Snore sound classification using image-based deep spectrum features

Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Amiriparian, Shahin, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. "Snore sound classification using image-based deep spectrum features." In *Interspeech 2017, 20-24 August 2017, Stockholm, Sweden*, edited by Francisco Lacerda, 3512–16. ISCA. https://doi.org/10.21437/interspeech.2017-434.

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Snore Sound Classification Using Image-based Deep Spectrum Features

Shahin Amiriparian[1,2], *Maurice Gerczuk*[1], *Sandra Ottl*[1], *Nicholas Cummins*[1], *Michael Freitag*[1], *Sergey Pugachevskiy*[1], *Alice Baird* [1], *Björn Schuller*[1,3]

[1]Chair of Complex & Intelligent Systems, Universität Passau, Germany
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3]Machine Learning Group, Imperial College London, U.K.

shahin.amiriparian@tum.de

Abstract

In this paper, we propose a method for automatically detecting various types of snore sounds using image classification convolutional neural network (CNN) descriptors extracted from audio file spectrograms. The descriptors, denoted as deep spectrum features, are derived from forwarding spectrograms through very deep task-independent pre-trained CNNs. Specifically, activations of fully connected layers from two common image classification CNNs, AlexNet and VGG19, are used as feature vectors. Moreover, we investigate the impact of differing spectrogram colour maps and two CNN architectures on the performance of the system. Results presented indicate that deep spectrum features extracted from the activations of the second fully connected layer of AlexNet using a viridis colour map are well suited to the task. This feature space, when combined with a support vector classifier, outperforms the more conventional knowledge-based features of 6 373 acoustic functionals used in the INTERSPEECH ComParE 2017 Snoring sub-challenge baseline system. In comparison to the baseline, unweighted average recall is increased from $40.6\,\%$ to $44.8\,\%$ on the development partition, and from $58.5\,\%$ to $67.0\,\%$ on the test partition.
Index Terms: convolutional neural networks, deep learning, snore sound, spectral features, computational paralinguistics

1. Introduction

Aside from negatively impacting the quality of life of those affected [1, 2], snoring can also be a marker of *Obstructive Sleep Apnea* (OSA) [3] which, after insomnia, has the highest prevalence of all sleep disorders, affecting approximately 3–7 % of the middle-aged men and 2–5 % of middle-aged women [4–6] in the general population. OSA is characterised by repetitive episodes of partial or complete collapses of the upper airway during sleep, causing impaired gaseous exchanges and sleep disturbance [7]. As a chronic condition that is caused by an obstruction of the upper airways during sleep, OSA can lead to an increased risk of cardiovascular and cerebrovascular diseases [8, 9]. An integral part of successful treatment is locating the site of obstruction and vibration [10], which is the subject of the INTERSPEECH 2017 ComParE Snoring sub-challenge [11]. The challenge requires participants to identify four different sources of vibration from audio snore samples: epiglottis, oropharyngeal lateral walls, tongue, and velum.

An audio perspective on the analysis of snoring has made use of, among others, amplitude [12], frequency [13] and wavelet features [14]. Agrawal et al. show that palatal (velum) and tongue-based snoring differ significantly in peak-frequency. While the former's median peak frequency was observed at 137 Hz, the latter's was located at 1 243 Hz [15]. Among their subjects, they also measured peak frequencies of snores originating from the tonsils (part of the oropharyngeal lateral walls) and the epiglottis at 170 Hz and 490 Hz, respectively. Furthermore, Xu et al. show that the audio spectra of snores after upper level and lower level obstructive apnea differ [16]. Similarly, Qian et al. performed *Snore Sound* (SnS) classification by fusing different acoustic features and found spectral features to be amongst the best performing [14].

As convolutional neural networks (CNNs) have become increasingly popular in machine learning research, their application has branched out from visual recognition tasks to other areas, including audio analysis [17, 18]. Schluter et al. used a CNN on spectrograms for musical onset detection [19], whilst Eghbal et al. used them for the detection and classification of acoustic scenes and events [20]. Similar to the work proposed herein, Huang et al. used spectrograms of speech together with a CNN to perform emotion recognition [21]. Results presented in [21], indicate their CNN proposed system is more robust to noise and other confounding factors than more established computational paralinguistic paradigms.

It is worth noting that these papers all used their own custom CNN architectures and train their nets for the task at hand. As training requires a considerable amounts of data, time and computational power, research efforts into how best to leverage pre-trained CNNs for other tasks has been undertaken [22]. In this context, descriptors extracted from large pre-trained deep CNNs have become a popular and effective choice for many *visual recognition systems* [23]. Sharif et al. investigated the performance of these descriptors on a series of problems that are minimally related to the object recognition task (specifically ImageNet [24]) used for training the CNN (OverFeat [25]) they employ [26]. To date, very little research has been undertaken exploring deep CNN feature representations for audio processing; to the best of our knowledge, they have only been used together with spectrograms for Music Information retrieval [27].

Our approach to the INTERSPEECH ComParE 2017 Snoring sub-challenge is therefore motivated by the following. First, spectral features have been found to be effective for SnS analysis. Second, CNNs have been used successfully in connection with speech and audio related tasks, and finally, CNN-descriptors are dominant image processing features. Thus, we propose a system that makes use of both spectrograms and pre-trained CNNs to produce features that can be used for SnS classification.

The rest of this paper is laid out as follows: In Section 2, we describe how our deep spectrum system is built. This is followed by a description of our experimental set-up and the results our system achieved in comparison to the challenge baseline in Section 3. Finally, we draw conclusions from our results and give a perspective on future research in Section 4.

http://dx.doi.org/10.21437/Interspeech.2017-434

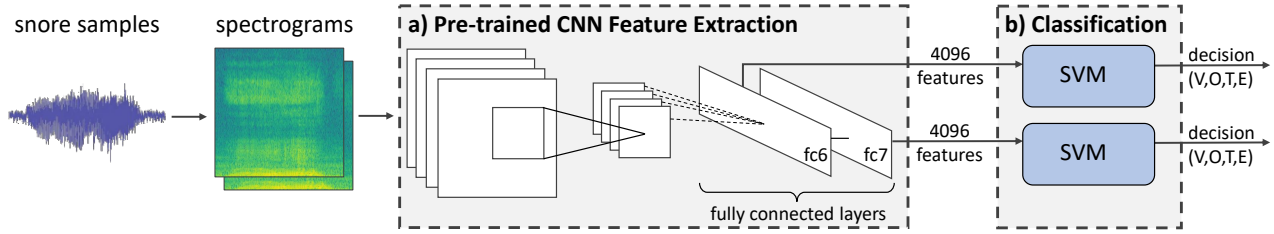


Figure 1: *Design of our proposed system. Spectrograms are generated from whole audio files and plotted with the Python matplotlib package. We then use these plots as input for pre-trained CNNs and extract the activations of fully connected layers as large deep spectrum feature vectors (a). Finally, we use these vectors to train linear support vector machine (SVM) (b).*

2. Proposed System

An overview of the proposed system is depicted in Figure 1. It consists of two main components: a) a pre-trained deep CNN which is used for the extraction of *deep spectrum* features from spectrogram plots, and b) support vector machine (SVM) classifiers which are trained for classification on the extracted feature vectors.

2.1. Spectrogram Creation

We transform the snore samples into a format that can be processed by the pre-trained CNNs by creating spectrograms of the audio files. We use Hanning windows of width 16 ms, and overlap 8 ms, and compute the power spectral density on the dB power scale. For this and the creation of the actual plots, we use the Python package matplotlib [28]. We analyse the impact of three 'standard' spectrogram colour mappings to find a good representation of the snore samples: *jet* which is the default colour map of matplotlib and varies from blue (low range) to green (mid range) to red (upper range); *gray* which is a sequential grey-scale mapping which varies from black (low range) to grey (mid range) to white (upper range); and finally, *viridis* which is a perceptually uniform sequential colour map varying from blue (low range) to green (mid range) to yellow (upper range). Further, the plots are scaled and cropped to square images without axes and margins to comply with the input needed by the CNN. Our spectrograms have an intermediate size of 387×387 pixels, and are then further scaled down to 224×224 (for VGG19) and 227×227 pixels (for AlexNet).

Example plots as used by our final system for each class of the Snoring sub-challenge are shown in Figure 2. It is worth noting that, even with the human eye, some clear distinctions between the spectrograms of different classes can be made.

2.2. Deep Spectrum Feature Extraction

Having created the spectrogram plots for the snore samples, we now use pre-trained CNNs to extract our deep spectrum features. We obtain the models and weights for AlexNet [29] and VGG19 [30] via the publicly available CAFFE [31] toolkit. AlexNet was utilised as it was the first large, deep CNN to be successfully applied to the ImageNet task in 2012[1]. VGG19 was chosen because of its popularity for creating CNN descriptors of images. Both deep CNNs were trained on approximately 1.2 million images from theImageNet corpus [29, 30].

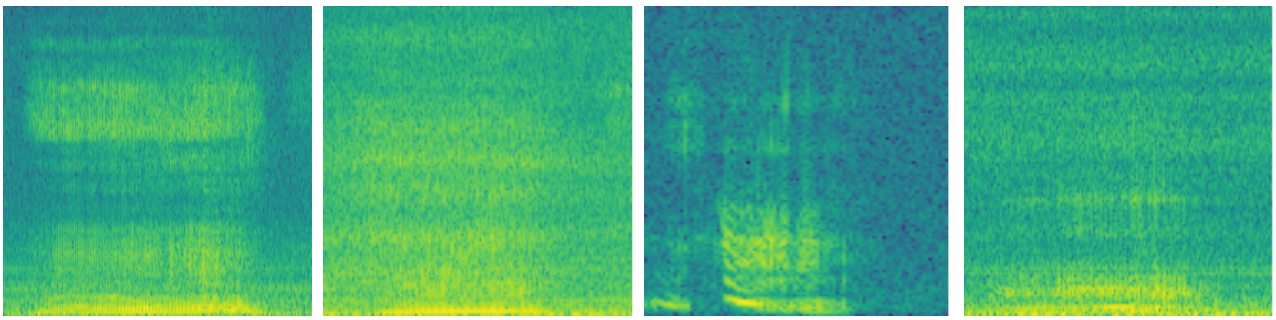
Whilst both of these nets are large deep CNNs that use a combination of convolutional, maxpooling, fully connected layers and recitified linear units [32] as activation functions, there are several key differences between their respective architectures. AlexNet consists of five convolutional layers, followed by three which are fully connected, of which the last is used to perform the 1 000-way classification on ImageNet by applying softmax. VGG19 on the other hand is made up of 19 layers that are grouped in five stacks of convolutional layers with maxpooling and, like in AlexNet, three fully connected layers. While AlexNet uses varying kernel sizes for each convolutional layer, VGG19 only uses small 3×3 kernels on all of them. An overview of similarities and differences of the two CNN architectures used in our work is given in Table 1.

[1] In both the classification and localisation tasks they secured the first place competing against traditional image analysis approaches: http://image-net.org/challenges/LSVRC/2012/results.html

Table 1: *Overview of the architectural similarities and differences between the two CNNs used for the extraction of deep spectrum features, AlexNet and VGG19.* conv *denotes convolutional layers and* ch *stands for channels.*

AlexNet	VGG19
input: RGB image	
1×conv size: 11; ch: 96; stride: 4	2×conv size: 3; ch: 64; stride: 1
maxpooling	
1×conv size: 5; ch: 256	2×conv size: 3; ch: 128
maxpooling	
1×conv size: 3; ch: 384	4×conv size: 3; ch: 256
	maxpooling
1×conv size: 3; ch: 384	4×conv size: 3; ch: 512
	maxpooling
1×conv size: 3; ch: 256	4×conv size: 3; ch: 512
maxpooling	
fully connected *fc6* 4 096 neurons	
fully connected *fc7* 4 096 neurons	
fully connected 1 000 neurons	
output: soft-max of probabilities for 1000 object classes	

For the deep spectrum feature extraction, our spectrogram plots are forwarded through the pre-trained networks and the activations from the neurons on the first and second fully connected layers (*fc6* and *fc7*) are extracted as feature vectors. The resulting feature set has 4 096 attributes - one for every neuron in the CNN's fully connected layer. Doing this for both networks and three different sets of spectrograms (one for each of the colour maps we used) results in $2 \times 2 \times 3 = 12$ distinct feature sets that can then be compared for their suitability to perform automated SnS classification [33].



(a) ***V*** *(devel_0001.wav)* (b) ***O*** *(devel_0007.wav)* (c) ***T*** *(devel_0144.wav)* (d) ***E*** *(devel_0008.wav)*

Figure 2: *Representative spectrograms for the different types of snore sounds using the best performing colour map* viridis*. Each one of the four classes relating to the point of vibration (V: velum, O: oropharyngeal lateral walls, T: tongue base, E: epiglottis) produces a unique spectral image. The samples from which these spectrograms have been extracted are given in parentheses.*

2.3. Classification

As per the INTERSPEECH 2017 ComParE Snoring sub-challenge requirements [11], we perform classification of the snore audio samples into the four classes which reflect the place of obstruction causing the snore: *Velum* (V), *Oropharyngeal* (O), *Tongue* (T), and *Epiglottis* (E). This is achieved by training linear SVM on the extracted feature sets. We use SVM because of its robustness to smaller amounts of training data.

3. Experiments

3.1. Database

The INTERSPEECH 2017 ComParE Snoring sub-challenge is based on the *Munich-Passau Snore Sound Corpus*, which contains 828 snore samples from four classes. Each one of these classes relates to one source of vibration (cf. Section 2.3). For the challenge, the corpus has been split equally into training, development, and test partitions [11].

The classes have uneven distribution, with substantially more *V* samples [11]. Therefore, we perform upsampling of our data, by replicating samples from the *O*, *T*, and *E* classes proportional to their relative frequency. We use the same upsampling factors as those used in the challenge baseline system. This results in all classes having approximately the same number of samples. For a detailed description of the corpus and the class distributions the reader is referred to [11].

3.2. Experimental Settings

We evaluate our deep spectrum features for all combinations of CNN-descriptors and spectrogram colour maps, resulting in 12 different configurations (cf. Table 2). Features are extracted from spectrograms of three different colour maps either by using fc6/7 activations of both AlexNet and VGG19. We use the LibLINEAR library with the L2-regularised L2-loss dual solver [34] via the WEKA machine learning toolkit [35] and optimise the SVM complexity parameter $C \in [10^{-6}; 10^{-1}]$ on the development partition. For each configuration, we present the two best results of adjusting C. As preliminary evaluations (results not given) indicated that normalisation and standardisation of our data negatively impacted on classifier performance; neither technique was applied. The performance of our configurations is scored using *Unweighted Average Recall (UAR)*, as specified in the challenge baseline [11].

Table 2: *Results for the snore sub-challenge using linear SVM on four different CNN-descriptors (AlexNet fc6, AlexNet fc7, VGG19 fc6, and VGG19 fc7) extracted from spectrograms with three different colour maps (gray, jet, and viridis). Unweighted Average Recall (UAR %) is used as measure and C is optimised on the development partition. The chance level is 25.0 % UAR. On the test set only five trials were allowed to be uploaded.*

CNN-descriptor	Colour Map	C	devel	test
AlexNet *fc6*	gray	10^{-1}	39.7	–
		10^{-3}	42.0	–
	jet	10^{-4}	36.2	–
		10^{-6}	36.8	–
	viridis	10^{-4}	43.5	–
		10^{-5}	41.6	–
AlexNet *fc7*	gray	10^{-1}	38.2	–
		10^{-2}	38.2	–
	jet	10^{-2}	37.4	–
		10^{-4}	38.8	–
	viridis	10^{-3}	**47.4**	63.3
		10^{-4}	44.8	**67.0**
VGG19 *fc6*	gray	10^{-3}	28.4	–
		10^{-4}	30.7	–
	jet	10^{-2}	31.2	–
		10^{-3}	31.4	–
	viridis	10^{-4}	38.5	–
		10^{-5}	37.4	–
VGG19 *fc7*	gray	10^{-2}	29.9	–
		10^{-3}	31.5	–
	jet	10^{-1}	31.7	–
		10^{-2}	31.3	–
	viridis	10^{-2}	39.5	
		10^{-3}	39.0	–

3.3. Fusion

Finally, we evaluate the effects of three different fusion scenarios on our system: First, we fuse the deep spectrum features extracted from the spectrograms of the different colour maps to investigate if a specific mapping contains important information that cannot be found in the others. Second, we perform fusion of the different CNN layers used for feature extraction. Third, descriptors of different CNN architectures are fused to analyse

Table 3: *Comparison of fusion strategies for our deep spectrum feature system. We fuse different colour mappings, layers and CNN architectures all on both feature (feat.) and decision (dec.) level. We use linear SVM for feature level fusion (optimising C on the development partition) and use the best SVM models obtained during development of the non-fusion configurations. For a detailed description of which colour maps, layers, and CNN architectures are fused the reader is referred to Section 3.3.*

Fusion Model	UAR [%]			
	devel		test	
	feat.	dec.	feat.	dec.
Colour-Map Fusion	38.1	42.2	–	–
Layer Fusion	43.8	46.1	–	63.8
CNN Fusion	44.7	46.4	57.4	62.0

if they complement each other. This leaves us with three models to evaluate: Fusing features extracted from AlexNet's *fc7* layer for all three colour maps (*Colour-Map Fusion*), combining the features extracted from both of AlexNet's fully connected layers *fc6* and *fc7* (*Layer Fusion*) and, finally, fusing features extracted from *fc7* of AlexNet and VGG19 (*CNN Fusion*). We use the best performing colour map *viridis* for layer and CNN fusion.

In all of these scenarios, both feature and decision level fusions are evaluated. We perform feature level fusions by concatenation and classification via linear SVM, and decision level fusions by majority voting of the best non-fused linear SVM configurations, i. e., we use the optimal value for C determined during development.

3.4. Results

The results of our experiments, across all 12 combinations of spectrogram colour map, pre-trained CNN, and extraction layer used to form the different deep spectrum representations, are shown in Table 2. Our best results are achieved with features extracted from AlexNet's *fc7* layer and the spectrogram colour map *viridis*. With $C = 10^{-4}$, we achieve an UAR of 44.8 % on the development, and 67.0 % on the test partition, outperforming the challenge baseline system (cf. Table 5). The confusion matrix of classification labels on the test set for this best performing system is displayed in Table 4.

Analysing the results produced by our different fusion configurations (see Table 3), we can see that fusing features extracted from spectrograms of different colour maps decreases performance for both feature and decision level fusion compared to only using the best colour map. While fusing features extracted from different layers reduces performance for early fusion, decision level fusion produces results similar to the respective single layer configuration. Here, the performance decrease might be caused by the increased feature size. Lastly, fusing the best performing features from AlexNet and VGG19 produces similar results: Decreased performance for feature level fusion and performance similar to the non fusion model for decision level fusion.

Our evaluation also shows two characteristics of interest from the extracted deep spectrum features. First, the features extracted from AlexNet perform better than those of VGG19. This is the opposite of results presented for the ImageNet task, in which VGG19 drastically outperforms AlexNet, achieving a top-1 validation error of 24.7 % and a top-5 validation error of 7.5 % compared to AlexNet's 40.7 % and 18.2 %, respectively [30]. Second, the choice of colour map in the spectrogram creation step has an observable impact on the performance of the whole system: In all but one configuration (AlexNet *fc6*) *viridis* increases UAR while a simple grey-scale mapping leads to improvements over the standard *jet* map only for AlexNet *fc6*. Since both nets are pre-trained on a large corpus of natural images, it seems to be intuitive that the choice of colour for an artificial image like a spectrogram plot would impact the models' ability to extract useful features.

Table 4: *Confusion Matrix of the best classification on the test set instances achieved by our approach with V: velum, O: oropharyngeal lateral walls, T: tongue base, E: epiglottis.*

		Actual			
	#	*V*	*O*	*T*	*E*
Prediction	*V*	96	19	1	0
	O	27	38	2	2
	T	17	3	10	2
	E	15	5	3	23
	Recall	61.9 %	58.5 %	62.5 %	85.2 %

Table 5: *Comparison of our deep spectrum based approach with the challenge baseline (functionals), and the end-to-end approach (CNN & LSTM) investigated in the baseline paper.*

Model	Ref.	UAR [%]	
		devel	test
Baseline CNN & LSTM	[11]	40.3	40.3
Baseline functionals	[11]	40.6	58.5
Deep Spectrum	Table 2	**44.8**	**67.0**

4. Conclusions

This paper proposed a method for classifying snore sounds that relies on the ability of large, deep pre-trained CNNs to extract useful information from spectrograms. Using our deep spectrum feature extraction method and linear SVM as a classifier, we were able to substantially outperform the baseline for the snoring sub-challenge which utilises classic knowledge-based audio features. In comparison to the baseline features, our system relies solely on spectral information and large, deep CNNs' ability to infer a higher level representation of arbitrary input images. In our experiments, we also found that both the choice of colour for the spectrogram plots, and the pre-trained CNN used for feature extraction has a substantial impact on performance.

Further research will include analysing which CNNs and spectrogram colour maps work best as feature extractors for computational paralinguistics tasks. We will also investigate the performance of our deep spectrum features when fused with conventional acoustic feature representations. Also of interest would be to experiment with adding a dense layer to our deep CNN to perform classification on the extracted features. Finally, we want to consider segmenting the audio files into chunks of equal length prior to generating the spectrograms, in this way providing input for long short-term memory networks.

5. Acknowledgements

This work was supported by the EU's 7^{th} Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu). and No. 688835 (RIA DE-ENIGMA)

6. References

[1] M. Armstrong, C. Wallace, and J. Marais, "The effect of surgery upon the quality of life in snoring patients and their partners: a between-subjects case-controlled trial," *Clinical Otolaryngology & Allied Sciences*, vol. 24, no. 6, pp. 510–522, 1999.

[2] R. Gall, L. Isaac, and M. Kryger, "Quality of life in mild obstructive sleep apnea." *Sleep: Journal of Sleep Research & Sleep Medicine*, 1993.

[3] M. S. Aldrich, *Sleep medicine*. Oxford University Press, 1999.

[4] I. Fietze, T. Penzel, A. Alonderis, F. Barbe, M. Bonsignore, P. Calverly, W. De Backer, K. Diefenbach, V. Donic, M. Eijsvogel *et al.*, "Management of obstructive sleep apnea in Europe," *Sleep medicine*, vol. 12, no. 2, pp. 190–197, 2011.

[5] T. Young, L. Evans, L. Finn, M. Palta *et al.*, "Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women," *Sleep*, vol. 20, no. 9, pp. 705–706, 1997.

[6] P. Jennum and R. L. Riha, "Epidemiology of sleep apnoea/hypopnoea syndrome and sleep-disordered breathing," *European Respiratory Journal*, vol. 33, no. 4, pp. 907–914, 2009.

[7] J. Lam, S. Sharma, and B. Lam, "Obstructive sleep apnoea: definitions, epidemiology & natural history." 2010.

[8] A. S. Shamsuzzaman, B. J. Gersh, and V. K. Somers, "Obstructive sleep apnea: implications for cardiac and vascular disease," *Jama*, vol. 290, no. 14, pp. 1906–1914, 2003.

[9] O. Parra, A. Arboix, J. Montserrat, L. Quinto, S. Bechich, and L. Garcia-Eroles, "Sleep-related breathing disorders: impact on mortality of cerebrovascular disease," *European Respiratory Journal*, vol. 24, no. 2, pp. 267–272, 2004.

[10] C. Croft and M. Pringle, "Sleep nasendoscopy: a technique of assessment in snoring and obstructive sleep apnoea," *Clinical Otolaryngology*, vol. 16, no. 5, pp. 504–509, 1991.

[11] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, ISCA. Stockholm, Sweden: ISCA, August 2017, 5 pages.

[12] P. Hill, B. Lee, J. Osborne, and E. Osman, "Palatal snoring identified by acoustic crest factor analysis," *Physiological measurement*, vol. 20, no. 2, p. 167, 1999.

[13] S. Miyazaki, Y. Itasaka, K. Ishikawa, and K. Togawa, "Acoustic Analysis of Snoring and the Site of Airway Obstruction in Sleep Related Respiratory Disorders," *Acta Oto-Laryngologica*, vol. 118, no. 537, pp. 47–51, 1998.

[14] K. Qian, C. Janott, V. Pandit, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller, "Classification of the Excitation Location of Snore Sounds in the Upper Airway by Acoustic Multi-Feature Analysis," *IEEE Transactions on Biomedical Engineering*, 2016.

[15] S. Agrawal, P. Stone, K. McGuinness, J. Morris, and A. Camilleri, "Sound frequency analysis and the site of snoring in natural and induced sleep," *Clinical Otolaryngology*, vol. 27, no. 3, pp. 162–166, 2002.

[16] H. Xu, W. Huang, L. Yu, and L. Chen, "Sound spectral analysis of snoring sound and site of obstruction in obstructive sleep apnea syndrome," *Acta Oto-Laryngologica*, vol. 130, no. 10, pp. 1175–1179, 2010.

[17] O. Abdel-Hamid, A. r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, Oct 2014.

[18] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, and B. Ramabhadran, "Deep Convolutional Neural Networks for Large-scale Speech Tasks," *Neural Networks*, vol. 64, pp. 39 – 48, 2015, special Issue on "Deep Learning of Representations".

[19] J. Schluter and S. Bock, "Improved musical onset detection with convolutional neural networks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Florence, ITA: IEEE, 2014, pp. 6979–6983.

[20] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU Submissions for DCASE-2016: A Hybrid Approach Using Binaural I-Vectors and Deep Convolutional Neural Networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[21] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proceedings of the 22nd ACM international conference on Multimedia*. Florida, US: ACM, 2014, pp. 801–804.

[22] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," in *International Conference on Machine Learning (ICML)*, vol. 32, Beijing, CHN, 2014, pp. 647–655.

[23] J. Deng, N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. Schuller, "The University of Passau Open Emotion Recognition System for the Multimodal Emotion Challenge," in *Chinese Conference on Pattern Recognition*. Singapore, SGP: Springer, 2016, pp. 652–666.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[26] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Ohio, US, 2014, pp. 806–813.

[27] G. Gwardys and D. Grzywczak, "Deep Image Features in Music Information Retrieval," *International Journal of Electronics and Telecommunications*, vol. 60, no. 4, pp. 321–326, 2014.

[28] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, vol. 25, pp. 1097–1105.

[30] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computing Research Repository (CoRR)*, vol. abs/1409.1556, 2014.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. Orlando, US: ACM, 2014, pp. 675–678.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, Haifa, ISR, 2010, pp. 807–814.

[33] M. Freitag, S. Amiriparian, M. Gerczuk, N. Cummins, and B. Schuller, "An 'End-to-Evolution' Hybrid Approach for Snore Sound Classification," in *Proceedings of INTERSPEECH*. Stockholm, SE: ISCA, 2017, 5 pages.

[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.