# A Bag-of-Audio-Words Approach for Snore Sounds' Excitation Localisation

*Maximilian Schmitt*[1], *Christoph Janott*[2], *Vedhas Pandit*[1], *Kun Qian*[1],
*Clemens Heiser*[3], *Werner Hemmert*[2], *Björn Schuller*[1,4]

[1]Chair of Complex & Intelligent Systems, Universität Passau, Germany
[2]Institute for Medical Engineering, Technische Universität München, Germany
[3]Department of Otorhinolaryngology/Head and Neck Surgery,
Klinikum rechts der Isar, Technische Universität München, Germany
[4]Machine Learning Group, Department of Computing, Imperial College London, UK
Email: maximilian.schmitt@uni-passau.de, c.janott@gmx.net

## Abstract

Habitual snoring and Obstructive Sleep Apnea are serious conditions that can affect the health of the snorer. For a targeted surgical treatment, it is crucial to identify the exact location of the vibration within the upper airways. As opposed to earlier work, we present the first unsupervised feature learning approach to this task based on bags-of-audio-words. Likewise, we cluster feature values within a given time-segment into acoustic 'words'. The frequency of occurrence per such word is then represented in a histogram per sound chunk to classify between four excitation locations. In extensive test runs based on snore sound data of 24 patients labelled by experts, we evaluated several feature sets as basis for audio word creation. In the result, we find audio words based on wavelet features, formants, and MFCC to be highly suited and outperform previous experiments based on the same data set.

**Keywords:** Snoring, Obstructive Sleep Apnea, Drug Induced Sleep Endoscopy, Bag-of-Audio-Words, Unsupervised Feature Learning

## 1 Introduction

Habitual snoring (chronic snoring almost every night) is a condition that affects approximately one third of the overall population [1]. Snoring severely affects the sleep quality of the bed partner [2]. Furthermore, it can be associated with Obstructive Sleep Apnea (OSA), a chronic disease that can severely affect health. OSA is defined as a syndrome with cessation or reduction of airflow during sleep due to complete (apnea) or partial (hypopnea) collapse of the upper airway for more than ten seconds and with five or more episodes per hour in sleep [3]. When untreated, OSA is an independent risk factor for cardiovascular diseases, stroke, hypertension, and myocardial infarction [4]. In more than 80 % of the cases, OSA is associated with snoring [5].

Snoring is caused by the vibration of soft tissue in the upper airways. The exact vibration location varies by patient depending on the individual anatomy. Typical areas of snoring noise generation include the soft palate, the uvula, the palatine tonsils, the base of the tongue, and the epiglottis. A variety of surgical options exist to treat snoring and OSA. The identification of the individual mechanism of snoring sound generation is vital for a targeted surgical approach. Drug induced sleep endoscopy (DISE) is increasingly used to identify the location and form of vibrations and obstructions [6]. However, DISE cannot be performed in natural sleep. Acoustic analysis could be an alternative to identify the vibration mechanisms within the upper airway.

The acoustic properties of snoring have been subject of research since the 1980s. The application of multi-feature acoustic analysis methods to determine the vibration or occlusion mechanisms has been proposed in [7]. Our group has applied machine learning models used for this purpose, comparing different acoustic feature sets for their performance in combination with frequently-used classifier models [8].

It is desirable to reduce the dimension of the feature vector used for classifier training in order to save computational effort, to achieve a robust representation of the classifier model, and to reduce the risk of overfitting (especially in smaller training databases). Aiming to achieve these goals and to further improve the classification performance, we apply an unsupervised feature representation known as the bag-of-audio-words (BoAW) approach.

The BoAW concept is inspired by an approach commonly used in text mining applications known as 'bag-of-words'. In audio signal processing, the bag-of-words concept is modified by representing audio features in the form of compact 'audio words', whereby each word corresponds to a combination of acoustic features. The classifier is then trained with a histogram representing the frequency of occurrence of the respective words, further reducing the complexity of the features used.

The efficiency of the BoAW method has been proven successfully in many audio recognition tasks, such as acoustic event detection [9–11], multimedia event detection [12–14], speech-based emotion recognition [15], and MIR [16]. Recently, the toolkit openXBOW has been introduced by part of the present authors [17], providing a comfortable way of combining arbitrary features of symbolic or numeric representations into a single bag-of-words.

The remainder of this article is organised as follows: Section 2 gives a detailed description of the data set used for the experiments; next, in Section 3, the acoustic features employed and the BoAW method are explained. Results are presented in Section 4 and discussed in Section 5. A final conclusion and an outlook on future research are given in Section 6.

## 2 Data

Our study is based on snore sounds from 24 subjects diagnosed with primary snoring or OSA who underwent drug induced sleep endoscopy (DISE) in order to determine adequate surgical intervention measures. The DISE investigation was performed using a flexible nasopharyngoscope; audio information was recorded in parallel using a headset microphone. Video and audio were synchronously stored in MP4-format.

From the video and audio recordings, the vibration sites

ISBN 978-3-8007-4275-2  

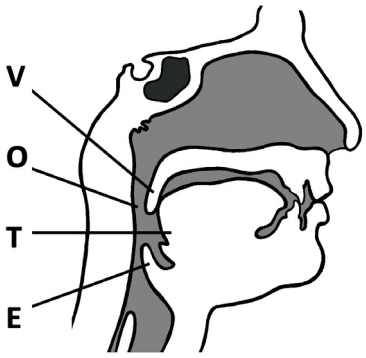


**Figure 1:** Corresponding positions of the VOTE classification in the upper airway. 'V' = soft palate level (velum). 'O' = oropharyngeal level. 'T' = tongue base level. 'E' = epiglottis level.

| Class | V | O | T | E | Total |
|---|---|---|---|---|---|
| Subjects | 14 | 4 | 2 | 5 | 24 * |
| Snore Episodes | 66 | 20 | 10 | 21 | 117 |

**Table 1:** Number of subjects and snore episodes per class. * One subject showed both E-type and V-type snoring episodes during the DISE-examination.

of snoring events were categorised by an ENT (ear, nose, and throat) expert. From each included subject, three to five snoring events have been manually selected, extracted from the audio data stream, and stored as separate audio files (sampling frequency: 16 kHz, resolution: 16 bit). Our sample set comprises 117 snoring episodes in total (length ranging from 0.31 s to 2.17 s, average 1.24 s). For details, see Table 1. Only snoring events that showed a clearly identifiable, single source of snoring have been included. Snoring events with unclear or mixed forms, e. g., several vibration sites, were excluded.

It must be noted that the 'site of vibration' generating the snore sounds and the 'site of obstruction' causing apnea are two different definitions, which may or may not coincide in individual patients. In our work, we exclusively focus on the determination of the site of vibration.

Classification of snoring sites is based on the 'VOTE' classification, introduced by Kezirian et al. for the standardisation of DISE evaluations [18]. Based on this classification, we distinguish between the velum, the oropharyngeal area, the tongue base, and the epiglottis level to distinguish different classes of snorers (see Figure 1).

In order to generate sufficient training and test instances from our data, we segmented the snore episodes into single instances of 200 ms length with an overlap of 50 % for neighbouring instances. The 24 patients were randomised into two groups for the purpose of cross-validation, i. e., each group contains the snore sound instances from 12 patients. The number of instances for each group is shown in Table 2.

| Class | V | O | T | E | Total |
|---|---|---|---|---|---|
| Group 1 | 376 | 132 | 18 | 125 | 651 |
| Group 2 | 434 | 111 | 46 | 141 | 732 |
| Total | 810 | 234 | 64 | 266 | 1383 |

**Table 2:** Number of snore instances per group per class.

# 3 Methods

## 3.1 Feature extraction

Three different kinds of acoustic features, which have been found suitable to classify snore sounds in previous experiments [8], were employed: Mel-frequency cepstral coefficients (MFCCs), formants, and wavelet-based features. All features have been computed as low-level descriptors (LLDs) over time, with a frame size of 25 ms and a hop size of 10 ms. For MFCCs and formants, a Hamming window was used for windowing.

**MFCCs** 1 to 12 and log-energy were computed using the feature extraction toolkit openSMILE [19]. A preemphasis with a coefficient of 0.97 was used in order to amplify the high frequencies.

**Formants** F1 to F3 were extracted in Matlab[1] following the method employed by Qian et al. [20]. Besides the frequencies, their amplitudes in the short-time FFT spectrum were used as features.

**Wavelet-based features** were extracted in Matlab. In [21], Khushaba et al. use 'fuzzy wavelet packet transform'-based features for monitoring of driver drowsiness from physiological signals. In our work, however, we used *Multiscale Wavelet Transform features* introduced by the same authors[2], including energy, variance, waveform length, and entropy of the decomposed signals. The wavelet decomposition is based on Matlab's Wavelet Toolbox. Only wavelets from the *Symlets* family were taken into account as they prove to work well for the task at hand [8]. This resulted in a frame-level feature vector of size 28.

As a preprocessing step, the LLD contours over time were normalised in order to ensure equal weight of all features in the audio word assignment step. This is especially important when features of different magnitudes are combined into one audio word, such as MFCCs and formant frequencies. In preliminary experiments, we found that normalisation to a range of 0 to 1 provides better results than standardisation to zero mean and unit variance. Normalisation has been done in an on-line approach, i. e., the minimum and the maximum of each LLD were derived from the respective training fold only, and then used for normalisation on both training and test fold.

## 3.2 Bags-of-audio-words

Instead of computing functionals [8], the BoAW method was employed. In Figure 2, the general process of generating a BoAW representation from a sequence of LLDs is shown. The LLDs are quantised according to a codebook learnt from the respective training fold in an unsupervised manner. Technically, any unsupervised learning scheme, such as kmeans, kmeans++ [22], and expectation maximisation clustering [10], or Deep Semi-NMF [23] can be used. It has been shown, however, that a crude random sampling of feature vectors to design a codebook competes with the computationally more expensive kmeans clustering very well [14]. Random sampling can be interpreted as the initialisation step of kmeans clustering. In our work, we used a random sampling which is alike the initialisation step of kmeans++, which we want to call 'random++'.

[1] http://mathworks.com

[2] http://rami-khushaba.com/, Matlab function: getmswtfeat()

ISBN 978-3-8007-4275-2

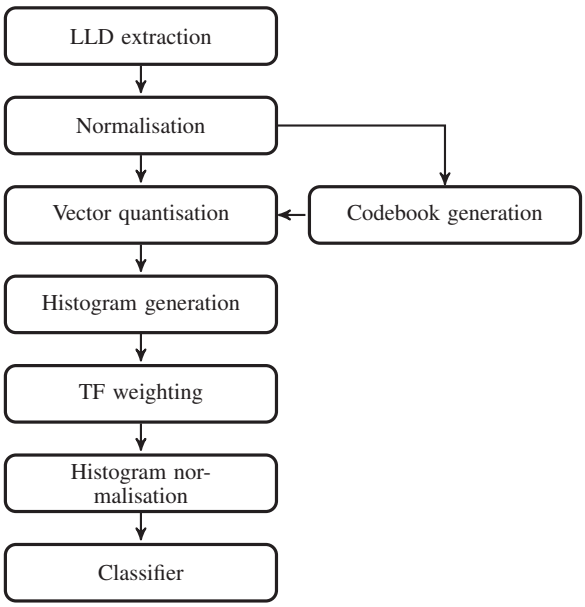


**Figure 2:** The BoAW workflow.

In fact, the generation of a proper codebook is a crucial point. As it might be the case that the codebook contains very similar audio words, a step of codebook reduction can help to make the codebook more robust and also smaller, which results in a lower-dimensional feature vector. In preliminary experiments, the linear correlation coefficient (CC) seemed to work very well in finding redundant words. Iteratively, for each audio word, the CC with all other audio words is computed and each time, the CC is above a defined threshold, the involved words are merged to create a new word, while the original words are discarded from the codebook.

In the vector quantisation step, every feature vector is now assigned to the audio word with the smallest Euclidean distance in the codebook. Finally, a histogram of the frequencies of occurrence of each word in the codebook is created for each audio segment. This kind of representation is equivalent to the bag-of-words approach known from natural language processing (NLP) [24].

Instead of taking only the closest audio word into account, each input feature may also be assigned to a certain number $N_a$ of closest words in the codebook, which we call 'multi-assignment'. It was shown that, in combination with a soft encoding of the audio words, where the term frequency in the bags is increased reciprocally proportional to the rank in distance, this can outperform the common hard encoding of only the closest audio word [14]. Another method to perform soft encoding by means of Gaussian encoding has been investigated by Pancoast and Akbacak [25].

The BoAW can be postprocessed with techniques from NLP, such as logarithmic term frequency weighting (log-TF), inverse document frequency (IDF) weighting, and histogram normalisation [24, 26]. The latter is crucial when the audio instances have different lengths, but it does not decrease the performance in case of equal-length input. In our experiments, each term frequency was divided by the number of input frames and then multiplied by the codebook size for numerical reasons.

The generation of BoAW was performed using our cross-modal bag-of-words toolkit openXBOW [17].

| Parameter | Values |
|---|---|
| $N_a$ | 1, 2, 5, 10, 20 |
| $C_s$ | 100, 200, 500, 1000, 2000 |
| $T_c$ | 0.8, 0.85, 0.9, 0.95, 1.0 |
| $C$ | $10^{-11}$, $10^{-10}$, $10^{-9}$, …, 1 |

**Table 3:** Considered ranges of the parameters for the exhaustive search.

### 3.3 Classification

Classification was done using a support vector machine (SVM) with linear kernel, where a fast implementation exists with LIBLINEAR [27]. Also the histogram intersection kernel has been tried, but the performance was worse. Optimisation of the complexity $C$ was done in the range between $10^{-11}$ and 1 (see Table 3).

Evaluations were performed with and without standardisation of the term frequencies. As we always applied histogram normalisation, which somehow restricts the ranges of the classifier input, the need of standardisation is not as significant as for feature vectors composed of different feature types.

## 4 Results

The parameters of the BoAW were optimised for each of the three acoustic feature sets and their combinations separately. The unweighted average recall (UAR) was used besides the weighted average recall (WAR) as a measure, as the 4 classes are highly imbalanced. The mean of the UARs and WARs achieved on both folds are reported in the following.

The best results using functionals by Qian et al. [8] were achieved using wavelet features. The reported results serve as a *baseline* (see Table 5) to prove the performance of our method:

| Measure | Maximum |
|---|---|
| UAR | 71.2 % |
| WAR | 78.2 % |

**Table 5:** Baseline results for the snore sounds database [8].

We did an exhaustive search in the parameter space of the number of assignments ($N_a$), (initial) codebook size ($C_s$), and threshold of correlation ($T_c$) to merge similar audio words. The considered values for each parameter are shown in Table 3. Besides, we evaluated the results with and without the weighting techniques log-TF and IDF.

For combinations of different feature types, we considered two types of fusion in BoAW. Firstly, different codebooks and BoAW are created for each feature type ('split codebooks'). The bags are then fused before putting them into the classifier. The given codebook sizes $C_s$ apply to each single codebook in this case. Secondly, a joint codebook is created for the whole feature vector consisting of several feature types.

It must be pointed out that the given codebook sizes are only the initial codebook sizes. In case of $T_c < 1.0$, the actual codebook will be smaller.

Our experiments have revealed that, performing standardisation of the term frequencies tends to provide better results for UAR, while WAR is usually higher without

ISBN 978-3-8007-4275-2

| Features | Split codebooks | log-TF / IDF | $N_a$ | $C_s$ | $T_c$ | C | UAR | WAR |
|---|---|---|---|---|---|---|---|---|
| MFCC | | log-TF | 10 | 500 | 0.8 | $10^{-3}$ | 72.5 ± 6.6 % | 75.4 ± 10.6 % |
| Formants | | | 2 | 500 | 0.95 | $10^{-5}$ | 76.4 ± 2.2 % | 78.0 ± 11.6 % |
| Wavelets | | log-TF | 5 | 500 | - | $10^{-5}$ | 73.7 ± 0.2 % | 75.5 ± 7.5 % |
| MFCC + Formants | yes | log-TF | 5 | 500 | - | $10^{-11}$ | 75.3 ± 6.0 % | 75.6 ± 12.1 % |
| MFCC + Formants | no | log-TF | 10 | 1000 | 0.9 | $10^{-3}$ | 78.3 ± 9.2 % | 78.9 ± 11.5 % |
| MFCC + Wavelets | yes | log-TF | 1 | 200 | - | $10^{-5}$ | 77.3 ± 0.3 % | 77.5 ± 8.0 % |
| MFCC + Wavelets | no | log-TF | 10 | 1000 | - | $10^{-5}$ | 78.8 ± 4.4 % | 78.2 ± 11.1 % |
| Formants + Wavelets | yes | log-TF | 5 | 500 | - | $10^{-5}$ | 78.1 ± 4.3 % | 77.4 ± 12.8 % |
| Formants + Wavelets | no | log-TF | 5 | 2000 | - | $10^{-11}$ | 78.3 ± 1.0 % | 78.7 ± 9.0 % |
| MFCC + Formants + Wavelets | yes | log-TF | 10 | 2000 | - | $10^{-6}$ | 77.9 ± 5.4 % | 77.5 ± 12.9 % |
| MFCC + Formants + Wavelets | no | log-TF | 5 | 1000 | 0.95 | $10^{-5}$ | **79.5 ± 1.2 %** | 79.7 ± 9.3 % |

**Table 4:** The best results (in terms of UAR) with corresponding standard deviation (over both folds), WAR, and configuration for the given feature sets.

standardisation. As our goal is to identify all classes of snore sounds equally well, results with standardisation are discussed in the following. In Table 4, the results with maximum UAR of all configurations for each feature type and each combination of feature types are displayed. The corresponding configurations, complexities of SVM, and the WAR are also shown.

# 5 Discussion

It is evident that using a combination of the three examined feature types performs better than using only a single type. The highest UAR of 79.5 % is achieved with a combination of all three feature types with only one codebook. It is interesting that a joint codebook seems to provide better results in all cases, even though the different distributions and properties of the LLDs cannot be taken into account so well, then. The optimum codebook sizes obviously tend to be larger in case of joint codebooks.

These findings were the same for the results without standardisation. However, performance of only MFCCs in terms of UAR was only 67.3 % and the optimum performance for a joint codebook of all feature types (UAR: 78.1 %) was achieved with a reduced codebook of a size of only 151 (fold 1) and 153 (fold 2) audio words, using a threshold of 0.95. Smaller codebooks usually have the advantage of a better generalisation, i. e., they are more robust in handling previously unseen data. The maximum WAR reached without standardisation was 81.1 %, with a joint BoAW of formants and wavelet-based features. Gaussian encoding did not have a meaningful effect on the performance, so all discussed results were achieved employing hard vector quantisation.

So far, our achievements outperform all results reported with this database in terms of both UAR and WAR. The improvement of the UAR is statistically significant with respect to the baseline and a one-sided z-test (level of significance: 0.001).

Table 6 shows the confusion matrix summed up over both folds for the best configuration with respect to UAR. Interestingly, the recall of type T (tongue) is 100 %, even though this class is the least represented in the data set. This shows that this type of snoring can be distinguished very well from the other types and that the found model does not overfit to the frequent classes too much. On the other hand, the class T-events come from only two individuals, and the results might be put into perspective when using a larger database with more tongue-base snorers.

| Predicted | → | V | O | T | E | Recall |
|---|---|---|---|---|---|---|
| Actual ↓ | V | 699 | 86 | 3 | 22 | 86.3 % |
| | O | 117 | 104 | 0 | 22 | 42.8 % |
| | T | 0 | 0 | 64 | 0 | 100 % |
| | E | 12 | 14 | 0 | 240 | 90.2 % |

**Table 6:** Confusion matrix for the best results in terms of UAR from Table 4. The sum over both folds is displayed.

Type V (velum) and Type O (oropharynx) are the two classes confused the most often. This can be explained with a view on the anatomy: the velum and the oropharyngeal area are located closely to each other within the upper airways and might therefore generate a similar frequency response.

It is an ongoing debate whether findings under DISE are comparable to natural sleep, as drug-induced sleep might induce different muscle relaxation patterns and in turn different vibration forms. However, it is supposed that the actual acoustic characteristics of a vibrating palate, tongue, or epiglottis are the same, no matter if they occur in natural or artificial sleep.

Our results are based on manually selected snoring events from a clearly identifiable, single source of vibration. In our future work, the classification task will be extended to unclear or mixed snoring forms. Further, an automated algorithmn needs to be employed to separate snoring events from non-snore sounds and periods of silence, in order to provide a feasible solution to complement current sleep analysis techniques.

# 6 Conclusions and outlook

We found that BoAW representations of wavelet-based features, formants, and MFCCs are suitable for the classification of snore sounds following the 4-class VOTE scheme. A UAR of 79.5 %, independent of the subject, could be reached, which outperforms the accuracy of previous classification experiments based on the same data set.

Future work will comprise evaluation on a newly recorded and independent database in order to show to which extent the trained model generalises. In addition, further types of acoustic features and automatic feature selection methods will be examined in the context of BoAW.

ISBN 978-3-8007-4275-2

## Acknowledgements

This work is supported by the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 645094 (IA SEWA) and by the China Scholarship Council (CSC).

## References

[1] M. Blumen, M. Quera, I. Vaugier, K. Leroux, M. d'Ortho, F. Barbot, F. Chabolle, and F. Lofaso, "Snoring intensity responsible for the sleep partner's poor quality of sleep?" *Sleep and Breathing*, vol. 16, no. 3, pp. 903–907, 2012.

[2] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," *New England Journal of Medicine*, vol. 328, no. 17, pp. 1230–1235, 1993.

[3] P. J. Strollo Jr and R. M. Rogers, "Obstructive sleep apnea," *New England Journal of Medicine*, vol. 334, no. 2, pp. 99–104, 1996.

[4] B. Mokhlesi, S. Ham, and D. Gozal, "The effect of sex and age on the comorbidity burden of osa: an observational analysis from a large nationwide us health claims database." *European Respiratory Journal*, vol. 47, no. 4, pp. 1162–1169, 2016.

[5] M. S. Aldrich, *Sleep medicine*. Transaction Publishers, 1999.

[6] M. El Badawey, G. McKee, H. Marshall, N. Heggie, and J. Wilson, "Predictive value of sleep nasendoscopy in the management of habitual snorers," *Annals of Otology, Rhinology & Laryngology*, vol. 112, no. 1, pp. 40–44, 2003.

[7] C. Janott, W. Pirsig, and C. Heiser, "Akustische analyse von schnarchgeräuschen," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 18, no. 2, pp. 87–95, 2014.

[8] K. Qian, C. Janott, Z. Zhang, C. Heiser, and B. Schuller, "Wavelet features for classification of vote snore sounds," in *Proc. 41st IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp. 221–225.

[9] A. Plinge, R. Grzeszick, and G. A. Fink, "A bag-of-features approach to acoustic event detection," in *Proc. 39th IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy: IEEE, 2014, pp. 3704–3708.

[10] R. Grzeszick, A. Plinge, and G. A. Fink, "Temporal acoustic words for online acoustic event detection," in *Proc. 37th German Conference on Pattern Recognition*, Aachen, Germany, 2015, pp. 142–153.

[11] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using lbp-hog based bag-of-audio-words feature representation," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3325–3329.

[12] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *Proc. of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 89–96.

[13] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proc. INTERSPEECH*. Portland, USA: ISCA, 2012, pp. 2105–2108.

[14] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust audio-codebooks for large-scale event detection in consumer videos," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 2929–2933.

[15] F. Pokorny, F. Graf, F. Pernkopf, and B. Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *Proc. 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015) held in conjunction with ACII 2015*, AAAC. Xi'an, P. R. China: IEEE, 2015, pp. 879–884.

[16] M. Riley, E. Heinen, and J. Ghosh, "A text retrieval approach to content-based audio hashing," in *Proc. 9th International Conference on Music Information Retrieval*, Philadelphia, PA, USA, 2008, pp. 295–300.

[17] M. Schmitt and B. W. Schuller, "openXBOW-introducing the passau open-source crossmodal bag-of-words toolkit," *preprint arXiv:1605.06778*, 2016.

[18] E. J. Kezirian, W. Hohenhorst, and N. de Vries, "Drug-induced sleep endoscopy: the vote classification," *European Archives of Oto-Rhino-Laryngology*, vol. 268, no. 8, pp. 1233–1236, 2011.

[19] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. 21st ACM International Conference on Multimedia*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[20] K. Qian, Z. Xu, H. Xu, Y. Wu, and Z. Zhao, "Automatic detection, segmentation and classification of snore related signals from overnight audio recording," *IET Signal Processing*, vol. 9, no. 1, pp. 21–29, 2015.

[21] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 121–131, 2011.

[22] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th annual ACM-SIAM symposium on Discrete Algorithms*. New Orleans, USA: SIAM, 2007, pp. 1027–1035.

[23] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. Schuller, "A Deep Semi-NMF Model for Learning Hidden Representations," in *Proc. 31st International Conference on Machine Learning*, vol. 32. Beijing, China: IMLS, June 2014.

[24] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. 10th European Conference on Machine Learning*. Chemnitz, Germany: Springer Berlin/Heidelberg, 1998, pp. 137–142.

[25] S. Pancoast and M. Akbacak, "Softening quantization in bag-of-audio-words," in *Proc. 39th IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy: IEEE, 2014, pp. 1370–1374.

[26] B. Schuller, A. E.-D. Mousa, and V. Vasileios, "Sentiment Analysis and Opinion Mining: On Optimal Parameters and Performances," *WIREs Data Mining and Knowledge Discovery*, vol. 5, pp. 255–263, September/October 2015.

[27] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

 