IDIAP RESEARCH REPORT

TOWARDS AUTOMATIC PREDICTION OF NON-EXPERT PERCEIVED SPEECH FLUENCY RATINGS

S. Pavankumar Dubagunta Edoardo Moneta
Eleni Theocharopoulos Mathew Magimai.-Doss

Idiap-RR-11-2021

Version of AUGUST 06, 2021

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

Towards Automatic Prediction of Non-Expert Perceived Speech Fluency Ratings

S. Pavankumar Dubagunta[1,2], Edoardo Moneta[3], Eleni Theocharopoulos[3], and Mathew Magimai.-Doss[2]

[1]Idiap Research Institute, Martigny, Switzerland

[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland

[3]speak & lunch S.A., Switzerland

July 2021

Abstract

Automatic speech fluency prediction has been mainly approached from the perspective of computer aided language learning, where the system tends to predict ratings similar to those of the human *experts*. Speech fluency prediction, however, can be questioned in a more relaxed social setting, where the ratings arise mostly from *non-experts*. This paper explores the latter direction, i.e., prediction of non-expert perceived speech fluency ratings, which has not been studied in the speech technology literature, to the best of our knowledge. Toward that, we investigate different approaches, namely, (a) low-level descriptor feature functionals, (b) bag-of-audio word based approach and (c) neural network based end-to-end acoustic modelling approach. Our investigations on speech data collected from 54 speakers and rated by seven non-experts demonstrate that non-expert speech fluency ratings can be systematically predicted, with the best performing system yielding a Pearson's correlation coefficient of 0.66 and a Spearman's correlation coefficient of 0.67 with the median human scores.

Keywords Perceived fluency, speech assessment, low level descriptors, bag of audio words, raw waveform modelling, zero frequency filtering, articulatory features.

1 Introduction

Speech fluency, i.e. a smooth flow of speech, is an important aspect of spoken language communication. Technologically, speech fluency estimation has been approached in the context of computer aided spoken language learning and testing. Several existing methods predict fluency automatically in a reference-based setting by comparing the utterance under test to a predefined reference in terms of its linguistic

This work was partially funded by Hasler Foundation through the project Flexible Linguistically guided Objective Speech aSsessment (FLOSS) and by Innosuisse through the Innocheque project 28464.1 INNO-ES and through the project Conversation Member Match (CMM) (38843.1 IP-ICT). e-mail: {pavankumar.dubagunta,mathew}@idiap.ch, eleni@speakandlunch.com

content and estimating a score: for example, using speech recognition system to estimate number of correct words per minute [1, 2] and phoneme-level goodness of pronunciation [3]. Vocal source characteristics have also been studied, such as comparing the prosody contour to a reference [4]. Fontan et al. [5] used automatic segmentation techniques and formant tracking to compute similar features without the use of speech recognition. In a no-reference setting, where only the expert scores of perceived fluency are available, Mao et al. [6] studied directly predicting the mean opinion scores using standard machine learning techniques on *fluency feature vectors*, which constitute pause durations, pause similarity scores based on their positions and durations w.r.t. a predefined set of references, estimated syllable speaking rate and pronunciation quality values.

Besides the language learning and testing perspective, speech fluency prediction can be questioned in a more informal or social settings. For instance, in spoken communication, perceived speech fluency may have an impact on the interaction and/or on other aspects such as, forming impressions about the person. Speech fluency prediction in such a context has certain differences when compared to language learning and testing. First, in language learning and testing, speech fluency prediction is a part of a broader aspect, more precisely, *proficiency* assessment, which also includes *linguistic accuracy*, i.e. the correctness of syntax and vocabulary [7]. Second, the assessment system is developed to predict a score that best correlates with *expert* ratings. In the literature, it has been found that native experts and non-experts tend to rate differently [7]. In particular, non-expert raters tend not to focus much on the linguistic accuracy aspects.

This paper focuses on predicting perceived speech fluency from non-expert ratings. Toward that, we collect a speech data set consisting of read speech and speech on a topic of interest to the volunteers; rate the speech fluency with non-expert raters; and investigate whether such non-expert ratings can be predicted in a consistent manner. To the best of our knowledge, this question has not been addressed before. So, with this research question in mind, we investigate different approaches that do not explicitly model linguistic information: (a) predefined set of acoustic low level descriptor (LLD) features-based, (b) unsupervised speech embeddings-based, and (c) end-to-end acoustic modelling-based.

The remainder of the paper is organised as follows. Sections 2 and 3 present the data collection and the investigated approaches respectively. Section 4 presents the experimental setup and results. Section 5 gives an analysis of the different approaches. Section 6 finally concludes the paper.

2 Data Collection

The data were collected in three different countries: Switzerland, Greece and the USA (city of New York). We mainly went in medium sized international companies as well as social gatherings and asked for volunteers to participate in the project. The volunteers who agreed to participate were provided with an informed consent form to sign. Each of the participants were then provided with an iPod or an iPhone with headphones and were asked to make audio or video recordings, as per their preference, of all the languages they spoke (whether fluent, intermediate or beginner level). They were asked to (i) make 4 recordings of minimum 15 seconds where they would speak about a topic of their choice, and (ii) read a phonetically balanced text, viz. the Northwind passage. Out of the 54 participants, 29 were women and 25 were men. The participants' age ranged between 25 and 75 years. They were from different nationalities, viz., Albanian, French, Greek, Italian, Mexican, Portuguese, Russian, Spanish, Swiss and Turkish.

The final collected data set comprises 187.36 minutes of data from 54 speakers, of which 144.14 minutes corresponds to English recordings, which we used in our analysis. On average, each speaker

Table 1: LLD features. (See [8] for detailed explanations.)

Source-related	*System-related*
Loudness	Alpha ratio
F0 semitone from 27.5 Hz	Hammarberg index
Jitter	Spectral slopes (0-500, 500-1500)
Shimmer	Spectral flux
HNR (dB)	F1 (freq, bw, ampLogRelF0)
logRelF0-H1-H2	F2 (freq, ampLogRelF0)
logRelF0-H1-A3	F3 (freq, ampLogRelF0)
	MFCC (1-4)

had about 2-4 minutes of speech. These data were then rated by seven raters (4 women and 3 men), aged between 37 and 75 years old. The raters were fluent English speakers, who are active professionals in the law and banking sector in the USA and Switzerland. The raters were asked to rate each audio or video on a 5-point Likert scale, with 1 being beginner and 5 being fluent. The Krippendorff's alpha coefficient for the ratings was found to be 0.584. The median values per each speaker were used as reference scores in our experiments.

3 Approaches

In this section, we motivate and present several approaches we used for automatic fluency prediction.

3.1 Using functionals of LLD features

We investigate the use of a generic set of LLD features, viz. the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [9, 8], that comprise several short-time features that correspond to the vocal source and tract, as listed in Table 1. Such features are typically used in paralinguistic and other tasks [9, 10, 11, 12, 13]. It is worth mentioning that the source-system classification of the features is arguable, for instance, regarding the spectral slope related features. In the first approach, their statistical properties, also called functionals, are computed at the utterance-level and are used to train standard linear support vector machine (SVM) classifiers to classify each utterance into the five rating categories. This approach is denoted as *Functionals (LLD) + SVM*.

3.2 Using BoAW representations

In the second approach, we compute histogram representations known as BoAW for each utterance using frame-level features. Such representations give the relative counts of events in each utterance that are determined by clustering the features. We also included time in seconds as an additional feature, so that the events clustered depend on the time of occurrence of the events. We investigate the use of two feature sets for the BoAW approach, viz. (i) eGeMAPS and (ii) wav2vec 2.0 representations, obtained by passing raw speech through multiple convolutional and self-attention layers, that are learned to predict quantised representations among a set of distractors [14]. Both these sets of representations could carry position dependent counts of pauses, disfluencies and syllables, which were shown to indicate

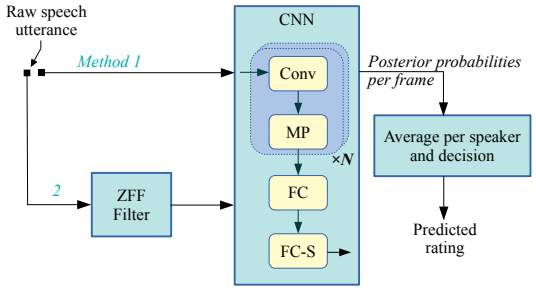


Figure 1: Proposed joint feature-classifier learning based on CNNs and using raw signal modelling. Conv: convolutional layer with rectified linear (ReLU) activation, MP: max-pooling, FC: fully connected layer with ReLU activation, FC-S: FC layer with softmax activation.

fluency [6]. These two approaches are denoted as *BoAW (LLD) + SVM* and *BoAW (wav2vec 2.0) + SVM*, respectively.

3.3 Using jointly learned feature-classifiers

In recent years, it was demonstrated that task-specific information can be automatically learned from raw waveforms using convolutional neural networks (CNN) [15, 16, 17, 18], as opposed to using hand-crafted features. Inspired from these works, we investigate how well this approach can predict speech fluency ratings. In this direction, following some of the recent works on speech recognition [15], speaker recognition [19] and paralinguistic speech processing [20, 21], as illustrated as Method 1 in Fig. 1, we train CNNs that take as input raw waveform and predict the probability for each fluency rating category, which are then averaged per speaker to make a decision about the fluency rating. Depending upon the length of the filters in the first convolution layer, two approaches can be distinguished, namely, (a) *subsegmental modelling* (subseg), where the filters span about 2 ms ($<$ 1 pitch period) and provide better time resolution and (b) *segmental modelling* (seg), where the filters span about 20 ms ($1-5$ pitch periods) and gives a better frequency resolution. This approach is denoted as *Raw SigProc*.

We also investigate methods where prior knowledge is integrated through signal processing or transfer learning. These two approaches are briefly presented below.

3.3.1 Voice source information-based approach

Voice source related information such as change in the fundamental frequency and energy over time could indicate speech fluency. However, separating the voice source related information from the vocal tract system related information and modelling it is not a trivial task. In recent works, it has been shown that glottal source activity can be characterised from speech signals through zero frequency filtering (ZFF) technique [22, 23] and can be modelled by CNNs for paralinguistic tasks such as sleepiness [21] prediction, dementia [24] and depression [20] detection. Briefly, ZFF involves passing raw speech through a cascade of two ideal digital resonators located at 0 Hz, and then removing the trend over a window spanning 1 to 2 pitch periods. As illustrated in Method 2 of Fig. 1, we take inspiration from these works to model zero frequency filtered signal for speech fluency rating prediction. This approach is denoted as *ZFF SigProc*.

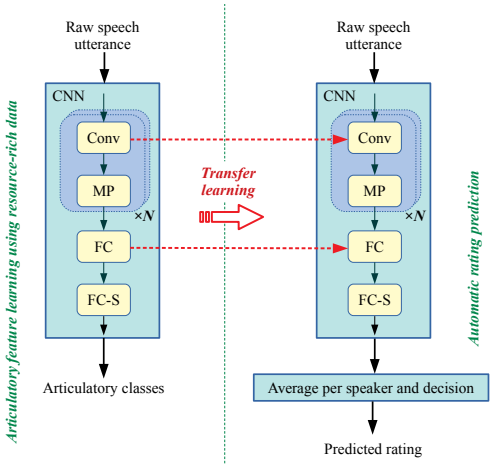


Figure 2: Transfer learning based on articulatory knowledge.

3.3.2 Implicit linguistic knowledge modelling-based approach

Linguistic accuracy has been studied in the literature in the prediction of fluency [1, 3, 2]. To investigate whether non-expert fluency rating prediction can benefit from such knowledge, we utilise transfer-learning to implicitly model articulatory feature information such as place and manner of articulation. In recent works, such implicit modelling of linguistic knowledge have been found beneficial for dialect identification [25] and degree of sleepiness prediction [21]. In this work, similar to [25, 21], as illustrated in Fig. 2, we initialise the speech fluency rating prediction network with the CNNs trained to predict articulatory features, by only modifying the output layer. There are four such pre-trained CNNs corresponding to manner of articulation, place of articulation, height of articulation and vowels, following a previous work on AF representations for speech recognition [26]. More information related to the pre-trained nets can be found in the Appendix A. This approach is denoted as *Artic*.

4 Experimental setup and results

We evaluated our systems based on 10-fold validation with non-overlapping speakers. Specifically, we split the speakers into 10 parts, where 9 parts are used for training and the 10^{th} part was used for testing. Models were trained using 10 such possible splits, and the performance was measured by computing Pearson's and Spearman's correlations between the predicted and the median human scores. eGeMAPS features and their functionals were extracted using OpenSMILE toolkit [27]. BoAW representations were extracted using OpenXBow toolkit [28]. Linear SVM classifiers were trained using scikit-learn [29] with the default parameters, without optimising the hyperparameters. For BoAW representations, the codebook size used was 50, as the data was limited and contained mostly read speech. We included the time information of the frame as an additional feature to the BoAW representations, as we found that this improves the performance. For the wav2vec 2.0 representations, we used the pre-trained *base* model provided by the authors [30], which was trained on LibriSpeech corpus [31].

CNNs for joint feature-classifier modelling were trained using Tensorflow [32, 33]. The terms *subseg* and *seg* refer to 30 sample *sub-segmental* and 300 sample *segmental* modelling respectively (see Table A1

Table 2: Results in terms of correlation coefficients, with p-values in parentheses.

		Pearson's	Spearman's
Functionals (LLD) + SVM		0.338 (6e-71)	0.356 (4e-79)
BoAW (LLD) + SVM		0.627 (7e-37)	0.641 (6e-39)
BoAW Source (LLD) + SVM		0.337 (4e-10)	0.347 (1e-10)
BoAW System (LLD) + SVM		**0.657** (2e-41)	**0.668** (2e-43)
BoAW (wav2vec 2.0) + SVM		0.556 (1e-27)	0.578 (3e-30)
Raw SigProc	subseg	0.431 (4e-16)	0.446 (3e-17)
	seg	0.569 (3e-29)	0.563 (1e-28)
ZFF SigProc	subseg	0.560 (3e-28)	0.576 (4e-30)
	seg	0.515 (2e-23)	0.545 (2e-26)
Artic	Manner	0.497 (1e-21)	0.527 (1e-24)
	Place	0.517 (1e-23)	0.528 (9e-25)
	Height	0.489 (6e-21)	0.499 (7e-22)
	Vowel	0.416 (5e-15)	0.437 (1e-16)
	Overall	0.493 (3e-21)	0.516 (2e-23)

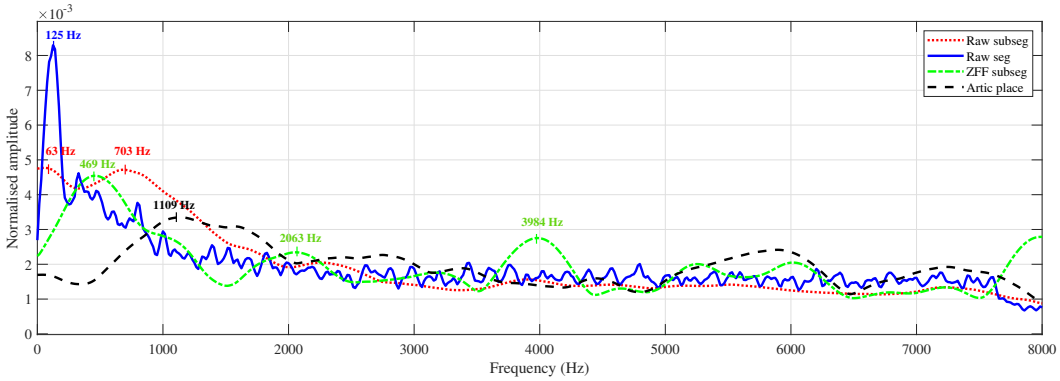
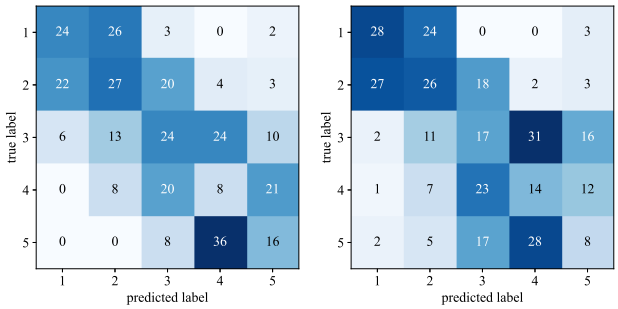


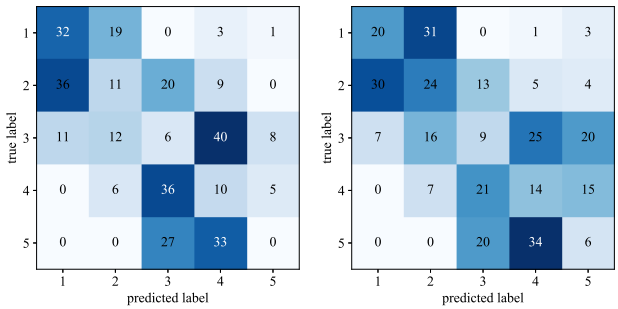
Figure 3: Frequency responses of the first convolutional layers of some systems.

in Appendix A for the architecture details). During training, all the five classes were ensured of equal representation in each epoch by duplicating some of the utterances presented. The input to the CNNs is a 250ms signal, overlapped by a 10ms shift. The networks were trained using cross-entropy loss with stochastic gradient descent. Learning rate was halved, in the range 10^{-2} to 10^{-6}, between successive epochs whenever the training-loss stopped reducing.

For the training procedure of the articulatory networks, the reader is referred to either Sec. A or [25]. It is worth noting that four AF CNNs were first pre-trained using AMI corpus [34] to individually predict the four AF categories: place, manner, height and vowel. Transfer learning for fluency prediction involved initialising 4 corresponding CNNs from the pre-trained ones, of the same architecture (*Artic*) except for the final layer, and fine-tuning them with the same training procedure as above. The posterior



(a) BoAW (system) (b) BoAW (wav2vec 2.0)



(c) raw seg (d) zff subseg

Figure 4: Confusion matrices of four best systems.

probabilities obtained from the 4 CNNs for each utterance were averaged before classification.

4.1 Results

Results are reported in Table 2 in terms of Pearson's correlation coefficient and Spearman's rank correlation coefficient. The p-values are provided in parentheses. For both evaluation measures, all the systems yield good correlation with a p-value well below 0.01, i.e., the results are statistically significant. We can observe that the BoAW approach modelling LLDs yields the best results. We also conducted experiments by considering only the source-related LLDs and vocal-tract system related LLDs. The vocal-tract system related features contribute the most. Having said that, a better performance of ZFF-based approach than BoAW approach modelling source-related LLDs indicates that the former is able to better model source-related information for speech fluency prediction. We can also observe that, in the subseg raw signal modelling based systems, initialising the neural network with articulatory feature information improves its performance, most prominently with the place of articulation. Finally, it is interesting to observe that BoAW approach with wav2vec 2.0 embeddings yields performance similar to subseg ZFF SigProc and seg Raw SigProc approach.

5 Analysis

Fig. 3 shows the cumulative frequency responses of the first convolution layer of the different CNN-based systems. It can be observed that most of the systems focus on the low frequency regions that are more related to the fundamental frequency and voice source related aspects [19, 20], which are more linked to fluency than the linguistic accuracy, corroborating with the finding in [7]. However, the articulatory feature initialised networks focus on low-to-mid frequencies, which are typically modelled by the CNNs when trained to classify phones and tend to model formant related information [35]. This suggests that the initialisation of subseg Raw SigProc approach with CNNs trained to classify AFs helped shift the focus of the network more towards linguistic unit related information and consequently improved the performance. Fig. 4 shows the performance of four best systems covering the different approaches. For all the systems, the predictions are centred around the true rating, indicating a systematic prediction of the speech fluency ratings. We have observed similar trend for other systems (see Fig. A1 in Appendix A).

6 Conclusion and Future Work

In this work, we investigated the prediction of perceived fluency from non-expert ratings. In this regard, we collected non-expert perceived fluency ratings of non-native speech and studied several approaches to automatically predict the human scores. Our investigations demonstrated the feasibility of predicting such scores using several approaches, and that BoAW based system modelling hand-crafted vocal-tract related LLD features performing the best. Automatic feature learning methods also obtained encouraging performance. In particular, end-to-end based acoustic modelling approach is able to better model the source related information than LLD-based or BoAW-based approach. Our future will focus along the following directions: (a) combining learned speech representations and hand-crafted feature representations, (b) fine tuning or adapting the wav2vec 2.0 models with the collected speech data for improved fluency prediction and (c) contrasting expert rating prediction with non-expert rating prediction.

References

[1] A. Loukina *et al.*, "Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead," in *Proceedings of Interspeech*, 2019, pp. 21–25. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2889

[2] A. C. Kelly *et al.*, "SoapBox Labs Fluency Assessment Platform for Child Speech," in *Proceedings of Interspeech*, 2020, pp. 488–489.

[3] C. Yarra, A. Srinivasan, S. Gottimukkala, and P. K. Ghosh, "SPIRE-fluent: A Self-Learning App for Tutoring Oral Fluency to Second Language English Learners," in *Proceedings of Interspeech*, 2019, pp. 968–969.

[4] Y. Xiao and F. K. Soong, "Proficiency assessment of ESL learner's sentence prosody with TTS synthesized voice as reference," in *Proceedings of Interspeech*, 2017, pp. 1755–1759. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-64

[5] L. Fontan, M. Le Coz, and S. Detey, "Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with japanese learners of french," in *Proc. Interspeech 2018*, 2018, pp. 2544–2548. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1336

[6] S. Mao, Z. Wu, J. Jiang, P. Liu, and F. K. Soong, "NN-based ordinal regression for assessing fluency of ESL speech," in *Proceedings of ICASSP*, 2019, pp. 7420–7424.

[7] K. Duijm, R. Schoonen, and J. H. Hulstijn, "Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach," *Language Testing*, vol. 35, no. 4, pp. 501–527, 2018. [Online]. Available: https://doi.org/10.1177/0265532217712553

[8] F. Eyben, *Acoustic Features and Modelling*. Cham: Springer International Publishing, 2016, pp. 9–122. [Online]. Available: https://doi.org/10.1007/978-3-319-27299-3_2

[9] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 02, pp. 190–202, Apr 2016.

[10] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. Interspeech 2017*, 2017, pp. 1263–1267. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-917

[11] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" in *Proc. Interspeech 2018*, 2018, pp. 147–151. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1238

[12] W. Xue, C. Cucchiarini, R. van Hout, and H. Strik, "Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech," in *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, 2019, pp. 48–52. [Online]. Available: http://dx.doi.org/10.21437/SLaTE.2019-9

[13] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2020.

[14] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

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

[16] 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. of ICASSP*, 2016.

[17] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *Proc. of Interspeech*, 2016.

[18] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection with raw waveform CLDNNS," in *Proc. of ICASSP*, 2017.

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

[20] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning voice source related information for depression detection," in *Proceedings of ICASSP*, 2019. [Online]. Available: http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP-2_2019.pdf

[21] J. Fritsch, S. P. Dubagunta, and M. Magimai.-Doss, "Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based cnns," in *Proceedings of ICASSP*, 2020.

[22] 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.

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

[24] N. Cummins *et al.*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *Proceedings of Interspeech*, 2020, pp. 2182–2186.

[25] S. P. Dubagunta and M. Magimai.-Doss, "Using speech production knowledge for raw waveform modelling based styrian dialect identification," in *Proceedings of Interspeech*, 2019.

[26] R. Rasipuram and M. Magimai.-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Computer Speech and Language*, vol. 36, pp. 233–259, 2016. [Online]. Available: http://publications.idiap.ch/downloads/papers/2015/Rasipuram_CSL_2015.pdf

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[28] M. Schmitt and B. Schuller, "openxbow – introducing the passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/17-113.html

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] A. Baevski *et al.*, https://github.com/pytorch/fairseq/tree/master/examples/wav2vec.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of ICASSP*, 2015, pp. 5206–5210.

[32] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," http://tensorflow.org/, 2015.

[33] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

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

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

A Supplementary Information

A.1 CNN architectures

This section contains additional information about the network architectures used in joint feature-classifier and articulatory models, listed as *SigProc* and *Artic* in the Table A1 respectively, and confusion matrices of more systems that were not included in the main content due to space constraints. The SigProc subseg and seg architectures were inspired from the previous work on modelling voice source information for depression detection [20].

Table A1: CNN architectures. N_f: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling.

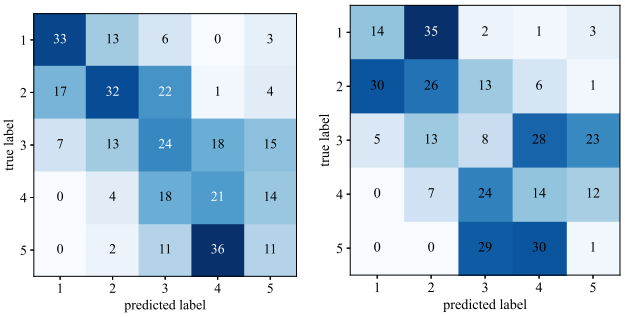
Model		**Layer**	**Conv**			**MP**
			N_f	**kW**	**dW**	
SigProc	subseg	1	128	30	10	2
		2	256	10	5	3
		3	512	4	2	-
		4	512	3	1	-
	seg	1	128	300	100	2
		2	256	5	2	-
		3,4	same as subseg			
Artic	subseg	1	80	30	10	3
		2,3	60	7	1	3

A.2 Off-the-shelf AF CNNs

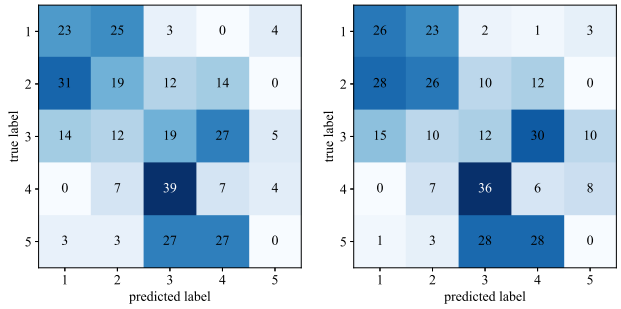
We used the articulatory feature (AF) CNNs originally trained for dialect identification study [25] and later used for sleepiness prediction [21]. Briefly, following the previous work on AF-based speech recognition [26], four AF CNNs, corresponding to manner of articulation, place of articulation, height of articulation and vowel, were trained on the independent headset microphone portion of the AMI data set [34], consisting of 77 hours of speech. The architecture of the AF CNNs is listed as *Artic* in Table A1. For further details related to the training of CNNs, the reader is referred to [25].

A.3 Confusion matrices

Fig. A1 provides the confusion matrices of a few more trained systems. It can be observed that the trends are similar to the results presented earlier in Section 5.



(a) BoAW (eGeMAPS) (b) zff seg



(c) Artic place (d) Artic manner

Figure A1: Confusion matrices of four additional systems.