Learning Voice Source Related Information for Depression Detection

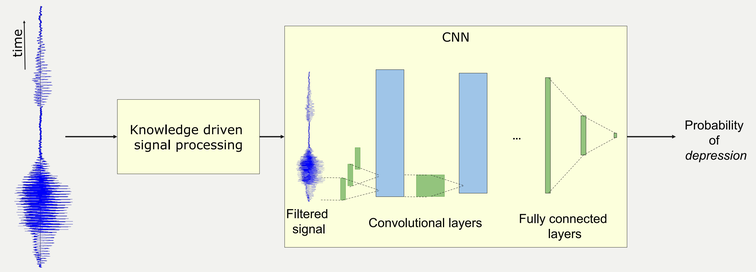
S. Pavankumar Dubagunta[1,2], Bogdan Vlasenko[3] and Mathew Magimai Doss[1]

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

1 Introduction

- Depression can cause neurophysiological changes, thereby may affect laryngeal control i.e. behaviour of the vocal folds.
- Characterising these changes from speech signals is non-trivial, since this involves reliable separation of the voice source information from them.
- Conventional hand-crafted feature sets used in speech based detection are related to:
  - **voice quality**: pitch frequency, jitter, shimmer, degree of breathiness,
  - **vocal tract shape**: formant locations, cepstrum based features,
  - **statistical properties** of features: low-level descriptors.
- Classifiers used: Support vector machines, neural networks (CNN, LSTM, etc.)
- Here we investigate:
  - knowledge-driven signal processing to extract voice-source related signals.
  - automatic feature learning from these raw signals using CNNs for depression detection.

2 Proposed approach



Knowledge-driven signal processing for extracting voice-source related information

- *Raw speech signals*: contain all the information.
- *Low pass filtered (LPF) speech*: contains F_0 and the first few formants.
- *Signals based on source-filter model*:
  - homomorphically filtered voice source signals (HFVS)
  - linear prediction residual (LPR)

Source → Filter → Speech

- *Zero-frequency filtered (ZFF) signals*: characterise glottal activity in terms of its excitation strengths and F_0.

Modelling approaches of the first convolutional layer

- *Sub-segmental*: 30 samples (<1 pitch period), for better time resolution.
- *Segmental*: 300 samples (2-3 pitch periods), for better frequency resolution.

3 Data set and experimental setup

- Speech data from the Depression analysis interview corpus - wizard of Oz (DAIC-WOZ), of the AVEC 2016 challenge, was used.
- Tools used were Keras/Tensorflow.
- Model architecture of the convolutional neural networks (CNNs):
  - input of length 250ms, where the signals were sampled every 10ms,
  - 4 convolutional layers with rectified linear (ReLU) activations and max-pooling,
  - 1 fully connected hidden layer with ReLU activations, and
  - a single output node with a sigmoid to predict the probability of depression.
- Training involved:
  - performing stochastic gradient descent with cross-entropy loss, and
  - repeating this on 10 models with different random initialisations.
- Testing involved:
  - predicting the depression probability from the individual segments,
  - estimating a speaker-level probability by averaging from all the 10 trained models, and
  - thresholding the probability for depression/control prediction.

4 Comparison of performances

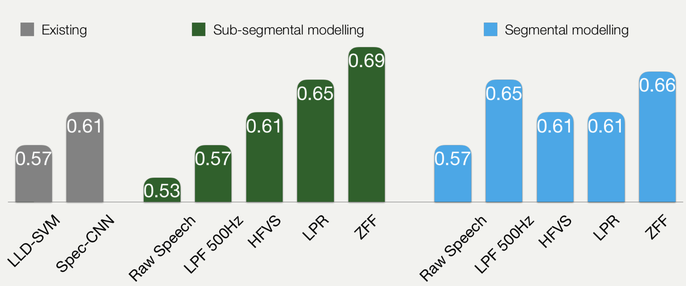
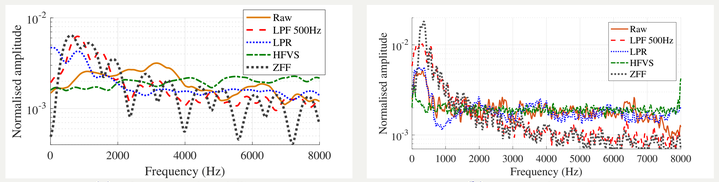


Figure: Unweighted average F_1 scores of the investigated approaches.

- *LLD-SVM*: Valstar et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in Proc. 6th Int. Workshop on AVEC, 2016, pp. 3–10, ACM.
- *Spec-CNN*: Ma et al., "DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification," in Proc. 6th Int. Workshop on AVEC, 2016, pp. 35–42, ACM.

5 Analysis: spectral characteristics of the first convolutional layer

- The overall magnitude spectrum of the first convolutional layer reveals the immediate focus of the CNN towards the detection task.



(a) sub-segmental modelling (b) Segmental modelling

Figure: Frequency response of various systems modelled at sub-segmental and segmental levels.

- Most of our systems, shown above, emphasise lower frequency regions that contain voice-source related information.
- The filters modelling at segmental level show better frequency resolution, as expected.

6 Relevance analysis

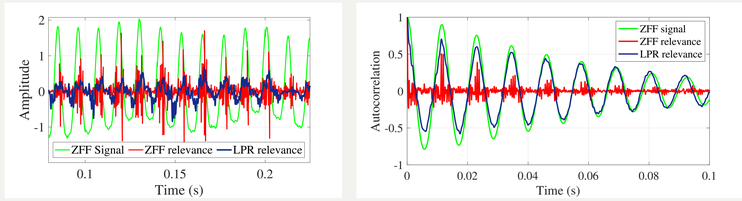


Figure: Relevance analysis on ZFF and LPR based sub-segmental models using a sustained vowel "uh".

- A relevance signal is computed by backpropagating the gradients of the CNN's output activation, for a given input signal.
- In this case, it indicates the input samples that are most informative for predicting depression.
- Our analysis using guided backpropagation showed that
  - the sub-segmental ZFF and LPR based CNNs focus on the ZFF signal's positive to negative zero-crossings, and
  - these correspond to the glottal closure instants.
- Autocorrelation of the relevance signals indicates that they retain F_0 information.

7 Conclusion

- We investigated directly modelling voice source related signals using CNNs for speech based depression detection.
- Our studies showed that filtering speech based on prior knowledge leads to effective depression detection than using raw speech or hand-crafted feature based systems.
- Furthermore, lower frequency regions, including F_0, and glottal closure instants were found to carry the most depression-related information.

Acknowledgements

This work was funded by the Hasler foundation under the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS) and was entirely carried out at the Idiap Research Institute.

Get in touch? pavankumar.dubagunta@idiap.ch https://linkedin.com/in/pavankumards

For full PDF, visit http://publications.idiap.ch or scan the QR code above.