A CNN-GRU APPROACH TO CAPTURE TIME-FREQUENCY PATTERN INTERDEPENDENCE FOR SNORE SOUND CLASSIFICATION

Jianhong Wang[1], *Harald Strömfelt*[1], *Björn W. Schuller*[1,2]

[1] Department of Computing, Imperial College London, U.K.
[2] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
Email: {jianhong.wang16; h.stromfelt17; bjoern.schuller}@imperial.ac.uk

ABSTRACT

In this work, we propose an architecture named DualConvGRU Network to overcome the INTERPEECH 2017 ComParE Snoring sub-challenge. In this network, we devise two new models: the Dual Convolutional Layer, which is applied to a spectrogram to extract features; and the Channel Slice Model, which reprocess the extracted features. The first amalgamates an ensemble of information collected from two types of convolutional operations, with differing kernel dimension on the frequency axis and equal dimension on the time axis. Secondly, the dependencies within the convolutional layer channel axes are learnt, by feeding channel slices into a Gated Recurrent Unit (GRU) layer. By taking this approach, convolutional layers can be connected to sequential models without the use of fully connected layers. Compared with other state-of-the-art methods delivered to INTERPEECH 2017 ComParE Snoring sub-challenge, our method ranks 5th on performance of test data. Moreover, we are the only competitor to train a deep learning model solely on the provided training data, except for Baseline. The performance of our model exceeds the baseline too much.

Index Terms— DualConvGRU Network, Dual Convolutional Layers, Channel Slice Model

1. INTRODUCTION

Snoring, whilst potentially being embarrassing for a snorer, can have greater health implications such that correct identification of its cause can save lives. For example, *Obstructive Sleep Apnea* (OSA), a disorder that results from blockage of upper airways during sleep, can provoke damaging short and long-term effects to a person's well-being and can even cause death [1, 2, 3]. Snoring is a commonplace symptom of OSA and, in light of this, the INTERSPEECH 2017 ComParE Snoring sub-challenge [4] presents an annotated database of snoring audio samples, to be classified based on which region of the upper airways' vibrations cause the specific type of snoring that is heard. Precise determination of the *Velum* (V), *Oropharyngeal* (O), *Tongue* (T) or *Epiglottis* (E) is important, as it provides a first step to successful medical prevention of OSA [5, 6, 7].

In order to perform the classification, INTERSPEECH have provided an acoustic feature set consisting of 6,373 features. Having this many features is beneficial in paralinguistics [8], but also exposes an algorithm to the 'curse of dimensionality' and can cause dramatic increase in computation time. To combat this, techniques such as Principle Component Analysis (PCA) [9, 10], feature quantisation [11, 12] and feature selection [13, 14] are sometimes used. The alternative option is to use End-to-End neural network models, making use of the raw acoustic samples themselves and learning the features. However, through the years INTERSPEECH ComParE has presented data that has been sampled in-the-wild, presenting the important challenge of class imbalance in the data, wherein the data does not present an evenly distributed sample set over the classes. This results in classification algorithms performing well only on majority classes. Coupled with having a large feature set, algorithms are also exposed to over-fitting on the under-represented classes and is why DL approaches require large annotated datasets, which paralinguistic corpora, such as this one, often cannot provide. It is worth noting that this is the first year that INTERSPEECH ComParE has provided an End-to-End DL baseline, alongside the more typical Support Vector Machine (SVM) based approaches [4], which can be favourable when faced with small data.

In this work, we propose an End-to-End Deep Learning (DL) method that can address the problem of data imbalance when learning a small portion of data. The benefit of using DL, is that it can enable the model to learn a generalised representation for each sample category based on an abstracted encoding of the data that presents better comparison. This results in strong performance on unseen data. Our contributions to DL methods are threefold. Firstly, we propose the "Dual Convolutional Layer", which combines information collected by two separate convolutions with equal time axis kernel width but different frequency axis kernel specification. Their outputs are then merged via element-wise average. Secondly, we propose the "Channel Slice Model", which

ISBN 978-90-827970-1-5 © EURASIP 2018

can be used as the connector between 2D convolutional networks and sequential models without needing fully connected layers as features. In this method, we directly slice the output tensor from the convolutional layer along with the channel axis and feed it to a GRU layer. Hence, the dependency that we wish to encode in the sequential model is channel dependency, rather than traditional time dependence along inputs. Finally, we combine these partitions together to construct the "DualConvGRU" Network.

Through our work, we aim to accelerate the progress of DL in sound classification tasks, even for those which contain small data. In Section 2, we survey relevant audio classification work. We then describe the data processing measures taken and the "DualConvGRU" network in Section 3, followed by a short description of data in Section 4, the results and discussion in Section 5 and finally our conclusions and proposed further work in Section 6.

2. RELATED WORK

Due to the high dimensionality of features and the data class imbalance, typical approaches to audio-sample classification often involve SVM, as they are robust under these constraints. For example, in the sub-challenge baseline both an End-to-End method, utilising CNN + LSTM fed with 40 ms raw waveform chunks, and an SVM, trained on functionals computed over Low-Level Descriptor (LLD) contours, were tested [4]. It was found that in classifying snore audio, SVM performed best. However, the quality of features impacts the performance of the algorithm and a number of approaches to the sub-challenge have employed their own feature extraction. Amiriparian et al. [15] used AlexNet and VGG19 pre-trained image recognition Convolutional Neural Networks (CNNs) to the sample power spectrogram. The extracted features were then fed to two dense layers in series. They then used two SVMs, each trained on one of the two dense layer activations, to perform the final classification. A similar approach by Freitag et al. [16] made use of the same model but instead used a single SVM trained on a feature subset chosen through "Competitive Swarm Optimisation" [17]. Gosztolya et al. took a different approach to the challenge, by extracting frame-level features with openSMILE [18] following an approach taken in the Vocalization Sub-Challenge of ComParE 2013 [19], dividing utterances into 10 equal length parts and averaging the features across them. They then trained an SVM on these new features and combined with it results obtained from a separate SVM trained on the challenge features. In an approach that does not make use of CNNs, Kaya & Karpov applied score level fusion between different classifiers [20]. They proposed two new classification methods, namely "Weighted *Kernel Extreme Learning Machine*" (W*KELM*) to deal with class imbalances in the data and "Weighted *Kernel Partial Least Squares*" (W*KPLS*) regression. Their highest performance, which won the Snoring sub-challenge but remains behind Amiriparian et al. who were not official participants, came from a fusion between *KPLS*, WKPLS and *KELM* trained with openSMILE features and WKELM trained on novel features, extracted via Fisher Vectors [21].

Whilst the above models performed well on their respective tasks, only the challenge baseline made use of an End-to-End architecture.

3. METHODOLOGY

In this section, we will describe the algorithms used to deal with data imbalance and signal padding. The architecture of our network will then also be explained. The common approach for the INTERSPEECH ComParE 2017 participants was to apply traditional 'handmade features' as training data for statistical models such as GMM and SVM. However, our approach makes use of an End-to-End DL model and is shown to perform comparably. It is our hope that this work will inspire further developments in end-to-end audio classification.

3.1. Spectrogram

To represent each signal we use spectrograms with 21.875ms Hanning window, 50% overlap and absolute value magnitude of each component produced by Short-Time Fourier Transform (STFT). Due to the symmetry of STFT at each time step, we employ $(window\_size + 2)/2$ discretised bands for the frequency axis. The length of each signal is empirically restricted to 2.75 s. Any signals with lower length are padded via the proposed "Repeat Padding" algorithm, which is described in the following section. To convert each spectrogram to an image, we use the viridis colour map.

3.2. Reproducing Data

To overcome the data imbalance, we select more samples from the classes with less data available on the training set. Samples of each category are duplicated according to their inverse count rate, such that the new corpus is comparatively balanced on the distribution of categories. In experiments, we only duplicate training data.

3.3. Dual Convolutional Layer

To extract spectrogram features, we propose the Dual Convolutional Layer, which combines information collected from two convolutional operations. We use kernel sizes of $(3, 4)$ and $(3, 2)$, and implement a stride size of $(2, 2)$ for both. When applied to a spectrogram image, the x and y kernel axes are mapped to the time and frequency axes, respectively. Intuitively, the kernel with size of $(3, 4)$ summarises global frequency axis information, whilst the kernel with size of $(3, 2)$

ISBN 978-90-827970-1-5 © EURASIP 2018

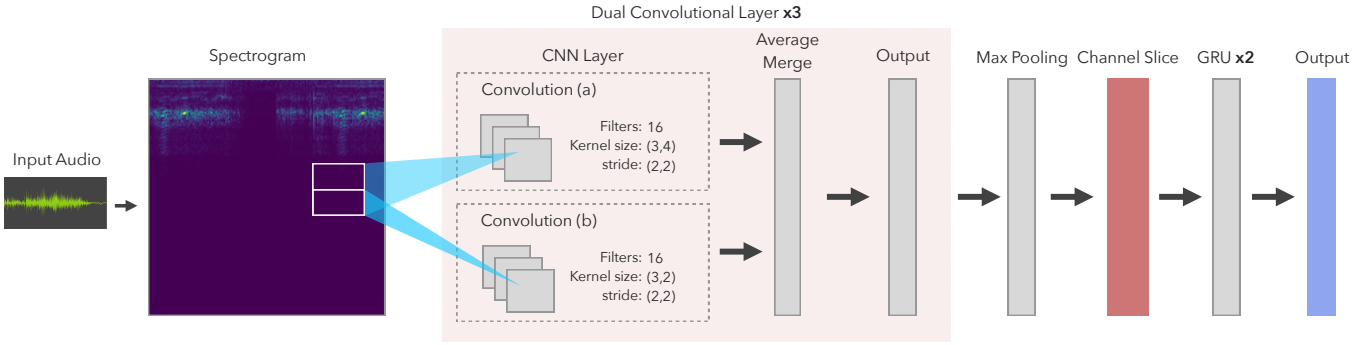

Fig. 1. This is the DualConvGRU architecture used in this paper. First, audio samples are converted into spectrograms. The Dual Convolutional Layer (shaded) then takes the spectrogram 3D tensor as input. This input is processed by two convolutions, with specifications as described in Section 3.3, and are element-wise averaged so as to create a new 3D tensor. This is repeated three times. Next we apply a max-pooling layer and use the proposed Channel Slice model to prepare data for entry into a series of two GRU layers. Finally, the output is computed as a softmax layer.

collects local frequency information. Different from similar methods, we restrict the time axis of the two convolutions' filters to be consistent and only consider variant information collected along with the frequency axis. The main reason for this decision is that we believe it is meaningless to gather information across different patches of signals. Instead, the problem should be concentrated on different representations along frequency domain of the same temporal signal segments. We can represent this convolutional operation mathematically as Equation (1) below, where Y is the output, F_1 and F_2 are two different filters, s is the stride and X is the input image.

$$Y = \frac{1}{2}(F_1 *_s X + F_2 *_s X) \tag{1}$$

The architecture of the Dual Convolutional Layer is shown in the shaded region of Figure 1. Each 3D tensor input is processed accordingly and the resultant tensors are then merged, to create a new 3D tensor. We use a small number of filters so that it can avert the problem of generalisation for small size corpora. For the same reason, we merge the tensors via element-wise average, as opposed to concatenation as used in Inception Networks [22].

3.4. Channel Slice Model

Since we need to connect Dual Convolutional Layers with a GRU layer, we propose the Channel Slice Model as an alternative to common methods, such as connecting fully connected embedding layers. The mechanism of this model is simple to understand. Instead of slicing along the time axis of signals, we directly slice the 3D tensor from Dual Convolutional Layers along the channel axis. We then flatten each feature map so that it becomes a 1D tensor, consisting of several segments of frequency features at each time step, which are concatenated according to time step sequence. Finally, we intuit that there should exist some *feature map* dependency across the whole sample, instead of frequency dependence across time segments within the sample. We thus cascade these 1D tensors so as to form a 2D tensor that can be passed to the two GRU layers, forming the overall DualConvGRU network as shown in Figure 1.

4. DATA DESCRIPTION

This work uses the *Munich-Passau Snore Sound Corpus*, which consists of 828 snore samples, separated across the four VOTE classes. The data is prepared into train, development and test sets of equal size as per the Interspeech 2017 specifications (see [4] for further details). However, the data per class is not equal, for example V class samples dominate the distribution. Therefore, data imbalance must be considered as is discussed in Sections 1 and 3.2.

5. EXPERIMENTS AND RESULTS

Consistent with past INTERSPEECH Paralinguistic Challenges, we use Unweighted Average Recall (UAR) as our performance measure [23, 24, 25]. All of the models and experiments are implemented by TensorFlow[1] and Keras[2]. Before experiments, we pre-process the data. First, we use bilinear interpolation to resample each spectrogram to mitigate some issues from noisy representations so as to reduce the probability of overfitting, but maintain their original dimensions. Then, we normalise the range of each element of the spectrograms from $0 - 255$ to $0 - 1$. Each spectrogram is computed by LibROSA[3] and converted to an RGB image by matplotlib[4]. We have attempted other image settings but

[1] https://www.tensorflow.org/
[2] https://keras.io/
[3] https://librosa.github.io/librosa/
[4] https://matplotlib.org/

ISBN 978-90-827970-1-5 © EURASIP 2018

Method	Ref	Devel (UAR%)	Test (UAR%)
Deep-Spectrum-SVM	[15]	44.8	**67.0**
End-to-Evolution	[16]	56.7	66.5
Fusion-Weighted-Kernel-Classifier	[20]	unknown	64.2
DNN-based Feature Extraction and Classifier Combination	[26]	unknown	64.0
DualConvGRU+l2-regulariser	/	51.7	63.8
DualConvGRU+dropout	/	46.8	61.1
Baseline Functionals	[4]	40.6	58.5
Baseline CNN & LSTM	[4]	40.3	40.3

Table 1. This is the table that shows the results of our methods and other state-of-the-art methods in INTERSPEECH ComParE 2017. The performance measure used in the table is UAR (i.e. added recalls per class divided by number of classes) in percentage. The bold method names are our methods, with dropout and l2-regulariser respectively.

#	V	E	O	T
V	68	25	41	21
E	1	21	0	5
O	25	6	30	4
T	0	2	0	14

Table 2. Confusion matrix of the results of each category of snores on test data, produced by DualConvGRU with l2-regulariser (best result of this contribution). In this table, the rows represent Actuals whereas the columns represent Predictions.

this setting performs best empirically. Next, we compare two strategies to avert over-fitting: adding a dropout layer of 0.5 after the output of a maxpooling layer and adding an l2-regulariser, which only considers the convolutional layer parameters to the objective function. For the objective function, we use KL-Divergence. In order to tune the model parameters, we use training data for training and development data for validation. We choose the final parameters based on the best validation performance during training. Afterwards, we use the official test data to evaluate the performance of the model. The evaluation results of our models and other works on the same dataset are shown in Table 1. We can see that the DualConvGRU with a l2-regulariser performs better than with dropout on both development data and test data. By contrast with results of other state-of-the-art methods, our highest performing model ranks 5th on test data, with a UAR of 63.8%.

The test data confusion matrix (Table 2) for the results of the DualConvGRU with the l2-regulariser shows that the performances on E and T are better than the other two categories. We believe that the V and O classes may be adversarial samples, compared with other categories. For this reason, accurate classification between these two categories may be a worthy concern for further work.

6. CONCLUSION

In this paper, we introduced a new End-to-End neural network model to tackle the INTERSPEECH 2017 ComParE Snoring sub-challenge, which presents the challenge of a small size dataset. We designed a new topology of convolutional layers which allows us to combine information from both local and global frequency scope, named Dual Convolutional Layers. We then provide a new strategy for connecting convolutional layers to sequential models, named Channel Slice Model, which extracts channel slice data as features from the convolutional layers. The results obtained show that our methods present a great improvement compared with the End-to-End challenge baseline on both the development data and test data. Moreover, contrary to most other challenge competitor methods, our approach is End-to-End and as such only depends on the raw snore sample spectrograms, needing no further acoustic features.

In further work, we would like to investigate what dependency is actually learnt in the Channel Slice Model and assess its performance on other datasets. Additionally, to further improve the problem of lacking data on audio processing, using deep learning methods, we propose to learn a generative model to represent the hidden space of each category of samples rather than repeating data samples directly.

7. REFERENCES

[1] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," *American Journal of Epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.

[2] A. Shamsuzzaman, B. Gersh, and V. Somers, "Obstructive sleep apnea: Implications for cardiac and vascular disease," *JAMA*, vol. 290, no. 14, pp. 1906–1914, 2003.

[3] T. Young, L. Finn, P. E. Peppard, M. Szklo-Coxe, D. Austin, F. J. Nieto, R. Stubbs, and K. M. Hla, "Sleep disordered breathing and mortality: Eighteen-year follow-up of the wisconsin sleep cohort," *Sleep*, vol. 31, no. 8, pp. 1071–1078, 2008.

[4] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3442–3446.

[5] D. L. Herath, U. R. Abeyratne, and C. Hukins, "HMM-based snorer group recognition for Sleep Apnea diagnosis," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, 2013, pp. 3961–3964.

[6] J. Fiz, J. Abad, R. Jane, M. Riera, M. Mananas, P. Caminal, D. Rodenstein, and J. Morera, "Acoustic analysis of snoring

ISBN 978-90-827970-1-5 © EURASIP 2018

sound in patients with simple snoring and obstructive sleep apnoea," *European Respiratory Journal*, vol. 9, pp. 2365–2370, 1996.

[7] J. Sola-Soler, R. Jane, J. A. Fiz, and J. Morera, "Automatic classification of subjects with and without sleep apnea through snoring analysis," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 6093–6096.

[8] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - State-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013.

[9] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, "Detecting Depression using Vocal, Facial and Semantic Communication Cues," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, Amsterdam, The Netherlands, 2016, pp. 11–18.

[10] C. J. C. Burges, "Dimension reduction: A guided tour," *Foundations and Trends in Machine Learning*, vol. 2, no. 4, pp. 275–365, 2010.

[11] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, San Franciso, USA, 2016, pp. 495–499.

[12] S. Amiriparian, J. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. Schuller, "Is deception emotional ? An emotion-driven predictive approach," in *Proceedings of INTERSPEECH*, San Francisco, USA, 2016, pp. 2011–2015.

[13] A. Ivanov and G. Riccardi, "Kolmogorov-Smirnov test for feature selection in emotion recognition from speech," in *Proceedings of ICASSP*, Kyoto, Japan, 2012, pp. 5125–5128.

[14] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Random discriminative projection based feature selection with application to conflict recognition," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, 2015.

[15] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3512–3516.

[16] M. Freitag, S. Amiriparian, N. Cummins, M. Gerczuk, and B. Schuller, "An 'End-to-Evolution' Hybrid Approach for Snore Sound Classification," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3508–3511.

[17] R. Cheng and Y. Jin, "A competitive swarm optimizer for large scale optimization," *IEEE Transactions on Cybernetics*, vol. 45, no. 2, pp. 191–204, 2015.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich Versatile and Fast Open-source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. Florence, Italy: ACM, 2010, pp. 1459–1462.

[19] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proceedings of INTERSPEECH*, 2013, pp. 148–152.

[20] H. Kaya and A. A. Karpov, "Introducing Weighted Kernel Classifiers for Handling Imbalanced Paralinguistic Corpora: Snoring , Addressee and Cold," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3527–3531.

[21] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Proceedings of INTERSPEECH*, Dresden, Germany, 2015, pp. 909–913.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 2818–2826.

[23] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proceedings of INTERSPEECH*, Singapore, 2014, pp. 427–431.

[24] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition," in *Proceedings of INTERSPEECH*, Dresden, Germany, 2015, pp. 478–482.

[25] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of INTERSPEECH*, San Franciso, USA, 2016, pp. 2001–2005.

[26] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based Feature Extraction and Classifier Combination for Child-Directed Speech , Cold and Snoring Identification," in *Proceedings of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3522–3526.

ISBN 978-90-827970-1-5 © EURASIP 2018