in frequency likewise exhibiting different temporal evolution. In order to capture these dynamic events in time and frequency, localized at different scales in time and frequency, a single spectrographic representation as obtained by a single branch CNN is clearly inadequate. This calls for a mechanism to generate time-frequency representations at different time-frequency resolutions, that is made possible by considering multiple branches in the CNN, each branch with a pre-specified but variable kernel size which is same for all the kernels in that branch.

Fig. 2 shows such a multi-branch CNN with 3 branches. Shown are branches 1, 2 and 3 with the corresponding kernel sizes 16 (1ms), 80 (5ms)and 160 (10ms). Such a multi-branch CNN will generate a spectrographic feature-map in 'each' of the 3 branches, each such feature map having its unique time-frequency trade-off determined by the kernel size used in the corresponding branch. For example, here, Branch 1 with kernel size 16 samples (1ms), will yield a very wide-band spectrogram (with a very fine time-resolution and poor frequency resolution), Branch 2 with kernel size 80 samples (5 ms) will yield a less wide-band spectrogram and Branch 3 with a kernel size 160 samples (10 ms) will yield a very narrow-band spectrogram. The 3 branches taken together will yield multi time-frequency resolution spectrographic feature maps, each of size 32 frequency channels $\times$ number of filter outputs decided by the stride of the convolution kernel in that branch (e.g. 32×400 for Branch 1 with stride of 1). Each of these are subject to max-pooling to reduce them to a feature-map of size $3 \times 32 \times 133$ which is further processed (as in Fig. 1) in the pipeline in Panel A by a '2-D Conv' block for efficient 2-D t-f representation learning and in Panel B by an aggressive pooling to generate a multi-scale feature vector as in Zhu's and E2E-SincNet systems.

V. Data Corpus

Our experiments were evaluated on the TIMIT speech corpus containing recordings (sampling rate = 16kHz) of phonetically-balanced read English speech. We worked using the standard train-dev-test split of the TIMIT database consisting of 3,696 training utterances (excluding SA utterances) from 462 speakers, 400 validation utterances by 50 speakers and 192 utterances by 24 test speakers. The validation set is used to determine the best performance model on which the results on the test dataset are reported. Raw speech waveform corresponding to each utterance is split into segments of 25ms (400 samples) and given as input to the system.

VI. Experiments and Results

In this section, we present the frame-wise accuracy and continuous phoneme decoding PER of our proposed multi time-frequency resolution CNN based architecture and compare the results with Zhu's multi-scale feature [15] and the E2E-SincNet [16] systems. For each variant, we compare the performance of single branch fixed kernel size CNN with multi-branch variable kernel-size CNN to understand and show the advantage of the multi time-frequency systems.

Frame-wise phoneme classification results: The output from the convolution layers is fed to 3-layer DNN containing 1024 hidden units followed by a softmax layer. Table I outlines the framewise accuracy of the three architectures considered here. We see that our proposed multi time-frequency resolution CNN architecture gives the best frame-wise accuracy of 63.03% which i) is up to 7% (absolute) superior to a single-branch system and ii) which offers an improvement of 3% (absolute) over Zhu's multi-scale feature based ASR and a 8% (absolute) over SincNet.

TABLE I
FRAME-WISE PHONEME CLASSIFICATION RESULTS ON TIMIT DATASET OBTAINED WITH THE PROPOSED (MULTI T-F RESOLUTION CNN), ZHU'S MULTI-SCALE FEATURE AND THE SINCNET SYSTEMS

Kernel size			Frame-wise accuracy		
1 ms	5 ms	10 ms	Proposed	Zhu	SincNet
96	0	0	56.3	60.3	56.5
0	96	0	56.8	60.7	54.1
0	0	96	59.6	60.5	51.8
32	32	32	63.03	60.8	55.5

E2E systems for continuous phoneme decoding: All the three architectural variants take in raw-audio blocks of 25ms (400 samples) directly as input and use an encoder-decoder architecture which is trained on joint CTC-attention loss function with the tunable hyper-parameter (weighing the two losses), $\lambda = 0.5$. The output from the convolution layers are fed to Bi-LSTM encoder with 4 layers consisting of 512 hidden units. The decoder used in the attention network is a one-layer LSTM of size 512 units. Table II reports the PER obtained by the three architectural variations discussed in our work. We observe that our proposed E2E multi time-frequency system gives the best PER of 20.4% and is superior to a single branch CNN architecture by as much as 2% (absolute) PER and the other two (Zhu and E2E SincNet) multi-branch systems. Our proposed system gives a 3% (absolute) performance gain compared to the other 2 variants.

TABLE II
PER FOR PROPOSED (MULTI T-F RESOLUTION CNN), ZHU'S MULTI-SCALE FEATURE AND THE E2E-SINCNET SYSTEMS

Kernel size			PER		
1 ms	5 ms	10 ms	Proposed	Zhu	E2E-SincNet
96	0	0	22	23.3	22.9
0	96	0	21.1	24.2	22.4
0	0	96	22.4	25.9	23.3
32	32	32	20.4	23.7	24.1

VII. Conclusions

We have proposed a multi time-frequency (t-f) resolution CNN architecture for E2E speech recognition from raw speech waveform - which works with unconstrained variable-length kernels and a back-end 2D-convolutional layers to process the t-f spectrographic feature maps. We have shown a performance gain of 3-8% accuracy (absolute gain) in the frame-wise classification task and 3% PER (absolute gain) in the continuous phoneme decoding task of the proposed multi t-f resolution architecture with a 2-D CNN back-end over two recent variants - a multi-scale feature based system and the E2E-SincNet.

References

[1] Y. LeCun et al.. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, vol. 1, pp. 541-551, 1989.

[2] Y. Bengio, A. Courville, P. Vincent Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, issue 8, pp. 1798-1828, Aug. 2013.

[3] Alex Krizhevsky, Ilya Sutskever, Hinton, Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks Communications of the ACM. 60 (6): 84–90, June 2017.

[4] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn and Dong Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Trans. on Audio, Speech and Language Processing, vol. 22, no. 10, pp. 1533- 1545, Oct. 2014.

[5] D. Palaz, R. Collobert, R. Magimai-Doss, Analysis of CNN based speech recognition system using raw speech as input. Proc. Interspeech '15, Dresden, 2015.

[6] Tara N. Sainath, Ron J. Weiss, Andrew W. Senior, Kevin W. Wilson and Oriol Vinyals Learning the speech front-end with raw waveform CLDNNs". Proc. Interspeech '15, Dresden, 2015.

[7] Wei Dai, Chia Dai, Shuhui Qu Juncheng Li Samarjit Da. Very deep convolutional neural networks for raw waveforms. Proc. ICASSP '17, New Orleans, LA, 2017.

[8] Tokozume, Y., Harada, T. Learning environmental sounds with end-to-end convolutional neural network. Proc. ICASSP '17. New Orleans, LA, 2017.

[9] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim and Nam, Juhan. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. Proc. 14th Sound and Music Computing Conference, pp. 220–226, Espoo, Finland, 2016.

[10] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim and Juhan Nam. SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification. Appl. Sci., 8, 150, 2018.

[11] Christian Szegedy et al. Going Deeper with Convolutions. Proc. CVPR 2014.

[12] Boqing Zhu, Changjian Wang, Feng Liu, Jin Lei, Zengquan Lu, Yuxing Peng. Learning Environmental Sounds with Multi-scale Convolutional Neural Network. Proc. IJCNN 2018, (also arXiv:1803.10219v1, Mar 2018)

[13] Boqing Zhu, Kele Xu, Dezhi Wang, Lilun Zhang, Bo Li, Yuxing Peng, Environmental Sound Classification Based on Multi-temporal Resolution Convolutional Neural Network Combining with Multi-level Features. arXiv:1805.09752v2, Jun 2018.

[14] T. Vijaya Kumar, R. Shunmuga Sundar, Tilak Purohit, V. Ramasubramanian. End-to-end audio-scene classification from raw audio: Multi time-frequency resolution CNN architecture for efficient representation learning. 2020 International Conference on Signal Processing and Communications (SPCOM), DOI: 10.1109/SPCOM50965.2020.9179600, IEEE, 2020.

[15] Zhenyao Zhu, Jesse H. Engel and Awni Hannun. Learning Multiscale Features Directly From Waveforms. Proc. Interspeech 2016, pp. 1305-1309, San Francisco, USA, Sep. 2016.

[16] Titouan Parcollet, Mohamed Morchid and Georges Linares. E2E-SincNet: Toward fully end-to-end speech recognition. Proc. ICASSP 2020, pp. 7714-7718, 2020.

[17] M. Ravanelli and Y. Bengio. Interpretable Convolutional Filters with SincNet. 32nd Conference on Neural Information Processing Systems (NIPS 2018) IRASL workshop, Montréal, Canada.

[18] M. Ravanelli and Y. Bengio. Speaker recognition from raw waveform with SincNet. Proc. of SLT, pp. 1021-1028, 2018.

[19] T. F. Quatieri. Discrete Time Speech Signal Processing. Prentice Hall, 2002