Unsupervised Representation Learning with Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech

Shahin Amiriparian
The Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg
shahin.amiriparian@informatik.uni-augsburg.de

Pawel Winokurow
The Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg
pawel.winokurow@informatik.uni-augsburg.de

Vincent Karas
The Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg
vincent.karas@informatik.uni-augsburg.de

Sandra Ottl
The Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg
sandra.ottl@informatik.uni-augsburg.de

Maurice Gerczuk
The Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg
maurice.gerczuk@informatik.uni-augsburg.de

Björn Schuller
The Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg
schuller@informatik.uni-augsburg.de

ABSTRACT

Motivated by the attention mechanism of the human visual system and recent developments in the field of machine translation, we introduce our attention-based and recurrent sequence to sequence autoencoders for fully unsupervised representation learning from audio files. In particular, we test the efficacy of our novel approach on the task of speech-based sleepiness recognition. We evaluate the learnt representations from both autoencoders, and conduct an early fusion to ascertain possible complementarity between them. In our frameworks, we first extract Mel-spectrograms from raw audio. Second, we train recurrent autoencoders on these spectrograms which are considered as time-dependent frequency vectors. Afterwards, we extract the activations of specific fully connected layers of the autoencoders which represent the learnt features of spectrograms for the corresponding audio instances. Finally, we train support vector regressors on these representations to obtain the predictions. On the development partition of the data, we achieve Spearman's correlation coefficients of .324, .283, and .320 with the targets on the Karolinska Sleepiness Scale by utilising attention and non-attention autoencoders, and the fusion of both autoencoders' representations, respectively. In the same order, we achieve .311, .359, and .367 Spearman's correlation coefficients on the test data, indicating the suitability of our proposed fusion strategy.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks.*

KEYWORDS

unsupervised representation learning, attention mechanism, sequence to sequence autoencoders, audio processing, driver safety

ACM Reference Format:
Shahin Amiriparian, Pawel Winokurow, Vincent Karas, Sandra Ottl, Maurice Gerczuk, and Björn Schuller. 2020. Unsupervised Representation Learning with Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech. In *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (MuSe'20), October 16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3423327.3423670

1 INTRODUCTION

Sleepiness is a state identified by reduced alertness that varies according to a circadian rhythm, i. e., with the time of the day [16, 25]. Its detection is important for safety applications, as it has been shown, for example, that sleepiness impacts driving performance, even more so than fatigue [31, 34]. Most systems that aim to detect a sleepy driver rely on signals derived from interaction with the vehicle, such as abnormal steering behaviour, failures in lane keeping or irregular use of the pedals [32]. Furthermore, in [20, 30] the authors have demonstrated that sleepiness is one of the supervening cause of an accident amongst the professional drivers. Various studies could find a correlation between short sleep and a range of disorders, such as breathing problems [37], obesity [22], and mental disorders [33]. Moreover, in [44], Van Der Helm et al. have shown that sleep deprivation negatively influences the ability to classify the intensity of human facial emotions. Their findings have been substantiated in [28].

Research dealing with automatic sleepiness recognition has investigated methods to derive the state based on different bio-signals. Performing visual analysis of a subject's face, e. g., measuring blinking, can serve in assessing sleepiness but may be negatively affected by changing environmental parameters, such as illumination [17]. While electroencephalography (EEG) has also been shown to be

MuSe'20, October 16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8157-4
https://doi.org/10.1145/3423327.3423670

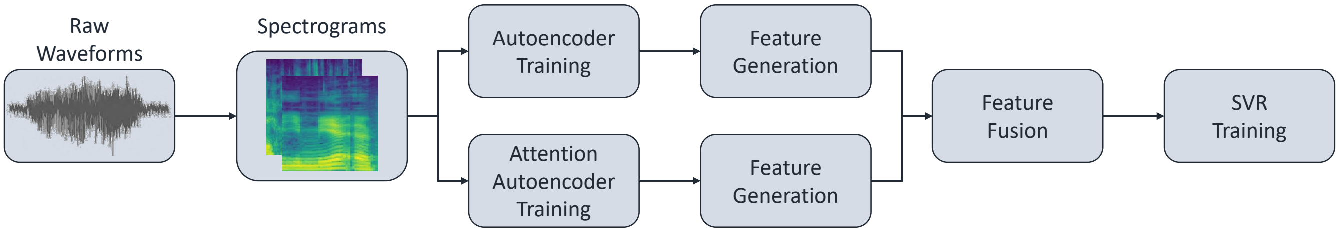

Figure 1: An overview of our framework for unsupervised representation learning with sequence to sequence recurrent autoencoders. Except for the SVR training, the approach is fully unsupervised. A detailed description of the procedure is given in Section 3.

a robust approach to the problem [13], it is far more intrusive and can only be achieved with special equipment and professional setup. In contrast to this, performing acoustic analysis of speech is non-obtrusive and does not need as much sensor application and calibration effort [27].

In order to define a suitable target for the automatic analysis of sleepiness, the state can be described by the Karolinska Sleepiness Scale (KSS) in terms of ratings ranging from 1 to 9, sometimes additionally extended by an extra category 10 for extreme sleepiness. A binary classification of sleepy speech has been previously performed as part of the INTERSPEECH 2011 speaker state challenge [39]. The sleepiness subchallenge of INTERSPEECH 2019's ComParE challenge deals with the detection of continuous sleepiness, and the sleepiness of a speaker is assessed as a regression problem [40]. In this challenge's baseline, the problem was approached in a number of ways, including a traditional acoustic feature extraction pipeline, Bag-of-Audio-Words (BoAW) [38], and a deep recurrent autoencoder framework. Here, the unsupervised sequence to sequence model auDeep [4, 18] achieved the strongest results.

A profound analysis of unsupervised representation learning techniques, including recurrent autoencoders and convolutional generative adversarial networks for speech and audio signal processing is given in [1]. Deep learning models, such as auDeep , that process raw or low level inputs yield state-of-the-art results for a wide range of machine learning problems [8, 21, 47]. Many of these approaches consider inputs as a whole, treating every part with the same importance. Often, however, some pieces of the input contain more information pertinent to solving the task at hand than others. A popular approach that takes this notion into consideration can be found with attention mechanisms, such as the one introduced by Bahdanau et al. [10] for machine translation. Compared to regular sequence to sequence autoencoders where all information is compressed into the last hidden state of the encoder, the dynamic context vector in the attention model retains information about all hidden states of the encoder and their alignment to the current decoding step. Since their introduction, attention mechanisms have also been adapted to speech recognition [11, 15], visual image captioning [45] or question answering [9], and speech emotion classification [24].

Motivated both by the effectiveness of recurrent autoencoders for acoustic analysis of sleepiness from speech as well as by the improvements to sequence to sequence models achieved with attention mechanisms, we evaluate the performance of our unsupervised recurrent approaches and analyse the impacts of combining them for the detection of continuous sleepiness on the respective 2019 edition of the INTERSPEECH ComParE sub-challenge.

The remainder of the paper is structured as follows. First, we describe the dataset used for the experiments in Section 2. Our autoencoder fusion framework is then introduced in Section 3 where we detail the feature learning process and both types of autoencoder models trained on the sleepiness challenge data. We further discuss our hyperparameter choices and experimental settings in Section 4. In Section 5, we present the results achieved during the evaluations before concluding our work and outlining future research directions in Section 6.

2 DATASET

We use a subset of the SLEEP Corpus that was employed in the 2019 edition of the INTERSPEECH Computational Paralinguistics Challenge (ComParE) [40]. The corpus contains speech recordings of 915 individuals (364 females, 551 males) at varying levels of sleepiness. The participants performed different pre-defined speaking tasks and read out text passages. Furthermore, spontaneous speech is included in the form of elicited narrative content. The sessions which lasted up to an hour per participant were further held between 6am to 12pm in order to capture high variability in the levels of perceived sleepiness. The resulting audio files have a sample rate of 16 kHz with a 16 bit quantisation. Each file is annotated with a KSS-score, ranging from 1 to 9, with 9 denoting extreme sleepiness. The labels were derived by averaging self-report with the scores of two external observers. The dataset is split into three partitions with 5 564, 5 328, and 5 570 samples.

3 APPROACH

A high-level overview of our proposed approach is depicted in Figure 1. First, Mel-spectrograms are extracted from audio signals. Recurrent autoencoders (AEs), both with and without attention mechanism, are then trained on the Mel-spectrograms to find compressed representations of the input data. Afterwards, the weights of the AEs are frozen and learnt representations are obtained from their hidden layers. As the final step, we fuse the representations of both types of recurrent AEs and classify them.

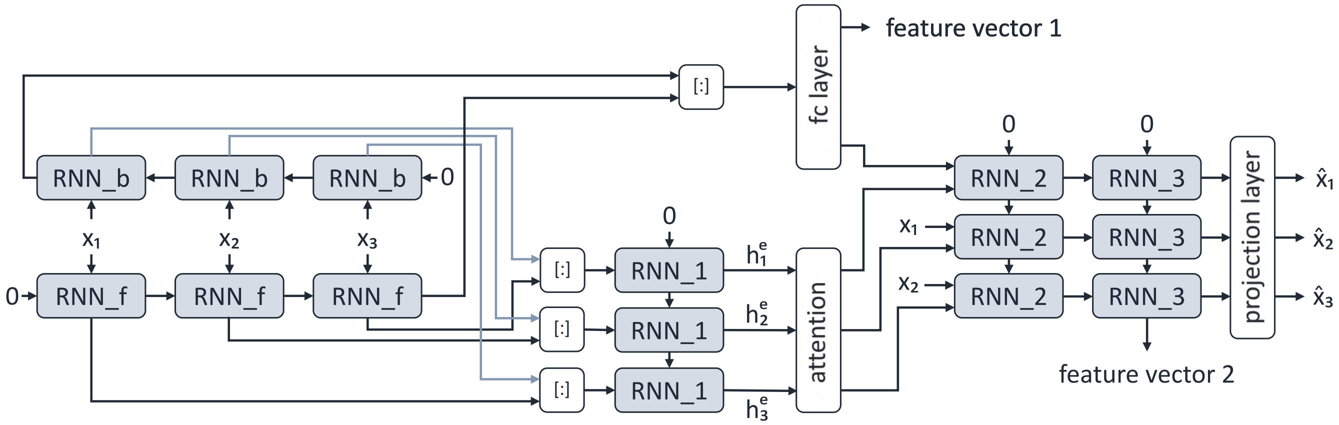

Figure 2: Schematic structure of our attention-based autoencoder with stacked encoder and decoder RNNs. Feature vectors 1 and 2 are extracted from the activations of the fully connected layer of the encoder RNN and the last hidden state of the decoder RNN, respectively. A detailed account of the proposed architecture is provided in Section 3.3.

3.1 Spectrogram Extraction

First, the Mel-spectrograms of audio recordings are extracted using periodic Hamming windows with width w and overlap $0.5w$. From these, a given number of log-scaled Mel-frequency bands are then computed. Finally, we normalise the Mel-spectra to have values in [-1; 1], since the outputs of the recurrent sequence to sequence autoencoders are constrained to this interval. These spectrograms are further treated by the autoencoders as sequences of one-dimensional frequency vectors, i. e., the networks process them as time series data.

3.2 Autoencoder Architecture without Attention

For this architecture, we utilise auDeep[1] [4, 18], our recurrent sequence to sequence autoencoder. For the representation learning with this framework, we can adjust a range of autoencoder parameters, including the direction (e. g., uni- or bidirectional) of the encoder and decoder RNNs, types of RNN cells, e. g., gated recurrent units (GRUs), or long short-term memory (LSTM) cells, and the number of hidden layers and units. To use LSTM-RNNs in the decoder, the LSTM cell is modified to work with a context vector similar to the GRUs in the encoder-decoder model proposed by Cho et al. [14]. The weight matrices C_i, C_f, C_o, and C_z are added to the input z_t, input gate i_t, forget gate f_t, and output gate o_t to enable an LSTM cell to work with the context vector:

$$
\begin{aligned}
z_t &= \tanh\left(W_z x_t + R_z y_{t-1} + C_z c + b_z\right) \\
i_t &= \sigma\left(W_i x_t + R_i y_{t-1} + C_i c + p_i \odot c_{t-1} + b_i\right) \\
f_t &= \sigma \quad W_f x_t + R_f y_{t-1} + C_f c + p_f \odot c_{t-1} + b_f \\
o_t &= \sigma\left(W_o x_t + R_o y_{t-1} + C_o c + p_o \odot c_t + b_o\right).
\end{aligned}
\tag{1}
$$

For each input sequence, the initial hidden state vector of the encoder is zero-padded. The last concatenated hidden state vector of the encoder $h_T^e = \left[\overrightarrow{h}_T \overleftarrow{h}_T\right]^T$ is then passed through a fully connected layer with tanh activation which has the same number of units as the decoder RNN. The output of this layer represents the context vector and is used as the first hidden state vector of the decoder h_0^d. During the feature extraction, the context vector also represents the feature vector. The outputs of the decoder are passed through a fully connected projection layer with tanh activation at each time step in order to map the decoder output dimensionality to the target dimensionality. The weights of this output projection are shared across time steps. For the network training, the teacher forcing algorithm [29] is applied. Following this method, instead of feeding the decoder with the predicted output at time step $t-1$ ($\hat{y}_{t-1}$), the expected decoder output at time step $t-1$ (y_{t-1}) is fed as an input to the decoder. This means that the decoder input is the same original spectrogram, only shifted by one step in time. Instead of the first step, the zero vector is inserted and the frequency vector at the last step is removed. During the training, the autoencoder learns to reconstruct the reversed input spectrogram [4, 6, 43]. Mean squared error (MSE) is used as the loss function to compare the reversed source spectrogram with the concatenated spectrogram obtained from the projection layer.

3.3 Autoencoder Architecture With Attention

In the second model, we add an attention mechanism to the autoencoder architecture. Here, encoder and decoder have almost the same structure as the baseline autoencoder. The problem of the sequence to sequence model is that the encoder must map all essential information of the input sequence to a fixed-length vector. This may not be enough to represent a long input sequence. To circumvent this, we adapt the attention mechanism introduced by Bahdanau et al. [10] for our sequence to sequence autoencoder architecture [4] (cf. Section 3.2). To the best of our knowledge, this is the first time, that such a mechanism has been directly applied for representation learning from the spectrograms of raw audio signals.

At each time step of the decoder, the attention mechanism, which includes dynamic computation of the context vector, enables to

[1] https://github.com/auDeep/auDeep

choose the hidden state vectors of the encoder that contain the most significant information to generate the context vector c_t which is used to generate the output y_t. This computation is based on all hidden state vectors of the encoder and the last hidden state vectors of the decoder. The context vector c_i is the linear combination of the hidden state vectors of the encoder h^e_j:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h^e_j. \tag{2}$$

Here, the weights α_{ij} are numbers between 0 and 1 and define which hidden states h^e_j have the biggest influence on y_i. α_{ij} is calculated with the softmax-normalised inner activation of the alignment model e_{ij}:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}. \tag{3}$$

The alignment model $e_{ij} = a(h^e_{i-1}, h^d_j)$ is a small feedforward neural network trained together with the sequence to sequence model using backpropagation and is defined as

$$v_{a_i}^{\top} \tanh \;\; W_a h^d_{i-1} + U_a h^e_j \;\; , \tag{4}$$

where $v_a \in \mathbb{R}^l$ is a weight vector, and $W_a \in \mathbb{R}^{l \times n}$ and $U_a \in \mathbb{R}^{l \times 2n}$ are weight matrices, and l is the number of units in the alignment model.

The difference to the autoencoder described in Section 3.2 is that the context vector is calculated dynamically using the alignment model and is not solely used as a feature vector. If we compute the feature vector from the last hidden state vector of the encoder, the attention mechanism will have less influence on the feature extraction. The feature vector, however, should contain information of all context vectors and, therefore, is extracted after their creation. For this reason, an additional set of feature vectors is extracted from the hidden state vector of the last decoder layer at the last time-step. The attention mechanism in the sequence to sequence model is visualised in Figure 2.

3.4 Fusion

As an optional final step before using the learnt features to predict sleepiness from speech, we perform feature level fusion. We take the representations extracted from trained recurrent AEs, with and without attention, and concatenate them along the feature axis. In this way, we want to evaluate if the information condensed in these representations is complimentary, i. e., if the change in focus introduced by the attention mechanism leads to different aspects of the input being expressed by the features.

4 EXPERIMENTAL SETTINGS

For all experiments, we train autoencoder models on the sleepiness spectrograms and then extract features for the three partitions. These feature vectors are used to train a linear support vector regressor (SVR) for which we optimise the complexity parameter on a logarithmic scale of 10^{-5} to 1. The complexity parameter is chosen based on the Spearman's correlation coefficient (ρ) achieved on the development partition.

Table 1: Performance comparison of the features obtained from the fully connected layer of the encoder (fc$_{enc}$) and the last hidden state of the decoder (state$_{dec}$) in our attention autoencoder. id: feature identifier, Dim.: feature dimensionality.

Parameters				fc$_{enc}$		state$_{dec}$	
id	Epoch	Cell	Dim.	ρ_{devel}	ρ_{test}	ρ_{devel}	ρ_{test}
1	20	GRU	512	.262	.308	.276	.294
2	25	GRU	512	.258	.308	.278	.298
3	30	GRU	512	.250	.314	.267	.292
4	35	GRU	512	.260	.312	.266	.298
5	40	GRU	512	.253	.307	.265	.288
6	20	LSTM	512	.293	.294	.318	.303
7	25	LSTM	512	.303	.276	**.324**	**.311**
8	30	LSTM	512	.298	.263	.322	.305
9	35	LSTM	512	.303	.285	.289	.304
10	40	LSTM	512	.307	.302	.288	.299

Our proposed autoencoder approaches contain a large amount of adjustable hyperparameters (cf. Sections 3.2 and 3.3), which prohibits an exhaustive exploration of the parameter space. For this reason, we choose suitable values for the hyperparameters in multiple stages, using the results of our initial experiments to bootstrap the process. In preliminary experiments, we test different configurations for the spectrograms that are used as input for the autoencoders. We arrived at using Mel-spectrograms with 160 and {128, 256} Mel-bands extracted from the audio samples for our autoencoders with and without attention, respectively. For the fast Fourier transform (FFT), we apply Hamming windows of 40 ms width and 20 ms overlap. We further experiment with the architecture of our recurrent autoencoder models. Here, we test one and two layer variants for both encoder and decoder using either GRU or LSTM cells with 128, 256, or 512 hidden units. Moreover, bidirectional and unidirectional encoders are compared. For the attention autoencoder, models with two-layer bidirectional encoders and two-layer unidirectional decoders worked best with 512 hidden units. For the results presented herein, we therefore settled on this architecture with the choice of either GRU or LSTM layers (cf. Table 1). Furthermore, we introduce an additional RNN layer in the decoder that serves to produce hidden states for the attention mechanism. The architecture of those models is visualised in Figure 2. We evaluate features extracted from both the last hidden state of the encoder and decoder. All attention models are trained using a batch size of 256 with the Adam optimiser [26] and the learning rate set to 10^{-4} for a maximum of 40 epochs. The model checkpoints at 20, 25, 30, 35, and 40 epochs further serve as feature extractors. For the auDeep experiments (cf. Section 3.2), we found the best configuration with 2 hidden layers each with 256 hidden units. We then optimise the direction of the encoder and decoder and adjust the RNN cell type. Additionally, we filter some of the background noise in the recordings by clipping amplitudes below {-40, -50, -60,

Table 2: Results obtained from our autoencoder without attention. id: feature identifier, window width: width of the Hamming window, Mel-Bands: number of Mel-bands, Clip-Level: clipped amplitudes below a certain threshold to filter some noise from audio, Dimension: dimensionality of each feature set, Direction: direction of the encoder-decoder RNN.

Autoencoders without attention								
id	window width [s]	Mel-Bands	Clip-Level	Cell Type	Dimension	Direction	ρ_{devel}	ρ_{test}
1	0.08	256	fused	GRU	4 096	uni-bi	.286	.338
2	0.08	256	-70 dB	GRU	1 024	bi-uni	**.283**	**.359**
3	0.06	256	-70 dB	GRU	1 024	bi-uni	.281	.357
4	0.08	256	-60 dB	GRU	1 024	uni-bi	.278	.331
5	0.04	256	-70 dB	GRU	1 024	uni-bi	.278	.340
6	0.06	256	fused	GRU	4 096	bi-uni	.277	.346
7	0.06	256	-70 dB	GRU	1 024	uni-bi	.277	.348
8	0.06	128	-70 dB	LSTM	1 024	uni-bi	.277	.317
9	0.08	128	-60 dB	GRU	1 024	bi-uni	.275	.324
10	0.04	256	-60 dB	GRU	1 024	uni-bi	.275	.336

-70} dB thresholds, and fuse them together resulting in five different feature vectors for each data partition. In [2, 7, 12], we have shown that our amplitude clipping approach can effectively eliminate unwanted audio effects and lead to a better overall performance of the machine learning system in various audio classification scenarios.

5 RESULTS AND DISCUSSION

All results obtained with our autoencoders and their early fusion are shown in Tables 1 to 3. In the attention model (cf. Table 1), features from the fully connected layer of the encoder RNN (fc$_{enc}$) generalise better when GRUs are applied (ρ_{devel} = .250, ρ_{test} = .314), and the features from the last hidden state of the decoder RNN (state$_{dec}$) perform better with LSTM cells (ρ_{devel} = .324, ρ_{test} = .311). From both autoencoder approaches, the recurrent model without attention shows the best performance on the test partition (ρ_{test} = .359), whilst the attention model achieves the highest results on the development partition (ρ_{devel} = .324). The result implies possible overfitting of the attention model on the development data. This issue is not strongly present in our model without attention, and we hypothesise that this is mainly because of the filtering of some of the background noise found in the audio data by clipping amplitudes below a certain threshold. In Table 2, we provide the highest achieved results with various thresholds and hyperparameter combinations. Furthermore, we fuse the best performing attention feature set on the development set (ρ_{devel} = .324, ρ_{test} = .311) with all non-attention (auDeep) features to analyse the complementarity of the learnt representations. The results in Table 3 demonstrate an improvement of all results after early fusion. The highest improvement on the test partition after fusion is achieved when the best attention feature set is combined with the fourth auDeep feature with GRUs, and the unidirectional encoder and bidirectional encoder trained on Mel-spectrograms with 256 Mel-bands (FFT window width of 80 ms and overlap of 40 ms) and -60 dB amplitude clipping (cf. Table 3). It is worth mentioning that the dimensionality of the attention features are either 1/2 or 1/8 of the auDeep features, leading to a faster classifier training. Moreover, the training process with attention autoencoders can be performed faster, as the encoder RNN is relieved from encoding all information in the whole input sequence of the Mel-spectrograms into a fixed-length vector [1]. We further compare our best performing approaches with the best challenge baselines [40], the winner of the challenge who combined Fisher vectors with baseline features [19], and the runner-up who utilised a fusion of convolutional neural networks (CNNs) and RNNs [46] (cf. Table 4).

Table 3: Results of our early fusion experiments with the best attention result (id$_{att}$ = 7) and all results provided in Table 2. id$_{att}$ and id$_{audeep}$: identifiers for the attention feature and auDeep features which are fused. C_{SVR}: Complexity of the SVR which is optimised on the development partition after fusion.

Early fusion				
id$_{att}$	id$_{audeep}$	C_{SVR}	ρ_{devel}	ρ_{test}
7	1	10^{-3}	.315	.359
7	2	10^{-2}	.336	.360
7	3	10^{-2}	.334	.365
7	4	10^{-1}	**.320**	**.367**
7	5	10^{-2}	.333	.349
7	6	10^{-3}	.319	.363
7	7	10^{-2}	.326	.361
7	8	10^{-2}	.333	.341
7	9	10^{-2}	.340	.351
7	10	10^{-2}	.339	.357

Table 4: Comparison of our best performing models with best performing challenge baselines and the challenge winner. *Dimension*: **feature dimensionality of each system,** ρ_{dev} **and** ρ_{test}: **Spearman's correlation coefficients on development and test partitions. *) It should be noted that the results provided by the challenge winner [19] only contain test results after the fusion with the challenge baselines ComParE and/or BoAW.**

System	Dimension	ρ_{dev}	ρ_{test}
Challenge Winner* [19]			
ComParE + BoAW + Fisher vectors	–	–	**.383**
Runner-up [46]			
CNNs and BLSTMs with attention	–	.373	.369
Best Challenge Baselines [40]			
ComParE	6 373	.251	.314
Bag-of-Audio-Words	500	.250	.304
Autoencoders	1 024	.243	.325
Late fusion of best	–	–	.343
Best of Our Proposed Approaches			
With attention	512	.324	.311
Without attention	4 096	.286	.338
Early fusion	4 608	**.320**	**.367**

6 CONCLUSIONS AND FUTURE WORK

In Section 3.3, we have introduced a novel attention mechanism for unsupervised representation learning from spectrograms of audio signals with recurrent sequence to sequence autoencoders[2]. We have demonstrated the suitability of two fully unsupervised representation learning techniques for continuous sleepiness recognition, and their superior performance to engineered, supervised features, such as ComParE and semi-supervised methods, such as Bag-of-Audio-Words (cf. Table 4). Furthermore, we have conducted a feature fusion strategy and demonstrated the complementarity of the learnt representations from both autoencoder architectures (cf. Section 5). The *autoencoder training* process in the attention architecture can be performed faster than auDeep as the encoder RNN does not need to encode all information in the whole input sequence of the Mel-spectrograms into a fixed-length representation. Moreover, the smaller dimensionality of the attention features (either 1/2 or 1/8 of the auDeep features) could lead to a faster *classifier training*. During the training process, we have observed that the attention-based autoencoder achieves a better performance on the development partition than auDeep, however, it is not highly distinguished by its performance on the unseen test set, implying its possible overfitting on the development data. Therefore, for the future work, we plan to add amplitude clipping into the attention autoencoder – similar to the auDeep architecture – to reduce the potential overfitting problems. We further plan to evaluate our systems over a wide range of audio recognition tasks, including speech emotion recognition [23, 42], sentiment analysis [35, 41], and music emotion classification [5]. We also want to utilise dimensionality reduction techniques [3] to cope with the high-dimensionality of our autoencoder features. Finally, we want to combine our methods with CNN-based representation learning systems, such as deep convolutional generative adversarial networks [36].

7 ACKNOWLEDGEMENTS

This research was partially supported by BMW AG, Deutsche Forschungsgemeinschaft (DFG) under grant agreement No. 421613952 (ParaStiChaD) and Zentrales Innovationsprogramm Mittelstand (ZIM) under grant agreement No. 16KN069455 (KIRun).

REFERENCES

[1] Shahin Amiriparian. 2019. *Deep Representation Learning Techniques for Audio Signal Processing.* Dissertation. Technische Universität München, München.

[2] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, Maurice Gerzcuk, Sergey Pugachevskiy, and Björn W. Schuller. 2018. A Fusion of Deep Convolutional Generative Adversarial Networks and Sequence to Sequence Autoencoders for Acoustic Scene Classification. In *Proc. EUSIPCO.* EURASIP, IEEE, Rome, Italy, 982–986.

[3] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. 2017. Feature Selection in Multimodal Continuous Emotion Prediction. In *Proc. International Workshop on Automatic Sentiment Analysis in the Wild (WASA).* AAAC, IEEE, San Antonio, TX, 30–37.

[4] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. 2017. Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio. In *Proc. DCASE.* "Tampere University of Technology. Laboratory of Signal Processing", Munich, Germany, 17–21.

[5] Shahin Amiriparian, Maurice Gerczuk, Eduardo Coutinho, Alice Baird, Sandra Ottl, Manuel Milling, and Björn Schuller. 2019. Emotion and Themes Recognition in Music Utilising Convolutional and Recurrent Neural Networks. In *Proc. MediaEval 2019 Multimedia Benchmark Workshop.* CEUR, Sophia Antipolis, France.

[6] Shahin Amiriparian, Jing Han, Maximilian Schmitt, Alice Baird, Adria Mallol-Ragolta, Manuel Milling, Maurice Gerczuk, and Björn Schuller. 2019. Synchronisation in Interpersonal Speech. *Frontiers in Robotics and AI, section Humanoid Robotics, Special Issue on Computational Approaches for Human-Human and Human-Robot Social Interactions* 6 (November 2019), 1–10. Article ID 116.

[7] Shahin Amiriparian, Maximilian Schmitt, Nicholas Cummins, Kun Qian, Fengquan Dong, and Björn Schuller. 2018. Deep Unsupervised Representation Learning for Abnormal Heart Sound Classification. In *Proc. EMBC.* IEEE, IEEE, Honolulu, HI, 4776–4779.

[8] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proc. ICML (ICML'16).* JMLR, New York, NY, USA, 173–182.

[9] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. tt. In *Proc. CVPR.* IEEE, Salt Lake City, UT, USA, 6077–6086.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs.CL]

[11] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *Proc. ICASSP.* IEEE, IEEE, Shanghai, China, 4945–4949.

[12] Alice Baird, Shahin Amiriparian, Miriam Berschneider, Maximilian Schmitt, and Björn Schuller. 2019. Predicting Biological Signals from Speech: Introducing a Novel Multimodal Dataset and Results. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP).* IEEE, IEEE, Kuala Lumpur, Malaysia, 1–5.

[13] Rodney Petrus Balandong, Rana Fayyaz Ahmad, Mohamad Naufal Mohamad Saad, and Aamir Saeed Malik. 2018. A Review on EEG-Based Automatic Sleepiness Detection Systems for Driver. *IEEE Access* 6 (2018), 22908–22919. https://doi.org/10.1109/ACCESS.2018.2811723

[14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[15] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-Based Models for Speech Recognition. In *Proc. NIPS* (Montreal, Canada) *(NIPS'15).* MIT Press, Cambridge, MA, USA, 577–585.

[2] Our audio-based attention framework (auttention), and all codes to reproduce the attention results are provided here: https://github.com/auttention/SleepyAttention

[16] William C. Dement and Mary A. Carskadon. 1982. Current perspectives on daytime sleepiness: The issues. *Sleep: Journal of Sleep Research & Sleep Medicine* 5, Suppl 2 (1982), 56–66. https://doi.org/10.1093/sleep/5.S2.S56

[17] F. Friedrichs and B. Yang. 2010. Camera-based drowsiness reference for driver state classification under real driving conditions. In *2010 IEEE Intelligent Vehicles Symposium*. IEEE, San Diego, CA, USA, 101–106.

[18] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. 2018. auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks. *Journal of Machine Learning Research* 18 (2018), 1–5.

[19] Gábor Gosztolya. 2019. Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds. In *Proc. Interspeech*. ISCA, Graz, Austria, 2413–2417.

[20] Gholam Hossein Halvani, Mehrzad Ebrahemzadih, and A Esmaeili. 2019. SLEEPINESS AND ACCIDENTS AMONG PROFESSIONAL DRIVERS. *Sigurnost* 61, 1 (2019), 15–25.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*. "IEEE", Las Vegas, NV, USA, 770–778.

[22] Jim Horne. 2008. Short sleep is a questionable risk factor for obesity and related disorders: statistical versus clinical significance. *Biological psychology* 77, 3 (2008), 266–276.

[23] M Shamim Hossain and Ghulam Muhammad. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion* 49 (2019), 69–78.

[24] Che-Wei Huang and Shrikanth Shri Narayanan. 2017. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, IEEE, Hong Kong, China, 583–588.

[25] William D.S. Killgore. 2010. Effects of sleep deprivation on cognition. In *Progress in Brain Research*, Gerard A. Kerkhof and Hans P.A. van Dongen (Eds.). Vol. 185. Elsevier, 105–129. https://doi.org/10.1016/B978-0-444-53702-7.00007-5

[26] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[27] J. Krajewski, A. Batliner, and M. Golz. 2009. Acoustic sleepiness detection – Framework and validation of a speech adapted pattern recognition approach. *Behavior Research Methods* 41 (2009), 795–804.

[28] Simon D Kyle, Louise Beattie, Kai Spiegelhalder, Zoe Rogers, and Colin A Espie. 2014. Altered emotion perception in insomnia disorder. *Sleep* 37, 4 (2014), 775–783.

[29] Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems* (Barcelona, Spain, 5.-10. Dec. 2016) *(NIPS'16, Vol. 1)*. Curran Associates Inc., 57 Morehouse Lane, Red Hook, NY, United States, 4601–4609.

[30] Mona Lichtblau, Daniel Bratton, Philippe Giroud, Thomas Weiler, Konrad E Bloch, and Thomas Brack. 2017. Risk of sleepiness-related accidents in Switzerland: results of an online sleep apnea risk questionnaire and awareness campaigns. *Frontiers in medicine* 4 (2017), 34.

[31] James M Lyznicki, Theodore C Doege, Ronald M Davis, Michael A Williams, et al. 1998. Sleepiness, driving, and motor vehicle crashes. *Jama* 279, 23 (1998), 1908–1913.

[32] Alina Mashko. 2015. Review of approaches to the problem of driver fatigue and drowsiness. In *2015 Smart Cities Symposium Prague (SCSP)*. IEEE, Prague, Czech Republic, 1–5. https://doi.org/10.1109/SCSP.2015.7181569

[33] Maurice M Ohayon, Malijai Caulet, Pierre Philip, Christian Guilleminault, and Robert G Priest. 1997. How sleep and mental disorders are related to complaints of daytime sleepiness. *Archives of internal medicine* 157, 22 (1997), 2645–2652.

[34] Pierre Philip, Patricia Sagaspe, Nicholas Moore, Jacques Taillard, André Charles, Christian Guilleminault, and Bernard Bioulac. 2005. Fatigue, sleep restriction and driving performance. *Accident Analysis & Prevention* 37, 3 (2005), 473–478. https://doi.org/10.1016/j.aap.2004.07.007

[35] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.

[36] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434 [cs.LG]

[37] O Resta, MP Foschino-Barbaro, G5 Legari, S Talamo, P Bonfitto, A Palumbo, A Minenna, R Giorgino, and G De Pergola. 2001. Sleep-related breathing disorders, loud snoring and excessive daytime sleepiness in obese subjects. *International journal of obesity* 25, 5 (2001), 669–675.

[38] Maximilian Schmitt and Björn Schuller. 2017. OpenXBOW: introducing the passau open-source crossmodal bag-of-words toolkit. *The Journal of Machine Learning Research* 18, 1 (2017), 3370–3374.

[39] Björn Schuller, Anton Batliner, Stefan Steidl, Florian Schiel, and Jarek Krajewski. 2011. The INTERSPEECH 2011 Speaker State Challenge. In *Proc. Interspeech*. ISCA, Florence, Italy, 3201–3204.

[40] Björn W. Schuller, Anton Batliner, Christian Bergler, Florian B. Pokorny, Jarek Krajewski, Margaret Cychosz, Ralf Vollmann, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, Alejandrina Cristia, Amanda Seidl, Anne S. Warlaumont, Lisa Yankowitz, Elmar Nöth, Shahin Amiriparian, Simone Hantke, and Maximilian Schmitt. 2019. The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Proc. Interspeech* (Graz, Austria). ISCA, ISCA, 2378–2382. https://doi.org/10.21437/Interspeech.2019-1122

[41] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. MuSe 2020 – The First International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop. arXiv:2004.14858 [cs.MM]

[42] Lukas Stappen, Nicholas Cummins, Eva-Maria Meßner, Harald Baumeister, Judith Dineley, and Björn Schuller. 2019. Context modelling using hierarchical attention networks for sentiment and self-assessed emotion detection in spoken narratives. In *Proc. ICASSP*. IEEE, IEEE, 6680–6684.

[43] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *CoRR* abs/1409.3215 (2014).

[44] Els Van Der Helm, Ninad Gujar, and Matthew P Walker. 2010. Sleep deprivation impairs the accurate recognition of human emotions. *Sleep* 33, 3 (2010), 335–342.

[45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML (ICML'15)*. JMLR.org, Lille, France, 2048–2057.

[46] Sung-Lin Yeh, Gao-Yi Chao, Bo-Hao Su, Yu-Lin Huang, Meng-Han Lin, Yin-Chun Tsai, Yu-Wen Tai, Zheng-Chi Lu, Chieh-Yu Chen, Tsung-Ming Tai, Chiu-Wang Tseng, Cheng-Kuang Lee, and Chi-Chun Lee. 2019. Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition. In *Proc. Interspeech*. ISCA, Graz, Austria, 2398–2402. https://doi.org/10.21437/Interspeech.2019-2110

[47] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* 13, 3 (2018), 55–75.