MODELLING GLOTTAL SOURCE INFORMATION FOR DEPRESSION DETECTION

D S Pavan Kumar Bogdan Vlasenko
Mathew Magimai.-Doss

Idiap-RR-13-2018

AUGUST 2018

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Modelling glottal source information for depression detection

D. S. Pavan Kumar[1,2], Bogdan Vlasenko[1], and Mathew Magimai.-Doss[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Switzerland
{pavankumar.dubagunta, bogdan.vlasenko, mathew}@idiap.ch

6th August 2018

Abstract

Detection of depression based on speech signal alone is a challenging problem. Inspired from recent works on direct modelling of raw waveform for speech processing, this paper investigates automatic modelling of glottal source information from speech signal using convolutional neural networks (CNNs) to detect depression. Since this task is challenging, in addition to raw speech modelling, we investigate modelling of linear prediction residual signal and zero frequency filtered (ZFF) signal. Experimental studies on AVEC 2016 challenge data set show that depression can be better modelled using ZFF signal with low complexity than using raw speech signal or linear prediction residual. Furthermore, the resulting system is better than or comparable to state-of-the-art system based on low level descriptors.

Keywords Convolutional neural networks, depression detection, zero-frequency filtering, glottal source signals.

1 Introduction

Humans express their emotions through vocal, linguistic and facial gestures that convey the person's mental state. Depression is one such phenomenon, whose detection and severity assessment have gained interest in the recent years [1, 2, 3, 4]. Automatic depression classification and severity prediction have been carried out in the literature by measuring parameters from sessions of patient clinical interviews through multiple modes: audio, video and text, and by using appropriate classification/regression tasks, for example [5, 6, 7, 8]. Valstar et. al. [4], as part of the audio-visual emotion challenge (AVEC) 2016, report that the detection scores are lower on purely speech based analyses as compared to those using multiple modes, indicating the need for further research in the field.

Depression is known to affect human speech production and cognitive processes. Specifically, depression impacts speech motor control [1, 9]. Depression, similar to many speech motor control disorders [10], can be identified by articulatory and phonetic errors, prosodic abnormalities. Speech signal analysis based depression detection focuses on these aspects. Sahu and Epsy-Wilson [11], Honig et al. [12] and Scherer et al. [13] have researched on voice quality features, such as degree of breathiness, jitter, shimmer. Simantiraki et al. used glottal source related features to detect depression: precisely, phase distortion deviation that is related to the shape of the glottal pulse [14]. Valstar et al. used prosodic, spectral and voice quality-related features [4]. Williamsons et al. used vocal tract correlation features for prediction of an individual's level of clinical depression [15]. i-vectors have also been studied for depression detection [16].

There have been also approaches motivated from speech emotion recognition research. Stasak et. al. used Geneva minimalistic acoustic parameter set (GeMAPS) [17] to detect depression [18]. Similarly, Gupta et. al. [19] used depression severity to predict affective states. Vlasenko et. al. used the extended GeMAPS (eGeMAPS) features and vowel formant location information to improve the detection, indicating the role of vocal tract system component in depression detection [20]. He and Cao [21] used a combination of features along with LLDs to predict depression severity, such as intermediate representations from CNNs trained to predict filter-bank representations from raw speech and to predict texture classification based features from spectrograms. Another method is to automatically learn features through machine learning methods, such as using CNNs [22]. These models can be further analysed for cues on their depression-related feature learning. Ma et. al. [23] proposed to predict depression using neural networks comprising convolutional and long-short term memory layers on log Mel filter-bank (LMFB) and magnitude-spectrogram features.

The present paper focuses on modelling glottal source information for depression detection. Depression can effect muscle tension and control, which in turn can affect vocal fold behaviour [1, 24]. The main challenge lies in accurately characterising the glottal source information from the speech signal, as it requires reliable

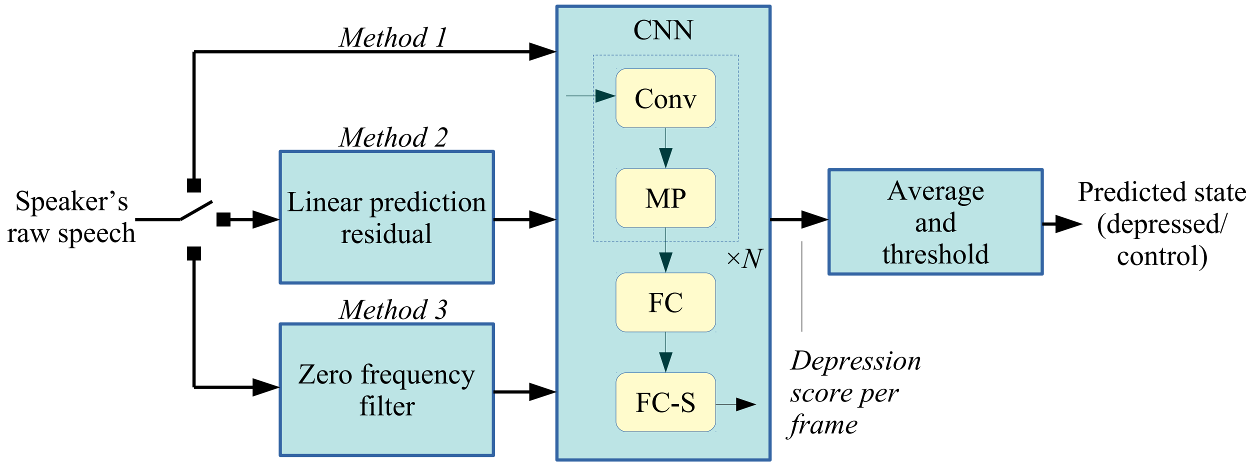


Figure 1: Proposed methods. CNN architecture: Conv: convolutional layer with ReLU activations, MP: max-pooling layer, FC: fully connected layer with ReLU activations, FC-S: fully connected layer with a single output node and sigmoid activation.

separation of vocal tract system information from the speech signal. In recent years, directly modelling raw waveforms for speech processing using neural networks have emerged [25, 26, 27, 28, 29, 30]. It has been found that these approaches are capable of modelling voice source and vocal tract system information with minimal assumptions [31, 30, 32, 33]. This paper investigates whether such methods can help in modelling better glottal source information for depression detection. We carry out three lines of investigation: (a) modelling of speech signal without any pre-processing, (b) modelling of linear prediction residual signal and (c) modelling of zero frequency filtered (ZFF) signal. Our studies show that modelling ZFF signal leads to systems with low complexity and better performance.

The rest of the paper is organised as follows. Section 2 details the proposed methods and motivates them. Section 3 presents the experimental setup. Section 4 presents results and analysis. Section 5 concludes the paper.

2 Proposed methods and motivation

We adopt the CNN-based approach to directly model raw speech signal that was first developed in the context of speech recognition [25], and then was later extended to presentation attack detection [34], speaker recognition [30, 32], paralinguistic challenges [35] and gender recognition [33]. As illustrated in Figure 1, the system consists of convolution layers followed by a multilayer perceptron, also referred to as a fully connected layer. In the remainder of the section, we briefly motivate the proposed methods that differ only in terms of input waveform. Figure 2 illustrates the three input waveform.

2.1 Speech signal

Speech is produced by excitation of the vocal tract system by the vibration of vocal cords. As speech signal contains voice source related information, one can ask whether deep learning methods can figure that information out automatically, given a few constraints in terms of speech processing. For instance, in the speaker recognition [30] and presentation attack detection [34] studies it was found that by letting the first convolution layer model *segmental speech* (speech of duration about 1-3 pitch periods) the CNN learns to capture low frequency information that could be related to speech/voice quality. Would such a modelling help depression detection? Similarly modelling of *sub-segmental speech* (speech of duration below 1 pitch period) enables capturing of vocal tract system related information [31, 32]. Glottal activity such as glottal closure instants (GCIs), glottal opening instants (GOIs) are temporal events that would need high time resolution. In that context, would sub-segmental speech modelling enable deciphering of those temporal events for depression detection?

2.2 Linear prediction residual signal

Linear prediction (LP) is a technique that is typically employed to deconvolve time varying vocal tract system information and excitation information in the speech signal [36]. More precisely, the residual signal carries glottal source information, and thus LP analysis forms one of the methods for glottal signal analysis [37, 38].

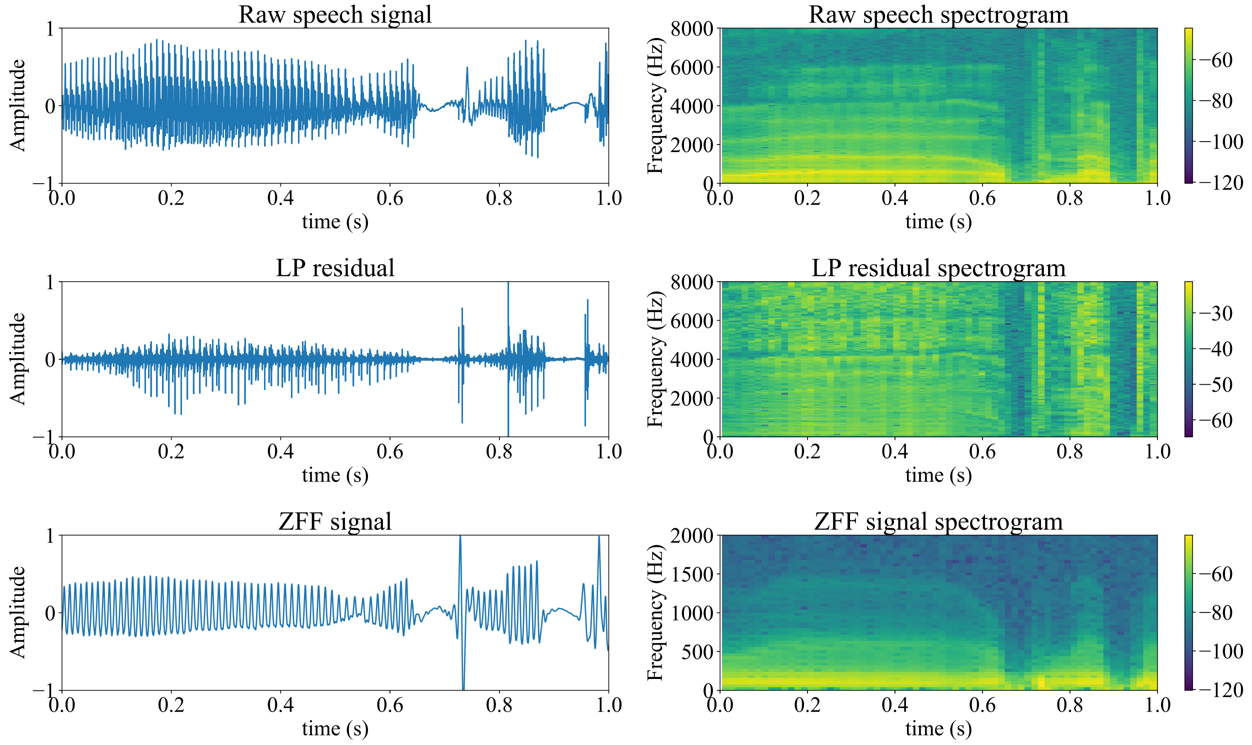


Figure 2: Illustration of a speech signal, 16^{th} order LP residual signal and ZFF signal and their respective spectrograms of an utterance *”to uh open up a...”*.

So can modelling of LP residual signal help in depression detection? Similar to speech signal modelling we can model LP residual signal at segmental level and sub-segmental level.

2.3 Zero frequency filtered signal

Zero frequency filtering is a technique that has recently emerged for characterising glottal source activity [39, 40]. It exploits the property of impulse-like excitation at glottal closure instance to detect GCIs. ZFF signal is obtained by passing the speech signal through a cascade of two ideal digital resonators located at 0Hz, and then removing the trend in the resulting signal by subtracting the average over a window of the size in the range of 1 to 2 pitch periods. In addition to GCIs, strengths of glottal excitation, fundamental frequency and glottal opening instants can be estimated from the ZFF signal [39, 41]. Furthermore, Kadiri et. al. have also investigated emotion recognition from these signals [42]. ZFF signal thus could be used for depression detection. Again we can consider modelling ZFF signal at segmental level and sub-segmental level.

Hypothesis: Among these three signals, we hypothesise that ZFF signal should lead to a better system. The reason being: (a) deciphering glottal source information directly from the speech signal may not be a trivial task, especially with limited amount of training data, (b) it is well understood that LP residual often contains effects due to resonances of the vocal tract system; this can affect glottal source processing [38], and (c) Zero frequency resonator reduces the effect of high frequencies significantly, and thus can be expected to not be affected by the time carrying characteristics of the vocal tract system [40]. In other words ZFF signal could be regarded as a clean signal-level representation of the glottal source information.

3 Experimental setup

3.1 Data set

The audio part of the Distress analysis interview corpus - wizard of Oz (DAIC-WOZ) corpus [43] was used for experimentation. The data set comprises audio-visual interviews of 189 participants who underwent evaluation of psychological distress. The interviews were carried out in English using an animated virtual interviewer [44], and each participant was assigned a self-assessed depression score through patient health questionnaire (PHQ-8) method [45].

Table 1: CNN architectures. N_f: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling.

Model (Input frame size)	Layer	Conv			MP
		N_f	kW	dW	
CNN2-subseg (250ms)	1	128	30	10	2
	2	256	10	5	3
CNN-subseg (250ms)	1,2	same as CNN2-subseg			
	3	512	4	2	-
	4	512	3	1	-
CNN-seg (250ms)	1	128	300	100	2
	2	256	5	2	1
	3,4	same as CNN-subseg			
CNN8-subseg (510ms)	1	64	30	10	2
	2	128	10	5	2
	3	256	3	1	-
	4	256	3	1	2
	5,7	512	3	1	-
	6	512	3	1	2
	8	512	3	1	3
CNN8-seg (510ms)	1	64	300	5	2
	2-8	same as CNN8-subseg			

We used the time labels provided in the data set to extract only the patient's speech recordings for experimentation. *We excluded the sessions 318, 321, 341 and 362 from the training set that had time-labelling errors.* We evaluated the proposed techniques on the dev set, since the test set was held out as part of the AVEC 2016 challenge [4].

3.2 Setup

The training data was split into 95% of speakers for training and 5% of speakers for cross-validation. The CNNs were trained using Keras deep learning library [46] with Tensorflow backend [47]. To ensure equal representation of both the control and depressed classes during training, we duplicated the depressed class utterances to a count matching as that of the control group. The architectures of CNNs used are listed in Table 1. The term *subseg* refers to *sub-segmental* modelling, where the filters in the first convolution layer models 30 samples (below 2 ms duration signal). Similarly, the term *seg* refers to *segmental* modelling, where the filters model 300 samples (about 20 ms signal). Note that CNN, unless specified, refers to a 4-convolutional layer network. "FC" in all the architectures contains 10 nodes. The input to the CNNs is 250 ms or 510 ms signal, which is shifted by 10 ms. All the frames of the depressed group were labelled 1, and the rest 0.

The CNNs were trained using cross-entropy loss with stochastic gradient descent. Learning rate was halved, in the range 10^{-1} to 10^{-6}, between successive epochs whenever the validation-loss stopped reducing. The frame level depression scores obtained were averaged per speaker to get speaker-level scores, which were thresholded to classify as depressed or control (not depressed). More precisely, the threshold was varied in steps of 0.01, and for each step the sum of the control and depressed F1 scores were computed. Finally the threshold that gave the maximum sum-F1-score was chosen and the results were reported accordingly. In order to ascertain that the systems are reproducible we repeated each experiment 10 times with different random initialisations.

We compare our results to existing works that have followed the same protocol, namely, (a) AVEC 2016 challenge support vector machine (SVM) based baseline system [4] using features extracted with COVAREP tool [48], (b) CNN-based system that learns to detect depression given mel filter bank energies and spectrogram as input [23] and (c) a system that models eGeMAPS features with SVM [20].

Table 2: Performance of various methods on AVEC 2016 dev set. Bold font marks the best system among the proposed methods in terms of the corresponding metric.

Experiment	F1 score		Precision		Recall	
	Depressed	Control	Depressed	Control	Depressed	Control
AVEC baseline [4] (COVAREP-SVM)	0.46	0.68	0.32	0.94	0.86	0.54
Ma et al. [23] (LMFB/Spec-CNN)	0.52	0.70	0.35	1.0	1.0	0.54
Vlasenko et al. [20] (eGeMAPS-SVM)	0.55	0.79	-	-	-	-
Raw speech with CNN-subseg	0.26 ± 0.01	0.79 ± 0.01	0.6 ± 0.09	0.69 ± 0.01	0.17 ± 0	**0.94 ± 0.03**
Raw speech with CNN-seg	0.57 ± 0.01	0.57 ± 0.01	0.43 ± 0.01	0.82 ± 0.02	**0.82 ± 0.03**	0.43 ± 0.01
LP residual with CNN-subseg	0.34 ± 0.01	0.78 ± 0.01	0.56 ± 0.05	0.7 ± 0.01	0.25 ± 0	0.89 ± 0.02
LP residual with CNN-seg	0.53 ± 0.02	0.57 ± 0.03	0.41 ± 0.02	0.77 ± 0.03	0.73 ± 0.03	0.45 ± 0.03
ZFF signal with CNN-subseg	**0.65 ± 0.02**	0.73 ± 0.02	0.54 ± 0.02	**0.87 ± 0.02**	0.81 ± 0.03	0.63 ± 0.03
ZFF signal with CNN-seg	0.52 ± 0.06	**0.8 ± 0.01**	**0.61 ± 0.03**	0.75 ± 0.02	0.45 ± 0.07	0.85 ± 0.02

Table 3: Effect of system complexity on performance.

Experiment	F1 score		Precision		Recall	
	Depressed	Control	Depressed	Control	Depressed	Control
Raw speech with CNN8-subseg	0.4 ± 0.04	0.73 ± 0.03	0.45 ± 0.05	0.7 ± 0.02	0.36 ± 0.04	0.77 ± 0.05
Raw speech with CNN8-seg†	0.56 ± 0	0.53 ± 0	0.42 ± 0	0.82 ± 0	0.83 ± 0	0.39 ± 0
ZFF signal with CNN2-subseg	0.59 ± 0.03	0.7 ± 0.03	0.5 ± 0.03	0.81 ± 0.02	0.72 ± 0.04	0.62 ± 0.04

†Training converged 4 out of 10 runs.

4 Results and analysis

Table 2 shows F1 scores, precision and recall of the proposed methods, shown as *mean* ± *standard deviation* of the 10 trials along with results of baseline systems. We can observe that ZFF signal consistently yields a better system in terms of F1 score than raw speech signal and 16^{th} order LP residual signal. When comparing sub-segmental modelling and segmental modelling, it is not obvious which one is better. For instance, sub-segmental modelling yields better F1 score for the depressed class, whilst segmental modelling yields better F1 score for the control class. A closer inspection reveals that this is due to low recall for the control class. When comparing the baseline systems with ZFF based system, we can observe that ZFF based system gains on the depressed class detection while competitive on control class detection. When modelling raw speech signal and LP residual signal, segmental modelling gives a good trade-off for both the classes, whilst sub-segmental modelling is able to detect well the control class than the depressed class. We also observe that LP residual is not providing any real advantage over modelling raw speech signal.

4.1 Impact of system complexity

As discussed earlier, speech signal contains multitude of information. One may need a more complex network to focus on glottal source information. In contrast, ZFF signal could be modelled with low complexity. In order to test this point, we trained:

1. Eight convolution layer CNN8-subseg and CNN8-seg system with raw speech signal as input.

2. Two convolution layer CNN2-subseg system with ZFF signal as input.

The details of the architecture can be found in Table 1.

Table 3 presents the results. It can be observed that CNN8-subseg improves slightly the depressed class detection in terms of F1 score. We can observe that CNN8-seg with raw speech signal as input does not always converge. Modelling ZFF signal with CNN2-subseg leads to slight drop in performance, but still is competitive to the state-of-the-art methods. These results show that ZFF signal modelling is easier than raw speech signal for depression detection.

4.2 Analysis of frequency response of first layer filters

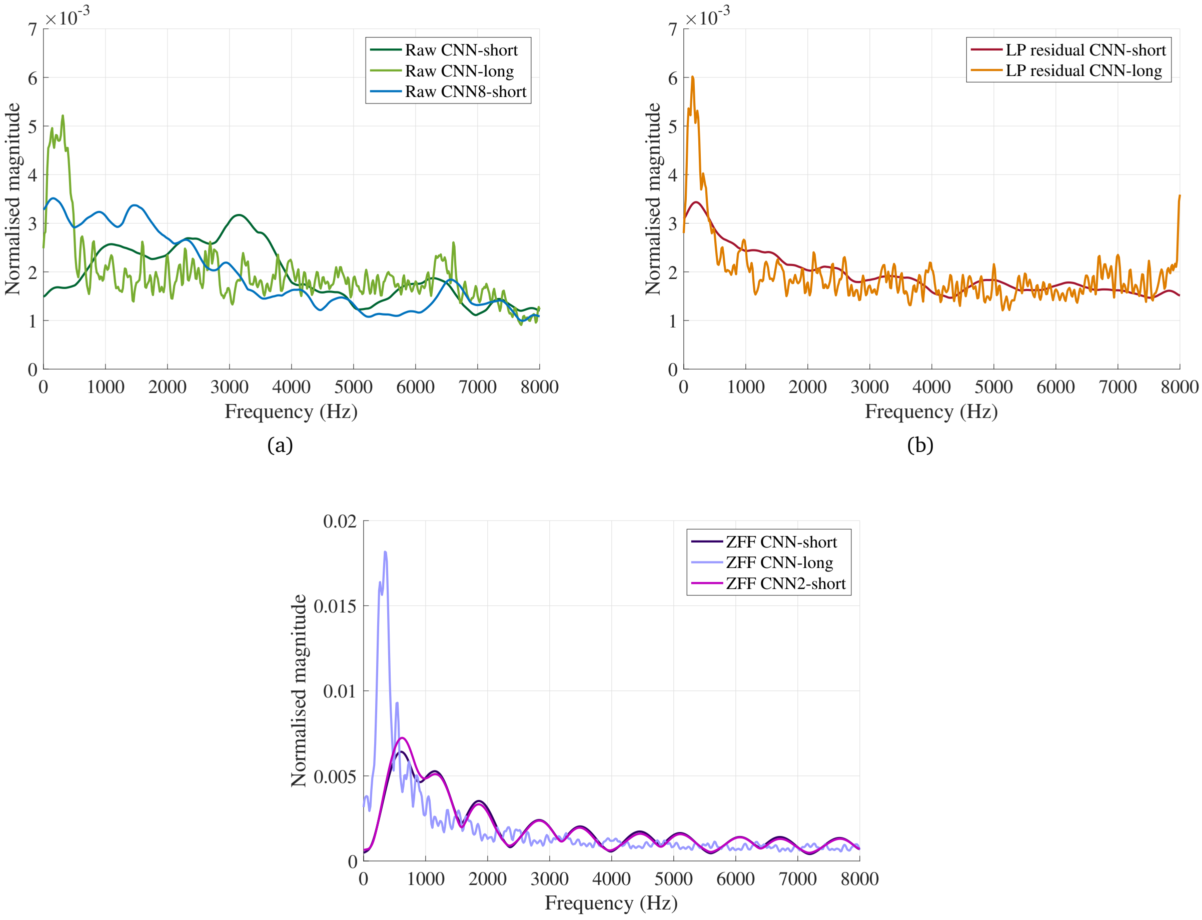
We visualised the cumulative frequency response of the filters in the first convolutional layer, which is computed as [31, 30]

$$F_{cum} = \sum_{k=1}^{N_f} F_k / \|F_k\|_2, \tag{1}$$

where N_f is the number of filters and F_k is the frequency response of filter f_k. Fig. 3 (a) shows the cumulative response in the case of modelling raw speech signal for CNN-subseg, CNN-seg, CNN8-subseg. Fig. 3(b) shows the cumulative frequency response in the case of modelling LP residual for CNN-subseg and CNN-seg. Figure 3(c) shows the cumulative frequency response in the case of modelling ZFF signal for CNN-subseg, CNN-seg and CNN2-subseg.

When modelling segmental speech, i.e. CNN-seg, in all the cases emphasis is given to very low frequency region that is mostly related to fundamental frequency. The main difference lies with ZFF signal based system where the filters by nature do not model high frequency regions, while with raw speech signal based system and LP residual signal based system the filters model high frequency regions, potentially related to vocal tract system. This could explain the difference between the CNN-seg systems reported in Table 2.

When modelling sub-segmental speech, i.e. CNN-subseg, in the case of modelling raw speech signal the filters give emphasis to high frequency regions (1000-4000 Hz). In the case of CNN8-subseg the emphasis shifts to low frequency regions (0-2000 Hz) while de-emphasising high frequency regions. This trend possibly explains slight gain in the depressed class detection with CNN8-subseg (see Table 3). In the case of residual signal modelling, the emphasis is given to low frequency region and a few high frequency regions. In the case



(a)

(b)

(c)

Figure 3: Comparison of the overall frequency responses of the first convolutional layers in various CNNs.

of ZFF signal modelling, the filters give emphasis to region below 1000 Hz. Change of network complexity, specifically CNN2-subseg, does not change much the frequency response of the filters. It can be observed again that ZFF signal based system by nature does not model any high frequency components.

5 Discussion and Conclusion

This paper investigated different ways to model glottal source information directly from raw waveform using CNNs. Our investigations on AVEC 2016 challenge data showed that filtering the speech signal with zero frequency resonator and then modelling the resulting ZFF signal is better than modelling raw speech signal or LP residual signal. Further, it was found that ZFF signal modelling can yield systems that are better than or comparable to state-of-the-art systems based on eGeMAPS features and deep neural network based systems modelling short-term spectral information. Our future work will focus along the following directions:

1. Analysis of the cumulative frequency response of filters learned in the first convolution layer suggests that it is beneficial to model low frequency regions (below 1000 Hz), while high frequency regions of speech signal could interfere with their modelling. A natural question that arises is: can modelling of raw speech signal after low pass filtering or modelling of LP residual after low pass filtering like done in SIFT algorithm [49] yield improved systems?

2. In our studies, we have observed that sub-segmental level modelling of ZFF signal helps in detecting the depressed class better, while segmental level modelling helps in detecting the control class better. In our future work, we aim to understand better the difference between these two modellings by gaining insight into what the neural network learns as a whole: more precisely, by utilising the recently proposed gradient-based visualisation approach to analysis raw waveform based neural network systems [50]. We believe that such an insight would not only help in understanding the specific glottal source information that is being modelled, but would also provide ways to exploit these two levels of modelling for improved performance.

Acknowledgements

This work was funded by Hasler foundation under the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS). The authors gratefully thank the Hasler Foundation for the financial support. We thank Dr. Vinayak Abrol for fruitful discussions.

References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10–49, 2015.

[2] B. N. Cuthbert and T. R. Insel, “Toward the future of psychiatric diagnosis: the seven pillars of RDoC,” *BMC Med.*, vol. 11, no. 1, pp. 1–8, 2013.

[3] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P Rosenwald, “Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses,” *Image Vis. Comput.*, vol. 32, no. 10, pp. 641–647, Dec 2013.

[4] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, Torres M., S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “AVEC 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proc. 6th Int. Workshop on AVEC*. ACM, 2016, pp. 3–10.

[5] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, “Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, Mar 2018.

[6] A. Jan, H. Meng, Y. F. A. Gaus, and F. Zhang, “Artificial Intelligent System for Automatic Depression Level Analysis through Visual and Vocal Expressions,” *IEEE Trans. Cognitive and Developmental Systems*, 2018 (to appear).

[7] M. Morales, S. Scherer, and R. Levitan, “A linguistically-informed fusion approach for multimodal depression detection,” in *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 13–24.

[8] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, “Decision tree based depression classification from audio video and language information,” in *Proc. 6th Int. Workshop on AVEC*. 2016, pp. 89–96, ACM.

[9] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, “Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 59–73, Jan 2016.

[10] R. D. Kent and Y. J. Kim, “Toward an acoustic typology of motor speech disorders,” *Clin. Linguist. Phon.*, vol. 17, no. 6, pp. 427–445, Jan 2003.

[11] S. Sahu and C. Y. Espy-Wilson, “Speech features for depression detection.,” in *Proc. Interspeech*, 2016, pp. 1928–1932.

[12] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, “Automatic modelling of depressed speech: Relevant features and relevance of gender,” in *Proc. Interspeech*, Singapore, 2014, pp. 1248–1252.

[13] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency, “Investigating voice quality as a speaker-independent indicator of depression and PTSD,” in *Proc. Interspeech*, Lyon, France, 2013, pp. 847–851.

[14] O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, and M. Cooke, “Glottal source features for automatic speech-based depression assessment,” in *Proc. Interspeech*, 2017, pp. 2700–2704.

[15] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B Yu, and D. D. Mehta, “Vocal biomarkers of depression based on motor incoordination,” in *Proc. 3rd Int. Workshop on AVEC*, Barcelona, Spain, 2013, pp. 41–48, ACM.

[16] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, “Multimodal and multiresolution depression detection from speech and facial landmark features,” in *Proc. 6th Int. Workshop on AVEC*. 2016, pp. 43–50, ACM.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Trans. Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[18] B. Stasak, J. Epps, N. Cummins, and R. Goecke, “An investigation of emotional speech in depression classification,” in *Proc. Interspeech*, 2016, pp. 485–489.

[19] R. Gupta, S. Sahu, C. Espy-Wilson, and S. Narayanano, “An affect prediction approach through depression severity parameter incorporation in neural networks,” in *Proc. Interspeech*, 2017, pp. 3122–3126.

[20] B. Vlasenko, H. Sagha, N. Cummins, and B. Schuller, “Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition,” in *Proc. Interspeech*, 2017, pp. 3266–3270.

[21] L. He and C. Cao, “Automated depression analysis using convolutional neural networks from speech,” *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.

[22] L. Yang, D. Jiang, W. Han, and H. Sahli, “DCNN and DNN based multi-modal depression recognition,” in *Proc. ACII*, Oct 2017, pp. 484–489.

[23] X. Ma, H. Yang, Q. Chen, D. Huang, and Yunhong Wang, "DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification," in *Proc. 6th Int. Workshop on AVEC*. 2016, pp. 35–42, ACM.

[24] T. N. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proc. Interspeech*, 2012, pp. 1059–1062.

[25] D. Palaz, R. Collobert, and Mathew Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proc. Interspeech*, 2013, pp. 1766–1770.

[26] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Proc. Interspeech*, 2014, pp. 890–894.

[27] Rubén Zazo, Tara N. Sainath, Gabor Simko, and Carolina Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *Proc. Interspeech*, 2016, pp. 3668–3672.

[28] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, 2016, pp. 5200–5204.

[29] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection with raw waveform CLDNNs," in *Proc. ICASSP*, 2017, pp. 4860–4864.

[30] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proc. ICASSP*. IEEE, 2018, pp. 4884–4888.

[31] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition," Idiap Research Report Idiap-RR-18-2016, Idiap, Jun 2016, `http://publications.idiap.ch/downloads/reports/2016/Palaz_Idiap-RR-18-2016.pdf`.

[32] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "On learning vocal tract system related speaker discriminative information from raw signal using CNNs," in *Proc. Interspeech*, 2018 (to appear).

[33] S. H. Kabil, H. Muckenhirn, and M. Magimai.-Doss, "On learning to identify genders from raw speech signal using CNNs," in *Proc. Interspeech*, 2018 (to appear).

[34] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "End-to-End Convolutional Neural Network-based Voice Presentation Attack Detection," in *Proc. IJCB*. IEEE, 2017, pp. 335–341.

[35] B. Vlasenko, J. Sebastian, D. S. Pavan Kumar, and Mathew Magimai.-Doss, "Implementing fusion techniques for the classification of paralinguistic information," in *Proc. Interspeech*, 2018 (to appear).

[36] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[37] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoustic Speech Signal Processing*, vol. 27, no. 4, pp. 309–319, Aug 1979.

[38] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana, "Glottal source processing: from analysis to applications," *Computer Speech and Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[39] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[40] B Yegnanarayana and Suryakanth V Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, no. 5, pp. 651–697, 2011.

[41] K. Ramesh, S. R. Mahadeva Prasanna, and D. Govind, "Detection of glottal opening instants using hilbert envelope," in *Proc. Interspeech*, 2013, pp. 44–48.

[42] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. Interspeech*, 2015.

[43] J. Gratch, R. Artstein, G. M Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, and S. Rizzo, "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*. ELRA, 2014, pp. 3123–3128.

[44] D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency, "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Proc. SigDial*. 2013, pp. 193–202, ACL.

[45] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disord.*, vol. 114, no. 1-3, pp. 163 – 173, 2009.

[46] François Chollet et al., "Keras," `https://github.com/fchollet/keras`, 2015.

[47] Martín Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," `http://tensorflow.org/`, 2015.

[48] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "COVAREP—a collaborative voice analysis repository for speech technologies," in *Proc. ICASSP*. IEEE, 2014, pp. 960–964.

[49] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio and Electroacoustics*, vol. 20, pp. 367 – 377, Jan 1973.

[50] H. Muckenhirn, V. Abrol, M. Magimai.-Doss, and S. Marcel, "Gradient-based spectral visualization of CNNs using raw waveforms," submitted to IEEE Workshop on Spoken Language Technology (SLT), `http://publications.idiap.ch/downloads/reports/2018/Muckenhirn_Idiap-RR-11-2018.pdf`, 2018.