ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413589

DEEP NEURAL NETWORK EMBEDDINGS FOR THE ESTIMATION OF THE DEGREE OF SLEEPINESS

José Vicente Egas-López[1], *Gábor Gosztolya*[1,2]

[1] University of Szeged, Institute of Informatics, Szeged, Hungary
[2] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

ABSTRACT

Estimating the degree of sleepiness from the human speech is an emerging research problem with straightforward applications. In this study, we employ the x-vector approach, currently the state-of-the-art in speaker recognition, as a neural network feature extractor to detect the level of sleepiness of a speaker. Besides using different corpora for fitting the x-vector DNN, we also experiment with adding noise and reverberation to the training samples. According to our experimental results for the publicly available Dusseldorf Sleepy Language Corpus, utilizing x-vector embeddings as features for Support Vector Regression consistently leads to competitive performance scores in sleepiness detection. In particular, we present the highest Spearman's correlation coefficient on the public corpus that was achieved by a single method.

Index Terms— computational paralinguistics, speech processing, sleepiness, x-vectors, DNN embeddings

1. INTRODUCTION

Excessive lack of sleep may lead to poor performance in daily activities, can contribute to accidents, and eventually lead to mortality. The most common causes of excessive daytime sleepiness (hypersomnia) are sleep deprivation and disorders like apnea (cessation of breathing) and insomnia (the inability to stay or fall asleep) [1]. The National Sleep Foundation of the United States, in their Sleep in America Poll for 2020 [1], found that almost half of Americans report feeling sleepy between three and seven days per week. The mentioned arguments make the detection and monitoring of sleepiness crucial for reducing the risks of having fatal accidents (e.g., when operating machinery or driving vehicles). Also, it may be beneficial for the early detection of specific neurological problems. Sleepiness is not a condition in itself; it is seen a symptom caused by an underlying problem such as a neurological disease [2, 3], to name an example. A non-invasive way to monitor and control the degree of sleepiness could be by using the speech of the subject, which could help in automatic risk detection while driving and in similar situations.

From the aspect of machine learning and feature extraction techniques utilized, estimating the degree of sleepiness of the speaker belongs to the general area of computational paralinguistics. The tasks in this field focus on modeling non-verbal latent patterns in the speech that go beyond the *linguistic* approach. The captured speaker-traits are used in various tasks. E.g., assessing the self-affect of individuals [4], screening neurological diseases such as Mild Cognitive Impairment (MCI) [5], and psychological disorders like depression [6]. Also, in miscellaneous tasks, like estimating the conflict intensity [7] or determining personality traits of a speaker [8].

Former state-of-the-art approaches for speaker recognition (e.g. i-vectors [9]) were also exploited in computational paralinguistics and give relevant performances for classifying the cognitive load [10], or estimating the speaker's age [11] using the speech. The current state-of-the-art technique for speaker recognition is the so-called *x-vector* approach [12], which employs a Deep Neural Network to map variable-length utterances to fixed-dimensional embeddings (i.e. x-vectors). A handful of previous studies exploited x-vector embeddings in computational paralinguistics tasks. For instance, to classify emotion from the speech of subjects [13], or to screen neuro-degenerative diseases like Alzheimer's Disease [14], and Parkinson's Disease [15].

In this study, we employ x-vectors to estimate the degree of sleepiness from the recorded speech of subjects. Based on the methodology outlined in [12], we will adopt the DNN architecture described there. We train the network from scratch employing the train and development sets of the SLEEP Corpus. The DNN extracts the final neural network embeddings, which are utilized by a Support Vector Regression (SVR) for the estimation. The methodology we present in this study gives the highest Pearson's correlation coefficient value obtained by a standalone method on the public SLEEP Corpus.

This research was partially supported by grant TUDFO/47138-1/2019-ITM of the Ministry for Innovation and Technology, Hungary, and by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program. G. Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-20-5.

[1] Link to the official summary of findings of the Sleep in America Poll 2020: https://www.sleepfoundation.org/wp-content/uploads/2020/03/SIA-2020-Q1-Report.pdf?x90960

2. DATA

Here, we make use of the SLEEP (Dusseldorf Sleepy Language) Corpus. The corpus comprises the recordings of 915 German speakers, 364 females, 551 males, from 12 to 84 years of age, with a mean age of 27.6. The utterances were recorded with 44.1 kHz and downsampled to 16 kHz, using a quantisation of 16 bit. The audios were made in quiet rooms with similar acoustic conditions. The subjects were asked to read passages and carry out speaking tasks. Likewise, the subjects were asked to speak about, for example, their last weekend or to describe a picture; this resulted in spontaneous narrative speech. It contains 5564, 5328 and 5570 utterances, training, development and test sets, respectively; all three subsets contain recordings of just below six hours, leading to 17 hours and 35 mins of speech overall.

The degree of sleepiness of the subjects was assessed using the Karolinska Sleepiness Scale (KSS) [16]. Each subject reported their sleepiness level on the Karolinska Sleepiness Scale (KSS): from 1 (extremely alert) to 9 (very sleepy). At the same time, two observers assigned posthoc observer KSS ratings. The average of both scores was the reference sleepiness value [17]. Later, this corpus was included in the Interspeech Computational Parainguistic Challenge in 2019 [17].

3. FEATURE EXTRACTION AND METHODOLOGY

3.1. Frame-level Representations

The well-known Mel-Frequency Cepstral Coefficients (MFCC) representations are utilized in this study. 20 MFCCs are extracted from the recordings using a frame-length of 25ms and a window step size of 10ms.

3.2. Deep Neural Network Embeddings

The x-vector approach can be thought as of a neural network feature extraction method that provides fixed-dimensional embeddings for variable-length utterances. Such a system can be viewed as a feed-forward Deep Neural Network (DNN) that computes such embeddings.

3.2.1. DNN Architecture

Table 1 describes the structure of the DNN. The *frame-level* layers have a time-delay architecture. Let us assume that t is the actual time step. At the input, the frames are spliced together; namely, the input to the current layer is the spliced output of the previous layer (i.e. input to layer *frame3* is the spliced output of layer *frame2*, at frames $t-3$ and $t+3$). Next, the *stats pooling* layer gets the T frame-level activations of the last frame-level layer (*frame5*), aggregates over the input segment, and computes the mean and the standard deviation. The mean and the standard deviation are nothing but segment-level statistics. These stats are concatenated and used as input for the next *segment6* and *segment7* layers, respectively. The last layer is the *softmax* output layer, which is discarded after training the DNN. The *x-vectors* embeddings can be extracted from any of *segment* layers. [12, 18]. Instead of predicting frames, the DNN is trained to predict speakers from variable-length utterances utilizing a multi-class cross entropy objective function (more details in [18]).

Table 1. DNN architecture of the x-vector system, consisting of five frame-level layers, a statistics pooling layer, two segment-layers and a final softmax layer. N represents the number of training speakers in the softmax layer. This architecture is based on the one described by Snyder et al. [18].

Layer	Layer context	Tot. context	In, Out
frame1	[t-2, t+2]	5	120, 512
frame2	{t-2, t, t+2}	9	1536, 512
frame3	{t-3, t, t+3}	15	1536, 512
frame4	{t}	15	512, 512
frame5	{t}	15	512, 1500
stats pooling	[0, T}	T	1500T, 3000
segment6	{0}	T	3000, 512
segment7	{0}	T	512, 512
softmax	{0}	T	512, N

3.2.2. The x-vector

The embeddings produced by this network capture information from the speakers over the whole audio-signal. Such embeddings are called *x-vectors* and they can be extracted from any *segment* layer; that is, either *segment6* or *segment7* layers (see Table 1). Normally, embeddings from the *segment6* layer give a better performance than those from *segment7* [12]. In this study, these type of representations can capture meaningful information from each utterance. This type of embedding may help us to discriminate better the utterances since the characteristics are acquired at the utterance level rather than at the frame-level. For this, we used the Kaldi Toolkit.

4. EXPERIMENTAL SETUP

4.1. DNN Training Data

We trained different x-vector Deep Neural Network models (i.e. extractors) using two distinct datasets. First, we used the data of the training and development sets of the SLEEP corpus combined (10892 utterances, 11 hours and 39 mins). Second, to experiment with the independence of the x-vectors from different recording and speaking conditions (e.g., language), we trained the extractor (DNN) on another corpus (also for speaker recognition). We used a subset of 60 hours (10636

Table 2. Results of the experiments on the SLEEP Corpus given in Spearman's Correlation Coefficient. We show the results of former studies as well. The * means that the scores were achieved by a fusion of the best configurations. In contrast, the rest of the scores were obtained by applying a single approach. The x-vectors scores are given in accord with the corpus used to train the DNN they were extracted with.

ComParE 2019 Features [17]	**Dev**	**Test**
ComParE Functionals	.251	.314
Bag-of-Audio-Words (BoAW)	.250	.304
AuDeep	.261	.310
Three-wise fusion*	—	.343
Former Studies		
Gosztolya* [20]	.367	.383
Yeh et al.* [21]	.373	.369
Amiriparian et al.* [22]	.320	.367
Wu et al.* [23]	.326	.365
Elsner et al. [24]	.290	.335
Fritsch et al.* [25]	.317	.325
DNN Embeddings (x-vectors)		
SLEEP Corpus train-dev (12h)	.303	**.365**
SLEEP Corpus train-dev (augmented)	.275	.324
BEA Corpus (60h)	.287	.313
BEA Corpus (augmented)	.256	.301
SWBD + SRE (pre-trained model, [12])	.300	.355

utterances) of the BEA Corpus, which contains Hungarian spontaneous speech (for more details, see [19]). This corpus has a relevant size (in comparison with the SLEEP Corpus), which is convenient when training DNNs.

4.2. Data Augmentation

It is a standard practice to employ data augmentation when training x-vector DNNs in order to improve the noise robustness of the model [12]. From the original training data, two augmented versions are added. From additive noises and reverberation, two of the following types of augmentation are chosen randomly: babble, music, noise, and reverberation. The first three types correspond to adding or fitting noise to the original utterances. The fourth one involves a convolution of room impulse responses with the audio, i.e. reverberation. The reader can see [12] for more details about the augmentation strategies used in this study. This process increased the DNN training sets to 52982 utterances (over 56 hours) and to 52636 utterances (293 hours), SLEEP and BEA corpora, respectively. Our goal is to evaluate the contribution of the augmentation techniques to the overall performance scores.

4.3. Deep Neural Network Embeddings

The *segment6* layer of the DNN is used to compute the 512-dimensional neural network embeddings, (i.e. *x-vectors*). In addition to the four x-vector training variations described above (*SLEEP Corpus train-dev*, *SLEEP Corpus train-dev (augmented)*, *BEA Corpus* and *BEA Corpus (augmented)*), we employ the publicly available, pre-trained x-vector model described by Snyder et al. [12]. The model was fitted on English speech, specifically, employing a combination of a portion of Switchboard (SWBD) with a subset of the NIST SRE corpus. This model can be downloaded from `https://kaldi-asr.org/models/m3`. Next, we also utilize this pre-trained model to extract x-vector embeddings from the SLEEP Corpus (*pre-trained x-vector DNN*). We aim to discover the differences amongst the DNN performances when using corpora that differ in both duration and language from the SLEEP Corpus.

4.4. Regression and Evaluation

Support Vector Regression (SVR) was utilized to estimate the degree of sleepiness of the speakers. DNN embeddings were standardized by removing the mean and scaling to unit variance before training the model. We relied on the libSVM implementation [26] with a linear kernel (nu-SVR method); the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}, based on the performance on the development set. Before rounding to the nearest integer in the $1 \ldots 9$ scale, first we linearly transformed the predictions to have the same mean and standard deviation as those of the labels of the training set; transformation parameters were set on the development set. Spearman's Correlation Coefficient is the performance metric employed in this regression task (see more in [17]).

5. RESULTS AND DISCUSSION

Table 2 outlines the Spearman's correlation coefficient scores got by the x-vectors embeddings. Overall, x-vector features extracted employing the SLEEP train-dev model gave better performances. These features achieved a .303 and a .365 of CC score on dev and test, respectively. However, using the augmented version of this model resulted in a decrease of the CC scores in both dev and test sets (.275 and .324). A similar situation occurred in the BEA Corpus model, namely, its augmented version led to a decrease in the CC scores. On dev, CC went from .287 to .256; and from .313 (no augmentation) to .301 (augmented) on the test set. Although augmentation gives more diversity to the original data and attempts to make the models more robust. Here, the results indicate that the DNN was able to capture more meaningful information from the non-augmented versions than from their noise-robust counterparts. That is, adding noises and reverberation to this particular datasets could have caused the DNN to learn from

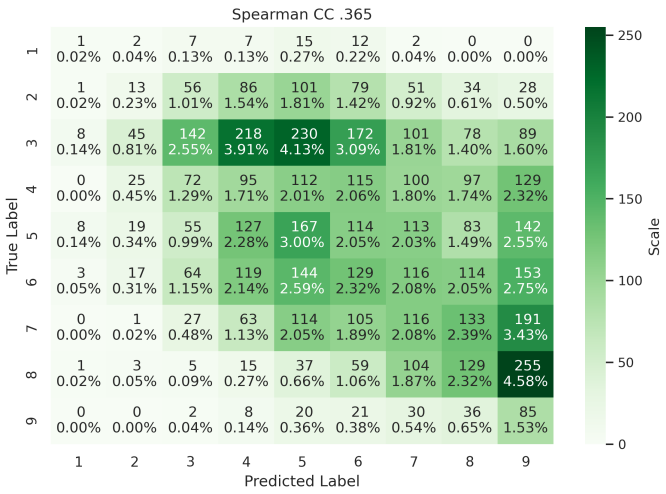

Fig. 1. Confusion matrix for the best results on the test set.

non-relevant information, resulting in a poorer mapping (i.e., x-vector embeddings) for the specified task.

Meanwhile, although the x-vector pre-trained model produced better results (.355 of CC on test), its performance could not reach that of the SLEEP train-dev extractor. This could be attributed to a language-dependant situation (i.e. the pretrained model was fitted using English corpora). It appears that, for this particular case, the model trained with in-domain data (i.e. using the SLEEP corpus) was able to generate better representations than the pre-trained model which was trained with huge data amounts of different domain data.

In Table 2 we also compare our performance scores with those of previous studies and official baselines on the same task. It can be seen that the proposed DNN embeddings were capable of outperform all the baseline scores of the Interspeech 2019 ComParE Challenge [17]. Moreover, it is evident that most of the former studies achieved their best results by relying on a *fusion* of the scores. In detail, in [20], a combination of Fisher Vectors, BoAW, and the ComParE functionals is carried out for their final CC scores (.383). (Note that this score was improved to .387 by training ensembles of classifiers [20].) In [21], the authors employ attention networks and adversarial augmentation, in the end, their best results (.369 of CC on test) are achieved by a fusion of neural network models. In [22], a .367 of CC was obtained by an early fusion of the learnt representations from attention and sequence to sequence autoencoders. Fisher Vector encodings were fused with the outputs of the ComParE Functionals in [23] to get a .365 of CC. In both [24] and [25], CNNs were exploited in an end-to-end deep learning approach: no fusion techniques are executed in the former study to get a .335 of CC; in the latter, a fusion of their CNN models was made to get a .325 of CC score. However, in our study, x-vector representations are still competitive and even outperform some of the former studies without the need for any kind of fusion strategy.

Fig. 1 displays the confusion matrix of our best configuration. The figure tells us that categories $3, 5, 6, 7, 8$ had similarly high accuracies. This means that the model was capable of distinguishing a large variety of categories including one of the extreme labels (8), the slightly extreme classes (3 and 7), as well as the middle categories (5 and 6). As for the extreme labels 1, 2 and 9, the scores are much lower. Perhaps this is due to the number of samples for these classes: these three categories represent approximately 13% of the number of samples in the dataset. Moreover, it seems that the model tends to overestimate the sleepiness level of the speaker, as we observed higher values mainly above the main diagonal.

6. CONCLUSIONS AND FUTURE WORK

This study investigated deep neural network embeddings for estimating the degree of sleepiness from speech. We employed five different DNN models to map utterances to fixed-sized representations (i.e. x-vectors). Utilizing the SLEEP and BEA corpora, two models were fitted using augmented data, and two with no augmentation. The fifth model used was the pre-trained DNN from [12]. Our findings indicate that the augmentation strategies applied on both corpora did not give any improvements: the quality of the embeddings extracted using the augmented models only reduced the final scores. Furthermore, it appears that making use of in-domain data causes the extractors (DNN models) to generate more meaningful features than just using out-of-domain data. In specific, we achieved the best performance employing the x-vector features computed via the SLEEP Corpus model.

In contrast to former studies, we did not rely on fusion strategies yet the results are competitive. More generally, we demonstrated that our methodology, besides surpassing the performances of various previous works, also produce the highest Spearman's CC score via a standalone (single) method for this particular task. In the future, we will further investigate the domain- and language-dependency of the DNN extractors using bigger datasets.

7. REFERENCES

[1] M.W. Johns, "Daytime Sleepiness, Snoring, and Obstructive Sleep Apnea: the Epworth Sleepiness Scale," *Chest*, vol. 103, no. 1, pp. 30–36, 1993.

[2] J.F. Pagel, "Excessive Daytime Sleepiness," *American Family Physician*, vol. 79, no. 5, pp. 391–396, 2009.

[3] B.J. Murray, "A Practical Approach to Excessive Daytime Sleepiness: a Focused Review," *Canadian respiratory journal*, vol. 2016, 2016.

[4] C. Montacié and M. Caraty, "Vocalic, Lexical and Prosodic Cues for the Interspeech 2018 Self-Assessed

Affect Challenge.," in *Proceedings of Interspeech*, 2018, pp. 541–545.

[5] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech," *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.

[6] Z. Liu, B. Hu, L. Yan, T. Wang, F. Liu, X. Li, and H. Kang, "Detection of depression in speech," in *Proceedings of ACII*, Los Alamitos, CA, USA, sep 2015, pp. 743–747, IEEE Computer Society.

[7] G. Gosztolya and L. Tóth, "DNN-based Feature Extraction for Conflict Intensity Estimation from Speech," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1837–1841, 2017.

[8] A.V. Ivanov and X. Chen, "Modulation spectrum analysis for speaker personality trait recognition," in *Proceedings of Interspeech*, 2012.

[9] N. Dehak, P. J Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[10] M. V. Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P Black, A. Potamianos, and Shrikanth S Narayanan, "Classification of Cognitive Load from Speech using an i-vector Framework," in *Proceedings of Interspeech*, 2014.

[11] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors," in *Proceedings of Interspeech*, 2016, pp. 1402–1406.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker verification," in *Proceedings of ICASSP*, 2018, pp. 5329–5333.

[13] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker verification," in *Proceedings of ICASSP*, 2020, pp. 7169–7173.

[14] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose Alzheimer's disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[15] L. Jeancolas, D. Petrovska-Delacrétaz, G. Mangone, B. Benkelfat, J. Corvol, M. Vidailhet, S. Lehéricy, and H. Benali, "X-vectors: New quantitative biomarkers for early Parkinson's Disease detection from speech," *arXiv preprint arXiv:2007.03599*, 2020.

[16] A. Shahid, K. Wilkinson, S. Marcu, and C. M Shapiro, "Karolinska Sleepiness Scale (KSS)," in *STOP, THAT and One Hundred Other Sleep Scales*, pp. 209–210. Springer, 2011.

[17] B.W. Schuller, A. Batliner, C. Bergler, F.B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S. Roelen, S. Schnieder, E. Bergelson, et al., "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity.," in *Proceedings of Interspeech*, 2019, pp. 2378–2382.

[18] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network embeddings for text-independent speaker verification," in *Proceedings of Interspeech*, 2017.

[19] T. Neuberger, D. Gyarmathy, T. E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *Proceedings of TSD*, Brno, Czech Republic, Sep 2014, pp. 424–431.

[20] G. Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," in *Proceedings of Interspeech*, 2019, pp. 2413–2417.

[21] S.L. Yeh, G.Y. Chao, B.H. Su, Y.L. Huang, M.H. Lin, Y.C. Tsai, Y.W. Tai, Z.C. Lu, C.Y. Chen, T.M. Tai, et al., "Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition.," in *Proceedings of Interspeech*, 2019, pp. 2398–2402.

[22] S. Amiriparian, P. Winokurow, V. Karas, S. Ottl, M. Gerczuk, and B. W. Schuller, "A Novel Fusion of Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech," *arXiv preprint*, 2020.

[23] H. Wu, W. Wang, and M. Li, "The DKU-LENOVO systems for the INTERSPEECH 2019 Computational Paralinguistic Challenge.," in *Proceedings of Interspeech*, 2019, pp. 2433–2437.

[24] D. Elsner, S. Langer, F. Ritz, R. Mueller, and S. Illium, "Deep Neural Baselines for Computational Paralinguistics," *arXiv preprint arXiv:1907.02864*, 2019.

[25] 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*. IEEE, 2020, pp. 6534–6538.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.