Snore Recognition Using a Reduced Set of Spectral Features

Enrique M. Albornoz[*], Leandro A. Bugnon[*], César E. Martínez[*†]
[*]Instituto de investigación en Señales, Sistemas e Inteligencia Computacional, sinc(*i*), UNL-CONICET, Argentina
[†]Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina
Email: emalbornoz@sinc.unl.edu.ar

***Abstract*—Snoring affects the sleep quality of the snorer itself and its social circle. Some types of snoring are related to sleep apnea, which leads to sleepiness during the day and to several health risks. Thus automatic detection of the different types of snoring may lead to more specific diagnosis and consequent treatment. In this work, we propose to use a reduced set of speech related features that includes spectral information, Mel-Frequency Cepstral Coefficients (MFCCs), prosodic values and bioinspired information. Extreme Learning Machines (ELM) are proposed to learn on the non-linear feature set. A well-known classifier as Support Vector Machines (SVM) is used as baseline. Several configurations for the feature sets and the ELM were evaluated. The bioinspired information shows promising results on the Munich-Passau Snore Sound Corpus (MPSSC) with respect to the baseline performance on the development partition.**

***Index Terms*—snore recognition, snoring classification, spectral features, auditory features**

I. Introduction

Snoring is produced by vibration of the respiratory upper way during the sleep. Beside this is not a harmful illness by itself, it leads to several affections. The sleep hygiene of the snorer is affected, leading to a poor performance during the day. Moreover, certain types of snoring are associated with the obstructive sleep apnea. Beyond reducing sleep quality the apnea causes cyclic reduction of blood oxygen saturation, which can rise several health risks as hypertension, strokes, and diabetes, among others [1]. In the social aspect, it can also disturb the bed-partner sleep and become a threat to their relationship.

Currently, there exists therapeutic and surgical methods to treat some cases of snoring. However, their effectiveness highly depends on the correct diagnosis of the snoring cause. The standard diagnostic procedure is the Drug Induced Sleep Endoscopy (DISE), in which the snoring cause is identified using video records. This technique is time consuming, requires anesthetics and an uncomfortable setting for the patient, thus finding new paths on the diagnostic tools is meaningful. Analyzing snoring using only sound records can provide a basis for studies during natural sleep. Moreover, by working on the recognition performance of snore sounds, simple pre-diagnostic evaluations could be made at home by the patient himself.

Objective measuring of snoring sounds have been already proposed to assess treatment evolution [1]. Authors stated that the pitch of snores is under 500 Hz and it relates to palatal vibration, whereas higher frequency come from other anatomic levels. Previous works mostly aim to detect snores as separate sounds from regular respiration. This detection was used mainly on sleep apnea pre-screening approaches [2], [3], [4], [5]. Others have modeled the airways obstructions as a filter and used its characteristics to identify apnea-related obstructions [3]. Several additional features have been explored such as: average power, zero-crossing rate and the lowest frequency and energy of the spectral peak with the lowest frequency [6]. A more recent work have proposed two extensive feature sets to identify snoring types [7]. The first one is a big set containing 6373 classical features obtained from statistics over low-level descriptor contours (FLLD). The second one is a bag-of-audio-words (BoAW) set, a codebook representation of audio low-level descriptor distributions [8], [9]. In addition, authors proposed an end-to-end method to avoid feature optimization. To this end, a long-short term memory neural network (LSTM) was trained using the windowed sound signal.

Besides snoring is not a speech event by itself, the snore sound is formed within the same tract, thus it define a particular speaker trait. Previous works have not explored advanced speech related features to model snore sounds. Among these features, auditive bioinspired features are robust to noise. Moreover, diagnostic information is often inferred from the bed-partners' reports about how they perceive the snore sound [10]. Therefore, the discriminative power of different bioinspired and prosodic features for snoring source identification can be further analyzed.

In this work we investigate the usability of several speech and auditory related features and propose a novel framework for snoring classification. For the feature extraction stage, we explore features that have been successfully exploited in speech and emotion recognition tasks. The sounds have common elements with paralinguistic communication as well that should be investigated [11]. Classical spectral features as MFCC, prosodic and spectral energy features of different spectral representations are evaluated. Additionally, we evaluate a bioinspired feature model extracted from the auditory model proposed in [12]. This model has been useful for feature extraction in robust speech and emotion recognition [13], [14], [15]. Using this model, we expect to mimic the auditory system and investigate if these features are useful for the recognition of the snoring type. Several combinations of these features are tested. For the classification stage, we propose the use of Extreme Learning Machines (ELM), an emerging classifier family that has shown interesting properties on several benchmarks [16]. The proposed methods are evaluated using the Munich-Passau Snore Sound Corpus (MPSSC), a challenging dataset that consist of four snoring types recorded on several subjects.

The organization of this work is as follows. In Section 2 we present the dataset used, the feature extraction and classifier methods proposed, and experimental protocol to evaluate the methods. In Section 3 we show and discuss obtained results. Finally, relevant conclusions are drawn in Section 4.

II. Materials and methods

A. Snoring dataset

The recording of snore sounds is difficult to replicate as placement of microphones, hardware and software are highly variable across different studies [1]. In an effort to improve this situation, the Munich-Passau Snore Sound Corpus (MPSSC) was publicly released [7]. This dataset consists of snore sounds recordings during the DISE procedure in different institutions. Experts labeled the samples in 4 classes according to the anatomical source of the snoring: V for Velum (palate), O for Oropharyngeal lateral walls, T for Tongue and E for Epiglottis. Public release of the dataset consist of two partitions: 282 samples for training and 283 for development. An additional test partition is provided but its use is restricted to avoid model overfitting. The partitions have highly imbalanced classes, which is consistent with clinical observations: 60% of samples from the class V, 25% class O, 11% class E and 4% class T. The snore sounds records have been already segmented, discarding events with non-stationary noise.

B. Acoustic features

The features proposed here have been of interest in different speech related tasks. Prosodic features have been extensively studied in emotion recognition [17], [18], [19]. The toolbox provided by Giannakopoulos and Pikrakis [20] was used to compute the prosodic features: zero crossing rate, energy, energy entropy, and fundamental frequency.

Additional spectral features have been included. Spectral entropy and MFCC are widely known in emotion recognition and speech analysis as well [21], [22], [23]. The first 13 MFCC were calculated within Hamming windows of 25 ms with a 25 ms frame shift. We also considered the mean of the log-spectrum (MLS) coefficients, defined as

$$S(k) = \frac{1}{N} \sum_{n=1}^{N} \log |v(n,k)|, \tag{1}$$

where k is a frequency band, N is the number of frames in the utterance, and $v(n,k)$ is the discrete Fourier transform of the signal in frame n. These were computed using spectrograms from non-overlapped Hamming windows of 25 ms. In speech related tasks, the MLS coefficients corresponding to lower frequencies ($0 - 1200$ Hz) contain the most useful information [24]. Therefore, only the first 30 MLS coefficients are extracted here.

We also propose a novel set of features based on an auditory spectrogram, using the neurophysiological model proposed by Yang et al. [25]. This model consists of two stages. The first one models an early auditory spectrum of the temporal signal at the auditory nerve level. The second stage mimics the way primary auditory cortex in mammalians process the spectrum. The first part of the model is composed of a bank of cochlear overlapping filters with center frequencies that are uniformly distributed along a logarithmic frequency axis. This process provides 128 coefficients representing the range of 0 to 4000 Hz, not equally distributed in frequencies. Given the low frequency component of the snore sounds, only the first 71 coefficients are used here. These coefficients correspond to the $[0 - 1200]$ Hz interval (same range than MLS), which proved to be satisfactory for discriminating important acoustic clues for emotion recognition [15].

We compute the mean of the log-spectrum as well from the auditory spectrogram (MLSa), which is defined as

$$S_a(k) = \frac{1}{N} \sum_{n=1}^{N} \log |a(n,k)|, \tag{2}$$

where k is a frequency band, N is the number of frames in the utterance and $a(n,k)$ is the k-th coefficient obtained by applying the auditory filter bank to the signal in frame n. The MLSa was computed using auditory spectrograms calculated for windows of 25 ms without overlapping. The sound representation in the auditory model was obtained using the Neural System Lab model[1]. All features were computed at frame level, and then the mean and standard deviation of all features over the whole utterance was calculated. Combinations of proposed features are obtained by concatenation.

C. Extreme Learning Machines

The primary implementation of ELM theory is a type of artificial neural network with one hidden layer [26]. The main differences with classical models are in the training algorithm. The hidden units are randomly generated, thus the parameter tuning of this layer is avoided. As a direct consequence, the training time is reduced significantly compared with other training methods that have to use more complex optimization techniques.

Starting with a feature set of N samples, their projection on the random hidden-layer gives the output $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^T$. The projected features are the input of a single-node layer with weights $\mathbf{W}$. In a matrix form, the network output can be written as

$$\tilde{\mathbf{Y}} = \mathbf{H}\mathbf{W}. \tag{3}$$

If the non-linear function of the nodes are infinitely differentiable and hidden weights are picked at random, it can be

[1] Neural Systems Lab., Institutes for Systems Research, UMCP. http://www.isr.umd.edu/Labs/NSL/

TABLE I
RESULTS OF SVMS CLASSIFIERS (BASELINES).

Feature set	# of features	UAR [%]
FLLD [7]	6373	40.60
BoAW [7]	500	46.60
Pros+MFCC	18	41.70
MLS	30	42.90
MLSa	71	40.60
MLS+MLSa	101	43.50
MLS+MLSa+(Pros+MFCC)	119	**48.10**

shown that the ELM can approximate the target $\mathbf{Y}$ as much as required by adjusting only the number of hidden units and the output weights [26]. The optimization problem for $\mathbf{W}$ can be written as

$$\underset{\mathbf{W}}{\text{minimize}} \quad ||\mathbf{H}\mathbf{W} - \mathbf{Y}||_2, \tag{4}$$

which is a least square optimization problem. The smallest norm solution is given by

$$\hat{\mathbf{W}} = \mathbf{H}^{\dagger}\mathbf{Y}, \tag{5}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose pseudoinverse [27]. The solution of this optimization problem is extremely fast when compared with other classical classifiers as SVM or backpropagation multi-layer perceptrons.

D. Experiments setup

To evaluate the proposed methods, different hyperparameters were used to train models with the training partitions defined in the dataset and evaluated on the development partition. The test partition was used to test only the optimal configurations as its use is restricted to avoid overfitting. The SVM classifier as well as FLLD and BoAW features were used as baseline methods.

The chosen performance measure is the Unweighted Average Recall (UAR),

$$\text{UAR} = \frac{\sum_{i}^{n_c} A_i}{C}, \tag{6}$$

where A_i is the class i recall and C is the number of classes. UAR is the average recall per class, which is more descriptive than the simple accuracy for highly class unbalanced distributions.

III. RESULTS

First, different sets of the proposed features were evaluated using the SVMs on the development partition. These baseline results are shown in TABLE I. The rows are the feature sets, where the first two are the features from [7] and the others are incremental concatenated sets as proposed in Section II-B. Columns show the size of the feature set and the UAR on development set. The FLLD seem to be outperformed by almost all of the proposed feature sets. In addition, the BoAW performance is overcome in 1.5% using all the proposed features.

TABLE II
BEST RESULTS USING THE PROPOSED FEATURE SETS ON ELM.$^{(*)}$BEST ELM CONFIGURATIONS.

Feature set	# of features	# of H.N.$^{(*)}$	UAR [%]
FLLD	6373	31865	34.68
BoAW	500	6500	42.70
Pros+MFCC	18	36	38.42
MLS	30	780	35.84
MLSa	71	142	**42.90**
MLS+MLSa	101	505	38.00
MLS+MLSa+(Pros+MFCC)	119	833	41.70

It is important to note that the proposed sets have a smaller dimension respect to the previous works. Results indicate that the proposed features (less than 2% of the FLLD dimension) improve the performance using in a similar baseline scheme. Respect to the BoAW performance, the improvement is of 1.5% using 20% of the number of inputs.

The next experiments were carried on using ELM as classifier. For every case, the number of hidden neurons ($H.N.$) was set as 1-to-30 times the input dimension. Unlike the SVM performance, the ELM had no benefit from upsampling or equalizing the training class balance. Consequently, the ELMs were trained using the unbalanced training set. Results show that the baseline is improved by using the features based on the auditory representation (TABLE II). In comparison with TABLE I, SVM get benefit from features with high dimensional space, thus the combination of all features seems to achieve better results. In the case of ELM, using only MLSa coefficients arrives to the best performance.

The MLSa was not optimized for snore sounds, but it was defined and adjusted for emotion recognition. As MLSa features have not been explored yet for the snoring task, we consider that the appropriate number of coefficients has to be optimized. Thus, the first 8 coefficients corresponding to lower frequencies were evaluated, then, the first 18 coefficients were evaluated and so on. The best results obtained for each set of MLSa coefficients are presented in TABLE III. The best results is reached using 98 coefficients giving a relative increase of 9% respect the best baseline provided, while several sets improved the FLLD performance presented in TABLE II,

In order to validate if 98 is the most suitable set size for MLSa, we reproduced the experiment but now training with the development partition and evaluating on the training partition. The same was performed for the baseline functionals and SVM. In Figure 1, the UAR scores are shown for each MLSa in comparison with the baselines. As can be seen in both experiments, the highest performance is reached for the same number of MLSa (98). The proposed features significantly improve the baseline for training-development case, while they perform well in development-training case but do not exceed the baseline.

After the previous analysis, we selected 2 models to evaluate its generalization capabilities on the testing partition, which was not used up to this point. The first model is the com-

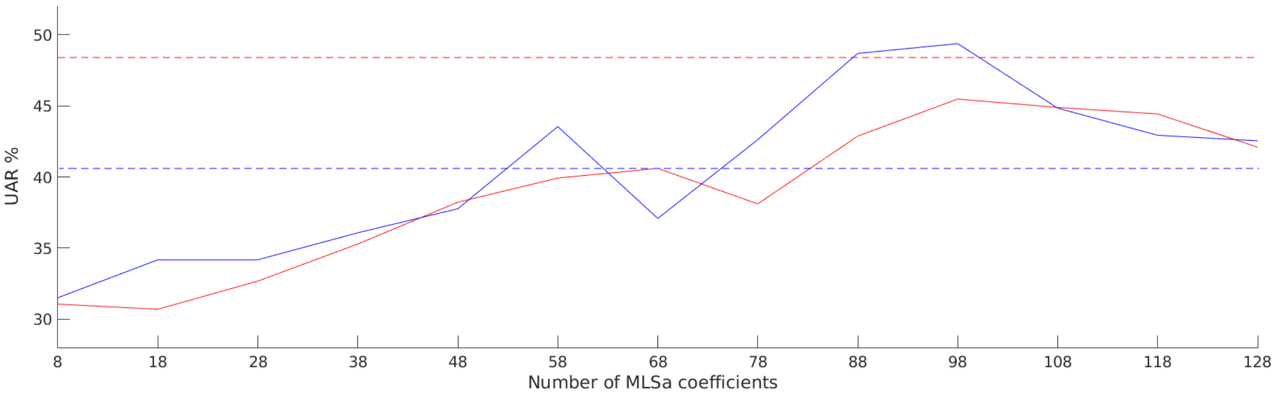

Fig. 1. UAR for Training-Development (blue) and Development-Training (red). Full lines are MLSa features with ELM, dotted lines are the best SVM baselines for reference.

TABLE III
Results for different number of MLSa coefficients on ELM (train-development). $^{(*)}$Best ELM configurations.

# of features	# of H.N.$^{(*)}$	UAR [%]
8	96	31.50
18	108	34.17
28	224	34.17
38	418	36.07
48	912	37.76
58	522	43.54
68	1360	37.08
78	234	42.62
88	2200	48.69
98	2450	**49.38**
108	756	44.84
118	3068	42.93
128	2688	42.54

plete set of proposed features MLS+MLSa+(Pros+MFCC) and SVM classifier, which obtained an UAR score of 37.89%. The difference with development results suggests that this classification framework is not generalizing well. The other model is ELM with the MLSa coefficients which were optimized in 98 coefficients. The optimal size of the random layer was chosen as the biggest number considering both previous experiments on development partition ($M = 2450$), expecting the model to be able to manage the data complexity. Results on test provided an UAR of 41.40%. This result outperform the state-of-the-art LSTM-based method presented in [7]. Other methods were reported to achieve an UAR=58.5% using FLLD and SVM [7]. However several models and their hyperparameters were tried on the test dataset and consequently the generalization of the approaches should be further investigated on independent datasets.

IV. Conclusions

In this work we proposed small feature sets based on spectral and auditory features for snore sound classification. The MLSa coefficients and ELM classifier achieved an improvement about 9% of UAR on the development partition with respect to the baseline. In addition, the size of this set is below 2% of the baseline feature set FLLD. ELM classifiers outperformed SVM when auditory features were used for this task. Moreover, ELM training and prediction algorithms are faster and require lower computer capabilities than SVM. This may lead to efficient implementations on mobiles and wearables devices for effective snoring detection.

On future works some automatic methods to reach optimised sets of features will be implemented. It would be useful to consider other models to extract features from the high frequency spectrum, related with the noisy characteristics of snore sounds.

Acknowledgements

The authors wish to thank to *Agencia Nacional de Promoción Científica y Tecnológica* (ANPCyT)(with PICT-2015-977), *Universidad Nacional del Litoral* (with PACT 2011 #58, CAI+D 2011 #58-511 and CAID-PJ-50020150100055LI) and *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET), from Argentina, for their support.

References

[1] D. Pevernagie, R. M. Aarts, and M. D. Meyer, “The acustics of snoring,” *Sleep Med Rev.;14(2):131-44*, pp. 39–58, 2010.

[2] W. D. Duckitt, S. K. Tuomi, and T. R. Niesler, “Automatic detection, segmentation and assessment of snoring from ambient acoustic data,” *Physiological Measurement*, vol. 27, no. 10, pp. 1047–1056, 2006.

[3] A. S. Karunajeewa, U. R. Abeyratne, and C. Hukins, “Multi-feature snore sound analysis in obstructive sleep apneahypopnea syndrome,” *Physiological Measurement*, vol. 32, no. 1, pp. 83–97, 2011.

[4] C. Doukas, T. Petsatodis, C. Boukis, and I. Maglogiannis, “Automated sleep breath disorders detection utilizing patient sound analysis,” *Biomedical Signal Processing and Control*, vol. 7, no. 3, pp. 256–264, 2012.

[5] E. Dafna, A. Tarasiuk, and Y. Zigel, “Automatic detection of whole night snoring events using non-contact microphone,” *PLoS ONE*, vol. 8, no. 12, 2013.

[6] “Snoring sounds variability as a signature of obstructive sleep apnea,” *Medical Engineering and Physics*, vol. 35, no. 4, pp. 479–485, 2013.

[7] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, “The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring,” in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017.

[8] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, “A bag-of-audio-words approach for snore sounds’ excitation localisation,” in *ITG Symposium on Speech Communication*, 2016.

[9] M. Schmitt, F. Ringeval, and B. W. Schuller, “At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech.” in *INTERSPEECH*, 2016, pp. 495–499.

[10] J. A. Fiz, R. Jan, J. Sol-Soler, J. Abad, M. . Garca, and J. Morera, “Continuous analysis and monitoring of snores and their relationship to the apnea-hypopnea index,” *The Laryngoscope*, vol. 120, no. 4, pp. 854–862, 2010.

[11] C. Janott, B. Schuller, and C. Heiser, “Akustische informationen von schnarchgeruschen,” *HNO;65(2)*, pp. 107–116, 2017.

[12] S. A. Shamma, R. S. Chadwick, W. J. Wilbur, K. A. Morrish, and J. Rinzel, “A biophysical model of cochlear processing: Intensity dependence of pure tone responses,” *The Journal of the Acoustical Society of America*, vol. 80, no. 1, pp. 133–145, 1986.

[13] C. Martínez, J. Goddard, D. Milone, and H. Rufiner, “Bioinspired sparse spectro-temporal representation of speech for robust classification,” *Computer Speech & Language*, vol. 26, no. 5, pp. 336–348, 2012.

[14] C. Martínez, J. Goddard, L. Di Persia, D. Milone, and H. Rufiner, “Denoising sound signals in a bioinspired non-negative spectro-temporal domain,” *Digital Signal Processing: A Review Journal*, vol. 38, pp. 22–31, 2015.

[15] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Feature extraction based on bio-inspired model for robust emotion recognition,” *Soft Computing*, pp. 1–14, 2016.

[16] G. Huang, G. B. Huang, S. Song, and K. You, “Trends in extreme learning machines: A review,” *Neural Networks*, vol. 61, pp. 32–48, 2015.

[17] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine belief network architecture,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 577–580, 2004.

[18] M. Borchert and A. Dusterhoft, “Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments,” *Proc. IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pp. 147–151, 2005.

[19] J. Adell Mercado, A. Bonafonte Cvez, and D. Escudero Mancebo, “Analysis of prosodic features : towards modelling of emotional and pragmatic attributes of speech,” 2005-09.

[20] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB® Approach*, 1st ed. Oxford, 2014.

[21] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1*, pp. 39–58, 2009.

[22] M. El Ayadi, M. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition, vol. 44, no. 3*, pp. 572–587, 2011.

[23] A. Batliner, S. Steidl, D. S. B. Schuller, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and W. N. Amir, “Searching for the most important feature types signalling emotion-related user states in speech,” *Computer Speech & Language, vol. 25, no. 1*, pp. 4–28, 2011.

[24] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,” *Computer Speech & Language, vol. 25, no. 3*, pp. 556–570, 2011.

[25] X. Yang, K. Wang, and S. A. Shamma, “Auditory representations of acoustic signals,” *IEEE Transactions on Information Theory, vol. 38, no. 2*, pp. 824–839, 1992.

[26] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[27] A. Ben-Israel and T. N. E. Greville, *Generalized inverses: theory and applications*, 2nd ed. Springer, 2001.