Multi-temporal end-to-end CNN: Audio-scene classification, speech emotion recognition and accent identification from raw speech signal

Pushpa Ramu[1], *T. Vijaya Kumar*[2], *R. Shunmuga Sundar*[2], *Anzar Zulfiqar*[2], *Rajeev Ranjan*[2],
Prashanth Narayan Jalli[1], *Tilak Purohit*[3], *V. Ramasubramanian*[3]

[1]Samsung Semiconductor India R&D, Bangalore(SSIR), India
[2]Samsung R&D Institute, Bangalore (SRIB), India
[3]International Institute of Information Technology - Bangalore (IIIT-B), Bangalore, India

{pushpa.r, vijay.t , sundar.rs, anzar.zulfi, rajeev.r, prashanth.nj,
tilak.purohit}@iiitb.org, v.ramasubramanian@iiitb.ac.in

Abstract

We propose a novel multi-temporal CNN architecture for end-to-end classification of raw speech signal for 3 tasks, namely, audio scene classification (ASC), speech emotion recognition (SER) and native language (L1) recognition (NLR) in accented English (L2) speech. Conventional CNNs use a fixed size kernel (whether for image or 1-d signal classification) which corresponds to applying a filter bank, where each filter has a fixed time-frequency resolution (i.e., fixed duration impulse response and a fixed band-width frequency response), importantly with a specific time-frequency trade-off. In contrast, in a way to allow for multiple time-frequency resolutions, we use a multi-temporal CNN architecture having multiple kernel branches (up to 12 branches) each of different lengths, thereby allowing for multiple filter banks with different time-frequency resolution to process the input raw speech signal and create feature-maps corresponding to different time-frequency trade-offs. Applying this architecture to end-to-end classification of the above 3 tasks is shown to offer consistent and significant performance enhancements - 11-15% absolute in accuracy for the ASC task (e.g. for up to 12 branches), 2-8% absolute in accuracy for the SER task (e.g. 3, 6 branches) and 1.2-11.6% absolute in accuracy and (Precision, Recall) of (1,1) with 6 branches for the NLR task) for the multi-temporal case over the conventional single-temporal CNN and in all 3 tasks and also outperform state-of-art results in the respective tasks by other methods in an end-to-end framework.

Index Terms: Multi-temporal CNN, end-to-end classification, audio-scene classification, speech emotion recognition, accent identification

1. Introduction

We propose a multi-temporal CNN architecture in an end-to-end configuration, with emphasis on the ability of this new architecture to perform enhanced representation learning from 1-dimensional signals such as audio and speech, and address three problems of interest in audio and speech recognition, namely, i) audio-scene classification (ASC), ii) speech emotion recognition (SER) and iii) native language (L1) recognition (NLR) in accented English (L2) speech as in our recent work [1, 2, 3]. 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 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 [4] 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 [5] in its lower convolution layers and the discriminative fully-connected higher layers for multi-class classification tasks such as from raw images [6], speech spectrographic images [7], speech-waveform [8], [9], audio-waveform [10], [11], music-waveform [12], [13].

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 (or windowed linear filtering) 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, and even more so in the 1-d signal processing community. While several related work does indeed come close to handling multi-scale properties [14], it is only the most recent work of [15], [16] that addresses this issue for the first time, and proposes 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. Within the similar notion of using parallel branches in the CNN, [14] considers a 'parallel CNN architecture' with two branches with two different 2-d kernels, each designed to capture temporal and frequency relationship in an image-like 80×80 input of a log-amplitude transformed Mel-Spectrogram with 80 Mel-bands spectral and 80 STFT frames temporal resolution, thus not addressing directly the issue of time-frequency trade-offs from raw 1-d waveform as done in [15], [16] and here.

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) [17], 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.

2. Focus and contributions

In this paper, our focus is along the following lines:

1. **Audio-Scene Classification (ASC):** We address the ASC problem within the DCASE (Detection and Classification of Acoustic Scenes and Events) setting [18], [19], using the DCASE-2017 and DCASE-2018 datasets for our work. CNN architectures and CNN based approaches are among the most popular techniques and systems submitted to DCASE challenges (e.g. as many as 19 submissions among 49 in 2018), and we point to the fact that neither these different contributions nor related work in speech recognition or image classification using CNNs have attended to or brought to the fore the issue of variable time-frequency analysis by means of variable convolution kernel sizes.
2. **Speech emotion recognition (SER):** The SER problem is one of the difficult classification tasks from the speech signal since the features that are percetually clear as an emotional category, are difficult to extract and quantify, given that such features are highly spread over time, partly in spectral information and partly in the suprasegmental aspects of prosody and voice-quality, without regard to the phonetic or linguistic content of the speech (which of course may also bear information) [20]. Hence the focus of much SER work has been on the definition and extraction of meaningful discriminative features, and this has been traditionally addressed using hand-crafted features followed by various classifiers, progressing to the current trend of 'feature bruteforcing' where a large (thousands) of spectral, prosodic and temporal features are extracted, selected and applied discriminatively to a a classifier [21]. This brings the focus on potential end-to-end approaches, which learn the features from the signal [22], [23]. In contrast to using conventional CNN architectures in such recent work on end-to-end SER, in this paper our emphasis is on the ability of the proposed new architecture to perform enhanced representation learning from 1-dimensional signals such as speech for emotion recognition
3. **Accent-identification or native language recognition (NLR):** The accent-identification problem is acknowledged to be one of the most difficult among speech-based recognition tasks [24], which states it as much less widely studied (than related problems of language identification). This is largely due to the fact the the L1 accent origin in an accented L2 speech is manifest as the 'L1 effect on L2' in several different ways, a large part in the acoustic realization of phones giving rise to phonological differences, in the pronunciation of words, and other suprasegmental aspects such as prosody of the speech. This makes the features that can identify and discriminate L1 accent very difficult to extract and more so if the task is definde as an end-to-end task, where hand-crafted features are not brought into effect. To this end, [24] also concludes '. . . there remains a broad variety of further information that is conveyed in the acoustics of speech and the spoken words themselves that have not been dealt with either at all or . . .'. In this paper, we emphasis on this aspect and the ability of the proposed new CNN architecture to perform enhanced representation learning from 1-dimensional signals such as speech for accent-identification.

In this paper, our contributions are along the following lines:

1. To extend and generalize the multi-temporal architecture of [15], [16] to a highly scaled number of multi-temporal branches, allowing for creating multiple spectrographic feature maps with a wide range of time-frequency resolution trade-offs, equivalent to a conventional notion of applying a set of filter-banks corresponding to feature-maps that range from very narrow-band to very broad-band spectrographic information. As a reference, a conventional CNN architecture is a 'single-branch' architecture, with some fixed kernel size, and will at best yield one spectrographic time-frequency feature map somewhere within the range of narrow- to wide-band analysis.
2. To show the very significant performance gain for all the 3 tasks, namely, a) 11-15% absolute with 12 branches for the ASC task, b) 2-8% absolute in accuracy for the SER task with 3, 6 branches and c) (1.2-11.6% absolute) by the multi-temporal architecture (with 6 branches), over a conventional single-branch CNN operating at any of the kernel sizes that is part of the multi-temporal architecture.
3. To point to the fact that the results for ASC surpass the early result of [16], which used up to 3 branches only, and thereby establish the intrinsic nature of the multi-temporal architecture to gain proportionately with an increase in the number of branches due to progressively enhanced representation learning ability (which potentially can extend even beyond the 12 branches we have experimented with, possibly for more complex tasks, e.g. more number of classes or other tasks such as acoustic-modeling for continuous phoneme-decoding or LVCSR).
4. To show that, for the ASC task, our best results (90% accuracy) surpass the top system of DCASE-2018 challenge (81%) by as much as 9% absolute [25], [26], [27].

3. Multi-temporal CNN architecture

The multi-temporal CNN architecture considered here is as shown in Fig. 1. This architecture is shown in two parts (as in Fig. 1a) and b)): a) Formation of the multi time-frequency spectrographic feature maps and b) From the feature maps to fully connected layers. Fig. 1a) is the essential contribution in this paper - namely, the multi-branch CNN architecture capable of processing the raw 1-d signal input (audio signal for ASC) to create multiple spectrographic feature maps with a wide range of time-frequency resolution trade-offs. The feature maps in 'C' are a stack of 32 individual spectrographic maps, each of length (66150, 13250, 6615, . . ., 441) corresponding to the 12 branches, and each of these are subject to max-pooling to reduce them to a feature-map of size $(M \times 32) \times 441$ or 384×441 for $M = 12$. This is shown in Fig. 1b). The feature map stack in 'C', on being reduced to a feature-map of size 384×441 for $M = 12$, as shown in Fig. 1b) is further processed by 4 convolutional layers, each with 64, 128, 256 and 256 filters each filter being a 3×3 kernel with a stride 1×1, yielding respectively 64 (128×40), 128 (64×20), 256 (32×10) and 256 (16×5) feature maps on suitable max-pooling at each stage. The final output of size $256 \times 16 \times 5$ from the fourth convolution layer is used directly as input to the fully-connected layer with an output layer with N soft-max outputs (corresponding to

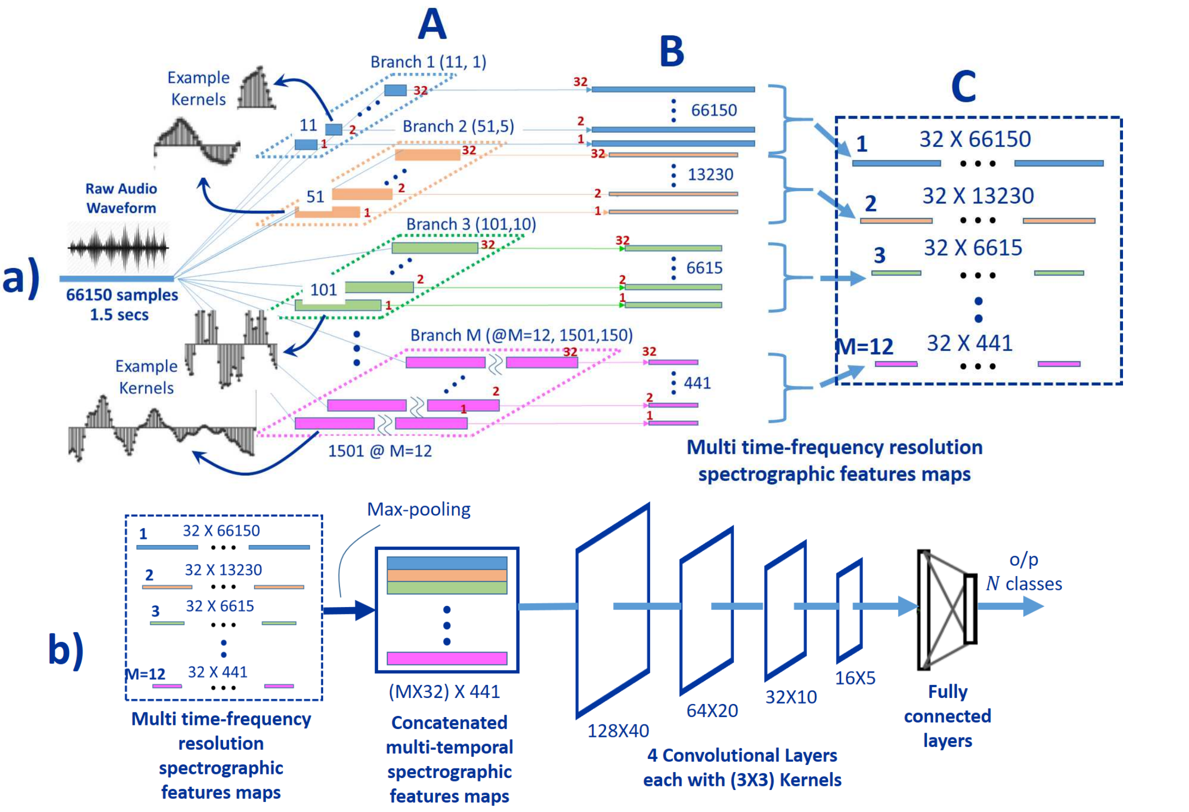

Figure 1: *Multi-temporal CNN architecture* - **a)** *formation of multi time-frequency spectrographic feature maps,* **b)** *from multi time-frequency spectrographic feature maps to fully connected layers*

N classes; e.g. $N = 15$ for the DCASE-2017 data-set; $N = 10$ for the DCASE-2018 data-set; $N = 5$ for the Speech Emotion Recognition task for 5 emotional classes in IEMOCAP data-set chosen here; $N = 6$ for the accent classes chosen from the Wild-cat data-set used here).

4. Experiments and Results

4.1. ASC task

4.1.1. Data Corpus

We have used the DCASE-2018 Task1(a) - Acoustic Scene Classification (ASC) dataset [19]. This challenge aims to classify a test recording into one of the 10 predefined classes that characterizes the environment in which it was recorded. The data-set used for this task is the 'TUT Urban Acoustic Scenes 2018' data-set. It consists of 10 acoustic environment scenes, such as for example: airport, urban park, travelling by an underground metro, indoor shopping mall etc. The sounds were collected from European cities; for a same scene, different locations were considered for recording the sounds. The dataset consists of 10-seconds audio segments from 10 acoustic scenes. Each acoustic scene has 864 segments (144 minutes of audio) comprising a total of 24 hours of audio. We have used a 70:30 train:test split and a 5-fold validation in all experiments. We now report results from experiments conducted with the proposed architecture on the ASC task, using the DCASE-2018 dataset [18], [19].

4.1.2. Results

First we show performance difference between the results obtained using a similar architecture by [15], Zhu2 on DCASE-2017 data-set and our architecture here for a 3-branch system (with kernels sizes 11, 51, 101). Fig. 2 shows the accuracy (%) of these systems. It can be noted that the proposed system offers an enhanced performance over the one reported by [15], [16]. More importantly, the all-3 branch system is significantly better than any of the single-branch (conventional CNN) system of [15], [16] on DCASE-2017, our system on DCASE-2017 and our system on DCASE-2018.

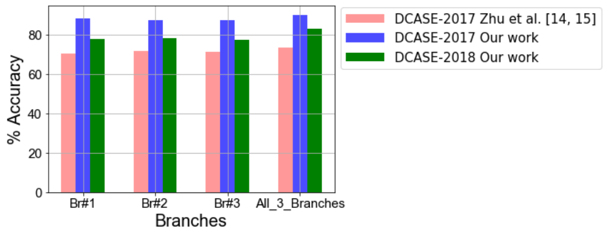

Figure 2: *ASC Task: Performance of proposed multi-temporal CNN architecture on DCASE-2017 and DCASE-2018 data-sets, and comparison to the related architecture of [15, 16] on DCASE-2017*

Fig. 3 shows the accuracy of ASC task for these cases (the four 'All_#_Branch' plots have symbols at the kernels sizes making up the All-branches). The following can be noted: The individual single-branch performances are considerably lower than all the multi-branch performances, with no particular

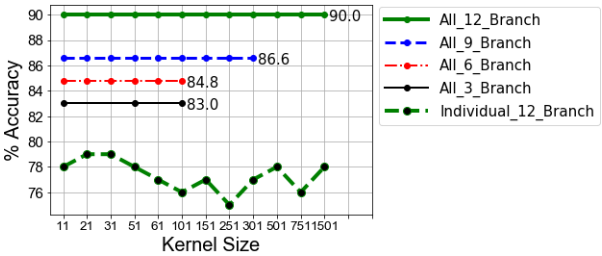

Figure 3: *ASC Task: Performance of proposed multi-temporal CNN architecture for varying number of multiple sub-sets of branches - 3, 6, 9 and 12 and comparison with single-temporal individual branch performances*

single-branch offering any specific advantage over others, while the joint multi 3-branch performance is 5% better than the best single-branch performance, validating the importance of having a joint time-frequency feature map. This performance increases significantly with increase in the number of branches M up to 12, with a performance gain of 11-15% (absolute) for the 12-branch system over the worst/best single-branch performances. This is a remarkable gain, considering most performance enhancements do not normally yield a quantum of the order of 11-15% (absolute)- which in this case is obtained by a direct consideration of the representation learning mechanism in the first layer of the CNN architecture.

4.2. SER task

4.2.1. Data Corpus

We have used the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database [**?**] for the work here. IEMOCAP is an acted, multimodal and multispeaker database, collected at SAIL lab at USC. It consists of dyadic sessions where actors perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions. IEMOCAP database is annotated by multiple annotators into categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance. We have used a single evalautor results to label the data as a particular class and merged all data so labeled as a class. Given the large variations in durations of each such class data (e.g. 17sec (disappointment), 17220 sec (frustration), we choose 5 classes namely, Anger, Frustation, Excited, Neutral and Sad (i.e., $N = 5$ in Fig. 1) that each each have 2460 sec of speech. We used a 70:30 (train, test) split and have carried out a 5-fold validation in all experiments.

4.2.2. Results

For the SER task, we consider the most important aspect of the multi-temporal architecture studied here, namely, the performance gain of a multi-temporal system over a single-branch (conventional CNN) system for increasingly larger number of branches. For this, we have considered number of branches $M = 1, 2, \dots, 12$, and obtained the i) individual branch performance for each of 12 single-branch architectures with kernel sizes 11, 51, 101, 151, 201, 251, 301, 501, 601, 751, 1001 and 1501 and, ii) the performance gain of multi-branch architectures for different $M = 3, 6, 9, 12$ - termed 3-branch, 6-branch, 9-branch and 12-branch architecture - across themselves (i.e., as the number of branches M increases in a M-branch system) and also with respect to the individual single-branch performances in (i). Fig. 4 shows the accuracy of SER task for these cases.

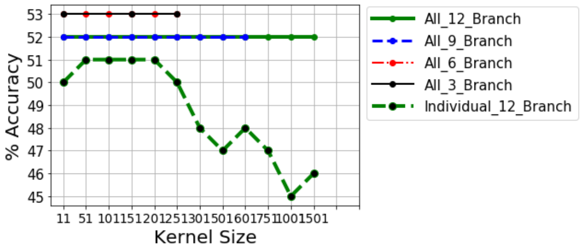

Figure 4: *SER Task: Performance of proposed multi-temporal CNN architecture for varying number of multiple sub-sets of branches - 3, 6, 9 and 12 and comparison with single-temporal individual branch performances*

The following can be noted: The individual single-branch performances show considerable variation (of over 6%, from 45% to 51%) in their performances across the 12 different kernel-sizes, with a few particular single-branches (at kernel sizes 51, 101, 151 and 201) offering a specific advantage over others. The multi 3-branch and 6-branch performances (coinciding at 53%) is better than this best single-branch performance by about 2%, but 8% better than the worst single-branch performance, validating the importance of having a joint time-frequency feature map, to combine the individual branch feature representation and to yield a consistent performance, which does not seem possible with the range of individual single-branch cases, where it becomes necessary to experiment with a range of such kernel-sizes to arrive at a best performance among a highly varied performance with kernel size. The multi-branch performance saturates here, with 9-branch and 12-branch offering 1% less than the best (for 3 and 6 branch cases) which is possibly due to the estimation difficulty in larger number of parameters with increase in the architecture complexity for the given training data conditions. The overall performance gain is therefore 2% (absolute) over the best individual single-branch case, and up to 8% (absolute) gain over the entire range of lesser performing single-branch cases. The results demonstrate the advantage of the multi-temporal architecture over a single temporal architecture (the conventional CNNs) - for which it is clear that the best kernel size has to be pre-determined for a given task, such as the kernel sizes of 51, 101, 151 offering the best performance for the SER task here, but which is obviated by the use of a multi-temporal architecture with even as small as 3 or 6-branches.

4.3. Accent identification task

4.3.1. Data Corpus

We have used the multi-accented dataset 'Wildcat Corpus of Native and Foreign-Accented English' [**?**] which has speakers from 12 different L1 (native language) origin speaking L2 accented English. These L1 groups are Chinese, Farsi, Indian, Italian, Japanese, Korean, Macedonian, Native, Russian, Spanish, Thai, Turkish. The scripted audio files with English as a target language (i.e. as the accented L2 language) were considered for this experiment. All speech files per speaker origin were merged and labeled by the speaker origin. Of these, considering minimum duration requirements per class, 6 origin (L1) classes - 'Chinese','Indian','Korean','Native','Spanish','Turkish' - were chosen, i.e. $N = 6$ in Fig. 1. The duration of speech data per class was 2210 secs. We have used a 70:30 (train,test) split of this data and performed a 5-fold validation in all experiments.

4.3.2. Results

Fig. 5 shows the accuracy of accent-identification task in terms of the performance gain of a multi-temporal system over a single-branch (conventional CNN) system for increasing number of branches, exactly as in the SER task.

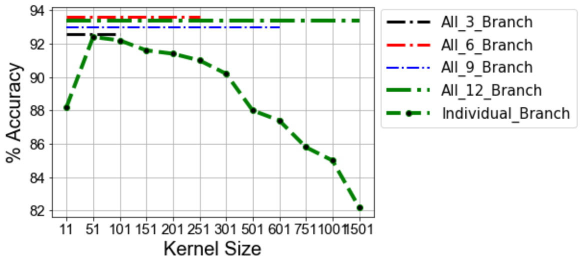

Figure 5: *Accent-identification task: Performance of proposed multi-temporal CNN architecture for varying number of multiple sub-sets of branches - 3, 6, 9 and 12 and comparison with single-temporal individual branch performances for M=1 to 12*

The following can be noted: The individual single-branch performances show considerable variation (over 10%) in their performances across the 12 different kernel-sizes, with only one particular single-branch (at kernel size 51) offering an advantage over others. The multi-branch performance increases significantly with increase in the number of branches M to 6, with a performance gain of 1.2% (absolute) over the 3-branch case and the best individual single-branch case, and up to 11.6% (absolute) gain over the entire range of poorly performing single-branch cases. The mutli-branch performance seems to saturate at 6 branches, with marginal variations for $M = 9$ and 12, with a slightly noticeable decrease, which is possibly due to the parameter estimation difficulty with increase in the architecture complexity for the given training data conditions. The results demonstrate the advantage of the multi-temporal architecture over a single temporal architecture (the conventional CNNs) - for which it is clear that the best kernel size has to be predetermined for a given task, such as the kernel size of 51 offering the best performance for the accent-identification task here, but which is obviated by the use of a multi-temporal architecture with even as small as 3 or 6-branches.

We also present the (Precision, Recall) measures in Fig. 6 for the multi-temporal architecture cases (as in Fig. 5 for % Accuracy) and the individual single-temporal architectures. As with % Accuracy in Fig. 5, the Precision, Recall measures are poor (in fact, much poorer than the % Accuracy) for the single-temporal cases, with no definitive indication of which kernel-size can offer the best performance, requiring these to be experimented with a priori before determining which kernel might be best for a given task. In comparison, the multi-temporal architecture, particularly $M = 6$ and above, offer remarkably high Precision, Recall, with $M = 6$ peaking at a Precision, Recall of 1.0.

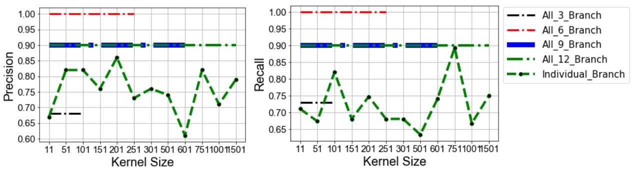

Figure 6: *Accent-identification task: Precision (left), Recall (right) of the multi-temporal architectures vs the individual single-temporal architectures*

5. Conclusions

We have proposed and studied a novel multi-temporal CNN architecture for three end-to-end tasks, namely, 'audio-scene classification' (ASC) from raw audio signal, 'speech emotion recognition' (SER) from raw speech signal and 'accent-identification' (or 'native language recognition' - NLR) from raw speech signal, as a generalization of a conventional single-branch CNN. The architecture is shown to offer consistent and significant performance enhancements - e.g. 11-15% absolute in accuracy for the ASC task (e.g. for up to 12 branches), 2-8% absolute in accuracy for the SER task (e.g. 3, 6 branches) and 1.2-11.6% absolute in accuracy and (Precision, Recall) of (1,1) with 6 branches for the NLR task) - for the multi-temporal case over the conventional single-temporal CNN and also outperforms state-of-art results for these tasks.

6. References

[1] T. Vijaya Kumar et al.. Multi-temporal end-to-end CNN: Audio-scene classification from raw speech. Submitted to Interspeech-2019, Graz, 2019.

[2] Anzar Zulfiqar et al.. Multi-temporal end-to-end CNN: Speech emotion recognition from raw speech. Submitted to Interspeech-2019, Graz, 2019.

[3] Rajeev Ranjan et al.. Multi-temporal end-to-end CNN: Accent identification from raw speech signal. Submitted to Interspeech-2019, Graz, 2019.

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

[5] 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.

[6] Alex Krizhevsky, Ilya Sutskever, Hinton, Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks Communications of the ACM. 60 (6): 8490, June 2017.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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. 220226, Espoo, Finland, 2016.

[13] 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.

[14] A. Schindler, T. Lidy, A. Rauber. Multi-temporal resolution convolutional neural networks for acoustic scene classification. Detection and Classification of Acoustic Scenes and Events 2017, November 2017, Munich, Germany.

[15] 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)

[16] 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.

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

[18] http://dcase.community/workshop2018/

[19] http://dcase.community/challenge2018/

[20] Bjorn W. Schuller. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. Commuications of the ACM, vol. 61, no. 5, pp. 90-99, May 2018.

[21] Bjorn Schuller, Stefan Steidl, Anton Batliner. The INTERSPEECH 2009 Emotion Challenge. Proc. Interspeech '09, pp. 312-316, Brighton, UK, 2009.

[22] Trigeorgis, G., Ringeval, F., Brckner, R., Marchi, E., Nicolaou, M., Schuller, B. and Zafeiriou, S.. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proc. ICASSP. pp. 5200-5204, Shanghai, P.R. China, 2016.

[23] Panagiotis Tzirakis, Jiehao Zhang, Bjorn W. Schuller. End-to-end speech emotion recognition using deep neural networks. Proc. ICASSP '18, pp. 5089-5093, 2018.

[24] Bjorn Schuller at al. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. Proc. Interspeech '16, pp. 2001-2005, San Francisco, USA, Sep. 2018.

[25] Yuma Sakashita, Masaki Aono. Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. Detection and Classification of Acoustic Scenes and Events 2018 Challenge, 2018 (http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a)

[26] http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a

[27] Plumbley, M. D., Kroos, C., Bello, J. P., Richard, G., Ellis, D. P. W., and Mesaros, A. (Eds.). Proc. of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018). Tampere University of Technology, 2018.