BIRD SOUNDS CLASSIFICATION BY LARGE SCALE ACOUSTIC FEATURES AND EXTREME LEARNING MACHINE

Kun Qian[1], *Zixing Zhang*[1,3], *Fabien Ringeval*[1,3], *Björn Schuller*[2,3]

[1]MISP group, MMK, Technische Universität München, Germany
[2]Machine Learning Group, Department of Computing, Imperial College London, UK
[3]Chair of Complex & Intelligent Systems, University of Passau, Germany

andykun.qian@tum.de, schuller@ieee.org

ABSTRACT

Automatically classifying bird species by their sound signals is of crucial relevance for the research of ornithologists and ecologists. In this study, we present a novel framework for bird sounds classification from audio recordings. Firstly, the p-centre is used to detect the 'syllables' of bird songs, which are the units for the recognition task; then, we use our openSMILE toolkit to extract large scales of acoustic features from chunked units of analysis (the 'syllables'). ReliefF helps to reduce the dimension of the feature space. Lastly, an Extreme Learning Machine (ELM) serves for decision making. Results demonstrate that our system can achieve an excellent and robust performance scalable to different numbers of species (mean unweighted average recall of 93.82 %, 89.56 %, 85.30 %, and 83.12 % corresponding to 20, 30, 40, and 50 species of birds, respectively).

Index Terms— Bird Sounds, p-centre, openSMILE, ReliefF, Extreme Learning Machine

I. INTRODUCTION

The regional activities and distributions of birds carry important information for ornithologists and ecologists measuring the biodiversity changes in a local area, which functions as an indicator to reflect the climate change [1] and habitat loss [2]. Classification of bird species by their sound singals could be a superior or essential supplementary monitoring method compared with traditional tools such as the telescope, specifically, when a bad weather condition is taken into account. Ornithologists could study the vocalisation of bird sounds for understanding of bird languages and distributions [3]. Therefore, with the advancement of signal processing and machine learning techniques, more work can be done in this promising area. Some early works were carried out by McIlraith and Card in 1997 [4]; they utilised backpropagation and multivariate statistics to get a performance ranging from 82 % to 93 % correct accuracy with 6 species of birds native to Manitoba. Kogan and Margoliash adopted Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs) to conduct a comparative study on automatic bird song recognition [5]. This method was not suitable when considering a noisy environment or with a short duration of bird sounds. A harmonic structure modelling technique was presented to be effective for classification of bird sounds in [6] and their approach appears computationally efficient. Different kinds of modelling and feature descriptor comparisons were considered by Chen et al. [7], Somervuo et al. [8], Selin et al. [9], and Lee et al. [10].Their results demonstrated a good performance in classification of some numbers of bird species by their sounds (accuracy rates are ranging from 70 % to 95 %). Ranjard and Ross considered varying characters of 'syllables'; therefore, they introduced a method based on unsupervised learning to analyse bird song evolution at different levels [11]. Large numbers of bird species were tested in the study by Lopes et al. [12]. They compared different classifiers with the same task of classification of bird sounds and scalable influence of classes (bird species) was evaluated. Neal et al. proposed an efficient method to detect syllables of bird sounds from noisy acoustic environments [13]. Graciarena et al. studied the unsupervised approach to obtain approximate note models from acoustic features extracted from bird sounds [14].

However, the works above are mainly focused on a limited scale of acoustic features and species less than 30 except [12], [14]. More recently, benchmark campaigns are further given, e. g., by the LifeCLEF Bird Identification Task [15]. This year, the work by Tan et al. proposed an algorithm that involved DTW and two passes of sparse representation (SR) classification [16], aiming to ease the problem by annotation of a large scale number of bird sounds. Related to our prior work in general acoustic event detection [17], [18], [19], we focus on the area of bird sounds classification in this paper. In our study, we evaluate a novel intelligent system for classification of up to 54 species of birds totally from a public database [20]. The main contributions of this paper are: 1) the p-centre [21] is used for detection of 'syllables' from bird sound audio recordings; 2) the openSMILE toolkit [22] is used to extract a *large* acoustic feature set; 3) Extreme Learning Machines (ELM) [23] with the ReliefF algorithm [24] are introduced for machine learning to classify bird sounds. In the following, we describe our framework and methods in the next section. Then, we show the experimental results in Section III and draw conclusions in Section IV.

II. FRAMEWORK AND METHODS

We separate our framework into the following parts: detection and segmentation of syllables (units) from audio recordings,

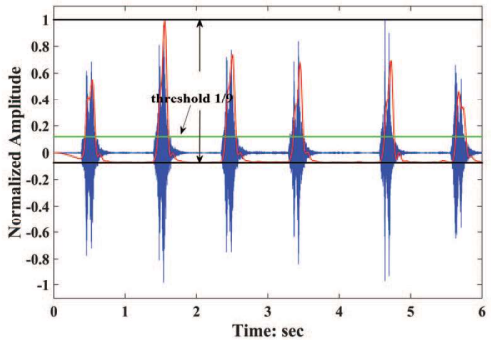

Fig. 1. *Setting the threshold for syllable detection by p-centre.*

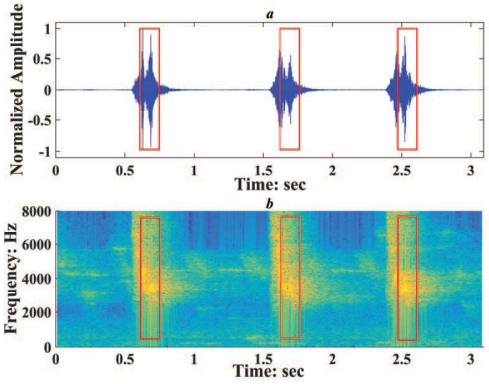

Fig. 2. *(a) Examplary detection of syllables by a p-centre-based detector from a bird sound audio recording. (b) Corresponding spectrogram to the example in (a).*

acoustic feature extraction from the units, feature space dimension reduction, and classifier training and modification.

II-A. Detection of syllables by p-centre

Similar to human speech, bird sounds consist of syllables [10], and the time durations vary from tens to hundreds of milliseconds [8]. As the basic elementary unit for further machine learning, syllables were detected by the p-centre tool, which originated in speech processing [25]. The p-centre is based on a low frequency (LF) signal filtering process, which makes it possible in the Fourier transform to estimate the values of entropy, the average frequency, and the centroid with the rhythmic envelope [25]. Compared to some classifier based methods (e. g., [13], where a Random Forest classifier was involved), p-centre based detection needs no data training phase and can be adaptive for individual audio recording conditions (e. g., the quality of audio signals, the background noise level, and the specific bird sound characters, etc.). We adopt the p-centre in our framework for the further need to process larger scales of data, rather than employing methods based on a classifier or a background noise-level estimation [8] for efficiency reasons and demonstrate its suitability and potential to this end.

The p-centre represents the prominent part of the audio signal as Figure 1 shows. We realise the required threshold setting by the p-centre, where the threshold could adaptively detect the bird song syllables within an audio recording. The audio recordings we consider are all in comparably high quality (the field noise and background interferences are in a low level condition); therefore, the threshold is set to be 1/9 (see Figure 1). To avoid some mistaken frames (e. g., some sudden background noise or interference episodes), we adopt a consecutive frames detection step (mentioned in our previous framework for snore related signals detection [26]) for this p-centre based detector. One sees in Figure 2 (*a*) there are still some missing parts both in the start and end point of the syllables, but the main energy parts of the syllables (see Figure 2 (*b*)), which could carry important information for distinction of bird vocalisations, are retained in the segmented episodes. Similar to the performance known in speech processing, by p-centre one can detect small units from bird sounds.

Table I. *Acoustic feature sets used in our experiments: Low-Level Descriptors (LLDs) and respective functionals.*

LLDs (16)	**Statistical functionals** (12)
MFCC 1–12	max, min, range, maxPos and minPos
RMS Energy	(absolue position of maximum/minimun
ZCR	value in frames), arithmetic mean, slope,
F0	offset and quadratic error for a linear approxi-
HNR	mation, standard deviation, skewness, kurtosis

II-B. Large acoustic feature space extraction

The typical machine learning in audio analysis relies on a compact, yet meaningful feature basis. Thus, acoustic features of bird sounds need to be extracted reflecting characteristics for the distinction of bird species [3]. Our open source toolkit openSMILE [22] is able to extract large spaces of acoustic features from units of analysis. We adopted the mature and frequently-used INTERSPEECH 2009 Emotion Challenge feature set [27], which contains 384 features as statistical functionals applied to low-level descriptors (LLDs). It includes the mel-frequency cepstral coefficients (MFCC) 1–12, the zero-crossing-rate (ZCR) of the time signal, the root mean square (RMS) frame energy, the pitch frequency (F0, normalised to 500 Hz) and the harmonics-to-noise ratio (HNR) by autocorrelation function. Detailed information about the features is given in Table I. For each of these, the *delta coefficients* are also computed, therefore, the whole number of attributes per feature vector is $16 \times 12 \times 2 = 384$. To our knowledge, existing research in bird sounds classification rarely uses larger spaces of acoustic features. While larger feature sets are available in openSMILE, the chosen one appears promising for the particular target task.

II-C. ReliefF algorithm for feature selection

Feature selection can be essential prior to the further machine learning [28]. In our case, we choose a the ReliefF algorithm for feature selection over reduction such as by PCA to retain the original features' physical meanings, which are significant for us in the further study. The ReliefF algorithm [24] was found efficient

for feature selection in our previous works [26] and [29]. It gives the ranking weights $W_{(i)}$ of the ith feature evaluated by an iterative process described in [30]. Here, we calculate the *contribution rate* as follows:

$$\textit{contribution rate} = \frac{\sum_{j=1}^{M} W_{(j)}^{+}}{\sum_{i=1}^{N} W_{(i)}^{+}}, \tag{1}$$

where W^{+} represents the descending sorted weights of features who give **positive** ranks for the subsequent machine learning, as evaluated by ReliefF. According to the the set of contribution rates, we select m features to obtain a subset of the original features.

II-D. Extreme Learning Machine (ELM)

The Extreme Learning Machine (ELM) has been demonstrated to provide high accuracies while remaining efficient as a classifier [23]. It has repeatedly been reported that it can achieve a higher recognition rate while being less time consuming when compared to similar classifiers such as Support Vector Machines (SVMs) or 'conventional' Neural Networks [23]. This fast and accurate method has been used widely in recent works [31], [32]. The ELM is a feedforward neural network with a single hidden layer, which randomly assigns the weights and biases of the nodes. The core idea of ELM can be simply described as a Three-Step Learning Model [23], in which the first layer (the input layer) with hidden node parameters is assigned randomly its values, thus allowing it to be regarded as an unsupervised feature mapping process [32]. Then, the hidden layer output matrix:

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}(\mathbf{x_1}) \\ \mathbf{h}(\mathbf{x_2}) \\ \vdots \\ \mathbf{h}(\mathbf{x_N}) \end{pmatrix} \tag{2}$$

is calculated. Finally, the output weights $\boldsymbol{\beta}$ are calculated. The output matrix can be thought to be 'subjective' to a supervised learning procedure [32]. The prediction of $\mathbf{x}$ is given by $\mathbf{Y} = \mathbf{h}(\mathbf{x})\boldsymbol{\beta}$, and the class with the maximum score in $\mathbf{Y}$ will be selected in multi-class classification [23], [31].

Due to the limited size of our training database, learning algorithms based on gradient descent will tend to converge to highly suboptimal solutions; however, an ELM can help overcome this effect by the use of a least squares constraint [31]. Thus, it is reasonable to choose an ELM for learning in our case rather than, e. g., a Deep Neural Network (DNN), which is rather suitable for large scales of data for training [33]. Detailed information and the theory of ELMs is provided in [23], which is beyond the scope and aim of this paper. Here, we simply propose some basic parameters in our ELM modelling: the number of hidden nodes are empirically selected (from 5 to 50 000) as 30 000, and the activation function 'radbas' [23] is chosen for its superior performance in our experiments.

III. RESULTS AND DISCUSSIONS

The dataset considered stems from a public bird sound database [18], which includes sound recordings from a total of 54 species of birds. Among the audio recordings, half of them are at a sampling rate of 44.1 kHz and the others are at 22.05 kHz. All files are 16 bit encoding and mono-channel. To avoid the impairment of audio quality, we make no changes of the original sampling rate. The time duration of the whole recordings is 1577.2 seconds (around 27 minutes, the maximum length is 237.6 seconds and the minimum is 3.3 seconds). The p-centre selected 2 135 syllables (units) totally for our training and testing sessions, in which the mean number of syllables is 39.54, with a maximum number of 326 and minimum number of 2. This is not a big database for bird sounds, however, it is sufficient to prove the feasibility and robustness of our framework, specifically, the numbers of bird species is 54 in total, which makes it a difficult classification task with several different targets related to the small scale of the dataset. We will gradually modify our system to different scales of bird sounds of consideration. Due to the limited number of training and testing syllables, we utilise a 10-fold cross validation strategy to make full use of this data. Also considering the different numbers of the syllables corresponding to each kind of birds, we conducted our evaluation method as unweighted average recall (UAR), which represents the accuracy in a dataset with equal class priors. This is especially important in our case where the class distribution is imbalanced and high accuracy could be achieved by picking the majority class. It is calculated by the sum of recall-values (class-wise accuracy) for all classes divided by the number of classes. This is the standard measure of the INTERSPEECH Computational Paralinguistics Challenge series [34]. In addition, we provide the accuracy (weighted) as complimentary results to evaluate the efficiency of our system.

Table II. *Mean UAR and Accuracy of different classifiers for 54 species of birds with all features.*

Classifiers	**UAR** %	**Accuracy** %
ELM	73.04	80.09
SVM	70.76	77.93
Ensemble	62.56	71.13
kNN	53.11	63.66

In a first experiment, we compare the performances of different frequently-used classifiers (e. g., SVM, Ensembles, kNN) with the proposed ELM-based classifier. The results of the 10-fold cross validation by mean UAR and accuracy are shown in Table II. Here, we use the following parametrisation: A one-versus-one multi-class SVM training; Bagging Decision Trees with 10 trees; the numbers of Nearest Neighbors, namely k, is set to be 10. It can be seen in Table II that ELM outperforms the frequently used SVM, a Tree-Ensemble, and a kNN classifier in our experiment. With feature selection, the ELM could achieve an enhancement of its UAR and accuracy: In our study, the ELM with ReliefF could improve by nearly 10 % absolute (from 73.04 % to 83.71 %) when compared to the baseline of only ELM training. The trend of mean UAR (with 54 species of birds) corresponding to different feature numbers (percent in total feature numbers) decided by the ReliefF algorithm is illustrated in Figure 3. It appears intuitive that, with the reduction of redundant features, the classifier could achieve a better

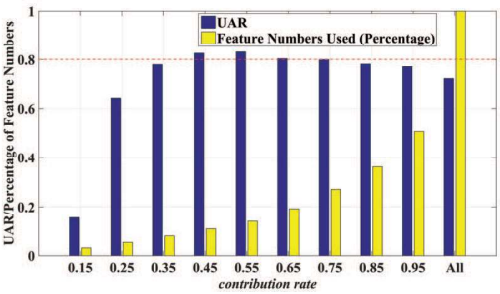

Fig. 3. *Mean UAR (10-fold cross validation) with different feature numbers selected by the ReliefF algorithm.*

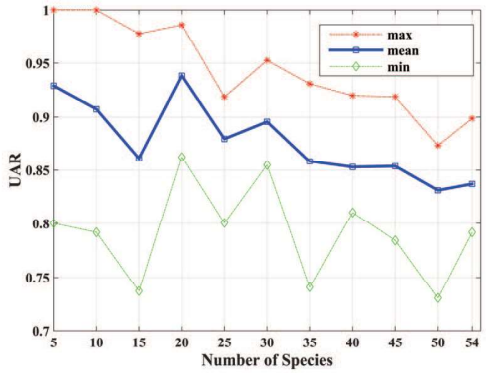

Fig. 4. *UAR for different numbers of bird species by 10-fold cross validation.*

performance in classification. We obtained higher than 85 % mean UAR with less than 50 features (less than 15 % of the total feature number, with a contribution rate of 0.55, shown in Figure 3). We set the target feature number by the ReliefF algorithm by ranging the *contribute rate* from 0.1 to 1.0 (step length is 0.05). For better visibility, we only present part of these results in Figure 3). Note that, the *contribution rate* at 1.0 (not seen in Figure 3) is the representation of all *positive* ranked features [24]. Figure 4 gives the results of UAR (maximum, mean, and minimum values of the 10-fold cross validation) of the ELM-based classifier with the ReliefF algorithm for different numbers of species of birds (ranging from 5 to 54). It can be observed that, our proposed system obtains a high performance with species below 45 (mean UAR from 85.30 % to 93.82 %). The results remain comparably high with the number of species of birds ranging up to 50 (83.12 %) and 54 (83.71 %) in total. Table III gives both, the mean UAR and Accuracy for 10, 20, 30, 40, 50 and 54 species of birds, which proves the efficiency of our framework. In our study, we found that **MFCC** and **ZCR** related attributes are ranked as highly contributing features.

IV. CONCLUSIONS

In this study, we propose a novel framework for automatic classification of bird sounds, which is composed of a p-centre detector, a larger space feature extraction and corresponding feature selection (ReliefF), and a state-of-the-art efficient machine learning classifier

Table III. *Mean UAR and accuracy for different species of birds (randomly selected).*

Species	UAR %	Accuracy %
10	90.74	94.71
20	93.82	93.91
30	89.56	89.56
40	85.30	89.03
50	83.12	85.60
54	83.71	86.57

– an extreme learning machline. We firstly utilised p-centre for segmentation. Then, with our openSMILE toolkit, we extracted a larger scale of acoustic features from bird sound syllables as units. Compared to the traditional state feature sets, this kind of statistic feature set applies functionals to Low-Lever Descriptors. After that, the feature selection phase improved the baseline of the trained classifier with a 10 % baseline enhancement and more than 85 % of the features reduced. The ELM has been chosen as classifier due to its fast and accurate performance. We compared the mean UAR and accuracy with often seen alternatives such as SVM, a Tree-Ensemble, and kNN – the results show the superior ability of the ELM. Overall, the experiments have demonstrated the efficiency of our methods and its robustness is shown by increasing the number of bird species (a maximum of 54 achieving a mean UAR of 83.71 %). In future studies, we will consider much larger datasets for our system training (up to several 100s of bird species with longer time durations of the audio recordings). Also, we will take low quality audio environments (with noise and field interference as well as presence of other species) into account, which is practical for real life use. In addition, a deeper investigation on the relevance and suitability of specific acoustic features which make a significant contribution to the distinction of the bird species is worth to explore, which could be helpful for ornithologists to conduct research related to the bird vocalisation mechanism and bird species evolution history.

V. ACKNOWLEDGEMENTS

This work is supported by China Scholarship Council (CSC), the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC Starting Grant iHEARu) and No. 645378 (RIA ARIA-VALUSPA).

VI. REFERENCES

[1] C. Parmesan and G. Yohe, "A globally coherent fingerprint of climate change impacts across natural systems," *Nature*, vol. 421, no. 6918, pp. 37–42, 2003.

[2] A. Balmford, R. E. Green, and M. Jenkins, "Measuring the changing state of nature," *Trends in Ecology & Evolution*, vol. 18, no. 7, pp. 326–330, 2003.

[3] C. K. Catchpole and P. J. Slater, *Bird song: biological themes and variations*. Cambridge, UK: Cambridge university press, 2003.

[4] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2740–2748, 1997.

[5] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study," *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2185–2196, 1998.

[6] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proc. of ICASSP*, Montreal, Canada, 2004, pp. 701–704.

[7] Z. Chen and R. C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2974–2984, 2006.

[8] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2252–2263, 2006.

[9] A. Selin, J. Turunen, and J. T. Tanttu, "Wavelets in recognition of bird sounds," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 141–141, 2007.

[10] C.-H. Lee, C.-C. Han, and C.-C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1541–1550, 2008.

[11] L. Ranjard and H. A. Ross, "Unsupervised bird song syllable classification using evolving neural networks," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4358–4368, 2008.

[12] M. T. Lopes, L. L. Gioppo, T. T. Higushi, C. A. Kaestner, C. Silla, and A. L. Koerich, "Automatic bird species identification for large number of species," in *Proc. of IEEE ISM*, Montreal, Canada, 2011, pp. 117–122.

[13] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 2012–2015.

[14] M. Graciarena, M. Delplanche, E. Shriberg, and A. Stolcke, "Bird species recognition combining acoustic and sequence modeling," in *Proc. of ICASSP*. Prague, Czech Republic: IEEE, 2011, pp. 341–344.

[15] M. Lasseck, "Large-scale identification of birds in audio recordings," in *Working notes of CLEF 2014 conference*, Sheffield, UK, 2014.

[16] L. N. Tan, A. Alwan, G. Kossan, M. L. Cody, and C. E. Taylor, "Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1069–1080, 2015.

[17] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in *Proc. of ICASSP*. Prague, Czech Republic: IEEE, 2011, pp. 337–340.

[18] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic recognition of emotion evoked by general sound events," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 341–344.

[19] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *Proc. of ICASSP*, Kyoto, Japan, 2012, pp. 333–336.

[20] "Gallery of living nature: Voices of birds," 2015. [Online]. Available: http://gallery.new-ecopsychology.org/en/voices-of-nature.htm

[21] S. Tilsen and K. Johnson, "Low-frequency fourier analysis of speech rhythm," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. EL34–EL39, 2008.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the munich versatile and fast open-source audio feature extractor," in *Proc. of ACM MM*, Firenze, Italy, 2010, pp. 1459–1462.

[23] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[24] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[25] F. Ringeval, M. Chetouani, and B. W. Schuller, "Novel metrics of speech rhythm for the assessment of emotion." in *Proc. of INTERSPEECH*, Porland, OR, 2012, pp. 2763–2766.

[26] 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.

[27] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *Proc. of INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.

[28] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[29] K. Qian, Z. Xu, H. Xu, and B. P. Ng, "Automatic detection of inspiration related snoring signals from original audio recording," in *Proc. of ChinaSIP*. IEEE, 2014, pp. 95–99.

[30] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[31] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. of the 4th International Workshop on Audio/Visual Emotion Challenge*, Orlando, FL, 2014, pp. 65–72.

[32] H. Kaya and A. A. Salah, "Combining modality-specific extreme learning machines for emotion recognition in the wild," in *Proc. of ICMI*, Istanbul, Turkey, 2014, pp. 487–493.

[33] N. Jaitly, P. Nguyen, A. W. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. of INTERSPEECH*, Portland, OR, 2012, no pagination.

[34] B. W. Schuller, "The computational paralinguistics challenge [social sciences]," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, 2012.