# Feature Learning from Spectrograms for Assessment of Personality Traits

Marc-André Carbonneau, *Member, IEEE,* Eric Granger, *Member, IEEE,* Yazid Attabi, *Member, IEEE,* and Ghyslain Gagnon, *Member, IEEE*

***Abstract*—Several methods have recently been proposed to analyze speech and automatically infer the personality of the speaker. These methods often rely on prosodic and other hand crafted speech processing features extracted with off-the-shelf toolboxes. To achieve high accuracy, numerous features are typically extracted using complex and highly parameterized algorithms. In this paper, a new method based on feature learning and spectrogram analysis is proposed to simplify the feature extraction process while maintaining a high level of accuracy. The proposed method learns a dictionary of discriminant features from patches extracted in the spectrogram representations of training speech segments. Each speech segment is then encoded using the dictionary, and the resulting feature set is used to perform classification of personality traits. Experiments indicate that the proposed method achieves state-of-the-art results with an important reduction in complexity when compared to the most recent reference methods. The number of features, and difficulties linked to the feature extraction process are greatly reduced as only one type of descriptors is used, for which the 7 parameters can be tuned automatically. In contrast, the simplest reference method uses 4 types of descriptors to which 6 functionals are applied, resulting in over 20 parameters to be tuned.**

***Index Terms*—Automatic personality perception, spectrograms analysis, sparse coding, dictionary learning, feature learning, speaker trait classification, acoustic and prosodic modeling.**

## I. Introduction

PEOPLE spontaneously infer the personality of others from a wide range of cues. These cues may be visual, like facial expressions or posture, and may also be aural, like intonation patterns, choice of words or voice timbre. This assessment of personality traits naturally influences the way we interact with each other [1]. The method proposed in this paper aims at performing this assessment automatically.

Being able to accurately predict the personality of an interlocutor is an important step toward better human-machine interactions. For example, people attribute personality traits to machines and interact differently with them depending on this perceived personality. For instance, extroverted people will interact longer with robots they perceive as extroverted [2]. Detecting and understanding a person's personality would enable a machine to adapt its behavior to the user. It can also be used in e-learning applications by giving appreciative feedback on the personality projected by a user to improve its leadership or sale skills.

All authors are affiliated with the École de technologie supérieure, Montréal, Canada, e-mails: marcandre.carbonneau@gmail.com, ghyslain.gagnon@etsmtl.ca and eric.granger@etsmtl.ca

Manuscript received August 12, 2016; revised March 4, 2017; revised September 20, 2017.

In the literature, five personality traits (the *Big-Five*) corresponding to psychological phenomenon are observable regardless of the situation and culture: openness, conscientiousness, extroversion, agreeableness and neuroticism [3]. These traits influence the way people act and speak. For instance, in [4] a correlation is established between openness and neuroticism and the probability of maintaining a blog. The choice of words by a subject based on his/her personality traits has also been studied in informal texts [5], conversations [6] and on social media [7].

In the 2012 edition of the Interspeech competition on paralinguistics, one of the challenges was personality traits assessment from speech. This has motivated the proposition of several methods for this task. The baseline systems for the competition were designed using support vector machine (SVM) and random forest (RF) classifiers trained with 6125-dimensional feature vectors [8]. They performed particularly well, and only two contestants were able to surpass their performance on the test set. It was observed that increasing the number of features tends to increase recognition performance [8], thus large feature sets were extracted in the hope of capturing more of the relevant discriminant information. Some of the features were redundant or non-informative which motivated some contestants to use feature selection on the set of 6125 features [9]–[11]. The winners of the competition [12] added 21760 spectral features to the baseline feature set before performing selection.

Since 2012, the Interspeech competition 6125-dimension feature set of the baseline system has grown even larger. In 2015, it had increased to 6373-dimension [13]. Many of these features are statistics on the usual prosody features such as pitch, formants and energy, as well as more complex features, such as log harmonics to noise ratio, harmonicity and psycho-acoustic spectral sharpness. All of these application-specific feature extraction techniques require a fair knowledge and experience in speech processing to tune their parameters, select thresholds, pre-process data, etc. Moreover, results may vary from one implementation to another which limits the reproductibility of the experiments.

Many practitioners use software tools to extract prosody features, which accelerates the design of recognition solutions. However, even if these tools contain complete implementations of feature extraction algorithms, expertise in speech processing is required to configure the several parameters and options of each module. For instance, in openSMILE [14], one must choose between the cPitchACF (4 parameters) object and the cPitchShs object (9 parameters) to extract pitch, which in turn

must be configured. The user may also use a pitch smoother, where four more parameters must be set. There are even more parameters to consider when extracting formants.

Aside from the complexity and variability of these feature extraction procedures, the use of large feature sets reduces the generalization capability of pattern recognition algorithms [15]. Indeed, the exponential growth of the search space increases the amount of data needed to obtain a statistically significant representation of the data [16]. This represents a problem in affective computing application where data is limited because collection is costly. Moreover, smaller feature sets are desirable because they allow for faster training and classification.

The difficulties described above have been discussed by several researchers in the affective speech recognition community. The CEICES (Combining Efforts for Improving automatic Classification of Emotional user States) initiative attempted to create a standardized set of feature for emotion recognition in speech [17]. The proposed set is a combination of 381 acoustic and lexical features selected from a pool of 4024 features that the authors have successfully used in their previous research. While the collection of features was standardized, the implementation of the feature extraction algorithms was not. Recently, another attempt has been made to reduce the size of the feature collection used for automatic voice analysis [15]. A minimal number of descriptors were selected based on theoretical and empirical evidence. While the minimal and extended sets are compact (62 and 88 features respectively) several different algorithms are used for the extraction of the descriptors. These algorithms require expertise when tuning their various parameters[1].

In this paper, a method inspired by the recent developments in feature learning and image classification is proposed to alleviate these design choices for automatic assessment of personality traits. The temporal speech signals are translated into spectrogram images. Small sub-images, called patches, are densely extracted from these spectrogram images, and used during training to learn a feature dictionary yielding a sparse representation. The dictionary is used to encode each of the local patches. Each spectrogram is thus represented as a collection of encoded patches, which are pooled to create a histogram representation of the entire spectrogram. These histograms are used to train a classifier. During testing, a new speech signal is represented by a histogram, using the same dictionary, before classification.

The proposed method of representation, which is based on local patches, allows to capture para-linguistic information compactly. Because it encodes raw parts of the spectrogram images, the representation is richer than methods which characterize speech signals with statistics on the whole signal [8], [15], [18]. For instance, these methods use the mean, the standard deviation, kurtosis, min and max of the pitch or spectrum and cepstrum bins, which discard the relevant cues for personality assessment that the local shape of the signal contains. Moreover, when compared to these methods, the proposed method has fewer parameters, which can be more easily tuned using standard automatic hyper-parameter optimization techniques (e.g. cross-validation). In addition, the method inherits the robustness to deformation and noise of local image recognition methods applied to spectrogram analysis [19], [20]. Finally, since the dictionary learning process is performed in an unsupervised manner, additional training examples from other speech application domains can be used to learn a richer representation.

In essence, the proposed method leverages the power of representation inherent to sparse modeling, which learns features from the data. This approach generally leads to a high level of accuracy [21]. The dimensionality of feature vectors needed for this level of performance is reduced by an order of magnitude when compared to the number of features used in the Interspeech challenges. Moreover, only one method is used for feature extraction which limits the number of parameters needing careful tuning. Finally, the proposed technique does not necessitate a feature selection stage which is usually time consuming during training.

The proposed method is compared to 6 reference methods on the SSPNet Speaker Personality Corpus used in the Interspeech 2012 competition. As stated in the overview of the challenge published in 2015 [22], research in automated recognition of speaker traits is still active, and still requires much exploration to isolate suitable features and models for this task. In this regard, the novel technique proposed in this paper aims to provide a simpler alternative for extraction of a compact set of features that achieve state-of-the-art results.

The rest of the paper is organized as follows: The next section provides background information on feature learning in the context of speech analysis. Section III describes the proposed method. Section IV presents the experimental data, protocol and reference methods. The results are analyzed in Section V-A.

## II. Feature Learning for Speech Analysis

Feature learning algorithms extract relevant features themselves, instead of relying on human-engineered representations, which are time consuming to obtain and are often sub-optimal. Feature learning has been used in several speech analysis applications. Some methods use deep neural networks, which intrinsically learn features, to perform automatic speech recognition (ASR) [23]–[25]. These systems are not suitable for personality trait recognition because they analyze local time series (e.g. a phoneme), and fail to capture the global information in a speech segment. Deep learning has also been used for automatic emotion recognition. In [26] a deep convolutional recurrent network learns a representation from the raw signal, while in [27], [28], the neural network learns a feature representation, not from the raw signal, but from a set of prosodic, spectral and video features. In [29], [30], utterances where represented using sparse auto-encoders to perform emotion recognition. In [31], base features were learned using independent component analysis on spectrograms. After a feature selection process, the selected features were combined in a higher hierarchical level, using non-negative sparse coding.

[1]The feature set has been made publicly available through the openSMILE toolkit [14].

These feature combinations were used with an hidden Markov model (HMM) to perform ASR. In [32] features called Sigma-Pi were extracted from spectrograms and chosen using feature selection process to perform ASR.

Feature learning can be performed on several types of signal representation. When a speech signal is represented as a spectrogram, (i.e. concatenation in time of windowed Discrete Fourier Transform (DFT)), it can be analyzed through image processing. It has been demonstrated by neuroscientists that the same parts of the brain can be used to process both visual and audio signals [33]. This has motivated several researchers to investigate the application of image recognition techniques to spectrograms to analyze and recognize sound and speech signals. For example, histograms of oriented gradients (HOG) were used to perform word recognition [34]. In [35], spectrograms amplitudes are quantized and mapped into a color coded image. Color distributions are then characterized and analyzed. This method is inspired by content-based image retrieval methods [36]. In [20], spectrograms and cochleograms are divided in frequency sub-bands and analyzed as visual textures using gray-tone spatial dependence matrix features [37] alongside cepstral features. Audio spectrograms were employed with a convolutional deep Bayesian network, typically used for image recognition, to perform speaker identification and gender classification [38] and with convolutional neural networks to perform emotion recognition on utterances [39], [40]. The representation achieved a higher recognition performance when compared to mel-frequency cepstral coefficients (MFCC) and raw spectrograms. The Gabor function (sinusoidal tapered by a decaying exponential), were found to be good models of receptive fields in the human visual cortex [41]. This has motivated several authors to apply log-Gabor filter banks to spectrograms [42], [43] to analyze paralinguistics.

A popular paradigm for image analysis is to extract features locally (instead of globally) from salient regions of an image, called patches. The set of patches, is used to represent an entire image. This type of approach, often called bag-of-words, have been successfully applied in numerous contexts for recognition in image [44], [45] and video [46], [47]. Using local features in image recognition may lead to an increased robustness to intra-class variation, deformation, view-point, illumination and occlusion [48]. When working with spectrograms, it translates to an increased robustness to noise [19], [35]. In [49] the SIFT descriptor was used to detect and encode key-points in spectrogram images of musical pieces to perform genre classification. Schutte proposed a deformable part-based model of local spatio-temporal features in speech recognition [19]. The method allowed to improve recognition performance over the HMM baseline system especially in the presence of noise.

Local-based methods in image recognition often exploit a set of predefined basis for decomposition such as wavelets, wedgelets and bandlets [50]. However, it has been shown that learning the basis directly on the data leads to a higher level of accuracy in several applications such as signal reconstruction [51] and image classification [52] and reconstruction [53]. Based on these results, several recently proposed spectrogram analysis methods learn representation on training data in order to benefit from the improved performance. For instance, in [54] the spectrograms are segmented at different scales, and each segment is encoded as the most resembling word in a dictionary learned using the *k*-means algorithm. In [55] the spectrograms of musical instruments are interpreted as visual textures. Sounds are represented by a vector encoding the resemblance between the spectrogram and a randomly constituted dictionary.

In the aforementioned dictionary-based methods, local descriptors are associated with the most representative codeword in the dictionary. Some algorithms use sparse coding to perform this association and learn a representation [51], [56]. Sparse coding is a type of feature learning which expresses a signal using a small number of basis from a learned set, usually called dictionary. Experiments have shown that encoding audio and visual signals using a sparse decomposition can lead to a high level of accuracy for various tasks such as acoustic event detection [57], speaker, gender and phoneme recognition [38]. Also, it was shown that a learned sparse representation of audio signals is akin to the early mammalian auditory system [58]. This is why several recent methods use sparse coding to learn the dictionary and encode signals.

In the context of personality assessment from speech, paralinguistic cues must be analyzed globally. A personality trait is something that endures throughout entire speech segments belonging to the same speaker. This is different from many other speech recognition problems, like emotion recognition, where the target events have a relatively short duration. Methods used in other speech analysis applications, such as ASR and emotion recognition, do not typically capture global information from long speech segments. In most existing methods for personality recognition, this is achieved using statistical operators on low-level features. Unfortunately, this results in a high dimensional representation, which is prone to the curse of dimensionality, and require fair signal processing expertise to extract the low-level features. The proposed method represents a complete speech segment as an image then uses image recognition techniques, and thus, can perform global analysis. Moreover, it uses a feature learning approach, which reduces the burden associated with feature engineering and yields a compact representation, and leads to increased recognition performance.

## III. Proposed Feature Learning Method

This section presents a new method for predicting personality traits in speech based on spectrogram analysis and feature learning. The main stages of the proposed method are depicted in Figure 1. Specific details regarding our proposed solution for feature extraction, classification and dictionary learning are described in the next sections. The upper part is the pipeline for training. At first, for each speech segment $\mathbf{F}$ in the data set, a spectrogram $\mathbf{S}$ is extracted by applying a Fourier transform on a sliding window, yielding a 2-dimensional matrix. Small sub-matrices, called patches $\{\mathbf{p}_1, ..., \mathbf{p}_k\}$ are then uniformly extracted from all the spectrogram matrices in the training set. A dictionary $\mathbf{D} = \{\mathbf{d}_1, ..., \mathbf{d}_m\}$ is learned from these patches, and at the same time, the patches are encoded as sparse vectors called code-words $\{\mathbf{c}_1, ..., \mathbf{c}_k\}$. A single $m$-dimensional feature vector representation $\mathbf{h}$ is obtained for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2763132, IEEE Transactions on Affective Computing



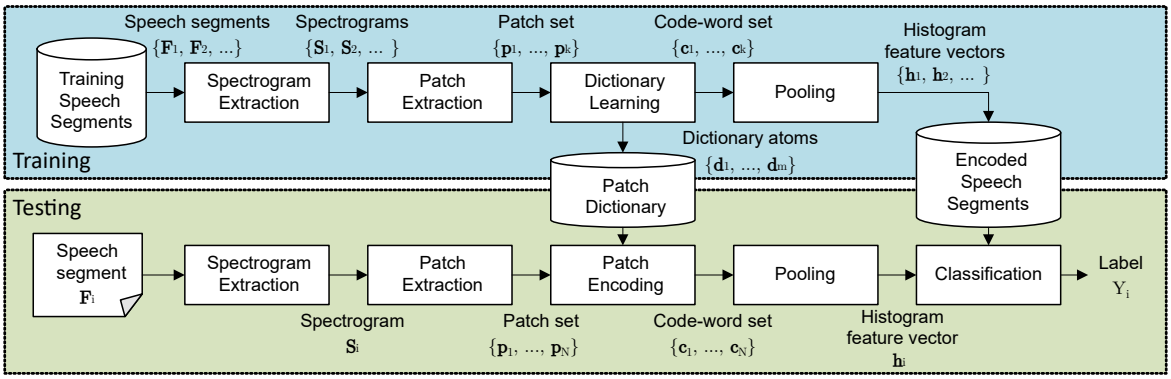



Fig. 1. Block diagram of the proposed system for the prediction of a personality trait. The upper part illustrates the operations performed during training. The lower part illustrates sequence of operations performed to process an input speech sequence in test.

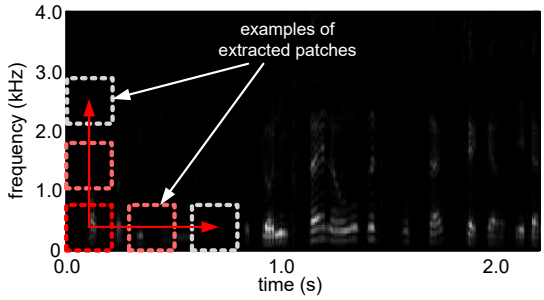



Fig. 2. Example of spectrogram extracted from a speech file in the SSPNet corpus. White indicates high values while black indicates low values.

each training speech sample by pooling together all code-words extracted from it. A two-class support vector machine (SVM) classifier is trained using these feature vectors for each personality trait.

The lower part is the pipeline used during testing, to predict a personality trait. Like in training, patches are extracted from the spectrograms. Each patch is encoded using the previously learned dictionary. The resulting code-words are then pooled to create a summarizing feature vector. This vector is used by a 2-class classifier which predicts if the speaker exhibits, or not, a specific personality trait.

### A. Feature Extraction

Given a speech segment $x(n)$, the spectrogram $\mathbf{S}$ is the concatenation in time of its windowed DFT:

$$\mathbf{S} = \{\mathbf{X}_0, ..., \mathbf{X}_t, ..., \mathbf{X}_{T-1}\}, \tag{1}$$

where $\mathbf{X}_t$ is a column vector containing the absolute amplitude of the DFT frequency bins for the frame at time $t$ and $T$ is the number of DFTs extracted from the signal. The absolute amplitude is favored over the log-amplitude as it has shown to yield better results for spectrogram image classification in [35] and in our own experiments. The spectrograms are normalized: each frequency bin is divided by the maximum amplitude value contained in a time frame. This normalization ensures a certain robustness to capture conditions. This process results in a 2-D matrix $\mathbf{S}$ which can be analyzed as a grey-scale image. An example of spectrogram extracted on the SSPNet Speaker Personality Corpus is illustrated in Figure 2.

From the matrix $\mathbf{S}$, small patches, or sub-images, of $p \times p$ pixels are extracted at regular intervals. A vector representation $\mathbf{p}_i \in \mathbb{R}^{1\times d}$ of each patch ($d = p \times p$) is obtained by concatenating the value of all pixels. The vector $\mathbf{p}_i$ is encoded into $\mathbf{c}_i$ using a previously learned dictionary $\mathbf{D}$ containing $m$ atoms (more details in Section III-C). These atoms are vector basis that are used to reconstruct the patches. The code-vector $\mathbf{c}_i$ corresponding to the patch $\mathbf{p}_i$ is obtained by solving

$$l(\mathbf{c}_i) \triangleq \min_{\mathbf{c}_i \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{p_i} - \mathbf{D}\mathbf{c}_i\|_2^2 + \lambda \|\mathbf{c}_i\|_1 \tag{2}$$

using the LARS-Lasso algorithm [59]. The loss function has two terms, each encoding an optimization objective, and $\lambda$ is a parameter used to adjust the relative importance of the two terms. The first term is the quadratic reconstruction error, while in the second term, the $\ell_1$ norm of the code vector is used to enforce sparseness. Once a code $\mathbf{c}_i$ is obtained for each patch $\mathbf{p}_i$, the absolute value of all the codes are summed to obtain a histogram $\mathbf{h}$ describing the entire spectrogram $\mathbf{S}$:

$$\mathbf{h} = \sum_i |\mathbf{c}_i| \tag{3}$$

These histograms represent the distribution of patches over speech segments. It is thus possible to directly compare segments of different length.

### B. Classification

The speech segments are represented by histograms and thus, appropriate distance measure should be employed. Several distance measures have been proposed to compare histograms. In this paper's implementation, the $\chi^2$ distance is used because it showed competitive performance for visual bag-of-words histograms [48]. The $\chi^2$ distance is given by :

$$d(\mathbf{g}, \mathbf{h}) = \sum_{i=1}^{m} \frac{(g_i - h_i)^2}{g_i + h_i}, \tag{4}$$

where $g_i$ and $h_i$ are the $i^{\text{th}}$ bins of histograms $\mathbf{g}$ and $\mathbf{h}$, and $m$ corresponds to the number of words in the dictionary.

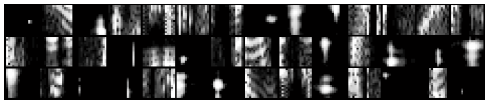


Fig. 3. Example of patches from a dictionary created with sparse coding.

In this paper $d$ is used in an SVM framework with an exponential kernel [60]:

$$k(\mathbf{g}, \mathbf{h}) = e^{-\gamma d(\mathbf{g},\mathbf{h})}, \tag{5}$$

where the parameter $\gamma$ controls the kernel size.

While the implementation of this paper employs the $\chi^2$ distance and an SVM classifier, the proposed methods is not bound to these choices, and other distance functions and classifiers can be used.

### C. Dictionary Learning

The objective of the dictionary learning phase is to generate a representative dictionary $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_m] \in \mathbb{R}^{d \times m}$ given the matrix $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_k] \in \mathbb{R}^{d \times k}$ containing patch vectors extracted from the training set. Generally, for image classification tasks, best results are obtained with over-complete ($m > d$) dictionaries [61].

A dictionary of atoms $\mathbf{D}$ and sparse code-words $\mathbf{C}$ can be obtained by minimizing the following loss function:

$$l(\mathbf{C}, \mathbf{D}) \triangleq \min_{\mathbf{C} \in \mathbb{R}^{m \times k}, \mathbf{D} \in \mathcal{C}} \frac{1}{2} \|\mathbf{P} - \mathbf{D}\mathbf{C}\|_2^2 + \lambda \|\mathbf{C}\|_1 \tag{6}$$

In this equation, $\lambda$ is the same as in (2) and is used to adjust the weight of the sparseness term in the loss equation. The convex set:

$$\begin{aligned} \mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{d \times m} \ &\textbf{s.t.} \ \forall i = 1, ..., m, \mathbf{d}_i^T \mathbf{d}_i \ \leq 1 \\ &\textbf{and} \ \forall i = 1, ..., m, \mathbf{d}_i \in \mathbb{R}_{\geq 0}\} \end{aligned} \tag{7}$$

enforces two constraints. The first is used to restrict the magnitude of the dictionary atoms. The second is used to make sure each element of each atom in the dictionary is positive. Since the spectrogram is purely positive, better results are obtained by enforcing this constraint. The joint optimization of $\mathbf{C}$ and $\mathbf{D}$ is not convex. However if one term is fixed the problem becomes convex. Thus, a common strategy is two alternate between updating $\mathbf{C}$ while $\mathbf{D}$ is fixed and updating $\mathbf{D}$ while $\mathbf{C}$ is fixed until a stopping criterion is met (e.g. max number of iterations) [62].

Figure 3 shows an example of dictionary atoms learned using the above described procedure. Some atoms encode short intonation patterns with ascending and descending linear patterns, while others encode more punctual accents which may help discriminate personalities based on speech energy variation. The audio files from the SSPNet Speaker Personality Corpus were used to learn the atoms. The same dictionary can be used for all traits.

## IV. Experimental Methodology

The SSPNet Speaker Personality corpus [18] is the largest and most recent data set for personality trait assessment from speech. It consists of 640 audio clips randomly extracted from French news bulletins in Switzerland. All clips have been sampled at 8 kHz and most of the clips are 10 seconds long, but some are shorter. Each clip contains only one of the 322 different speakers. Eleven judges performed annotation on each clip by completing the BFI-10 personality assessment questionnaire [63]. From the questionnaire a score is computed for each of the *Big-Five* personality traits. Precautions were taken to avoid sequence and tiredness effects in the annotation process. The judges did not understand French and therefore were not influenced by linguistic cues. In [18] the assessment of the judges were considered as positive if the score was greater than 0 and negative otherwise. The labeling scheme was refined for the competition [8]. In this case, an assessment was considered positive if the score given by a judge was higher than the average score given by this particular judge for the trait. In both cases, the final label for an instance was obtained by a majority vote from all of the 11 judges. Preliminary experiments showed a 1$\sim$2% difference in accuracy performance between the two labeling schemes. The results reported in this paper were obtained using the competition's labeling scheme.

The metric used to compare accuracy is the unweighted average recall (UAR), which is the same as in the competition. The UAR is the mean of each class accuracy, and thus is unaffected by class imbalance. To assess performance, a 3-fold cross-validation procedure was used to provide a better estimation of accuracy than with the hold out validation procedure used in the challenge. Precautions were taken to make sure that all samples belonging to the same speaker are grouped in the same fold. In the Interspeech 2012 Speaker Trait challenge, the results obtained for the conscientiousness trait with the development partition are lower than the results obtained with the test partition. For instance, the baseline method using SVM obtained a UAR of 74.5% in training, but increased to 80.1% in testing [8]. The same phenomenon was observed with the random forest classifier (74.9% to 79.1%). This suggests that the test data may have been easier to classify than the average data. This hypothesis is supported by the fact that the results obtained using a cross-validation procedure in [18] were also closer to 70% than 80%. Nested cross-validation [64] was used to optimize the hyper parameters for all classifiers and the dictionary learning parameters (dictionary size and $\lambda$). In nested cross-validation, an outer cross-validation loop (3 folds) is used to obtain the final test results, and an inner loop (5 folds) is used to find the best hyper parameter via grid search. Hyper-parameter optimization is thus performed for each of the 3 test folds separately.

For the proposed method, spectrograms were extracted using a short-time Fourier transform with a 128 sample Hamming window. This translates into 16 ms segments at the sample rate (8 kHz) of the SSPNet Speaker Personality corpus. There was a 75% overlap between two successive speech segments to capture short-term characteristics of vocal behavior as done in [18]. The extracted patches were 16$\times$16 *pixels*, yielding 256-dimensional feature vectors. A new patch was extracted each 8 time steps and each 4 frequency bins. All of these 5 parameters (FFT window size and overlap, window type, patch size and stride) were selected based on

TABLE I
PERFORMANCE ON THE SSPNET SPEAKER PERSONALITY CORPUS AND PARAMETER COMPLEXITY OF THE METHODS.

| | Unweighted Average Recall (%) | | | | | | Number of | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Algorithm** | O | C | E | A | N | Avr. | Features | Descriptors | Functionals | Parameters |
| Mohammadi & Vinciarelli (LR) [18] | 56.1 | 69.6 | 72.4 | 55.7 | 67.4 | 64.2 | 24 | 4 | 6 | >20 |
| Mohammadi & Vinciarelli (SVM) [18] | 57.7 | 68.0 | 74.3 | 57.4 | 65.5 | 64.6 | 24 | 4 | 6 | >20 |
| Interspeech Challenge Baseline (SVM) [8] | **58.7** | 69.2 | 74.5 | 62.2 | 69.0 | 66.7 | 6125 | 21 | 39 | >200 |
| Interspeech Challenge Baseline (RF) [8] | 52.9 | 69.0 | **77.5** | 60.1 | 68.2 | 65.5 | 6125 | 21 | 39 | >200 |
| GeMAPS (SVM) [15] | 56.3 | 72.2 | 74.9 | 61.9 | 68.9 | 66.8 | 62 | 13 | 10 | >100 |
| eGeMAPS (SVM) [15] | 53.7 | **72.5** | 75.1 | 62.0 | 66.6 | 66.0 | 88 | 16 | 12 | >100 |
| SAE 1-Layer (SVM) [29] | 57.1 | 64.3 | 69.2 | 62.0 | 65.8 | 63.7 | 100-800 | 1 | 1 | >30 |
| SAE 2-Layers (SVM) [29] | 57.3 | 63.6 | 69.0 | 60.3 | 61.9 | 62.4 | 100-800 | 1 | 1 | >30 |
| Proposed Method | 56.3 | 68.3 | 75.2 | **64.9** | **70.8** | **67.1** | 200-800 | 1 | 1 | 7 |

preliminary experiments and were not subsequently optimized. Only 2 parameters, the dictionary size $\in \{100, 200, 400, 800\}$ and $\lambda \in \{0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$, were optimized in the experiments using the aforementioned cross-validation scheme. An importance weighting scheme was used to deal with class imbalance [65]. This was achieved by attributing different misclassification cost in the SVM hinge loss function to the target classes. The cost for the positive class was multiplied by a factor corresponding to the class imbalance ratio. The SPAMS toolbox [66] was used for dictionary learning and encoding and LIBSVM [67] was used for the SVM implementation.

Three reference methods were selected to compare performance. The methods were chosen because they are well documented and can be reproduced without ambiguity. The first method was proposed by Mohammadi & Vinciarelli in [18]. Prosody features were extracted using Praat [68], the same software used in the original paper. The low-level feature extracted were pitch, first two formants, energy of speech, and length of voiced and unvoiced segments. The features were extracted using 40 ms long windows at 10 ms time steps. The features were whitened based on means and standard deviations estimated on the training folds. Four statistical properties were then estimated from the 6 prosody measures yielding a 24-dimensional feature vector for each speech file. The statistical features were the minimum, maximum, mean and the entropy of the differences between consecutive feature values. As in [18], an SVM and a logistic regression (LR) were used for classification. The logistic regression implementation of the MATLAB Statistic and Machine Learning Toolbox was used. For the SVM, the LIBSVM implementation was used with the linear and the radial basis function (RBF) kernels.

The second method is the baseline used in the Interspeech 2012 speaker trait challenge [8]. The 6125 low-level features were extracted using the openSMILE software [14] with the preset named after the challenge. The features were whitened based on means and standard deviations estimated on the training folds. For the linear SVM, the LIBSVM implementation [67] was used which performs sequential minimal optimization, the optimization algorithm used in the challenge baseline. The use of Gaussian kernel was also explored but did not yield better results. For the RF classifier, MATLAB implementation from the Statistic and Machine Learning Toolbox was used. This method was selected because it yield state-of-the-art performance. Only 2 of the methods proposed in the challenge outperformed the baseline by a margin of 0.1% for [69] and of 1% for [12] on UAR, which is not significant.

The third and most recent benchmark method uses the features prescribed in the Geneva minimalistic acoustic parameter set (GeMAPS) [15]. The minimalistic set can be extended (eGeMAPS) by including MFCC coefficients, spectral flux and additional formant descriptors. The features were extracted using the preset supplied in openSMILE. Classification was achieved by a linear SVM using the LIBSVM implementation. The hyper-parameters were optimized in the same way as for the Interspeech method. This method was selected because it is intended to reduce the complexity of the feature extraction stage in paralinguistic problems, same as the proposed method.

Finally, we replaced the feature learning algorithm in the proposed method by sparse auto-encoders (SAE) and stacked sparse auto-encoders using an implementation similar to [29]. The topology and loss function parameters were optimized using random search as prescribed in [70] because the number of hyper-parameters is too high to perform grid search in reasonable time. The number of neurons on each layer ranges from 50 to 800. A sample pool of 200k patches were used for training the SAE. Sparseness and regularization weights and parameters were sampled from log-uniform distributions.

## V. RESULTS

### A. Accuracy

Performance of the proposed and baseline methods on the SSPNet Speaker Personality corpus is reported in Table I. The best average UAR was obtained using the proposed method. However, the results obtained when using the challenge features and GeMAPS with an SVM classifier are comparable. The method proposed by Mohammadi and Vinciarelli yields slightly lower accuracy than the other methods, although the difference in performance in most cases is small and may be negligible. Particularities in the data set and the type of classifier, as well as its implementation, are most likely the reason for these variations in performance. For instance, using the same features and a different classifier, the Interspeech 2012 challenge baseline [8] obtains a UAR of 58.7% (SVM) and 52.9% (RF) for the openness trait. The performance gap between the proposed method and SAE is due in part

to the way sparsness is enforced in the optimization loss function. SAE use the Kullback–Leibler divergence [29] of the neuron activation proportion and a fixed parameter, while the proposed method uses the $\ell_1$ norm of the code vector. SAE represents complex intonation patterns with a combination of more generic patches while the proposed method tends to encode these complex patterns with single patches. The sum pooling process hides the discriminative information of intonation patterns represented as a composition of generic patches.

There are differences between the representations. For instance, the proposed method is not well adapted to represent pitch nor speech rate. Estimating the pitch is difficult because once the patches are extracted, their location is discarded. In contrast, all reference methods explicitly extract pitch and compute statistics on the measure. Speech rate is also difficult to represent by the proposed method since patches encode local information while speech rate is more of a global measure. All reference methods capture speech rate better because they extract statistics on the length and proportion of voiced and unvoiced segments. This slightly impedes the proposed method for the recognition of the openness trait, for which pitch and speech rate have been identified as markers [6], [71]. It could explain the 2.4% and 1.4% difference between the proposed and reference methods using SVM. However, these two markers are also indicative of neuroticism [6], and the proposed method performs well on this class. This could be explained by its ability to capture voice timbre and short intonation patterns. The proposed method uses raw chunks of the sound spectrogram as representation, and thus can capture this kind of information with high fidelity.

### B. Complexity

While accuracy is generally similar for all methods, the main advantage of the proposed method is the important reduction of effort and design choices needed for its implementation. The amount of human expert intervention is different for all methods. In the proposed method, only 1 feature extraction algorithm was used instead of 4 for [18], more than 10 for GeMAPS and over 20 in [8]. In addition, in these reference methods, a set of functionals were applied to the extracted features. Some of these functionals were simple measures like mean, min/max and standard deviation, but others were more complex and parametrizable. For instance, functionals relying on peak distance need a peak detector that has to be fine-tuned. These feature extraction algorithms require parametrization which must be performed by a signal processing expert. A similar argument applies to SAE. These models necessitate a fair amount of expertise and experience to choose the appropriate topology and loss function, to tune the numerous hyper-parameters and to configure the optimization algorithm. Also, when compared to the baseline of the Interspeech challenge, the feature set used in the proposed method is much smaller (at most 800 features instead of 6125). Smaller feature sets are desirable because they reduce algorithmic complexity, and are less subject to problems associated with the curse of dimensionality.

During training, the time complexity of the proposed method is higher than for the other methods because of the dictionary learning phase. However, at test time, less operations are required than for all other methods except SAE. In the proposed method, two main operations are performed – spectrogram extraction and patch encoding. Spectrogram extraction must be performed with all other methods. Then, methods [8], [15], [18] need to perform various operations like pitch extraction, power ratios, peak detection, linear regression, Viterbi-based smoothing, RASTA [72], etc. While many of these algorithms have a complexity that scales linearly with the sample length ($\mathcal{O}(n)$), some of them, like pitch extraction relying on autocorrelation, have a complexity of $\mathcal{O}(n \log n)$. In contrast, the proposed method needs to solve an optimization problem using the LARS-lasso algorithm which has the same computational complexity as regular least-square regression [59]. The complexity of the LARS-lasso algorithm grows linearly on the size of the dictionary ($\mathcal{O}(m)$). Some other techniques [73] for sparse coding are even faster ($\mathcal{O}(\sqrt{m})$) for applications where a larger dictionary is needed, which is not the case here. Moreover, the encoding of patches can be parallelized which allows for faster processing at test time. Of course, the time complexity of the proposed method also grows linearly on the length of the speech segment ($\mathcal{O}(n)$). The fastest model during testing is SAE because it only performs weight matrix multiplication to obtain the patch representation. Finally, one could argue that more memory is required by the proposed method to store the dictionary ($\mathcal{O}(md)$). However, a 800 word dictionary of 16×16 *pixel* patches requires a storage of about 1.6 MB in double-precision floating-point format, which is manageable with modern computers.

## VI. Conclusion

This paper presents a new method for automated assessment of personality traits in speech. Speech segments are represented using spectrograms and feature learning. The proposed representation is compact and is obtained using a single algorithm requiring minimal expert intervention, when compared to reference methods. Experiments conducted on SSPNet corpus indicate that the proposed method yields the same level of accuracy as state-of-the-art methods in paralinguistics that employ more complex representations, while remaining simpler to use.

As explained in Section V-A, the method is not properly equipped to capture pitch and speech rate. Research should be conducted to include these signal characteristics in the representation. In addition, experiments on different paralinguistic problems should be conducted to validate the applicability of the proposed method in different contexts. Experiments should also be conducted where the sparse dictionary learning and classifier algorithms used in our implementation is replaced by other methods enforcing group sparsity and discrimination. Finally, given the unsupervised nature of the feature learning process, experiments should be conducted to assess the potential benefits of using a larger number of examples from other speech data sets.

## REFERENCES


[1] J. S. Uleman, L. S. Newman, and G. B. Moskowitz, "People as flexible interpreters: Evidence and issues from spontaneous trait inference," *Adv. Exp. Soc. Psy.*, vol. 28, pp. 211–280, 1996.

[2] A. Tapus and M. J. Mataric, "Socially Assistive Robots: The Link between Personality, Empathy, Physiological Signals, and Task Performance." in *AAAI Spring Symp. on Emotion, Personality and Social Behavior*, 2008.

[3] J. M. Digman, "The curious history of the five-factor model." *The Five-Factor Model of Personality*, p. 20, 1996.

[4] R. E. Guadagno, B. M. Okdie, and C. A. Eno, "Who blogs? Personality predictors of blogging," *Computers in Human Behavior*, vol. 24, no. 5, pp. 1993–2004, Sep. 2008.

[5] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in *Joint Annu. Meeting of the Interface and the Classfication Soc. of North America*, 2005.

[6] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *J. Artif. Int. Res.*, vol. 30, no. 1, pp. 457–500, Nov. 2007.

[7] L. Qiu, H. Lin, J. Ramsay, and F. Yang, "You are what you tweet: Personality expression and perception on Twitter," *J. of Res. in Personality*, vol. 46, no. 6, pp. 710–718, Dec. 2012.

[8] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The Interspeech 2012 Speaker Trait Challenge," in *INTERSPEECH*, 2012.

[9] C. Chastagnol and L. Devillers, "Personality Traits Detection Using a Parallelized Modified SFFS Algorithm," in *INTERSPEECH*, 2012.

[10] D. Wu, "Genetic algorithm based feature selection for speaker trait classification," in *INTERSPEECH*, 2012.

[11] J. Pohjalainen, S. Kadioglu, and O. Räsänen, "Feature Selection for Speaker Traits," in *INTERSPEECH*, 2012.

[12] V. Ivanov and X. Chen, "Modulation Spectrum Analysis for Speaker Personality Trait Recognition," in *INTERSPEECH*, 2012.

[13] B. Schuller, S. Steidl, A. Batliner, S. Hantke, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The Interspeech 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition," in *INTERSPEECH*, 2015.

[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor," in *ACMMM*, 2013.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affective Computing*, vol. 7, no. 2, pp. 190–202, apr 2016.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, USA: Springer, 2006.

[17] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and Others, "Combining efforts for improving automatic classification of emotional user states," in *IS-LTC*, 2006.

[18] G. Mohammadi and A. Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features," *IEEE Trans. Affective Computing*, vol. 3, no. 3, pp. 273–284, Jul. 2012.

[19] K. T. Schutte, "Parts-based Models and Local Features for Automatic Speech Recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, USA, 2009.

[20] R. V. Sharan and T. J. Moir, "Subband Time-Frequency Image Texture Features for Robust Audio Surveillance," *IEEE Trans. Inf. Forens. Security*, vol. 10, no. 12, pp. 2605–2615, dec 2015.

[21] R. B. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-Invariance Sparse Coding for Audio Classification," *UAI*, 2007.

[22] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Computer Speech & Lang.*, vol. 29, no. 1, pp. 100–131, jan 2015.

[23] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Trans. on Audio, Speech, and Lang. Proces.*, vol. 20, no. 1, pp. 7–13, jan 2012.

[24] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Lang. Proces.*, vol. 20, no. 1, pp. 14–22, jan 2012.

[25] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[26] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.

[27] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *ICASSP*, may 2013.

[28] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *ICASSP*, 2011.

[29] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition," in *ACII*, 2013.

[30] S. Ghosh, E. Laksana, L. Morency, and S. Scherer, "Learning representations of affect from speech," *arXiv preprint arXiv:1511.04747*, 2015.

[31] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Commun.*, vol. 53, no. 5, pp. 736–752, May 2011.

[32] M. Kleinschmidt, "Robust speech recognition based on spectro-temporal processing," Ph.D. dissertation, Oldenburg, 2002.

[33] L. von Melchner, S. L. Pallas, and M. Sur, "Visual behaviour mediated by retinal projections directed to the auditory pathway," *Nature*, vol. 404, no. 6780, pp. 871–876, Apr. 2000.

[34] T. Muroi, R. Takashima, T. Takiguchi, and Y. Ariki, "Gradient-based acoustic features for speech recognition," in *ISPACS*, 2009.

[35] J. W. Dennis, "Sound Event Recognition in Unstructured Environments using Spectrogram Image Processing," Ph.D. dissertation, Nanyang Technological University, 2014.

[36] L.-H. C. J.-L. Shih, "Colour image retrieval based on primitives of colour moments," *IEE Proc. Vision, Image and Signal Process.*, vol. 149, pp. 370–376, dec 2002.

[37] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, nov 1973.

[38] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *NIPS*, 2009.

[39] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec 2014.

[40] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *APSIPA*, 2016.

[41] S. Marĉelja, "Mathematical description of the responses of simple cortical cells," *J. Opt. Soc. Am.*, vol. 70, no. 11, pp. 1297–1300, nov 1980.

[42] Y. Gu, E. Postma, and H.-X. Lin, "Vocal Emotion Recognition with Log-Gabor Filters," in *AVEC*, 2015.

[43] H. Buisman and E. Postma, "BNAIC: The log-gabor method: Speech classification using spectrogram image analysis," in *INTERSPEECH*, 2012.

[44] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[45] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV*, 2004.

[46] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[47] M.-A. Carbonneau, A. J. Raymond, E. Granger, and G. Gagnon, "Real-time visual play-break detection in sport events using a context descriptor," in *ISCAS*, 2015.

[48] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *Int. J. of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2006.

[49] T. Matsui, M. Goto, J.-P. Vert, and Y. Uchiyama, "Gradient-based musical feature extraction based on scale-invariant feature transform," in *EUSIPCO*, 2011.

[50] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd ed. Academic Press, 2008.

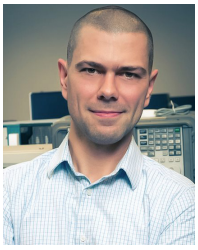
[51] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, dec 2006.

[52] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *ICML*, 2007.
[53] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, nov 2006.
[54] R. F. Lyon, "Machine Hearing: An Emerging Field [Exploratory DSP]," *Signal Process. Magazine, IEEE*, vol. 27, no. 5, pp. 131–139, sep 2010.
[55] G. Yu and J.-J. Slotine, "Audio classification from time-frequency texture," in *ICASSP*, 2009.
[56] G. Peyré, "Sparse Modeling of Textures," *J. of Math. Imaging and Vision*, vol. 34, no. 1, pp. 17–31, 2009.
[57] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *WASPAA*, 2011.
[58] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, feb 2006.
[59] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[60] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Network*, vol. 10, no. 5, pp. 1055–1064, Sep. 1999.
[61] I. Tosic and P. Frossard, "Dictionary Learning," *Signal Process. Magazine, IEEE*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
[62] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2006.
[63] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *J. of Res. in Personality*, vol. 41, no. 1, pp. 203–212, feb 2007.
[64] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. of the Royal Statistical Soc. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974.
[65] A. Rosenberg, "Classifying Skewed Data: Importance Weighting to Optimize Average Recall." in *INTERSPEECH*, 2012.
[66] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online Dictionary Learning for Sparse Coding," in *ICML*, 2009.
[67] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, may 2011.
[68] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2001.
[69] C. Montacié and M.-j. Caraty, "Pitch and Intonation Contribution to Speakers ' Traits Classification," in *INTERSPEECH*, 2012.
[70] J. Bergstra and Y. Bengio, "Random Search for Hyper-parameter Optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 281–305, feb 2012.
[71] D. W. Addington, "The relationship of selected vocal characteristics to personality perception," *Speech Monographs*, vol. 35, no. 4, pp. 492–503, 1968.
[72] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, oct 1994.
[73] T. Ge, K. He, and J. Sun, "Product sparse coding," in *CVPR*, 2014.

**Marc-André Carbonneau** received a B. Eng. degree in electrical engineering in 2010. He received the Ph.D. degree from École de technologie supérieure (Université du Québec), Montreal, in 2017 for his in multiple instance learning. His research interests include machine learning, weakly supervised learning, active learning, reinforcement learning, computer vision, action recognition and signal processing.

**Eric Granger** earned Ph.D. in EE from École Polytechnique de Montréal in 2001, and worked as a Defense Scientist at DRDC-Ottawa (1999-2001), and in R&D with Mitel Networks (2001-04). He joined the École de technologie supérieure (Université du Québec), Montreal, in 2004, where he is presently Full Professor and director of LIVIA, a research laboratory on computer vision and artificial intelligence. His research focuses on adaptive pattern recognition, machine learning, computer vision and computational intelligence, with applications in biometrics, face recognition and analysis, video surveillance, and computer/network security.

**Yazid Attabi** received a computer engineering degree in 1994 from the Universite des Sciences et de la Technologie Houari Boumedine, Algeria, and the master's of science and Ph.D. degree in software engineering in 2009 from École de technologie supérieure, Montréal. He is with Centre de recherche informatique de Montréal. His research interests include machine learning applied to emotion recognition from speech.



**Ghyslain Gagnon** received the Ph.D. degree in electrical engineering from Carleton University, Canada in 2008. He is now an Associate Professor at École de technologie supérieure, Montreal, Canada. He is an executive committee member of ReSMiQ and Director of research laboratory LACIME, a group of 10 Professors and nearly 100 highly-dedicated students and researchers in microelectronics, digital signal processing and wireless communications. Highly inclined towards research partnerships with industry, his research aims at digital signal processing and machine learning with various applications, from media art to building energy management