Department

Model ensemble for predicting heart and respiration rate from speech

Stavros Ntalampiras
University of Milan

***Abstract*—Stress levels comprise a significant source of information in assessing human wellbeing including both mental and physical health. Interestingly, speech signals can be indicative of stress and may be used to infer related physiological markers, such as heart rate and respiration cycles. To this end, this work proposes a non-intrusive, low-cost, and automatic stress monitoring framework facilitating timely activation of stress relief methods and/or stress prevention. Initially, we design a multidomain speech feature extraction scheme able to reveal complementary stress-related characteristics. Subsequently, these are modeled by a synergistic framework able to encode both linear and non-linear relationships via suitably-learned support vectors and recurrent neural network. We employed an appropriate corpus encompassing recordings of job interviews, constructed based on a standardized experimental protocol. Importantly, such an approach outperformed the state of the art by 12.3% and 33.3% in predicting heart rate and respiration respectively.**

UNFORTUNATELY, the ever-increasing pace of modern life could rise the stress levels, thus having a direct negative effect on both mental and physical health [1]. At the same time, there are methods aiming at reducing stress levels, while maintaining an acceptable efficiency in the workplace [2], so that burnout state is avoided [3]. In this context, speech signals can be employed to monitor human wellbeing [4] including physiological indicators of stress, i.e. heart rate and respiration cycles [5]. As this could comprise an important input for pervasive healthcare technologies, this work presents a cooperative framework exploiting knowledge available in speech signals to predict physiological parameters. Importantly, such a solution is completely non-intrusive and at the same time, highly cost-effective. Interestingly, such stress monitoring frameworks can assist the timely activation of methods designed to lower stress levels with significant applications on the entire branch of speech-based Human Computer Interaction.

Being a relatively new problem, the literature includes a limited amount of works addressing it. To the best of our knowledge, there are two related articles: 1) in [6] the authors extract EGEMAPS speech features (designed to address

 Published by the IEEE Computer Society

generic speech-based applications) in combination with AdaBoost and a decision tree regressor, and 2) in [7] the authors employ the same feature set and a long short-term memory neural network trained for regression. Interestingly, in this case the employed dataset is available for research purposes and thus used in the present work as well, in order to achieve directly comparable results. It should be mentioned that there are works focusing only heart rate prediction, e.g. [8], where, unfortunately, the considered dataset is unavailable [9].

The framework proposed in this work introduces the following novel aspects:

- rigidly formalizes the specific problem,
- proposes a suitable multidomain characterization of the speech signals aiming at capturing stress-related aspects, and
- designs an ensemble of models able to efficiently capture existing linear and non-linear relationships.

More specifically, we employ features coming from the cepstral domain (Mel-Frequency Cepstral Coefficients), the wavelet domain (Perceptual Wavelet Packets), and descriptors based on the Teager energy operator [10]. Subsequently, they are fed to an ensemble composed of a Support Vector Machine and an Echo State Network explaining both linear and non-linear relationships existing across modalities. We carried out thorough experiments on a publicly available dataset of synchronized speech, heart rate (measured in beats per second), and respiration cycles (chest displacement measurements at a range of -10 to +10 mV). Interestingly, the proposed approach provides promising results outperforming the state of the art.

The rest of this paper is organised as follows: the next section formalizes the present problem. Subsequently, we present the multidomain feature set along with the model ensemble. We then describe the experimental protocol and analyze the obtained results. In the final section, we draw our conclusions.

Problem Formulation

Let us consider a human speech signal over time s^t, the heart rate h^t, and continuous respiration r^t. Without loss of generality, we assume these signals are synchronized. We aim at a model $\mathcal{M}$ learning to predict physiological indicators of stress levels, i.e. h^t and r^t, from the associated speech signal s^t, i.e. $[h^t, r^t] = \mathcal{M}(s^t)$. In other words, $\mathcal{M}$ learns to transfer knowledge available in one modality to predict a different one. Given a series of k values representing the speech signal

$$s^1, s^2, ..., s^k,$$

where $t \in [t^s, t^e]$, t^s denotes the starting point in time and t^e and final one. The goal of the prediction model is to estimate the values

$$h^1, h^2, ..., h^l \text{ and } r^1, r^2, ..., r^n$$

while trying to minimize a figure of merit on the pointwise reconstruction error. The quantity of the points to be predicted depends on the sampling rate of each signal which typically are different and depend on the characteristics and variability of each physiological indicator [11].

The proposed solution

This section presents the pipeline of the proposed methodology describing the two fundamental processing stages, i.e. feature extraction and model ensemble learning.

Feature extraction

We employed three feature sets capturing diverse aspects of speech signals produced under stress.

- Mel-Frequency Cepstral Coefficients: the first feature set originates from speech/speaker recognition while it has been proven to be efficient in a wide range of generalized sound classification tasks including respiratory sounds [12]. MFCCs essentially comprise a compacted version of the spectrogram; to extract the coefficients, the spectrogram is first passed through the Mel-filterbank so that the frequency scale matches better the human hearing system; then, the log-operator is applied to adequately space the data, which is finally summarized using the Discrete Cosine Transform. The first 13 coefficients are kept along with their velocity since their evolution over time carries significant importance when processing speech.
- Teager Energy Operator autocorrelation envelope: the second feature set is based on

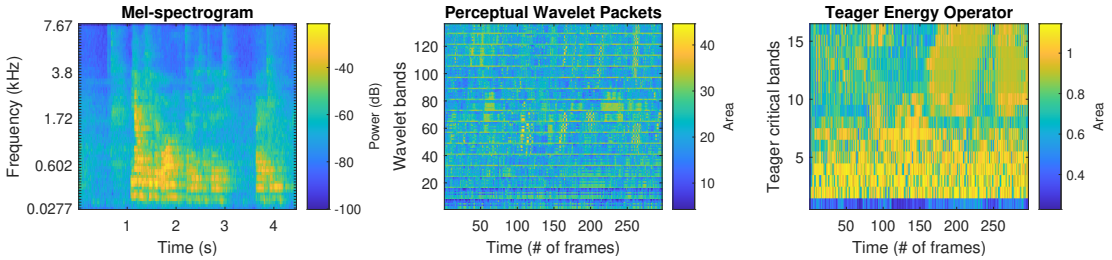

Figure 1. Illustration of the three feature sets, i.e. Mel-spectrogram, Perceptual Wavelet Packets, and Teager Energy Operator Autocorrelation Envelope, extracted from a speech segment.

the analysis carried out via the Teager energy operator [13]. It has been designed to process speech produced under stressful conditions by suitably modeling the associated airflow patterns. As such, it includes information which may not be captured by the MFCCs and could be useful for the scope of the present analysis. To extract the TEO autocorrelation envelope, the audio signal is initially filtered into sixteen critical bands. Subsequently, the area under the autocorrelation is calculated for each frame and divided by half the frame length for normalization purposes. The obtained dimensionality is sixteen, i.e. equal to the number of critical bands.

- Perceptual Wavelet Packet features: the third feature set aims at capturing complementary information with respect to the previous ones, thus it is based on wavelet analysis based on functions which are not smooth and symmetrical, such as those used in Fourier-based analysis. To this end, the Discrete Wavelet Transform (DWT) is employed which splits the audio signal into shifted and scaled versions of the Daubechies 1 (or Haar) function. We exploit both low and high frequency information as extracted via a three-stage filtering of the respiratory sound. Given that such signals may exhibit irregularities, DWT is able to process highly non-stationary audio signals at several diverse frequencies. Wavelet analysis of three levels is used, while after each level, we downsample by a factor of 2 each DWT coefficient following the Nyquist theorem. Then, for each frame we calculate the autocorrelation envelope area and normalize it by half the frame size. The dimensionality is equal to the number of critical bands times the number of the wavelet coefficients, i.e. 136. The specific feature set provides a complete description of the speech content in critical spectral areas which are perceptually motivated [14]. It reflects upon the degree of variability of a specific wavelet coefficient within a frequency band.

Fig. 1 illustrates the above-explained feature sets extracted from a speech segment. We observe that they are significant differences among them since they capture different characteristics of the audio structure. More specifically, a) the Mel-spectrogram captures the way the energy is distributed on the Mel-bands, b) PWP demonstrates how each wavelet coefficient changes based on an analysis carried out in three levels, while c) TEO approximates the alternations that the airflow path undergoes when the subject is under stress. Overall, they provide complementary information and their combination may be effective in addressing the task at hand.

Model ensemble

Given the complex nature of the relationships under approximation, we propose to exploit a relatively recent type of modeling, namely *ensemble modeling* [15]. The aim of the specific modeling logic is to effectively combine the strengths of independent models towards a finer estimation of the underlying relationship. To model the relationship between the produced speech s^t and the associated physiological parameters h^t and r^t, we rely on a combination of a linear and non-linear modeling types, i.e. a linear SVM and an ESN. Naturally, this set of models can be straightforwardly extended to consider additional

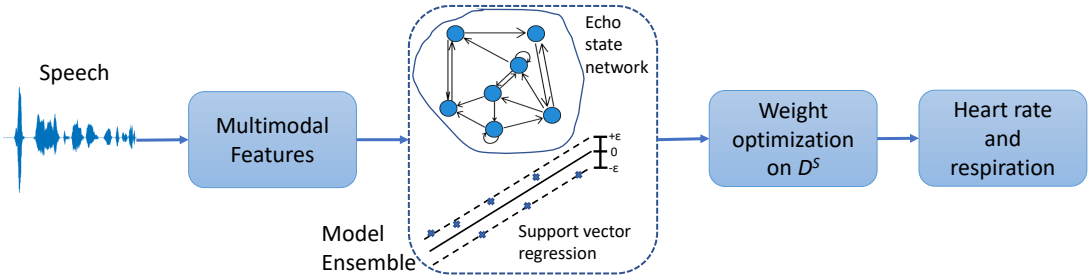

Figure 2. Block diagram of the proposed solution for predicting physiological parameters from speech signals.

and potentially diverse models, such as state space, extreme learning machines, deep models, etc.

Support Vector Regression

Similar to the respective classifier, SVR aims at identifying the hyperplane fitting the maximum amount of points. In this case, the term *support vectors* refers to the points within the training data which are relevant during the approximation of the optimal hyperplane. After early optimization experiments and based the related work presented in [7], we employed an SVR with a linear kernel denoted as $\mathcal{R}^s$. Such an SVR estimates the hyperplane using the training set, optimized on the development set and, when applied on the testing one, provides predictions of the physiological parameters, $\hat{h_s^t}$ and $\hat{r_s^t}$, i.e.

$$[\hat{h_s^t}, \hat{r_s^t}] = \mathcal{R}^s(s^t) \quad (1)$$

These estimates are then used by the ensemble producing the final predictions.

Echo State Network

Given the complex nature of the relationship under examination, we include a neural network based regressor able to capture non linear relationships. Among the wide variety of neural networks, we chose Echo State Networks, which have provided state of the art performance in many diverse tasks [16] while being inexpensive to both train and test. Interestingly, an ESN includes a reservoir network composed of neurons with randomly-generated, i.e. not trained, weights. Such neurons include non linear activation functions, while recurrent connections are allowed. Interestingly, in learning ESNs the only part which requires training are the weights of the output layer, i.e. the connections to physiological parameters, which are learned by means of linear regression. Thus, the ESN outputs $\hat{h_e^t}$ and $\hat{r_e^t}$, i.e.

$$[\hat{h_e^t}, \hat{r_e^t}] = \mathcal{R}^e(s^t) \quad (2)$$

The estimated values are then fed to the ensemble and contribute to the final predictions. The parameters of the ESN are optimized via a grid search carried out on the developments set as described in the experiments section.

Aggregating the individual models

The ensemble of models employs the estimations of independent models in order to achieve a finer prediction, i.e. $\mathcal{E} = \{\mathcal{R}^s, \mathcal{R}^e\}$. The aggregation process has a significant impact on the generalization abilities of $\mathcal{E}$ [15].

A straightforward way of aggregating is by means of a weighted average, i.e.

$$\hat{h^t} = \omega_e^h \times \hat{h_e^t} + \omega_s^h \times \hat{h_s^t}, \quad (3)$$

$$\hat{r^t} = \omega_e^r \times \hat{r_e^t} + \omega_s^r \times \hat{r_s^t}, \quad (4)$$

where ω_j^i represents a weight associated with physiological parameter i and model type j. Such weights are constrained as follows $\omega_e^h + \omega_s^h = 1$ and $\omega_e^r + \omega_s^r = 1$. Naturally, weight values play a significant role as regards to the performance offered by the ensemble.

There are several methodologies for determining the weights; they might range from being simply the average, i.e. 1/2 in this case, to Akaike's values [17]. Given that in the specific case, there is no closed-form approach providing a globally-optimum solution, we designed an estimation

Table 1. Root Mean Square Errors of the proposed approach and the contrasted one on predicting Heart Rate. The minimum errors per set are emboldened.

Approach	Development set	Test set
Multidomain+SVR	19.6	20.1
Multidomain+ESN	19.11	20.52
Multidomain+Ensemble	**18.24**	**19.9**
MFCC+Ensemble	19.02	20.71
TEO+Ensemble	19.56	20.52
PWP+Ensemble	19.12	21.18
EGEMAPS+LSTM [7]	19.32	22.7

scheme based on the ensemble reconstruction errors,

$$[\bar{\omega}_e^h, \bar{\omega}_s^h] = \arg\min_{\omega_e^h, \omega_s^h} \frac{1}{2} \sum_{k=1}^{|D^s|} (h^t - \hat{h^t})^2 \quad (5)$$

$$[\bar{\omega}_e^r, \bar{\omega}_s^r] = \arg\min_{\omega_e^r, \omega_s^r} \frac{1}{2} \sum_{k=1}^{|D^s|} (r^t - \hat{r^t})^2 \quad (6)$$

on the development set of data D^s. As such, weights ω_j^i are determined as those minimizing the quadratic errors $(h^t - \hat{h^t})^2$ and $(r^t - \hat{r^t})^2$ on D^s.

Experimental set-up and Results

This section describes the 1) employed dataset, 2) parameterization of the proposed method, 3) experimental results and analysis.

The dataset

We employed the Ulm-Trier Social Stress Test dataset where subjects undergo the well-accepted Trier Social Stress Test [18]. The considered scenario was a job interview where the speaking language is German, while recordings were carried out in southern Germany. There are 69 subjects with a male to female ratio equal to 20:49 and average age 25.06 years. The total duration is 5 hours and 47 minutes sampled at 16KHz, while the dataset is divided into 41 sessions for training (T_r^s), 14 for development (D^s), and another 14 for testing (T^s) purposes. The average sample duration is 6.24 ± 3.22 seconds across all subjects. The interested reader is referred to [19], [20] for more information where a standardized experimental protocol including suitable figures of merit.

Parameterization of the proposed solution

After early experimentation, features are extracted on windows of 100ms with 50% overlap. As regards to the SVM, the following parameters are tuned via a an exhaustive search using D^s: a) C: a parameter determining the focus placed on minimizing the fitting error on T_r^s (search space $[0.1, 1, 10, 100, 1000]$), and b) γ: it regulates the curvature of the fitting hyperplane (search space: $1/|f_v| \times var(f_v)$, scale: $1/|f_v|$, where f_v denotes the feature vector).

The ESN is parameterized in a) reservoir size [50:10:1000], b) input scaling factor [0.1, 0.5, 0.7, 0.95, 0.99], and c) spectral radius [0.7, 0.8, 0.9, 0.95, 0.99]. The combination of parameters minimizing the root mean square error on D^s is ultimately selected. It should be mentioned that deep architectures with recurrent connections have also been explored but were prone to overfitting caused by the unavailability of a significant amount of training data.

Results

In order to be fully comparable with the related literature, the proposed method respected the predetermined dataset division. Our method was contrasted with the one presented in [7] employing a long short-term memory (LSTM) network architecture trained on EGEMAPS features. The employed figures of merit are Root Mean Square Error for the h^t and Normalized Root Mean Square Error in (-10:10) for r^t. It should be mentioned that RMSE is suitable for evaluating the prediction of continuous-time signals, e.g. heart rate, and provided the range of values characterizing the respiration signals, the normalized version was employed.

The achieved results and the state of the art are tabulated in Tables 1 and 2 for heart rate and respiration prediction respectively. Overall, we observe that independent models, i.e. SVR and ESN, outperform the [7] which demonstrates the importance of the multidomain feature set able to consider diverse aspects of the audio structure. It is worth noting that the values of the figures of merit computed on the development and testing sets of data are close meaning that the models are able to generalize well over unknown subjects. Interestingly, the leading performance is offered by the ensemble which is able to efficiently com-

Table 2. Normalized Root Mean Square Errors of the proposed approach and the contrasted one on predicting Respiration. The minimum errors per set are emboldened.

Approach	Development set	Test set
Multidomain+SVR	0.09	0.09
Multidomain+ESN	0.08	0.09
Multidomain+Ensemble	**0.08**	**0.08**
MFCC+Ensemble	0.10	0.10
TEO+Ensemble	0.09	0.10
PWP+Ensemble	0.08	0.09
EGEMAPS+LSTM [7]	0.12	0.12

bine the independent models, thus offering the lowest errors with respect to both tasks. Overall, the obtained performance shows that predicting such physiological markers of stress from speech signals comprises a challenging task.

Ablation study

In order to gain better understanding of the obtained performance, an ablation study was carried out assessing the contribution of each feature set in predicting each physiological parameter.

In Table 1, we observe that each individual feature set offers satisfactory performance, while the MFCCs achieve the lowest RMSE. However, their combined representation offered a significant boost on both development and test sets. As regards to predicting respiration (Table 2), we observe that MFCCs and TEO autocorrelation envelope reach similar performance levels, while the lowest NRMSE are offered by the PWP-based feature set.

The prediction results of physiological parameters confirm that the combination of feature sets capturing such diverse aspects of the audio structure is beneficial.

Conclusions and Future Work

This article presented a solution for inferring physiological indicators of stress from the emitted speech signals. It was shown that speech encompasses relevant information for the effective prediction of the associated heart and respiration rate. Importantly, the proposed method is able to consider diverse aspects of speech signals while combining the benefits of heterogeneous models via a suitably-designed synergistic ensemble.

Audio signals constitute a non-intrusive and low-cost way for predicting stress markers. Such a research path may comprise a fruitful direction is assessing stress in real world conditions and potentially providing relevant information for a vast gamut of healthcare technologies. Nonetheless, given the great variability with which stress manifests in humans, constructing models capable of generalizing properly over novel data constitutes a challenging task. We argue that there is significant room for improvement and a fertile path could be the investigation of subject-depended modeling approaches, where deep architectures may be considered as long as a suitable amount of data becomes available.

Acknowledgment

This work was carried out within the project entitled “Advanced methods for sound and music computing” funded by the University of Milan.

REFERENCES

1. F. Hutmacher, “Putting stress in historical context: Why it is important that being stressed out was not a way to be a person 2, 000 years ago,” *Frontiers in Psychology*, vol. 12, Apr. 2021. [Online]. Available: https://doi.org/10.3389/fpsyg.2021.539799
2. S. Betti, R. M. Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. D. Ry, and F. Cavallo, “Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers,” *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1748–1758, 2018.
3. J. C. Fendel, J. J. Bürkle, and A. S. Göritz, “Mindfulness-based interventions to reduce burnout and stress in physicians: A systematic review and meta-analysis,” *Academic Medicine*, vol. 96, no. 5, pp. 751–764, Apr. 2021. [Online]. Available: https://doi.org/10.1097/acm.0000000000003936
4. X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, “Modeling of physical characteristics of speech under stress,” *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1801–1805, 2015.
5. A. J. A. A. Guyon, R. Cannavò, R. K. Studer, H. Hildebrandt, B. Danuser, E. Vlemincx, and P. Gomez, “Respiratory variability, sighing, anxiety, and breathing symptoms in low- and high-anxious music students before and after performing,” *Frontiers in Psychology*, vol. 11, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2020.00303
6. A. Jati, P. G. Williams, B. Baucom, and P. Georgiou, “Towards predicting physiology from speech during stressful conversations: Heart rate and respiratory sinus arrhythmia,” in *2018 ICASSP*, 2018, pp. 4944–4948.

7. A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, L. Stappen, J. Konzok, S. Sturmbauer, E.-M. Meßner, B. M. Kudielka, N. Rohleder, H. Baumeister, and B. W. Schuller, "An evaluation of speech-based recognition of emotional and physiological markers of stress," *Frontiers in Computer Science*, vol. 3, Dec. 2021. [Online]. Available: https://doi.org/10.3389/fcomp.2021.750284
8. J. Smith, A. Tsiartas, E. Shriberg, A. Kathol, A. Willoughby, and M. de Zambotti, "Analysis and prediction of heart rate using speech features from natural speech," in *2017 ICASSP*, 2017, pp. 989–993.
9. G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, 2022.
10. H. Beyramienanlou and N. Lotfivand, "An efficient teager energy operator-based automated QRS complex detection," *Journal of Healthcare Engineering*, vol. 2018, pp. 1–11, Sep. 2018. [Online]. Available: https://doi.org/10.1155/2018/8360475
11. K. Fujiwara, E. Abe, K. Kamata, C. Nakayama, Y. Suzuki, T. Yamakawa, T. Hiraoka, M. Kano, Y. Sumi, F. Masuda, M. Matsuo, and H. Kadotani, "Heart rate variability-based driver drowsiness detection and its validation with eeg," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1769–1778, 2019.
12. S. Ntalampiras, "Collaborative framework for automatic classification of respiratory sounds," *IET Signal Processing*, vol. 14, no. 4, pp. 223–228, Jun. 2020. [Online]. Available: https://doi.org/10.1049/iet-spr.2019.0487
13. G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.
14. S. Ntalampiras and I. Potamitis, "Transfer learning for improved audio-based human activity recognition," *Biosensors*, vol. 8, no. 3, p. 60, Jun. 2018. [Online]. Available: https://doi.org/10.3390/bios8030060
15. C. Alippi, "Learning in nonstationary and evolving environments," in *Intelligence for Embedded Systems*. Springer International Publishing, 2014, pp. 211–247.
16. M. Lukoševičius, "A practical guide to applying echo state networks," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 659–686. [Online]. Available: https://doi.org/10.1007/978-3-642-35289-8_36
17. Z. Zhao, Y. Zhang, and H. Liao, "Design of ensemble neural network using the akaike information criterion," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 8, pp. 1182–1188, Dec. 2008. [Online]. Available: https://doi.org/10.1016/j.engappai.2008.02.007
18. C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993. [Online]. Available: https://doi.org/10.1159/000119004
19. L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Meßner, E. Cambria, G. Zhao, and B. W. Schuller, "The MuSe 2021 multimodal sentiment analysis challenge," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. ACM, Oct. 2021. [Online]. Available: https://doi.org/10.1145/3475957.3484450
20. L. Stappen, E.-M. Meßner, E. Cambria, G. Zhao, and B. W. Schuller, "MuSe 2021 challenge," in *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, Oct. 2021. [Online]. Available: https://doi.org/10.1145/3474085.3478582

Stavros Ntalampiras is an Associate Professor at the Department of Computer Science, University of Milan, Italy. He received the engineering and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2006 and 2010, respectively. He has carried out research and/or didactic activities at Politecnico di Milano, the Joint Research Center of the European Commission, the National Research Council of Italy, and Bocconi University. Currently, he is an Associate Editor of IEEE Access, PLOS One, IET Signal Processing and CAAI Transactions on Intelligence Technology, as well as member of the IEEE Computational Intelligent Society Task Force on Computational Audio Processing. His research interests include content-based signal processing, machine learning, audio pattern recognition, bioacoustics, and cyber-physical systems.