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

Julian Fritsch[1,2]*, S. Pavankumar Dubagunta*[1,2]*, Mathew Magimai.-Doss*[1]

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

ABSTRACT

Speech-based degree of sleepiness estimation is an emerging research problem. This paper investigates an end-to-end approach, where given raw waveform as input, a convolutional neural network (CNN) estimates at its output the degree of sleepiness. Within this approach, we investigate constraining the first layer processing and integration of speech production knowledge through transfer learning. We evaluate these methods on the continuous sleepiness corpus of the Interspeech 2019 Computational Paralinguistics (ComParE) Challenge and demonstrate that the proposed approach consistently yields competitive systems. In particular, we observe that integration of speech production knowledge aids in improving the performance and yields systems that are complementary.

Index Terms— Paralinguistic speech processing, sleepiness, end-to-end acoustic modeling, convolutional neural networks, articulatory features.

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 US$138 billion 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 [1]. Although sleepiness 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) [2], which ranges from 1 (extremely alert) to 9 (very sleepy) in steps of one. This paper focuses on developing objective or automatic methods to predict sleepiness.

In the literature, estimating sleepiness has been addressed by investigating acoustic factors. Traditionally, baseline systems used a large number of general-purpose low-level descriptors (LLDs) such as short-term energy, short-term spectrum, voice-related features and their functionals, as in [3]. In [4], Schuller et al. reviewed contributions to the Interspeech 2011 Speaker State Challenge on sleepiness estimation, which is labeled in terms of KSS. Sleepiness is considered a medium term speaker state, meaning effects that usually last a few hours. It is expected to generally affect motor coordination processes and cognitive processing of speech. This manifests in terms of changes in prosody such as monotonic and flattened intonation, in shifted speech rate [5, 6], in articulation, such as slurred, less crisp pronunciation, mispronunciation [7] and in speech quality such as tensed, nasal, or breathy speech [8]. In [9], Hönig et al. analyzed the LLDs extracted from the Interspeech 2011 Speaker State Challenge sleepiness data. They found that male sleepiness correlated more with spectral changes such as less canonical pronunciation, whilst female sleepiness correlated more with lowered F_0.

More recently, as part of the Interspeech 2019 ComParE challenge, 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 were studied [3]. In [10], Gosztolya created utterance-level Fisher vectors by training a GMM on frame-level MFCCs, which are used for classification with SVM. Similarly, Wu et al. [11] investigated extraction of Fisher vectors from a large variety of acoustic features. In [12], the authors investigated raw waveform CNNs including data augmentation, such as inputting reverse samples, adding noise or using noisy labels. In [13], Yeh et al. presented a system that uses frame-level eGeMaps features that were input to a BLSTM-CNN network with attention. For data augmentation, an adversarial auto-encoder was used to generate synthetic samples. Additionally, border cases, e.g. samples with low and high KSS scores, that are intuitively more relevant to detect, were selected for an additional classifier to be used for score fusion. Wu et al. [14] aimed to address the ordinality of the KSS labels and introduce an ordinal triplet loss that is used to train binary classifiers for each label individually. Vijay et al. [15] used between-frame entropy, a measure that correlates with speech rate, to detect outliers, and creating utterance-level iVectors from voice quality features.

The above mentioned contributions investigated many relevant acoustic aspects and address issues such as the ordinality of the KSS labels or the imbalance of a data set to solve sleepiness estimation as a classification/regression problem. However, although relevant, acoustic-phonetic changes in sleepy speech have not yet been considered. Therefore, our goal is to study whether speech production differences from a phonetic perspective can be captured for degree of sleepiness estimation. We aim to address this by applying raw waveform based CNNs in an end-to-end manner. It was shown in recent years that this approach is able to learn task-related information without any short-term spectral processing [16, 17, 18, 19, 20, 21, 22]. We investigate constraining how the first convolution layer processes the speech signal (Section 2.1). Moreover, we incorporate prior knowledge by integrating speech production knowledge through transfer learning, inspired by Dubagunta et al. ([23]), described in Section 2.2. We validate these methods on the continuous sleepiness sub-challenge of the ComParE 2019 challenge [3] in Section 3 and present our conclusions in Section 4.

This work was partially funded through European Union's Horizon H2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network (MSCA-ITN-ETN) project TAPAS under grant agreement No. 766287 and through the HASLER Foundation under the project FLOSS.

2. PROPOSED SYSTEMS

We used the raw waveform based CNN framework originally developed for speech recognition [24], and later extended to other tasks such as speaker verification [21], gender identification [25], presentation attack detection [20] or depression detection [22]. 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. During training, both convolutional layer and MLP parameters are estimated using a cost function based on cross entropy. We used a decaying learning schedule which halves the learning rate between 10^{-3} and 10^{-7} whenever the validation loss stopped reducing. Similar to the baseline system studies reported in [3], 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. In each case, 5% of the data was used for cross-validation.

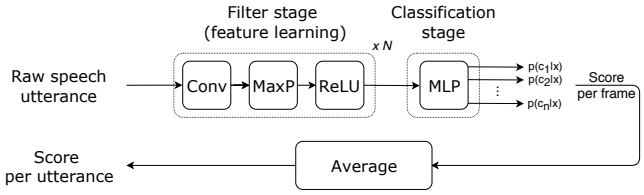

Fig. 1. Illustration of the proposed CNN architecture.

2.1. Raw waveform CNN architectures

We trained randomly-initialized CNNs to predict the degree of sleepiness. 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 [26, 21], 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 signal of about 2 ms of length (sub-segmental), the first convolution layer tends to model vocal tract system related information, such as formant information.

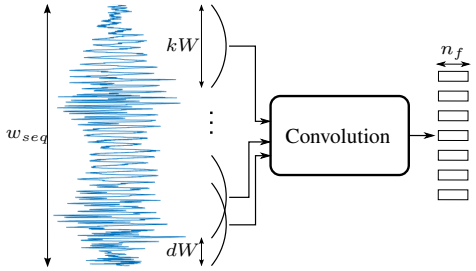

Fig. 2. Illustration of the first convolution layer processing.

Input to the CNN w_{seq} was a 250ms length speech segment (determined by the frame-level accuracy on the cross-validation set), which was shifted by 10ms. Table 1 presents the 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 over $2ms$, 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 AF-CNN architecture uses sub-segmental modeling, see Section 2.2. The classification stage consists of one hidden layer with 100 units.

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	-
AF-CNN	1	80	30	10	3
	2	60	7	1	3
	3	60	7	1	3
seg	1	128	300	100	2
	2	256	5	2	-
	3	512	4	2	-
	4	512	3	1	-

2.2. 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 slurred speech, less crisp or incorrect pronunciation. In order to integrate articulatory information into our models, we investigated a transfer learning framework where the CNN is first trained to predict articulatory features (AFs) within four broad categories, namely, manner of articulation (e.g. degree of constriction), place (of constriction), height (of the tongue) and vowel. These AFs are inspired by a recent work on articulatory feature based speech recognition [27]. 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 train those models. Figure 3 summarizes this procedure. Knowledge from the 4 AF categories is utilized to initialise 4 separate CNNs, which are fine-tuned on the sleepiness data. We hypothesize that such an initialization helps to exploit articulatory differences due to sleepiness.

AF predictors are trained based on knowledge that maps phones to AFs. With such mapping, one can train acoustic-to-AF predictors by using an alignment of transcribed speech. The challenge data is not transcribed, so we used the AMI corpus [28], which consists of 77 hours of speech. 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 and the phone-to-AF mappings were then used to train the above mentioned four AF-CNNs. The model architecture is similar to sub-segmental architecture and is described in Table 1 as AF-CNN, 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.

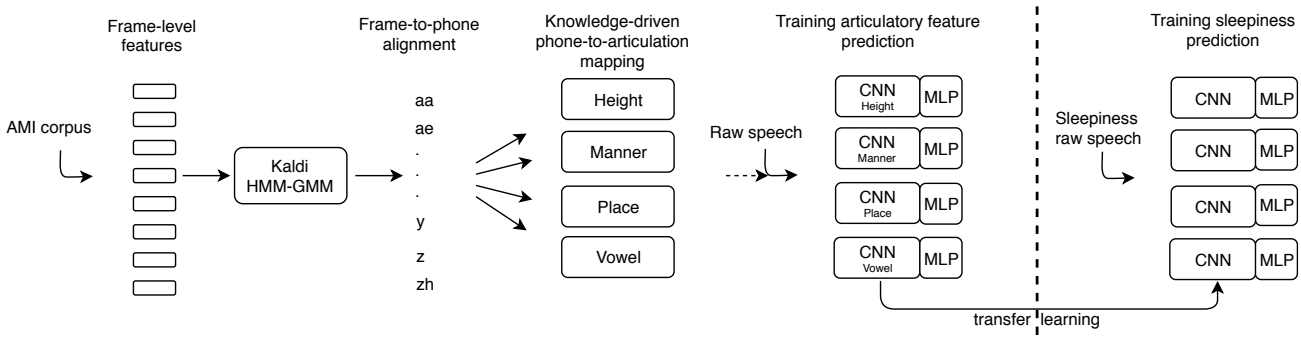

Fig. 3. Overview of transfer learning for sleepiness prediction from CNNs that were initially trained to predict articulatory features.

2.3. Posterior vector fusion with an MLP

We also investigated combining different systems. For that we used an MLP to fuse scores from different systems. The MLP had one hidden layer with 128 nodes with ReLU activation, a dropout layer with 10% and the output layer predicts the nine sleepiness categories.

3. EXPERIMENTAL RESULTS & ANALYSIS

3.1. Data and experimental protocol

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 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. For further details, the reader is referred to [3].

3.2. Results

Table 2 compares the performance of the proposed systems with the baseline systems provided as part of the challenge and systems reported as part of the challenge. It is important to mention that the challenge allowed only five trials on the test set, hence only five test 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 that the proposed raw waveform modeling methods perform comparable to the best baseline systems and systems reported as part of the ComParE challenge. 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 the raw waveform CNNs not necessarily generalize well. However, the AF-CNN and fusion systems generalize well. This shows that integrating speech production knowledge is indeed aiding in predicting degree of sleepiness and yields comparable systems.

Besides the proposed systems, Elsner et al. [12] and Wu et al. [14] investigated modeling raw waveform using CNNs for the sleepiness challenge. In [14], a system based on CNN-BLSTM yielded significantly poor results. In [12], it was found that a CNN-based system using a considerably longer window of speech input, more precisely 1.5 s speech, without data augmentation yielded a competitive system. In our case, the raw waveform based CNNs without modeling speech production knowledge model 250 ms of speech at the input. This difference could possibly explain low performance on the test set. However, when integrating speech production knowledge, although the CNN hyper parameters were chosen from previous speech recognition studies, we can observe that with 250 ms speech input we yield competitive systems. This suggests that raw waveform CNNs and AF-CNNs are modeling different information.

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

ComParE 2019 Baseline systems	**Dev**	**Test**
$ComParE_{2013}$ [3]	.251	.314
$COMPARE_{2013}BoAW_{500}$ [3]	.250	.304
$S2SAE_{-70dB}$ [3]	.261	.310
3-best Fusion [3]	-	.343
Competition systems		
Elsner et al. [12]	.290	.335
Yeh et al. [13]	.373	.369
Wu et al. [14]	.343	-
Ravi et al. [15]	.300	.331
Gosztolya [10]	.367	.383
Wu et al. [10]	.326	.365
Proposed raw waveform CNNs		
$subseg$	.280	.201
seg	.274	.222
Proposed AF-CNNs		
$height$	.267	-
$manner$	.292	-
$place$	.262	-
$vowel$	.295	.312
Proposed fusion		
$manner + place + vowel$	304	-
$manner + place$	.311	-
$manner + vowel$	.317	.325
$manner + seg$	.315	-
$manner + vowel + seg$	.319	-
$manner + seg + ComParE$	.329	-
$manner + seg + BoAW_{500}$	.344	.321

3.3. Analysis

We performed a confusion matrix analysis of the results obtained in the first experimental setup. Figure 4 shows the confusion matrix of our system $manner + vowel$. Unlike the baseline system [3], 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. Moreover, the highest number of predictions are reasonably spread along the diagonal. KSS label 1 is not correctly classified, presumably because of a lack of samples – at least 5 times less in both training and development set than KSS labels 2 to 8. In general, we found similar trends in other systems that we investigated.

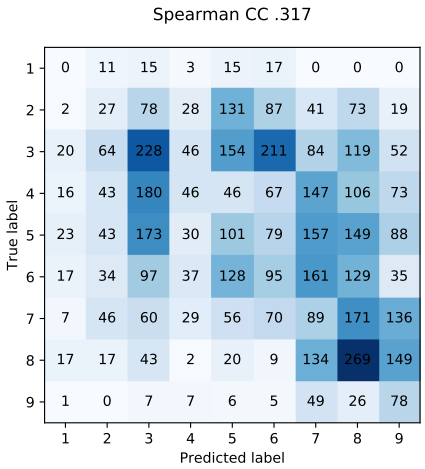

Fig. 4. Confusion matrix of the score fusion from the CNNs $manner$ and $vowel$.

To get an impression of what frequency regions the first convolutional layer is focusing on, we computed the cumulative frequency response (CFR) as follows [26]:

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

N_f denotes the number of filters and F_k is the frequency response of filter f_k. Figure 5 compares the CFR for raw waveform based systems. In both $subseg$ CNN and seg CNN frequency regions around 1000 Hz or below are given emphasis.

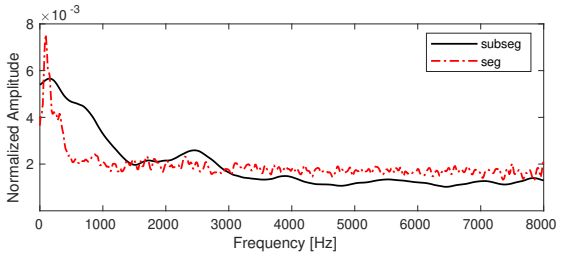

Fig. 5. Cumulative frequency responses of first convolutional layer from raw waveform CNNs.

Figure 6 shows the CFR for AF-CNNs after adaptation/training on sleepiness challenge data to estimate the degree of sleepiness. It can be observed that there are differences in the information modeled by the CNNs for different AFs. However, in general, the emphasis of frequency regions is similar to CNNs trained for speech recognition task [26]. Furthermore, when compared to raw waveform CNNs (Figure 5), the CFRs are very different, i.e. emphasis is given to frequencies above 1000 Hz that are associated with the articulation aspect of speech. This indicates that indeed the raw waveform CNNs and AF-CNNs are focusing on different information. In addition, it also explains the performance gains obtained when fusing these systems.

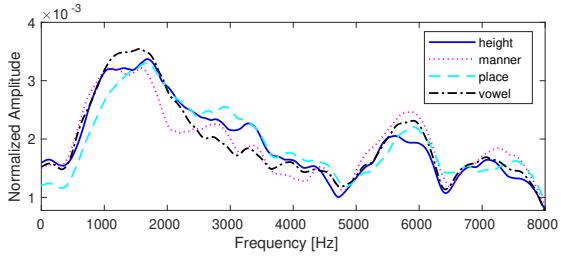

Fig. 6. Cumulative frequency responses of first convolutional layer from AF-CNNs.

4. CONCLUSIONS

This paper investigated how to estimate the degree of sleepiness from raw waveforms by integrating speech production knowledge into a CNN, by initially training it to predict articulatory features. We evaluated our methods on the ComParE 2019 continuous sleepiness challenge data. Our investigations showed that integrating this knowledge yields better systems, when compared to simply modeling raw waveforms. Among the AF-CNNs, the manner CNN and vowel CNN yield the best systems. First convolution layer analysis shows that raw waveform CNNs and AF-CNNs focus on different frequency information, hence capture complementary information. This could be exploited through score fusion. Finally, our experimental studies show that the proposed end-to-end approach can yield systems comparable to the conventional short-term speech processing based approaches.

From the performance point of view, the CNNs initialized to predict the $manner$ and $vowel$ categories seem to be more relevant to predict sleepiness in speech. Our future work will focus on understanding the speech sound-specific changes that are relevant for degree of sleepiness estimation.

5. REFERENCES

[1] Marco Hafner, Martin Stepanek, Jirka Taylor, Wendy M. Troxel, and Christian Van Stolk, "Why sleep matters — the economic costs of insufficient sleep: A cross-country comparative analysis," 2016, `www.rand.org/pubs/research_reports/RR1791.html`, Accessed: 2019-10-20.

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

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

[4] Björn Schuller et al., "Medium-term speaker states—a review on intoxication, sleepiness and the first challenge," *Computer Speech & Language*, vol. 28, no. 2, pp. 346–374, 2014.

[5] Jarek Krajewski et al., "Detecting fatigue from steering behaviour applying continuous wavelet transform," in *Proceedings of Measuring Behaviour*, 2010, pp. 326–329.

[6] Adam P. Vogel, Janet Fletcher, and Paul Maruff, "Acoustic analysis of the effects of sustained wakefulness on speech," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3747–3756, 2010.

[7] Daniel Bratzke, Bettina Rolke, Rolf Ulrich, and Maren Peters, "Central slowing during the night," *Psychological Science*, vol. 18, no. 5, pp. 456–461, 2007.

[8] Barbara E Kostyk and Anne Putnam Rochet, "Laryngeal airway resistance in teachers with vocal fatigue: A preliminary study," *Journal of Voice*, vol. 12, no. 3, pp. 287–299, 1998.

[9] Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski, "Acoustic-prosodic characteristics of sleepy speech–between performance and interpretation," in *Proceedings of Speech Prosody*, 2014, pp. 864–868.

[10] Gábor Gosztolya, "Using fisher vector and bag-of-audio-words representations to identify styrian dialects, sleepiness, baby & orca sounds," *Proceedings of Interspeech*, pp. 2413–2417, 2019.

[11] Haiwei Wu, Weiqing Wang, and Ming Li, "The dku-lenovo systems for the interspeech 2019 computational paralinguistic challenge," *Proceedings of Interspeech*, pp. 2433–2437, 2019.

[12] Daniel Elsner, Stefan Langer, Fabian Ritz, Robert Mueller, and Steffen Illium, "Deep neural baselines for computational paralinguistics," *Proceedings of Interspeech*, pp. 2388–2392, 2019.

[13] Sung-Lin Yeh et al., "Using attention networks and adversarial augmentation for styrian dialect continuous sleepiness and baby sound recognition," *Proceedings of Interspeech*, pp. 2398–2402, 2019.

[14] Peter Wu, SaiKrishna Rallabandi, Alan W Black, and Eric Nyberg, "Ordinal triplet loss: Investigating sleepiness detection from speech," *Proceedings of Interspeech*, pp. 2403–2407, 2019.

[15] Vijay Ravi, Soo Jin Park, Amber Afshan, and Abeer Alwan, "Voice quality and between-frame entropy for sleepiness estimation," *Proceedings of Interspeech*, pp. 2408–2412, 2019.

[16] Dimitri Palaz, Ronan Collobert, and Mathew Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proceedings Interspeech*, 2013.

[17] Tara N Sainath et al., "Learning the speech front-end with raw waveform cldnns," in *Proceedings of Interspeech*, 2015.

[18] George Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of ICASSP*, 2016.

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

[20] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcel, "End-to-end convolutional neural network-based voice presentation attack detection," in *International Joint Conference on Biometrics*, 2017.

[21] Hannah Muckenhirn, Mathew Magimai-Doss, and Sébastien Marcell, "Towards directly modeling raw speech signal for speaker verification using cnns," in *Proceedings of ICASSP*, 2018, pp. 4884–4888.

[22] S Pavankumar Dubagunta, Bogdan Vlasenko, and Mathew Magimai.-Doss, "Learning voice source related information for depression detection," in *Proceedings of ICASSP*, 2019, pp. 6525–6529.

[23] S. Pavankumar Dubagunta and Mathew Magimai-Doss, "Using speech production knowledge for raw waveform modelling based styrian dialect identification," in *Proceedings of Interspeech*, 2019, pp. 2383–2387.

[24] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *Proceedings of Interspeech*, pp. 1766–1770, 2013.

[25] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai-Doss, "On learning to identify genders from raw speech signal using cnns.," in *Proceedings of Interspeech*, 2018, pp. 287–291.

[26] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15–32, Apr. 2019.

[27] Ramya Rasipuram and Mathew Magimai-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Computer Speech & Language*, vol. 36, pp. 233–259, 2016.

[28] Jean Carletta et al., "The AMI meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.