IDIAP RESEARCH REPORT

ESTIMATING THE DEGREE OF SLEEPINESS BY INTEGRATING ARTICULATORY FEATURE KNOWLEDGE IN RAW WAVEFORM BASED CNNS

Julian Fritsch S. Pavankumar Dubagunta
Mathew Magimai.-Doss

Idiap-RR-06-2020

FEBRUARY 2020

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

Learning to Estimate the Degree of Sleepiness from Raw Waveforms using CNNs

Anonymous Author(s)

ABSTRACT

Speech-based degree of sleepiness estimation is an emerging research problem. In the literature, this problem has been mainly addressed through modeling of low level of descriptors. This paper investigates an end-to-end approach, where given raw waveform as input, a neural network estimates at its output the degree of sleepiness. Through an investigation on the continuous sleepiness subchallenge of the INTERSPEECH 2019 Computational Paralinguistics Challenge, we show that the proposed approach consistently yields performance comparable or better than low level descriptor-based, bag-of-audio-words-based and sequence-to-sequence autoencoder feature representation-based regression systems. Furthermore, a confusion matrix analysis on the development set shows that, unlike the best baseline system, the performance of our approach is not centering around a few degrees of sleepiness, but is spread across all the degrees of sleepiness.

KEYWORDS

Paralinguistic speech processing, Sleepiness, End-to-end acoustic modeling, Convolutional neural networks

ACM Reference Format:

Anonymous Author(s). 2019. Learning to Estimate the Degree of Sleepiness from Raw Waveforms using CNNs. 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Assessing sleepiness is relevant in scenarios, such as in preventing accidents or in evaluating when to recommend a break. Furthermore, sleep deprivation increases the mortality risk. To put this relevance into perspective: in 2016, the American think tank RAND reported an estimated 138 billion US $ damage to Japanese economy (2.92% of its GDP) caused by sleepiness at work, which is why companies, among other things, offer incentives to sleep more than six hours per night [6]. Although sleep is a multi-modal phenomenon, speech is one of the cheapest modalities that can be captured, most notably in a non-intrusive manner. Sleepiness can be subjectively assessed on the Karolinska Sleepiness Scale (KSS) [21], which ranges from 1 (extremely alert) to 9 (very sleepy). This paper focuses on developing objective or automatic methods to predict sleepiness.

In [19], Schuller et al. review contributions to the Interspeech 2011 Speaker State Challenge in two sub-challenges on alcohol intoxication and sleepiness. These medium term speaker states, meaning effects that usually last a few hours, are expected to generally affect motor coordination processes and speech of cognitive processing. This manifests in terms of changes in prosody such as monotonic and flattened intonation, shifted speech rate [10, 23], in articulation, such as slurred, less crisp pronunciation, mispronunciations [2] and in speech quality such as tensed, nasal, or breathy speech [9]. In addition, there could be linguistic dimensions, such as word repetition. The baseline system for the challenge extracted low-level descriptors (LLDs) such as short-term energy, short-term spectrum, voice related features and their functionals and classified them using support vector machine to predict sleepiness. In [7], Hönig et al. analyzed the LLDs extracted from the Interspeech 2011 Speaker State Challenge sleepiness data, which is labeled in terms of KSS. They found different trends for male and female. More precisely, male sleepiness correlated more with spectral changes such as less canonical pronunciation, whilst female sleepiness correlated more with lowered pitch.

In the recent years, with the advances in deep learning, approaches have emerged where task-related information are directly learned from raw speech signals using convolutional neural networks (CNNs) in an end-to-end manner, i.e. without any short-term spectral processing [5, 11, 12, 14, 18, 22, 24]. This paper investigates the use of such an approach for the degree of sleepiness estimation. We investigate different methods to incorporate prior knowledge: (a) constraining how the first convolution layer processes the speech signal, (b) filtering the speech signal to mainly model voice source related information and (c) integrating speech production knowledge through transfer learning. We validate these methods against the state-of-the art LLD-based, bag-of-audio-words-based and sequence-to-sequence autoencoder feature representation-based approaches on the continuous sleepiness subchallenge of the INTERSPEECH 2019 Computational Paralinguistics Challenge [20].

The remainder of the paper is organized as follows: Section 2 gives an overview of the used data and the baseline systems. Section 3 presents the proposed methods. Section 4 summarizes our results and presents an analysis of the trained neural networks. Finally, Section 5 concludes the paper.

2 EXPERIMENTAL PROTOCOL AND BASELINE SYSTEMS

The continuous sleepiness sub-challenge corpus consists of 5564 utterances (5hours 59 minutes) in the training set, 5328 utterances (5hours 44 minutes) in the development set and 5570 utterances (5hours 58minutes) in the test set from a total of 915 subjects (364 females, 551 males). No speaker IDs or speaker genders information are provided. Speech data consists of different reading and speaking tasks as well as narrative speech. According to the KSS scale, the

, ,

https://doi.org/10.1145/1122445.1122456

labels range from 1 to 9. True labels were averaged between self-assessment and two expert ratings. Spearman's cross-correlation coefficient, denoted as ρ, is used as the evaluation metric.

The challenge provided three support vector regression systems based different features, namely, the ComParE 2013 feature set, which contains 6373 LLDs and functionals, histogram representations of clustered LLDs, known as bag-of-audio-words (BoAW) and feature representations from sequence-to-sequence auto-encoders (S2SAE) trained on Mel-spectrograms. For each of these feature sets, systems with different configurations were built. The best baseline results on the development and the test set are shown in Table 2. For further details, the reader is referred to [20].

3 PROPOSED SYSTEMS

We used the raw-speech based CNN framework originally developed for speech recognition [15], and later extended to other tasks such as speaker verification [11], gender identification [8], presentation attack detection [12] or depression detection [5]. In this framework, as illustrated in Figure 1, the network consists of N convolution layers (Conv), maximum pooling (MaxP) and ReLU activations followed by a multilayer perceptron (MLP). At the output, the CNN predicts the posterior probability of the classes per frame. Frame-level posterior probabilities are then averaged to get per-utterance posterior probabilities.

Figure 2 illustrates the processing at the first convolution layer. kW denotes the kernel width in samples, dW denotes the stride or kernel shift in samples, W_{seq} in seconds is the segment of speech that is processed at one time frame and n_f is number of filters in the convolution layer. In [11, 16], it has been found that by modifying kW different information related to the speech production mechanism can be learned. More precisely, if kW covers a signal length of about 20 ms (segmental), the first convolution layer tends to model voice source related information. Similarly, If kW covers a lengthsignal of about 2 ms of length (sub-segmental), the first convolution layer tends to model vocal tract system related information, such as formant information.

During training, both convolutional layer and MLP parameters are estimated using a cost function based on cross entropy. During testing, the frame level posterior probabilities are averaged and a decision is made.

3.1 Raw-speech CNN

We trained randomly initialized CNNs to predict the degree of sleepiness. The input to the CNN w_{seq} was 250ms length signal, which was shifted by 10ms. Table 1 presents the two architectures used based on the first convolution layer kernel width. Depending upon the length of the filters in the first convolutional layer, we distinguish (a) sub-segmental modelling (subseg), where $kW = 30$, span $2ms$, equivalent to less than 1 pitch period, and (b) segmental modelling (seg), where $kW = 300$ spanning $20ms$, equivalent to 1 to 5 pitch periods. The classification stage consists of one hidden layer with 100 units.

3.2 Zero-frequency filtered signals

As discussed in Section 1, sleepiness leads to changes in vocal source characteristics, such as changes in fundamental frequency.

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

Model	Layer	Conv			MP
		N_f	kW	dW	
subseg	1	128	30	10	2
	2	256	10	5	3
	3	512	4	2	-
	4	512	3	1	-
seg	1	128	300	100	2
	2	256	5	2	-
	3	512	4	2	-
	4	512	3	1	-

In order to incorporate that information, we took inspiration from a recent work that showed that voice source related information for depression detection can be modeled by filtering the speech signal through a zero frequency filter (ZFF) [13] and feeding the filtered processed signal as input to CNNs [5].

3.3 Integrating speech production knowledge

As discussed in Section 1, sleepiness can induce changes in the articulation process, i.e. in the speech production process resulting in slurry speech or less crispy pronunciation or mispronunciation. In order to integrate articulatory information, we investigated a transfer learning framework where the CNN is first trained to predict broad phonetic features or articulatory features (AFs), namely, manner of articulation (degree of constriction), place of articulation (place of constriction), height of articulation (height of the tongue or roundedness) and vowel. These AFs are inspired from a recent work on articulatory feature based speech recognition [17]. To predict the degree of sleepiness, we use the AF-initialized CNNs, replace the output layer by an output layer consisting of the nine sleepiness categories and the CNN is trained again. Figure 3 summarizes this procedure.

In order to train the articulatory features, transcription of the speech signal is required. In the challenge data, no transcription has been provided so we used the AMI corpus [3], which consists of 77 hours of speech, from which we use a 70 hour clean subset of the training set, which is a standard practice in the Kaldi recipe. From this data, we used Kaldi to train HMMs for context-dependent phones, where the HMM states were jointly modelled by using subspace GMMs. The corresponding frame-to-phone alignments were then used to train four different AF-CNNs for each AF category (height, manner, place and vowel). The model architecture is the same as the sub-segmental architecture described in Section 3.1, except in this case, the single hidden layer MLP contains 1024 hidden units. We then adapted the resulting four AF-CNNs on the sleepiness data.

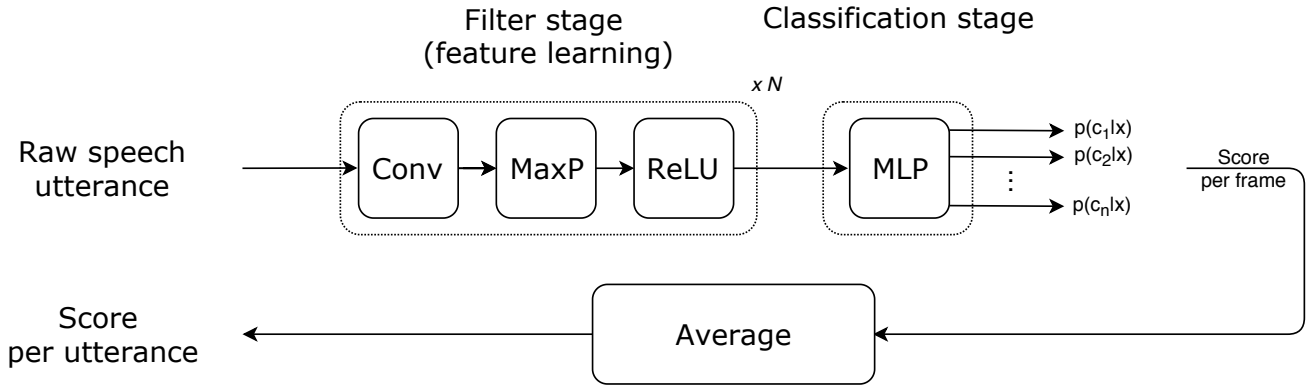

Figure 1: Illustration of our CNN architecture.

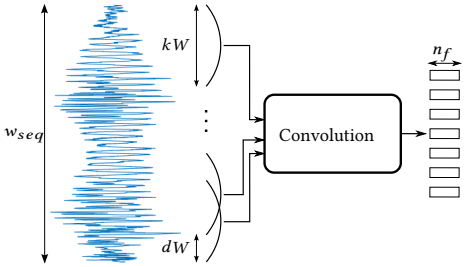

Figure 2: Illustration of first convolution layer processing.

3.4 Fusion of scores with an MLP

We also investigated combining different systems. For that we used an MLP to fuse scores from different systems. The MLP has one hidden layer with 128 nodes with ReLU activation, a dropout layer with 0.1% and the output layer predicts the nine sleepiness categories.

Similar to the baseline system studies reported in [20], we conducted studies with two experimental setups: (a) train the CNNs on the training data and test on development data and (b) train the CNNs on both training and development data and test on the test set. The networks were implemented with the Keras deep learning library [4] with TensorFlow as backend [1]. In each case, 5% of the data was used for validation. We used a decaying learning schedule which halves the learning rate between $10e-3$ and $10e-7$ whenever the validation loss stopped reducing.

4 RESULTS & ANALYSIS

Table 2 compares the performance of the proposed systems with the baseline systems. It is important mention that the challenge allows only five trials on the test set. So only five results for the proposed systems are reported.

On the first experimental setup i.e. training on the training set and evaluating on the development set, it can be observed the proposed raw waveform modeling methods perform comparable or better than the best baseline systems, except for sub-segmental modeling of ZFF signal. We can observe that score fusion leads to improvement in performance. Thus, indicating that different CNNs are capturing complementary information. When comparing on the second experimental setup, i.e. training on train and development set and evaluating on the test set, we can see that raw waveform CNNs not necessarily generalize. However, the AF-CNN

Table 2: Results of all the presented CNNs on the INTERSPEECH 2019 sleepiness sub-challenge data in Spearman's cross-correlation coefficient ρ. A + denotes a fusion using the MLP.

	Dev	Test
Baseline systems		
ComParE 2013	.251	.314
$COMPARE2013BoAW_{500}$	.250	.304
$S2SAE_{-70dB}$	.261	.310
Fusion (3-best Majority Vote)	-	.343
Raw waveform CNNs		
$subseg_{raw}$	.280	.201
seg_{raw}	.274	.222
$subseg_{zff}$	.244	-
seg_{zff}	.252	-
AF-CNNs		
height	.267	-
manner	.292	-
place	.262	-
vowel	.295	.312
Fusion		
$subseg_{raw} + seg_{raw} + seg_{zff}$	.303	-
manner + *place* + *vowel*	304	-
manner + *place*	.311	-
manner + *vowel*	.317	.325
manner + seg_{raw}	.315	-
manner + *vowel* + seg_{raw}	.319	-
manner + seg_{raw} + *ComParE*	.329	-
manner + seg_{raw} + $BoAW_{500}$	.344	.321

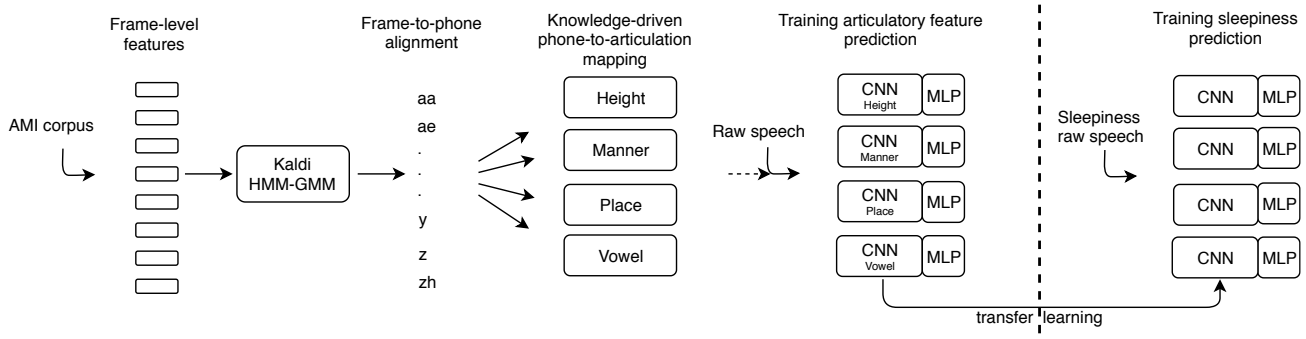

Figure 3: Overview of the procedure of transfer learning for sleepiness prediction from CNNs that where initially trained to predict articulatory features on the AMI corpus.

and fusion systems generalize well. This shows that integrating speech production knowledge is indeed aiding in predicting degree of sleepiness.

We performed a confusion matrix analysis of the results obtained in the first experimental setup. Figure 5 shows the confusion matrix of our system *manner + vowel*. Unlike the baseline system [20], it can be observed that classifications are spread over all degrees of sleepiness. We have highest accuracy for KSS rating of 3 and 8, meaning that our system is able to differentiate the extreme sleepiness categories well, whereas accuracy is lower for KSS ratings between 4 and 6, which are naturally difficult to distinguish. Figure 5 shows the confusion matrix of our system *manner* + seg_{raw} + $BOAW_{500}$. The confusion matrix has less values along the diagonal meaning less correct classifications. However the overall Spearman's correlation is higher.

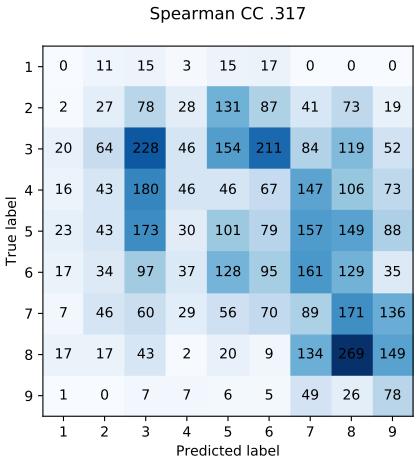

Figure 4: Confusion matrix of our fusion of the CNNs *manner* and *vowel*.

To get an impression of what frequency regions the first convolutional layer is focusing on, we computed the cumulative frequency

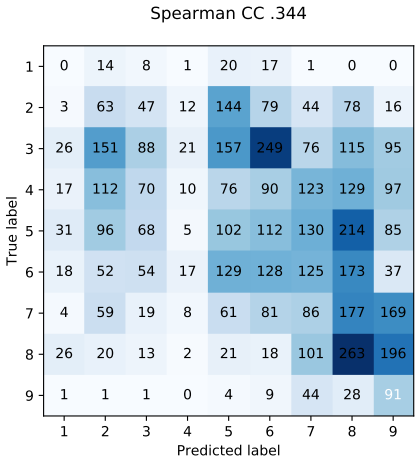

Figure 5: Confusion matrix of our best system: A fusion of the CNNs *manner* and seg_{raw} and the SVR output from $BOAW_{500}$.

response (CFR) as follows:

$$F_{cum} = \sum_{k=1}^{N_f} F_k / ||F_k||_2 \quad (1)$$

N_f denotes the number of filters and F_k is the frequency response of filter f_k. Figure 6 compares the CFR for raw waveform based systems. When modeling raw sub-segmental speech, stronger emphasis is given to frequencies around 1000 Hz, whilst the segmental system's CFR shows a few distinctive peaks well below 1000 Hz that could be related to fundamental frequency and formants. In the case of ZFF signals, the sub-segmental system has emphasis around 500 Hz, whereas the segmental system emphasizes very low frequencies.

Figure 7 shows the CFR for AF-CNNs. It can be observed that there are differences in the information modeled by the CNNs for different AFs. Furthermore, when compared to raw waveform CNNs (6), the CFRs are very different, i.e. emphasis is given to frequencies

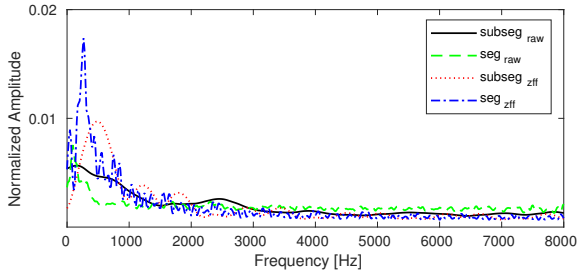

Figure 6: Cumulative frequency responses of first convolutional layer raw waveform CNNs.

above 1000 Hz. These differences explain the performance gains obtained in the fusion systems.

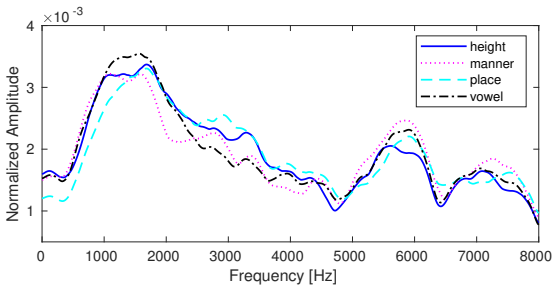

Figure 7: Cumulative frequency responses of first convolutional layer from AF-CNNs.

5 CONCLUSIONS

This paper investigated different ways to estimate the degree of sleepiness from raw waveform by implementing a CNN to model relevant information directly from the speech signal. We evaluated our methods on the continuous sleepiness sub-challenge of the INTERSPEECH 2019 Computational Paralinguistics data. Our investigations showed that integrating speech production knowledge by modeling articulatory features yields better systems, when compared to simply modeling raw waveforms. The raw waveform CNNs and the AF-CNNs focus on different frequency information, which could be exploited through score fusion. Among the AF-CNNs, the manner CNN and vowel CNN yields the best systems. The score fusion studies and the first convolution layer analysis shows that the AF-CNNs capture complementary information. Finally, our experimental studies show that the proposed end-to-end approach can yield systems comparable to or better than the conventional short-term speech processing based approaches.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 265–283.

[2] Daniel Bratzke, Bettina Rolke, Rolf Ulrich, and Maren Peters. 2007. Central slowing during the night. *Psychological Science* 18, 5 (2007), 456–461.

[3] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 28–39.

[4] Francois Chollet. 2018. *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG.

[5] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss. 2019. Learning Voice Source Related Information for Depression Detection. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6525–6529.

[6] Marco Hafner, Martin Stepanek, Jirka Taylor, Wendy M. Troxel, and Christian Van Stolk. 2016. Why sleep matters âĂŤ the economic costs of insufficient sleep: A cross-country comparative analysis. www.rand.org/pubs/research_reports/RR1791.html.

[7] Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. 2014. Acoustic-prosodic characteristics of sleepy speech–between performance and interpretation. In *Proc. of Speech Prosody*. Citeseer, 864–868.

[8] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai-Doss. 2018. On Learning to Identify Genders from Raw Speech Signal Using CNNs.. In *Interspeech*. 287–291.

[9] Barbara E Kostyk and Anne Putnam Rochet. 1998. Laryngeal airway resistance in teachers with vocal fatigue: A preliminary study. *Journal of Voice* 12, 3 (1998), 287–299.

[10] Jarek Krajewski, Martin Golz, Sebastian Schnieder, Thomas Schnupp, Christian Heinze, and David Sommer. 2010. Detecting fatigue from steering behaviour applying continuous wavelet transform. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*. ACM, 24.

[11] Hannah Muckenhirn, Mathew Magimai Doss, and Sébastien Marcell. 2018. Towards directly modeling raw speech signal for speaker verification using CNNs. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4884–4888.

[12] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcel. 2017. End-to-End Convolutional Neural Network-based Voice Presentation Attack Detection. In *International Joint Conference on Biometrics*.

[13] K Sri Rama Murty and B Yegnanarayana. 2008. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 8 (2008), 1602–1613.

[14] Dimitri Palaz, Ronan Collobert, and Mathew Magimai.-Doss. 2013. Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks. In *Proc. of Interspeech*.

[15] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss. 2013. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *Proc. Interspeech* (2013), 1766–1770.

[16] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert. 2019. End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition. *Speech Communication* 108 (April 2019), 15–32.

[17] Ramya Rasipuram and Mathew Magimai Doss. 2016. Articulatory feature based continuous speech recognition using probabilistic lexical modeling. *Computer Speech & Language* 36 (2016), 233–259.

[18] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals. 2015. Learning the speech front-end with raw waveform CLDNNs. In *Proc. of Interspeech*.

[19] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, Jarek Krajewski, Felix Weninger, and Florian Eyben. 2014. Medium-term speaker statesâĂŤA review on intoxication, sleepiness and the first challenge. *Computer Speech & Language* 28, 2 (2014), 346–374.

[20] Björn W. Schuller et al. 2019. The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. submitted to Interspeech. [Online; accessed 15th June 2019].

[21] Azmeh Shahid, Kate Wilkinson, Shai Marcu, and Colin M Shapiro. 2011. Karolinska sleepiness scale (KSS). In *STOP, THAT and One Hundred Other Sleep Scales*. Springer, 209–210.

[22] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. of ICASSP*.

[23] Adam P Vogel, Janet Fletcher, and Paul Maruff. 2010. Acoustic analysis of the effects of sustained wakefulness on speech. *The Journal of the Acoustical Society of America* 128, 6 (2010), 3747–3756.

[24] Rubén Zazo, Tara N. Sainath, Gabor Simko, and Carolina Parada. 2016. Feature learning with raw-waveform CLDNNs for voice activity detection. In *Proc. of Interspeech*.