LEARNING VOICE SOURCE RELATED INFORMATION FOR DEPRESSION DETECTION

S. Pavankumar Dubagunta[1,2], *Bogdan Vlasenko*[3], *Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland
[3]VIMA, Martigny, Switzerland

ABSTRACT

During depression neurophysiological changes can occur, which may affect laryngeal control i.e. behaviour of the vocal folds. Characterising these changes in a precise manner from speech signals is a non trivial task, as this typically involves reliable separation of the voice source information from them. In this paper, by exploiting the abilities of CNNs to learn task-relevant information from the input raw signals, we investigate several methods to model voice source related information for depression detection. Specifically, we investigate modelling of low pass filtered speech signals, linear prediction residual signals, homomorphically filtered voice source signals and zero frequency filtered signals to learn voice source related information for depression detection. Our investigations show that *subsegmental* level modelling of linear prediction residual signals or zero frequency filtered signals leads to systems better than the state-of-the-art low level descriptor based systems and deep learning based systems modelling the vocal tract system information.

Index Terms— Convolutional neural networks, depression detection, zero-frequency filtering, glottal source signals.

1. INTRODUCTION

Humans convey their mental state through vocal, linguistic and facial gestures. Depression is one such phenomenon, whose automatic detection and severity assessment have gained interest in the recent years [1, 2]. These tasks have been carried out in the literature by measuring parameters from patient interview sessions using multiple modes: audio, video and text, and by using appropriate classification/regression tasks [3, 4]. Purely speech based analyses continue to perform worse than multi-modal techniques [2], indicating the need for further research in the field.

Various speech features have been shown to be indicative of depression. Depression is known to affect human speech production and cognitive processes: it impacts speech motor control [1, 5], and can be identified by prosodic abnormalities and articulatory and phonetic errors [6]. Voice quality has been shown to be affected [7, 8, 9, 10, 11] in terms of features such as the shape of the glottal pulse, degree of breathiness, jitter and shimmer. Since depression can sometimes be associated with negative emotions, there have been features motivated from speech emotion recognition research such as [12, 13]. However expressing negative emotions is very different from having a depressed mental condition. Multiple works have used statistics of features, called low level descriptors (LLD), that are related to both the vocal-source and vocal-tract to improve the systems [12, 4, 14]; however not all the statistical properties contribute to the improvements. Despite these advances, there seem to be no concurred set of features for detecting depression from speech signals; and moreover, the performances of all these systems may be limited by the choice of features and their statistics. More recently, deep learning methods have been investigated. For instance, Ma et al. proposed predicting depression using neural networks comprising convolutional and long-short term memory layers on log Mel filter-bank (LMFB) and magnitude-spectrogram features [15].

Recent studies have hypothesised that during depression neurophysiological changes can occur, which in turn may affect the laryngeal control and its dynamics, i.e. the behaviour of the vocal folds [16, 17, 18, 19, 1]. Following these studies, various voice source related features such as jitter, shimmer and glottal flow characterisation have been proposed as speech-based biomarkers for depression detection [18, 19]. As discussed in [1], extracting and modelling these features for depression detection is a non-trivial task for reasons such as, (a) lack of a standardised approach to extract these features, (b) susceptibility to errors due to differing sound pressure levels between and within individuals, (c) difficulty in analysing and extracting these features from continuous speech in a reliable manner. In this paper, rather than extracting voice source related features from speech signals and then modelling them through a classifier, we develop methods to directly learn voice source related information in an end-to-end manner for depression detection. This is motivated from recent works that have shown that CNNs can learn task dependent information from raw speech signals in an end-to-end manner [20, 21, 22, 23]. Specifically, we show that, by combining prior knowledge based signal processing and the capability of CNNs to learn task dependent information from raw signals, depression can be effectively detected from the voice source information.

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

2. PROPOSED METHODS

We adopt the CNN-based framework described in [20], which was initially developed for speech recognition and later extended to other tasks [21, 24]. As illustrated in Fig. 1, the proposed system takes as input a fixed length signal (determined through cross validation) and processes it through multiple convolutional layers followed by fully connected layers and outputs the probability of detecting depression. The parameters of the system are optimised using cross entropy cri-

This work was funded by the Hasler foundation under the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS) and was entirely carried out at Idiap research institute. The authors gratefully thank the Hasler Foundation for the financial support. We thank Dr. Vinayak Abrol for fruitful discussions. e-mail: (see http://www.idiap.ch/en/people/directory).

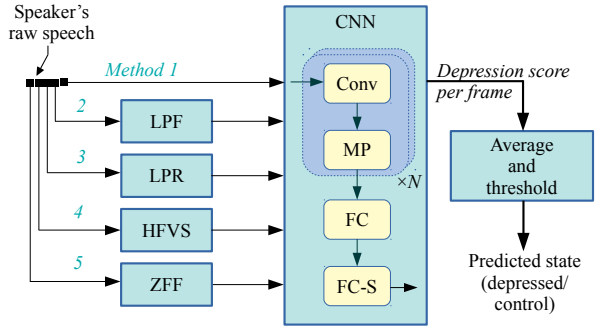


Fig. 1. The proposed methods. CNN architecture: Conv: convolutional layer with ReLU activations, MP: max-pooling layer, FC: fully connected layer with ReLU activations, FC-S: fully connected layer with a single output node and sigmoid activation.

terion. During testing, the scores obtained on multiple signals of each speaker are averaged to get a per-speaker score, which is later thresholded to get a binary classification (control/depressed). Depending upon the length of the filters in the first convolution layer, we distinguish two approaches, namely, (a) *subsegmental modelling* (subseg), where the filter spans about 2 ms ($<$ 1 pitch period) and yields a clear time resolution and (b) *segmental modelling* (seg), where filter spans about 20 ms ($1 - 5$ pitch periods) and gives a better frequency resolution.

We investigated the following signal inputs to detect depression based on voice source related information:

1) *Original raw speech* signal, Method 1 in Figure 1. Original speech signals contain information about the vocal source and the vocal tract system. Nevertheless, the motivation for this comes from a recent speaker recognition study [21], where it was found that, when the filters in the first convolutional layer process about 20 ms speech (1-3 pitch periods), they learn to model the fundamental frequency information and low frequency information that could be related to the voice quality. However, we also study the subsegmental approach, since information related to glottal pulses is present locally in time and may require time resolution.
2) *Low pass filtered* (LPF) speech signal, Method 2 in Figure 1. One way to enable the CNNs to effectively learn voice source related information is to low pass filter the input signals. This has indeed been observed in a recent study on CNN-based glottal closure instant detection [25].
3) *Linear prediction residual* (LPR) signal, Method 3 in Figure 1. LPR signals carry glottal source information, and thus LP analysis forms one of the methods for glottal signal analysis [26, 27]. LPR signals contain not only the excitation information but also the modelling errors of the vocal tract system due to the assumptions on the LP order [28]. One way to handle this issue is through low pass filtering the speech signals before extracting the residual. This is akin to simple inverse filter tracking method [29], which was proposed for fundamental frequency estimation. In our studies, the LPR signals are estimated from LPF signals.
4) *Homomorphically filtered voice source* (HFVS) signal, Method 4 in Figure 1. Complex cepstrum is a well-studied domain that allows transforming convolutive components of a time-domain signal into additive components. Here we employ a simple high-pass cepstral lifter to approximately remove the time varying vocal tract component from a speech signal and to retain the excitation source component [30]. Since the complex cepstrum transform is invertible, a corresponding time domain signal can be constructed from the liftered cepstrum. We perform this analysis using a sliding window on each raw speech utterance, and overlap-add the resultant excitation source signals to get the corresponding HFVS signal of each utterance.
5) *Zero frequency filtered* (ZFF) signal, Method 5 in Figure 1. Zero frequency filtering is a technique that has recently emerged for characterising the glottal source activity [31, 32]. It exploits the property of an impulse-like excitation at the glottal closure instance to detect glottal closure instants (GCIs). ZFF signals are obtained by passing speech signals through a cascade of two ideal digital resonators located at 0Hz, and then removing the trend in the resulting signals by subtracting the average over a window of the size in the range of 1 to 2 pitch periods. In addition to the GCIs, the strengths of the glottal excitations, the fundamental frequency and the glottal opening instants can be estimated from the ZFF signals [31, 33].

3. EXPERIMENTAL SETUP

3.1. Data set

We used Speech from the Distress analysis interview corpus - wizard of Oz (DAIC-WOZ) database [34] to validate the methods. The data set comprises of audio-visual interviews of 189 participants who underwent evaluation of psychological distress. Each participant was assigned a self-assessed depression score through patient health questionnaire (PHQ-8) method [35]. We used the time labels provided in the data set to extract only the participants' 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 [2].

3.2. Setup

The proposed signals to be investigated were extracted using multiple tools. For LPF, Kaiser windowed sinc filters of the SoX tool were used. LPR signals were extracted through 8-order LP modelling using COVAREP tool [36] with the default parameters, from the LPF signals. HFVS signals were extracted with a 40ms Hanning window, shifted by 20ms, using the standard complex cepstrum tools of MATLAB, and using a 50 sample quefrency cut-off.

The systems were trained using Keras deep learning library [37] with Tensorflow backend [38]. For each experiment, the training data were split into 95% of speakers for training and 5% of speakers for cross-validation. To ensure equal representation of both the

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

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

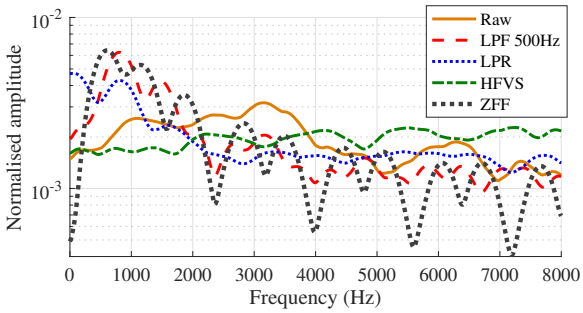
Table 2. Performances of various methods on the AVEC 2016 dev set. *D* indicates *depressed*, *C* indicates *control* and *O* indicates the *overall* score by un-weighted average over the two classes. Bold font marks the best system among the proposed methods in terms of the overall F1 score.

Experiment	*F1 score*			*Precision*		*Recall*	
	O	*D*	*C*	*D*	*C*	*D*	*C*
Feat - SVM [2]	0.57	0.46	0.68	0.32	0.94	0.86	0.54
LLD - LSTM [4]	-	0.50	-	0.71	-	0.38	-
Spec - CNN [15]	0.61	0.52	0.70	0.35	1.00	1.00	0.54
MFCC - DNN	0.52	0.42	0.61	0.37	0.68	0.49	0.56
Raw speech - subseg	0.53	0.26	0.79	0.60	0.69	0.17	0.94
Raw speech - seg	0.57	0.57	0.57	0.43	0.82	0.82	0.43
LPF 500Hz - subseg	0.57	0.56	0.59	0.43	0.81	0.79	0.46
LPF 500Hz - seg	0.65	0.61	0.69	0.50	0.84	0.79	0.59
LPR - subseg	0.65	0.60	0.70	0.50	0.82	0.74	0.61
LPR - seg	0.61	0.50	0.72	0.48	0.75	0.54	0.70
HFVS signal - subseg	0.61	0.52	0.70	0.47	0.75	0.58	0.65
HFVS signal - seg	0.61	0.54	0.68	0.46	0.77	0.64	0.61
ZFF signal - subseg	**0.69**	0.65	0.73	0.54	0.87	0.81	0.63
ZFF signal - seg	0.66	0.52	0.80	0.61	0.75	0.45	0.85

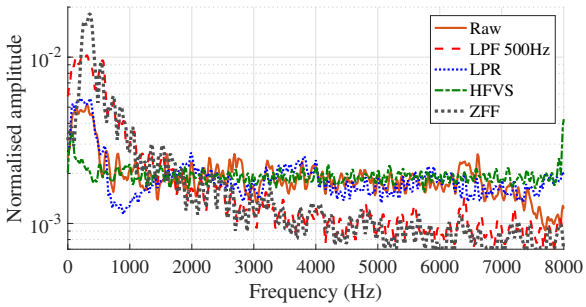
control and the depressed classes during training, we duplicated the depressed class utterances to a count matching as that of the control group. For the proposed methods, Table 1 lists the CNN architectures used. The term *subseg* refers to *sub-segmental* modelling, where the filters in the first convolution layer model 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). The number of convolutional layers in the CNNs is 4. "FC" layers in all the architectures contain 10 nodes. The input to the CNNs is a 250 ms signal, overlapped with a 10 ms shift. All the frames of the depressed group were labelled 1, and the rest 0.

The networks 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. We trained 10 networks for each experiment, starting with a different random initialisation, in order to ascertain the systems are reproducible. We evaluated them primarily by the average *F1 score* of both the classes computed from all the 10 networks trained. We additionally report precision and recall scores. To fix a threshold on the speaker-level scores for the binary classification, F1 scores were computed by varying the threshold in steps of 0.01. The threshold that gave the best unweighted average F1 score across all the 10 systems was then chosen, and the results were reported accordingly.

We compare our results with a few existing works that followed the same protocol, viz., (a) support vector machine (SVM) based baseline system from the AVEC 2016 challenge [2] that used features related to both the vocal tract and source, extracted using COVAREP tool [36], (b) long short term memory (LSTM) recurrent network based system that additionally used LLDs computed from the above features, and (c) CNN-based systems that detected depression from either spectrogram features or mel filter bank energies [15]. In addition, we trained a 3-hidden layer deep neural network (DNN) baseline system that models mel frequency cepstral coefficients (MFCC) to emulate a vocal tract system information based system.



(a) subseg modelling



(b) seg modelling

Fig. 2. Comparison of the overall frequency responses of the first convolutional layers in various CNNs.

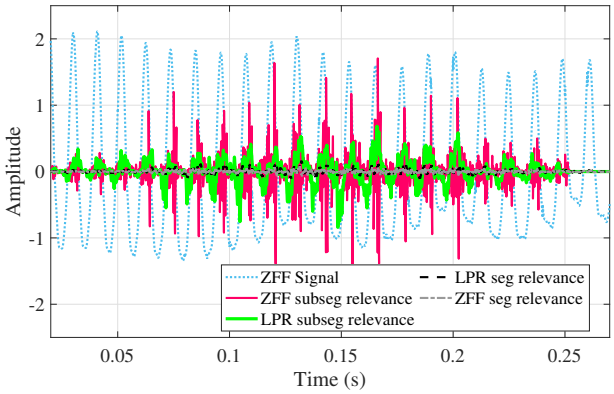
4. RESULTS AND ANALYSIS

Table 2 shows the F1 scores, precision and recall of the proposed methods along with the results of the baseline systems. It is worth mentioning that in the AVEC 2016 challenge the systems were ranked based on the F1 scores of both the classes. Except for the results from the existing works, each value shown indicates the *mean* performance obtained by training the DNN or CNN 10 times. We did this to ensure that the proposed methods are not sensitive to initialisation of DNN or CNN and the results are truly reproducible. The standard deviation of the performance of the systems were between 0.0 and 0.1. It can be observed that except raw speech modelling, the proposed methods of detecting depression based on voice source related information perform comparable to or better than the existing works. In particular, ZFF signals consistently yield better systems in terms of the overall F1 score than all the other methods. If we compare the systems based on F1 score for depression *D*, the proposed methods perform comparable or outperform existing methods, except in the case of subsegmental modelling of original raw speech signals.

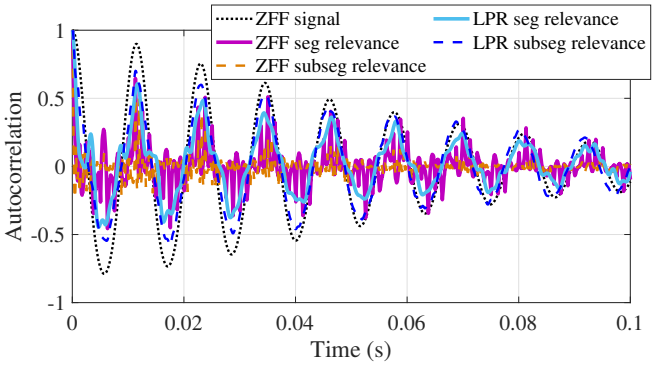
4.1. Analysis of frequency response of the first layer filters

To better understand the spectral information being modelled by the CNNs, we analysed the cumulative frequency response of the first convolutional layer filters, as done in [39, 21]:

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



(a) Relevance signals.



(b) Autocorrelation computed from relevance signals.

Fig. 3. Illustration of relevance signals and their autocorrelation signals. The example shown is part of the sustained vowel *uh*.

where N_f is the number of filters and F_k is the frequency response of filter f_k. Fig. 2a shows the cumulative responses of the CNNs modelling the proposed methods at the subsegmental level modelling (filters of length about 2 ms). As expected, for ZFF, LPF and LPR the emphasis is on low frequencies. For HFVS the response is almost flat across the frequencies. For raw speech signals, the emphasis is more on the high frequencies between 2 kHz - 4kHz, which is more related to the vocal tract system information.

Fig. 2b compares the cumulative frequency responses of the CNN filters with segmental level modelling for the proposed methods. It can be seen for all the signals, including raw speech, that the emphasis lies in the low frequency regions. It is interesting to observe that, except for the HFVS case, the low frequency region being emphasised is similar.

4.2. Relevance analysis

To gain insight about what the CNNs as a whole are learning, we applied a recently developed *guided backpropagation* based visualisation method [40]. In simple terms, given an input signal and the output class, the technique measures how a small variation or perturbation of each sample value will impact the prediction score. This corresponds to measuring the importance of each input sample value for the prediction. This process yields a relevance signal. Using this method, we contrasted the CNNs trained on ZFF signals with those on trained on LPR signals.

Fig. 3a shows the relevance signals computed for the subsegmental and segmental level modelling on both the types of signals, overlaid on the input ZFF signal, of a sustained vowel /uh/ of duration 250 ms from the database. In the case of subsegmental modelling, we observe that for both ZFF and LPR relevance signals there is a sharp focus at the positive-to-negative zero-crossings of the ZFF signals, which corresponds to the glottal closure instants (GCIs) [31]. This suggests that the subsegmental CNN is focusing on the GCI information for depression detection. In the case of segmental modelling, the relevance signal does not have such a sharp focus, indicating that all the samples are given importance. Fig. 3b shows the autocorrelation of the above signals. It can be observed that all the relevance signals are preserving the periodicity, i.e. F0, information.

Together these analyses reveal that the segmental level modelling of ZFF and LPR signals is mainly focusing on the F0 variation, whilst the subsegmental level modelling is focusing on time local events related to the voice source, viz. GCIs, similar to jitter and shimmer feature extraction as well as the F0 variation. This could be the reason why subsegmental level modelling of ZFF and LPR signals yields better system than segmental level modelling. Understanding these aspects further along with the analysis of LPF and HFVS CNNs is part of our future work.

5. CONCLUSION

This paper investigated methods to model voice source related information using CNN-based raw signal modelling techniques for depression detection. Our studies on the AVEC 2016 challenge data showed that, instead of modelling raw speech signals as they are, filtering them based on prior knowledge, such as low pass filtering to filter out the high frequency vocal tract system related information or ZFF leads to effective depression detection. More precisely, the systems based on ZFF signals and LPR signals yield better than the state-of-the-art LLD based systems and vocal tract system feature (LMFB, MFCC) based systems.

6. REFERENCES

[1] N. Cummins et al., "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[2] M. Valstar et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop on AVEC*. ACM, 2016, pp. 3–10.

[3] 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.

[4] Tuka Al Hanai, Mohammad Ghassemi, and James Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720.

[5] 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.

[6] 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.

[7] Amber Afshan et al., "Effectiveness of voice quality features in detecting depression," in *Proc. Interspeech*, 2018, pp. 1676–1680.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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.

[16] Christina Sobin and Harold Sackeim, "Psychomotor symptoms of depression," *American Journal of Psychiatry*, vol. 154, pp. 4–17, 1997.

[17] M. P. Caligiuri and J. Ellwanger, "Motor and cognitive aspects of motor retardation in depression," *Journal of Affective Disorders*, vol. 57, no. 1–3, pp. 83–93, 2000.

[18] Asli Ozdas, Richard G. Shiavi, Stephen E. Silverman, Marilyn K. Silverman, and D. Mitchell Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Trans. Biomed. Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] Shuai Yang, Zhiyong Wu, Binbin Shen, and Helen Meng, "Detection of glottal closure instants from speech signals: A convolutional neural network based method," in *Proc. Interspeech*, 2018, pp. 317–321.

[26] 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.

[27] 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.

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

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

[30] Thomas Drugman, Barış Bozkurt, and Thierry Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proc. Interspeech*, 2009, pp. 116–119.

[31] 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.

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

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

[34] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*. ELRA, 2014, pp. 3123–3128.

[35] K. Kroenke et al., "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.

[36] 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.

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

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

[39] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition," Tech. Rep. Idiap-RR-18-2016, Idiap research institute, Jun 2016.

[40] Hannah Muckenhirn, Vinayak Abrol, Mathew Magimai.-Doss, and Sébastien Marcel, "Gradient-based spectral visualization of CNNs using raw waveforms," Tech. Rep. Idiap-RR-11-2018, Idiap research institute, Jul 2018.