



When Old Meets New: Emotion Recognition from Speech Signals

Keith April Araño¹ · Peter Gloor² · Carlotta Orsenigo¹ · Carlo Vercellis¹

Received: 15 October 2020 / Accepted: 12 April 2021 / Published online: 19 April 2021
© The Author(s) 2021

Abstract

Speech is one of the most natural communication channels for expressing human emotions. Therefore, speech emotion recognition (SER) has been an active area of research with an extensive range of applications that can be found in several domains, such as biomedical diagnostics in healthcare and human–machine interactions. Recent works in SER have been focused on end-to-end deep neural networks (DNNs). However, the scarcity of emotion-labeled speech datasets inhibits the full potential of training a deep network from scratch. In this paper, we propose new approaches for classifying emotions from speech by combining conventional mel-frequency cepstral coefficients (MFCCs) with image features extracted from spectrograms by a pretrained convolutional neural network (CNN). Unlike prior studies that employ end-to-end DNNs, our methods eliminate the resource-intensive network training process. By using the best prediction model obtained, we also build an SER application that predicts emotions in real time. Among the proposed methods, the hybrid feature set fed into a support vector machine (SVM) achieves an accuracy of 0.713 in a 6-class prediction problem evaluated on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, which is higher than the previously published results. Interestingly, MFCCs taken as unique input into a long short-term memory (LSTM) network achieve a slightly higher accuracy of 0.735. Our results reveal that the proposed approaches lead to an improvement in prediction accuracy. The empirical findings also demonstrate the effectiveness of using a pretrained CNN as an automatic feature extractor for the task of emotion prediction. Moreover, the success of the MFCC-LSTM model is evidence that, despite being conventional features, MFCCs can still outperform more sophisticated deep-learning feature sets.

Keywords Speech emotion recognition · Machine learning · Deep learning

Introduction

Sentiment analysis and affective computing have been receiving a surge of interest from both the academic and business communities in recent years due to the proliferation of opinion-rich social media data and their increasing applications in different use cases. The research in this field has been traditionally focused on analyzing textual data.

However, the limitations of interpreting emotions from text features alone have led to studies devoted to predicting emotional states from other forms of modalities, such as speech [1]. Speech signals are the most natural, intuitive and fastest means of interaction for humans and are known to carry much more information than text or spoken words [2].

Emotion recognition from speech signals has been studied for decades. However, finding effective features and training machine learning models that can generalize well to real-world applications are still challenging tasks. This undertaking, along with the emergence of speech-based virtual assistants that provide readily available platforms for voice-based emotion recognition systems [3], drives a growing interest in the speech emotion recognition (SER) body of research [2].

With advancements in technology and the heightened optimism in deep learning, SER applications are expanding rapidly in different domains. In the healthcare sector, for instance, an SER system can be used as a medical diagnostic tool in which the acoustic properties of speech are analyzed

✉ Keith April Araño
keithapril.arano@polimi.it

Carlotta Orsenigo
carlotta.orsenigo@polimi.it

Carlo Vercellis
carlo.vercellis@polimi.it

¹ Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Milan 20156, Italy

² Center for Collective Intelligence, Massachusetts Institute of Technology, Boston 02139, USA

as indicators of depression and suicide risk [4]. In [5], the authors developed a real-time SER for online learning. Additionally, the automotive industry can benefit from SER for the purpose of automatically recognizing the emotional state of the driver and responding intelligently in order to improve the driving experience and encourage better driving [6].

A variety of approaches have been proposed in terms of preprocessing, feature extraction and classification algorithms using different SER databases. Traditional SER systems involve the calculation of statistical functions for low-level descriptors such as pitch, zero-crossing rate (ZCR), energy and mel-filter bank features, which are fed into traditional supervised learning algorithms such as support vector machines (SVMs) [7]. Recently, the growing interest in deep learning has led to more advanced approaches that use, for instance, deep neural networks (DNNs) to automatically learn speech features [8]. Various DNN architectures for classifying emotions have also been presented and have exhibited the viable results [9].

A major challenge in fully exploiting the potential of DNNs for the task of emotion recognition is the lack of a large number of emotion-labeled speech datasets, which inhibits the training of a deep network from scratch. For this reason, traditional speech features such as mel-frequency cepstral coefficients (MFCCs) are still widely used due to their remarkable performance in SER experiments [3]. In this paper, we leverage the strengths of both traditional and deep learning approaches by combining the classical MFCC features with image spectrogram features extracted by means of a pretrained DNN.

In terms of classification algorithms, SVM is considered one of the most popular options in SER [7] and is among the best known classical methods for image classification [10]. A long short-term memory (LSTM) network, on the other hand, is one of the most powerful algorithms for modeling sequential data, which are inherent in speech features [8]. Hence, we carried out various experiments on the above-mentioned features using the two classification algorithms SVM and LSTM. In particular, the main contributions of this paper are as follows:

1. We evaluated the predictive power of different feature sets represented by handcrafted MFCCs and deep-learned image features by carrying out various experiments using two classification algorithms, namely SVM and LSTM. To the best of our knowledge, the combinations of the aforementioned input features and supervised learning methods have not been considered in previous studies of this subject.
2. We investigated the use of a hybrid feature set to classify emotions from speech, specifically, a fusion of MFCCs and deep-learned features that were extracted from images depicting speech spectrograms by using a

pretrained convolutional neural network (CNN) model, namely ResNet50. To the best of our knowledge, this is the first attempt to fuse such feature sets for SER tasks.

3. We developed a real-time SER application for recognizing six emotional categories by using the best prediction model generated in our experiments.

Related Work

Affective Computing and Sentiment Analysis

Emotion recognition and polarity detection are the basic tasks in affective computing and sentiment analysis [11]. Earlier works in the field have been focused on the task of polarity detection, which often involves a binary classification problem that predicts “positive” versus “negative” sentiments [12, 13]. Emotion recognition, which aims to categorize different emotional states, has also been widely studied. One of the most influential emotional theories was introduced by Lazarus [14], who distinguishes three phases in the execution of an emotion. First is cognitive appraisal, where the individual assesses the event that triggers the emotion. Second are physiological changes, such as changes in the heart rate or hormones, and third is action, where the individual feels the emotion and reacts to it. Different emotion categorization models have been commonly used by the research community, such as the Hourglass of Emotions [15], the Circumplex Model of Affect [16] and Paul Ekman’s emotion model [17], among others.

The Hourglass of Emotions uniquely combines the tasks of polarity detection and emotion recognition by inferring the polarity of a sentence directly from the conveyed emotions [11]. The Circumplex Model of Affect, on the other hand, is a two-dimensional model in which valence describes the range of negative and positive emotions and arousal characterizes the level of alertness. High valence and high arousal, for example, represent a pleasant feeling with high activation, which describes emotions such as happiness and excitement. One of the most widely applied emotion categorization frameworks is Paul Ekman’s emotion model, where he classifies emotions into six basic categories: anger, happiness, fear, surprise, disgust and sadness.

Various machine learning methods have been proposed for recognizing sentiment and emotional states, which are normally applied to textual data. Supervised learning techniques such as Naive Bayes, maximum entropy and SVM, which rely on annotated corpora, are among the most successful approaches [18]. However, the lack of labeled datasets, especially in cross-domain and cross-lingual contexts, paved the way for using semisupervised methods, which train classifiers on both annotated and unannotated corpora in order to reduce the dependence on annotated data. In

[19], for example, the authors proposed a semisupervised approach that first determines the sentiment ambiguity through a spectral clustering algorithm and then adopts an ensemble of transductive SVMs to train a final classifier on labeled data. Extreme learning machines, which are primarily applied to supervised learning problems, have also been extended to semisupervised tasks in social data analysis [20].

Clearly, ensemble methods gained popularity in sentiment analysis because they take advantage of multiple approaches employed together. Recently, Akhtar et al. [21] proposed a stacked ensemble method for emotion and sentiment intensity prediction by combining the outputs obtained from several deep learning (i.e., LSTM) and classical feature-based (i.e., support vector regression) models. Another key challenge in resorting to machine learning methods is the use of black-box algorithms that often generate the uninterpretable results. An attempt to address this challenge is leveraging both subsymbolic (i.e., machine learning) and symbolic (i.e., knowledge base) approaches. In [22], for example, a new version of SenticNet [23] was proposed by integrating logical reasoning with deep learning networks to build a knowledge base.

The aforementioned studies on polarity detection and emotion recognition are limited to the analysis of textual data. When analyzing texts, however, the algorithms deal only with words, phrases and relationships, which are often not sufficient to interpret affective content [24], especially in more challenging tasks such as polarity disambiguation [25] and sarcasm detection [26, 27]. One approach to overcoming these problems is to use fuzzy logic [28], which depends on approximate reasoning. Chaturvedi et al. [29], for instance, proposed a sentiment classifier that integrates deep learning networks and fuzzy logic. Such challenges can also be addressed by looking at other forms of modalities, which is often referred to as multimodal analysis. One of the most notable works in the field of multimodal sentiment analysis proposed a technique that is able to continuously interpret affect over time by combining natural language text and facial expressions, termed sentic blending [30]. In [31] and [32], the authors combined audio, visual and textual data for real-time multimodal sentiment analysis.

While multimodal sentiment analysis is certainly effective in predicting sentiments and emotions, the precise detection of affect in each modality is a critical component for the success of the entire multimodal model. Since each modality is deemed important, distinct bodies of research dedicated to each type of modality have arisen, such as natural language processing for the study of natural language, computer vision for analyzing image data and SER for speech signals.

Speech Emotion Recognition

Psychological studies of emotions suggest that vocal parameters, especially pitch, intensity, speaking rate and voice quality, play a significant role in the recognition of emotions and sentiments [33]. SER has been an active area of research for decades primarily because speech is one of the principal communication channels for expressing human emotions and is therefore fundamental in the field of emotion recognition. Similarly to other approaches in sentiment analysis, the development of an SER system involves three main aspects: training data preprocessing, feature extraction and classification.

Due to the laborious nature of labeling emotions from speech, emotion-labeled speech datasets with a large number of samples are inherently limited. Moreover, most publicly available datasets are collected in a controlled environment containing clean data that are free of noise and may not accurately represent conditions in the real world [3]. These obstacles are often addressed by applying preprocessing techniques such as augmenting audio data through various random transformations such as time stretching, pitch scaling and noise addition [34, 35].

Emotion recognition heavily relies on the effectiveness of the features used for the task of classification. As a consequence, feature extraction plays a prominent role, and significant research has therefore been focused on identifying the speech features that best represent different emotional states. Classical feature extraction approaches involve handcrafted acoustic features such as MFCCs, pitch, energy and ZCR [36, 37]. The recent breakthroughs in deep learning, however, have led more studies to propose end-to-end DNNs to automatically extract speech features [38, 39]. For example, in [40], a context-aware emotional feature extractor was proposed by combining a CNN with LSTM in order to automatically learn the best speech representation from raw audio signals. In another study, Lee and Tashev [8] proposed a powerful learning method with a bidirectional LSTM to extract a high-level representation of emotional states from speech.

While such end-to-end DNNs may provide the valuable results, the limitation due to the scarcity of emotion-labeled speech datasets hinders the training of DNNs from scratch. A relevant number of studies, therefore, still employ traditional handcrafted speech features, particularly MFCCs, which are reportedly one of the most conventional and effective feature sets [3, 41]. In [42], for example, MFCCs achieved notable performance on the Audio Video Emotion Challenge 2016. Other classical audio features, including MFCCs, were used in another study by AlHanai and Ghassemi [43] to predict emotional states in real time.

In terms of classification algorithms, SVM is one of the most popular and accurate classifiers in SER studies [7], where it has proven to be on par with or better than many

other advanced approaches [42]. In particular, deep-learned features combined with SVM have been shown to achieve state-of-the-art performance [44]. DNNs, such as LSTM, are also widely used for emotion recognition in speech [8, 45]. In [46], Satt et al. successfully proposed an SER system based on LSTM applied directly to raw spectrograms.

Recognizing emotions from speech spectrograms is a relatively new approach in SER as most studies have resorted to traditional low-level descriptors and, more recently, to deep-learned features from raw audio signals. Motivated by the well-established success of classical MFCCs as well as by the recent optimism on the use of spectrogram images, we propose a new SER approach by concatenating MFCC features with the spectrogram image features extracted by a pretrained CNN (i.e., ResNet50). Note that the use of a pretrained model for automatic feature extraction addresses the issue of the scarcity of training data and eliminates the lengthy and resource-intensive end-to-end network training process.

Methodology

Dataset

One of the main objectives of the present study is to develop a real-time SER system that can cover a wide range of application areas in emotion recognition. A number of theories have been proposed that categorize different types of emotions; among these, one of the most widely used in emotion recognition research is Paul Ekman's model [17], in which emotions are classified into six basic categories: *happiness*, *sadness*, *fear*, *anger*, *disgust* and *surprise*. Our experiments were therefore based on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [47] dataset, as

it is one of the few publicly available SER datasets labeled with the six emotion types listed above. This dataset has been used extensively in similar studies focused on emotion recognition from speech [9, 48, 49], therefore allowing us to compare our results with previous research findings.

The RAVDESS dataset contains recordings from 24 professional actors (12 females and 12 males) vocalizing each emotion type at two levels of emotional intensity. The emotion content of the dataset was validated by 247 participants, where each utterance was rated 10 times by 10 different raters. In our tests we used a total of 1880 audio-only files (16 bit, 48 kHz .wav) that covered the six basic emotions. Specifically, for *happiness*, *sadness*, *anger* and *fear*, we included 376 instances each; for *disgust* and *surprise*, we analyzed 192 instances.

Proposed Method

The general framework of the proposed approach is depicted in Fig. 1. As illustrated, the feature extraction process involves two parallel tasks: The first generates the MFCC feature set from the raw speech signals, and the second builds the image features from the spectrograms.

Mel-frequency cepstrum (MFC) is a special kind of cepstrum analysis originally designed to mimic how the human auditory system processes sounds. Whereas classical cepstrum analysis equally weights different frequency ranges, MFC assigns more weight to lower frequencies. This is in line with the well-known dependence on frequency of the human ear's critical bandwidths. Indeed, the range of human hearing varies generally between 20 Hz (lowest pitch) and 20 kHz (highest pitch), but its ability to distinguish individual tones is markedly greater at low frequencies. As a consequence, the distance between 100 Hz and 200 Hz is perceived by a human as being much more relevant than that

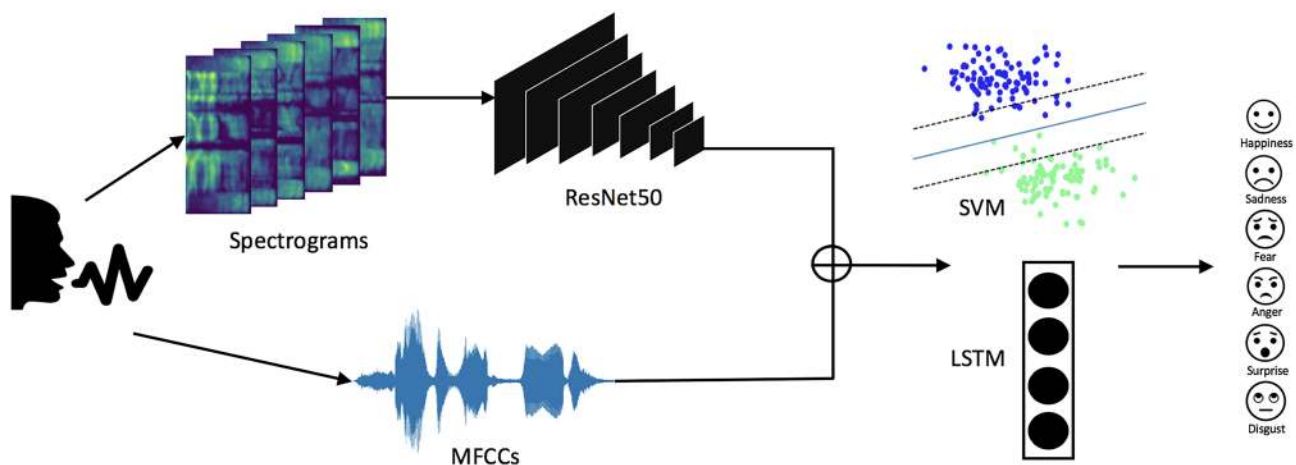


Fig. 1 Diagram of the proposed speech emotion recognition system

between 10 kHz and 10.1 kHz, even though their pairwise distances are the same.

MFCCs are the result of a cosine transform of the logarithm of the short-term spectrum expressed on a mel-frequency scale. MFCC feature extraction has been the dominant approach in speech recognition for a long time and has proven to be rather effective and robust under different conditions. Its success has also been due to the ability of emphasizing the perceptually significant characteristics of the audio spectrum and of providing a compact representation of it.

The process of generating MFCC features encompasses some major steps, where each step is motivated by perceptual or computational issues. Specifically, the raw audio signal is first passed through a high-pass filter (*pre-emphasis*) to amplify the high frequencies and is cut into overlapping frames to capture local spectral properties (*framing*). *Windowing* is then performed on the frames to reduce the amplitude of the discontinuities at the boundaries and therefore minimize the spectral distortion. A discrete Fourier transform is then applied to transform each frame into the frequency domain and to generate the short-term power spectrum. This is subsequently warped along the mel-frequency scale (*mel-frequency warping*), which can be defined as a perceptual scale of pitches. This warping is applied with the aim of reflecting the response of the human ear, which does not perceive pitches in a linear manner but is more sensitive to lower-frequency components. The mel-warped spectrum is then segmented into a number of critical bands by means of a filter bank, which typically consists of overlapping triangular filters. Finally, to generate the MFCC vector, a discrete cosine transformation is applied to the logarithm of the K filter bank outputs, E_k , $k = 1, \dots, K$, so that the d th MFCC can be computed as

$$MFCC_d = \sum_{k=1}^K E_k \cos[d(k - 0.5)\pi/K], \quad d = 1, \dots, D. \quad (1)$$

To create the MFCC vector, we used the feature extraction module of *librosa*, a Python package for music and audio analysis [50]. Specifically, we extracted the audio time series data from the audio recordings by means of a standard sampling rate (i.e., number of samples per second) of 44.1 kHz and used the default parameters pertaining to the short-time Fourier transform (STFT). In particular, we fixed the STFT window (i.e., the number of samples per window) to 2048 and the hop length (i.e., the number of samples between successive frames) to 512 samples. This package implements the Hann window function, which has a sinusoidal shape similar to that of the most common Hamming window but, in contrast to it, touches zero at both ends, thereby removing any discontinuity. In our experiments, a vector of 40 MFCCs was extracted ($D = 40$) by excluding the zero-order coefficient, which indicates the average power of the input signal and has

been shown to carry limited speech information [51]. Notice that in MFC, the lower-order coefficients generally contain most of the information on the overall spectral shape, and few of them, approximately 12–20, are usually retained for speech analysis. Based on our tests, however, we deemed it appropriate to include higher-order coefficients, which represent increasing levels of spectral detail.

To generate the RGB spectrograms from the speech signals, we used the same audio package, which performs a short-time Fourier transform to calculate the frequency spectrum over short, overlapping windows based on the following equation:

$$X(\omega, m) = STFT(x(n)) = \sum_{n=0}^{R-1} x(n - m)w(n)e^{-i\omega n} \quad (2)$$

where $x(n)$ is the input signal at time n and $w(n)$ is the window function of length R . Also, in this case, we adopted the Hann window and set the hop length to 512 samples. The image features were then extracted from the spectrograms by using a pretrained CNN model similar to that of [52], where the authors explored a transfer learning approach to determine the relevance of using spectrograms for the purpose of emotion prediction. In particular, we performed experiments on image feature extraction by trying different ImageNet pretrained models, namely ResNet50, VGG19, InceptionV3 and Xception. Among these deep learning neural networks, ResNet50 provided the best performance and was therefore selected as our image feature extractor. For this reason, the spectrogram images were resized to 224 x 224 pixels consistently with the input shape required by the Res-Net50 model. Figure 2 shows some sample spectrograms for each of the six emotion types in the RAVDESS dataset.

For the purpose of generating a single hybrid feature vector, we combined the MFCCs with the spectrogram image features according to the *early fusion* method, which creates a joint representation of input features of multiple modalities and can therefore be considered a viable way to perform multimodal learning. Indeed, early fusion has proven to be rather effective in the area of multimodal emotion recognition, where features extracted from various modalities, such as image, audio and text, are fused into a single vector for analysis [1]. The joint representation adopted here is based on the most commonly used criterion, which consists of concatenating the features. In this way, the subsequent classifier is expected to learn the interactions between the low-level features of each modality. In our study, the early fusion method gave rise to the following hybrid feature vector:

$$Hybrid = (MFCC_1, \dots, MFCC_D, IF_1, \dots, IF_L) \quad (3)$$

where the first sub-vector contains 40 MFCCs and the second contains 100352 image features derived from the spectrograms. In case of the hybrid representation, the final

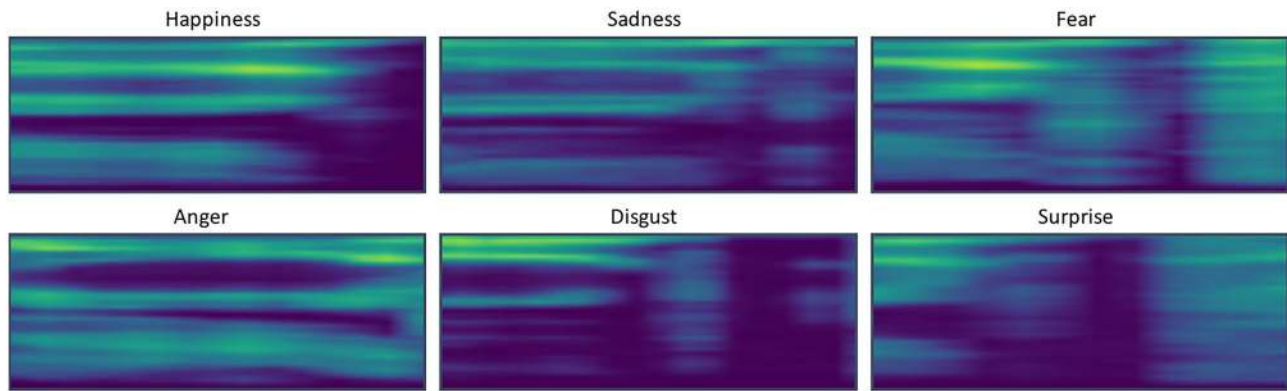


Fig. 2 Spectrograms of the six emotions

dataset was therefore described in terms of 100392 predictors, which were provided as input to the selected classifiers for the purpose of emotion recognition.

To evaluate the usefulness of the proposed methods, we investigated and compared the performance of MFCCs and spectrogram-based features, denoted as SpecsResNet, taken individually as inputs, as well as the performance of their early fusion combination in forming hybrid features (called Hybrid). In particular, we trained and evaluated six different models by using SVM and LSTM as classifiers. The first three models, relying on SVM and fed the alternative feature sets, were denoted as *MFCC-SVM*, *SpecsResNet-SVM* and *Hybrid-SVM*. The remaining models, employing the same LSTM architecture, were instead denoted as *MFCC-LSTM*, *SpecsResNet-LSTM* and *Hybrid-LSTM*.

The performance of all methods was evaluated according to the following scheme: The original dataset was divided into a training and a test set composed of 70% and 30% of the available instances, respectively. In doing this, a stratified random sampling strategy was adopted so that the training and test sets had approximately the same class distributions. Moreover, on the training set, a preprocessing step was performed with the purpose of fine-tuning some model parameters. In particular, threefold cross-validation was applied to select the kernel type for SMV-based models and

the size and the activation function for LSTM. Based on this first round of experiments, SVM with a linear kernel turned out to be the best choice for SpecsResNet and the Hybrid feature sets, whereas a polynomial kernel with degree 2 was selected for the MFCC-based model. For LSTM, instead, the architecture used in combination with the three sets of input features was composed of one input layer, three hidden layers (i.e., one dropout layer and two dense layers) and one dense output layer with a softmax activation function. Finally, to solve the multicategory problem with SVM, we used a one-versus-all framework.

Results

The classification results of the competing techniques are illustrated in Table 1, which contains the F1-score on the test set for each emotional state as well as two global measures represented by the overall F1-score and the overall accuracy. Here, the term *overall* refers to the average of the outcomes across the different prediction tasks. In particular, the global measures were computed as the mean over the six F1-score and accuracy values weighted by the number of samples belonging to the different classes. The use of the F1-score and the weighted accuracy as performance metrics ensures

Table 1 Performance of all the models on the test set across the six emotions

Model	F1-Score						Accuracy	
	Happiness	Sadness	Fear	Anger	Disgust	Surprise	Overall	Overall
MFCC-LSTM	0.769	0.748	0.672	0.791	0.715	0.690	0.736	0.735
Hybrid-SVM	0.646	0.689	0.633	0.791	0.838	0.736	0.710	0.713
SpecsResNet-SVM	0.657	0.621	0.575	0.771	0.724	0.767	0.674	0.676
MFCC-SVM	0.667	0.694	0.655	0.779	0.643	0.662	0.689	0.690
SpecsResNet-LSTM	0.523	0.601	0.490	0.682	0.680	0.586	0.586	0.582
Hybrid-LSTM	0.444	0.505	0.460	0.636	0.642	0.458	0.519	0.524
Average F1-Score	0.618	0.643	0.581	0.742	0.707	0.650		

that the class imbalance in the dataset is accounted for. To refine the performance evaluation, we also combined the individual outcomes in Table 1 across the emotional states. Specifically, for each method, we computed a score given by the ratio between its error rate, defined as 1 minus the F1-score, and the lowest error rate on a given emotion. For a model, the sum of the scores on all emotions can be seen as a measure of its ability to generate predictions that are close, if not equal, to the best ones in terms of accuracy. The results of this analysis are depicted in Fig. 3.

From the outcomes of Table 1 and Fig. 3, some empirical conclusions can be drawn. By focusing on the use of SVM as the classifier, we observe that the proposed hybrid input feature vector outperformed MFCCs and spectrogram-based features taken individually. In particular, compared to MFCC-SVM and SpecsResNet-SVM, Hybrid-SVM led to an improvement in the overall F1-score and accuracy of at least 2% and 2.3%, respectively. Note that

the results of Hybrid-SVM are fairly superior to the accuracy achieved on the same RAVDESS dataset in prior studies that use SVM-based methods for classification. For example, Zhang et al. [48] investigated the usefulness of alternative groups of acoustic features represented by different types of low-level descriptors. Among these, in the emotion recognition task, MFCCs derived by all (spoken and sung) utterances and combined with SVM achieved an accuracy of 0.538, which dropped to 0.487 when only spoken utterances were used. Our computational tests, therefore, support the effectiveness of a hybrid representation relying on MFCCs and spectrogram-based features, where the latter descriptors enhance the predictive ability of the former when they are used in conjunction with classical learning methods such as SVM. In particular, our findings shed light on the effectiveness of a pretrained CNN model, namely ResNet50, used as a feature extractor from speech spectrogram images. For the task of emotion prediction,

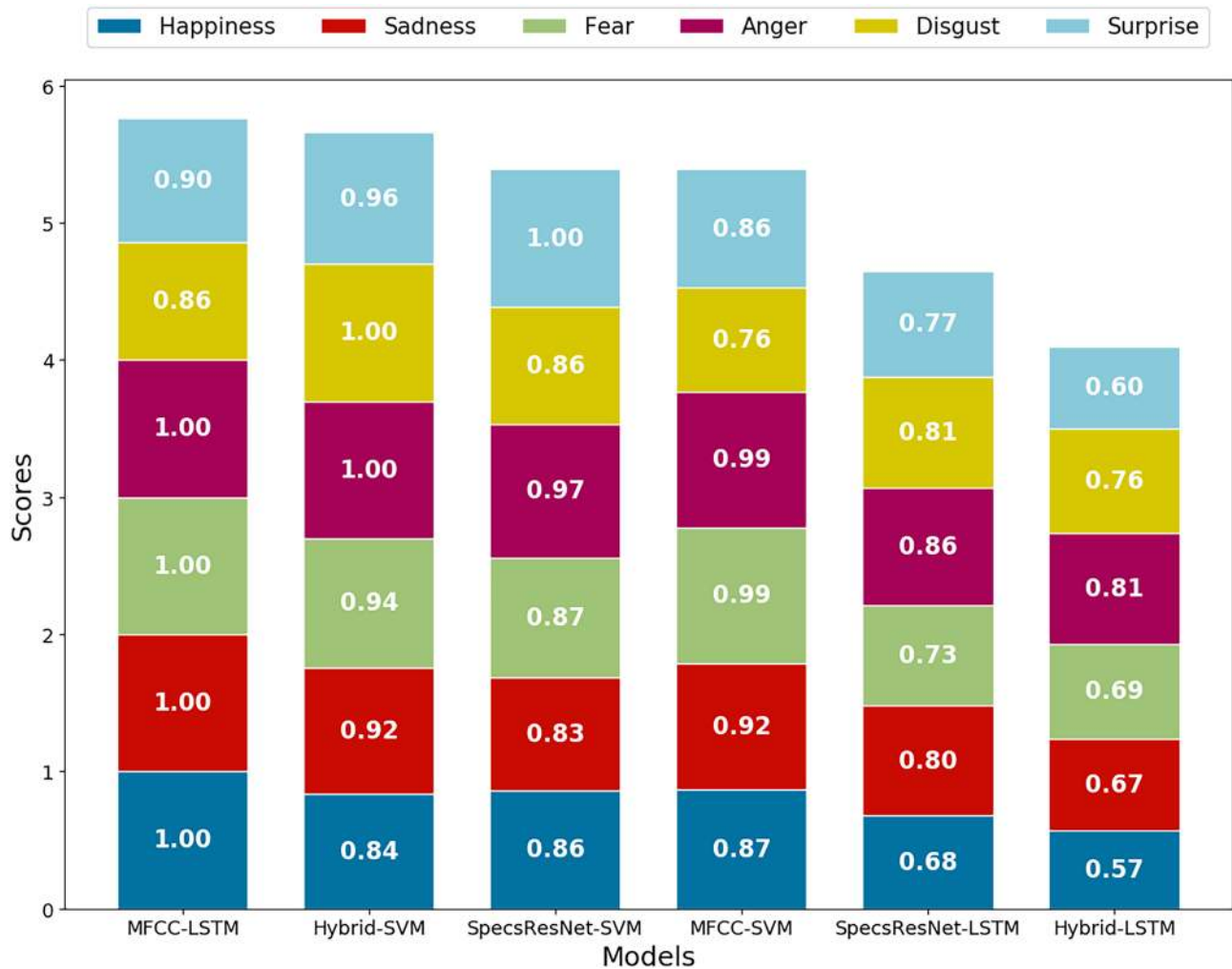


Fig. 3 Prediction scores based on the F1-scores of the different models across the six emotions

indeed, this approach also showed the interesting results on certain emotional states (i.e., *surprise*) when it was exploited to generate a unique input feature vector to feed an SVM classifier (SpecsResNet-SVM model). Note that this method eliminates the lengthy and resource-intensive process of network training from scratch. Therefore, it deserves attention as a promising novel approach for emotion classification from speech data when this task is formulated as an image classification problem, as suggested in [53]. Finally, our experiments also confirmed the suitability of SVM when hybrid representations are involved as input for classification. Indeed, our results agree with prior studies that demonstrated the effectiveness of SVM for emotion prediction when the learning is based on multimodal or fused data sources [54].

When deep learning algorithms (i.e., LSTM) were exploited for classification, we observed the superiority of MFCCs, which, taken as unique input, were able to outperform to a large extent the spectrograms and hybrid features combined with LSTM and to achieve a slightly higher accuracy compared to that of Hybrid-SVM. This noteworthy behavior is confirmed by the score analysis. Indeed, MFCC-LSTM reached the highest total score, being associated with ratios that were more densely distributed around 1 compared to those of the other techniques (Fig. 3). The success of MFCC-LSTM shows that, despite being traditional speech descriptors, MFCCs can still dominate more sophisticated deep-learning features. We observe that a known benchmark on the RAVDESS dataset was achieved in [49], where the authors reached an accuracy of 0.787 by using MFCCs and variants of artificial neural networks. Their result, however, is not directly comparable with our outcome for certain reasons. In particular, they cast the problem as a 5-class prediction task by including the unified *anxiety/stress* emotion in place of *fear* and *surprise*, where the former is the most difficult emotion to detect, as reported by the last row in Table 1, which indicates, for each emotion, the average F1-score over all methods. Moreover, they built the model on a larger training set by selecting 25% of the examples for testing. Nevertheless, the result achieved in our more difficult 6-class prediction problem can be judged to be at least on par with the benchmark obtained for the 5-class learning task. Furthermore, it is worth noting that the superiority of MFCC-LSTM over Hybrid-SVM is evaluated based on the global F1-score and the global accuracy, which both account for the class imbalance. Consequently, the difference in performance between MFCC-LSTM and Hybrid-SVM, where the latter prevailed over the former for *disgust* and *surprise* and was equally accurate on *anger*, can be significantly reduced by including more samples for *disgust* and *surprise*, which are precisely the underrepresented classes in the dataset.

By looking closely at the outcomes for the six emotion types, we observed that *fear* and *happiness* emerged as the most difficult to predict, while *anger* was consistently well detected by all models (Fig. 3). These findings are aligned with previous studies that showed that *happiness* is difficult to recognize due to its specific characteristics (i.e., it relies on distinct contextual information more than other emotion types) [9], whereas *anger* is often associated with the highest prediction accuracy and is therefore clearly the easiest emotion to predict from speech features [2, 55].

The direct comparison between the two top-ranked models, MFCC-LSTM and Hybrid-SVM, also revealed that the performance gap largely depends on *happiness*, which is more heavily misclassified by the latter. To investigate this discrepancy we analyzed the confusion matrices (Fig. 4), which showed that, for models exploiting spectrogram-based features, *happiness* is more often confused with *sadness*, *fear* and *anger*. This behavior suggested that a possible cause could be sought in the spectrograms of the aforementioned emotions, which make their mutual distinction more difficult. For this reason, we further analyzed the spectrograms by evaluating their similarity, based on the idea that features extracted from highly similar spectrograms may hinder the classification of the related emotions in the subsequent prediction task. To this end, we computed two indices between the spectrograms for each pair of emotions, represented by the cosine similarity and the structural similarity (SSIM) index [56], which accounts for image luminance l , contrast c and structural similarity s , as follows:

$$SSIM(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha \cdot [c(\mathbf{x}, \mathbf{y})]^\beta \cdot [s(\mathbf{x}, \mathbf{y})]^\gamma \quad (4)$$

where \mathbf{x} and \mathbf{y} are two image signals and α , β and γ are non-negative parameters used to regulate the importance of the three components, which are all set to 1 in our tests.

The results of this analysis are depicted in Fig. 5, which contains the average pairwise cosine similarity of the six emotions (part (a)) and a graph-based representation of the SSIM index (part (b)). These results support our conjecture regarding the misclassification of *happiness* by spectrogram-based models, which can be misleading due to the high degree of spectrogram similarity between *happiness* and *sadness*, *fear* and *anger*, as emerges clearly from Fig. 5. The same analysis highlights also a substantial similarity between *fear* and *sadness*. This may explain the performance on these two emotions of the spectrogram-based methods, which labeled examples as sad and fearful in at least 15.9% of cases. Hence, as a general remark, spectrogram-related features can be useful for the prediction of some emotions (i.e., *sadness*, *anger*, *disgust*, and *surprise*), but not all. This paves the way for further investigation of the effectiveness of such descriptors by

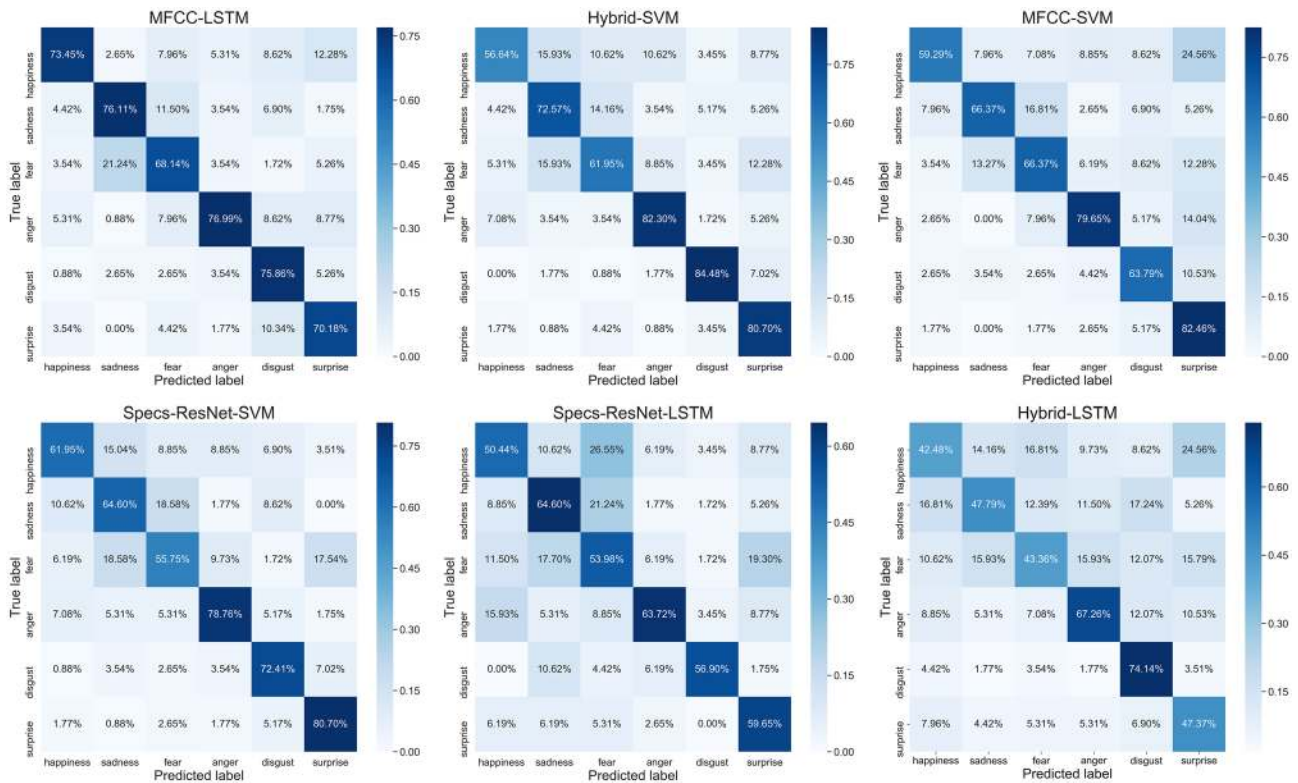


Fig. 4 Confusion matrices of the six models

devising novel features to be extracted that are able to better grasp and reflect the inherent diversity among the visual representations of the spectrum of frequencies of emotions' signals.

By adopting the best classification model based on our experiments (i.e., MFCC-LSTM), we built a real-time SER tool that takes speech signals as input and makes instantaneous

emotion predictions. In particular, to increase the generalizability of the application, we retrained the final model on a larger sample obtained by combining different publicly available datasets: the Surrey Audio-Visual Expressed Emotion [57] dataset, Crowd-Sourced Emotional Multimodal Actors Dataset [58], Interactive Emotional Dyadic Motion Capture [59] dataset, Toronto Emotional Speech Set [60] and Berlin Database of

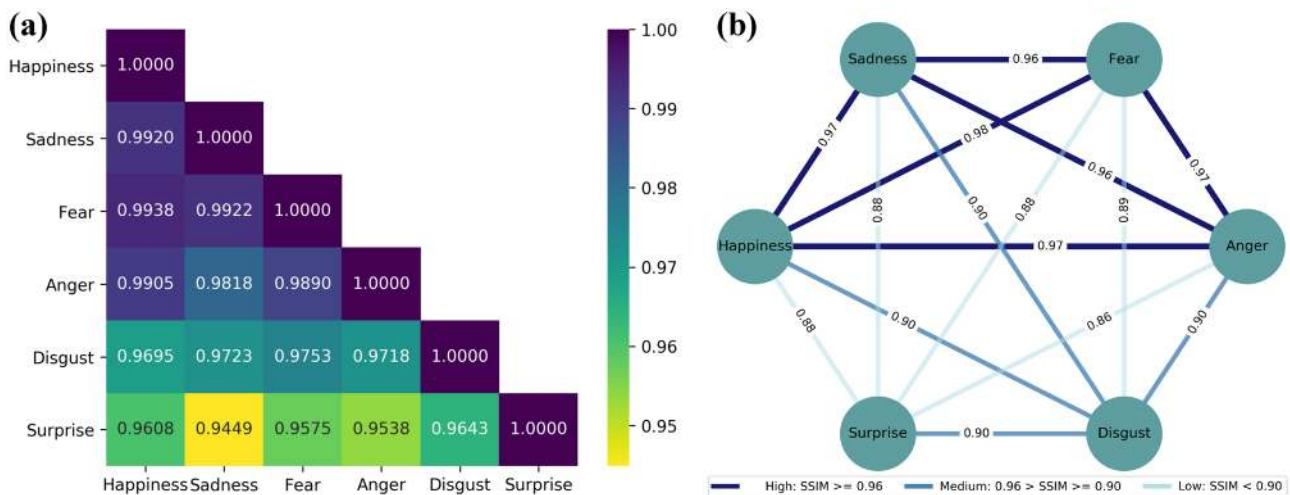


Fig. 5 Spectrogram similarities between all pairwise emotion combinations: (a) average pairwise cosine similarity (b) graph-based representation of the SSIM index

Emotional Speech [61]. The collection of these data provided a total of 14,309 instances of 2–5 seconds of speech.

Figure 6 illustrates the graphical user interface of the SER application, which displays the instantaneous probabilities for each of the six emotions as well as the overall emotion probabilities of the speech stream over time. These probabilities are derived from the softmax layer of the LSTM network. In particular, the overall predicted emotion \hat{e} was computed over the entire length of the input speech signals by means of a hard majority voting scheme. Thus, it was defined as the most frequent emotion label assigned to the entire speech stream from time t_1 to time t_n , according to the following expression:

$$\hat{e} = \text{mode}\{e_{t_1}, e_{t_2}, \dots, e_{t_n}\} \quad (5)$$

where e_{t_i} is the predicted emotion at time t_i , $i = 1, \dots, n$.

The performance of real-time SER systems is subject to the length of time necessary to process the speech features.

The use of our MFCC-LSTM model enables the real-time feasibility of our tool primarily because the extraction and preprocessing of the MFCC features require minimal computational effort. As shown in Table 2, the process of feature extraction and preprocessing is on the order of milliseconds, allowing for real-time prediction of emotional states. Note that this analysis was performed on a machine with 8 GB RAM and a 4-core CPU.

The proposed SER tool, through the related application programming interface (API), can be effectively framed within intelligent systems in order to automatically recognize emotions from speech, contribute to a better interpretation of verbal communication and enhance their capacity to interact with humans in a natural way. This plays a fundamental role currently, when it is believed that artificial systems need to become more sensitive toward human emotions in order to be more user friendly, trustworthy and effective. The developed SER application is available upon request.

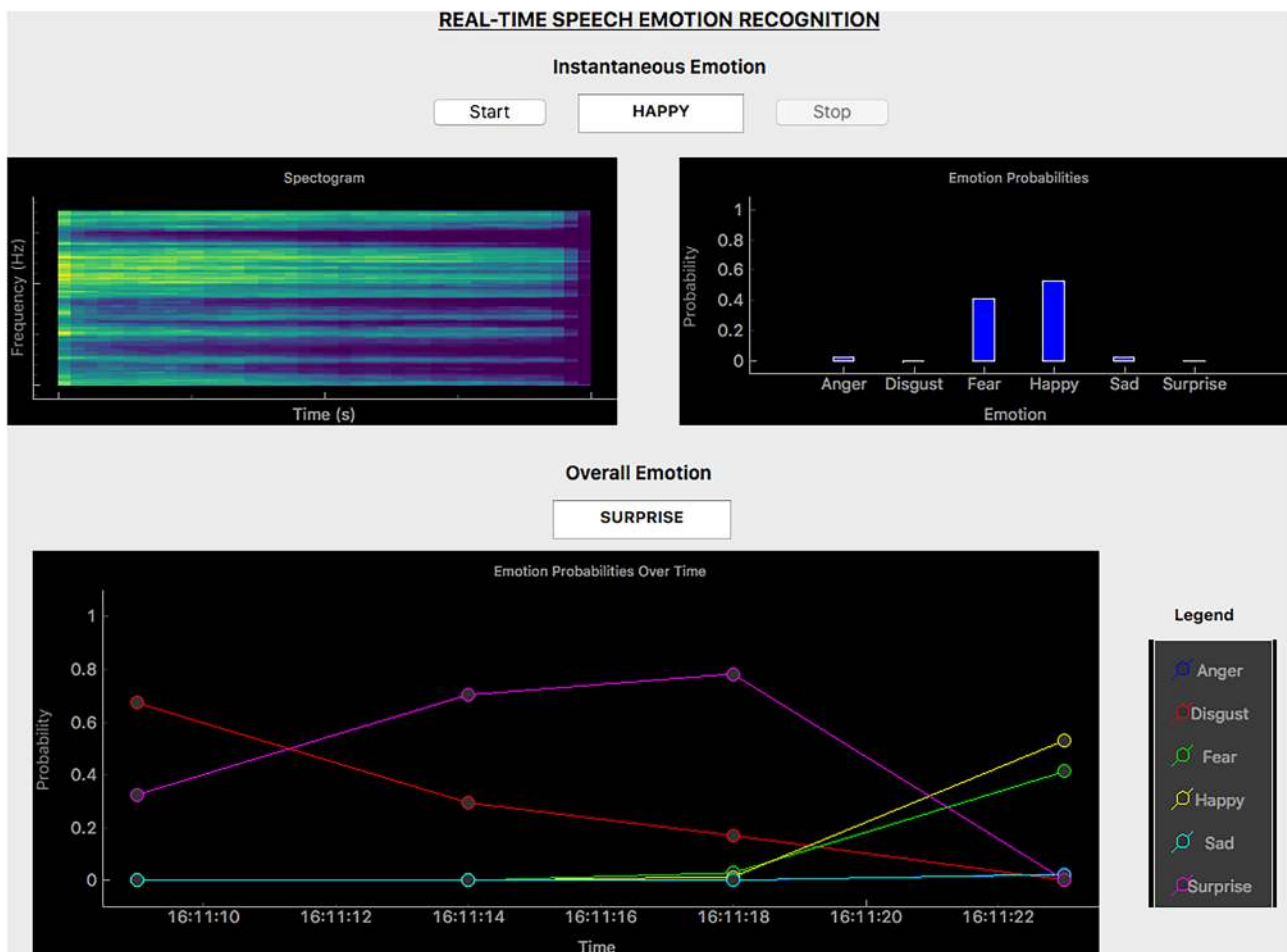


Fig. 6 Real-time SER system

Table 2 Time analysis of the real-time SER tool on a machine with 8 GB RAM and a 4-core CPU

Time stamp (H:M:S)	Feature extraction (milliseconds)	Feature preprocessing (milliseconds)	Emotion prediction (milliseconds)	Total time (milliseconds)
14:25:54	100.868	0.857	2.290	104.015
14:25:59	97.278	0.862	2.627	100.767
14:26:04	100.782	0.860	2.632	104.274
14:26:08	99.688	0.878	2.602	103.168
14:26:13	101.934	0.853	2.181	104.968
14:26:17	99.250	0.855	2.370	102.475
14:26:22	99.512	0.928	2.445	102.885
14:26:26	96.400	0.8500	2.358	99.608
14:26:31	96.706	0.866	2.566	100.138
14:26:36	98.009	0.857	2.736	101.602
Average Time	99.043	0.867	2.481	102.390

Conclusions and Future Work

In this paper, we investigated the fusion of speech MFCCs and image features extracted from signal spectrograms for the task of emotion recognition. In particular, we evaluated the effectiveness of different combinations of supervised learning algorithms, namely SVM and LSTM, and feature sets, represented by MFCCs, spectrogram images and their hybrid fusions, on the benchmark RAVDESS dataset. By using the best-trained model, we built a real-time SER tool that takes in input speech signals and predicts instantaneous and overall emotions over time.

Our results demonstrated that with a suitable classifier, in the hybrid fusion approach, MFCCs and spectrogram images do complement each other, resulting in an improvement in the prediction accuracy. Concerning the classification algorithms, our experiments revealed that classical SVM is recommended for handling such hybrid features, whereas LSTM provides the most favorable results when the more conventional MFCC features are taken as unique input. In regard to the six basic emotions, *happiness* emerged as one of the most difficult to predict when using speech features. Indeed, as noted previously in the literature, it relies more on context information than the other emotion types [9]. To build a real-time SER tool, we used the MFCC-LSTM model, which was shown to be the best in our experiments. It is worth noting that the extraction and preprocessing of the MFCC features in real time require less computational effort than the hybrid approach. This enhances the feasibility of the real-time system. By means of the related API, our SER application can be integrated into any intelligent system and can help practitioners who are working on the development of real-time emotion recognition tools. The developed SER software can be useful in several areas. For example, it could effectively support activities in a call center, where emotions stemming from phone conversations could be

monitored in real time. This would enable agents to better handle phone calls and ultimately increase customer satisfaction [62, 63]. Team collaboration measurement tools, such as the Happimeter [64], are other promising use cases for our work.

Despite the novel contributions of this study, we are aware of certain challenges and limitations. First, the limited number of emotion-labeled speech datasets prevented us from building and evaluating our models on larger and more diverse datasets. The generalizability of our prediction models could be further enhanced by using a larger speech training dataset collected from diverse populations in terms of language or culture, when such a dataset becomes available. Second, we found it challenging to predict *happiness* from speech features. Future works can therefore focus on finding novel descriptors that best represent this emotional state. For example, *happiness* could be better detected by means of other forms of modalities such as text, where the context can be better captured, or facial expressions and physiological characteristics such as changes in heart rate, which have been studied in multimodal emotion recognition research [42, 54]. Finally, our work evaluated two well-known algorithms for classification. Our findings, however, encourage future research directions devoted to exploring more sophisticated ensemble learning algorithms as well as alternative fusion methods.

Acknowledgements This work was supported by Philips Lighting as part of an overall MIT Grand Challenges in Lighting research grant.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflicts of Interest The authors declare that they have no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*. 2017;37:98–125.
2. Badshah AM, Rahim N, Ullah N, Ahmad J, Muhammad K, Lee MY, Kwon S, Baik SW. Deep features-based speech emotion recognition for smart affective services. *Multimed Tools Appl*. 2019;78(5):5571–89.
3. Marechal C, Mikołajewski D, Tyburek K, Prokopowicz P, Bougueroua L, Ancourt C, Węgrzyn-Wolska K. Survey on ai-based multimodal methods for emotion detection. In: Kołodziej J, González-Vélez H, editors. *High-performance modelling and simulation for big data applications: selected results of the COST Action IC1406 cHiPSet*. Cham: Springer International Publishing; 2019. p. 307–24.
4. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans Biomed Eng*. 2000;47(7):829–37.
5. Cen L, Wu F, Yu ZL, Hu F. Chapter 2 - a real-time speech emotion recognition system and its application in online learning. In: Tettegah SY, Gartmeier M, editors. *Emotions. Technology, Design, and Learning, Emotions and Technology*. San Diego: Academic Press; 2016. p. 27–46.
6. Jones CM, Jonsson IM. Performance analysis of acoustic emotion recognition for in-car conversational interfaces. In C. Stephanidis, editor. *Universal access in human-computer interaction. Ambient interaction*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. p. 411–420.
7. Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2017. p. 2227–2231.
8. Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*. 2015. p. 1537–1540.
9. Wu X, Liu S, Cao Y, Li X, Yu J, Dai D, Ma X, Hu S, Wu Z, Liu X, Meng H. Speech emotion recognition using capsule networks. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2019. p. 6695–6699.
10. Thai LH, Hai TS, Thuy NT. Image classification using support vector machine and artificial neural network. *International Journal of Information Technology and Computer Science*. 2012;4:32–8.
11. Cambria E, Das D, Bandyopadhyay S, Feraco A. *A practical guide to sentiment analysis*. 1st ed. Incorporated: Springer Publishing Company; 2017.
12. Liu B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*. 2012;5(1):1–167.
13. Pang B, Lee L. Opinion mining and sentiment analysis. *Found Trends Inf Retr*. 2008;2(1–2):1–135.
14. Lazarus RS, Lazarus BN. *Passion and reason: making sense of our emotions*. New York: Oxford University Press; 1994.
15. Susanto Y, Livingstone AG, Ng BC, Cambria E. The hourglass model revisited. *IEEE Intell Syst*. 2020;35(5):96–102.
16. Posner J, Russell JA, Peterson BS. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev Psychopathol*. 2005;17(3):715–734.
17. Ekman P. Basic emotions. In *Handbook of Cognition and Emotion*, John Wiley & Sons, Ltd; 1999. p. 45–60.
18. Sun S, Luo C, Chen J. A review of natural language processing techniques for opinion mining systems. *Information Fusion*. 2017;36:10–25.
19. Dasgupta S, Ng V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics. 2009. p. 701–709.
20. Huang G, Song S, Gupta JND, Wu C. Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*. 2014;44(12):2405–17.
21. Akhtar MS, Ekbal A, Cambria E. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. *IEEE Comput Intell Mag*. 2020;15(1):64–75.
22. Cambria E, Li Y, Xing FZ, Poria S, Kwok K. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery. New York, NY, USA, 2020. p. 105–114.
23. Cambria E, Poria S, Hazarika D, Kwok K. Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2018. p. 1795–1802.
24. Poria S, Hussain A, Cambria E. *Multimodal sentiment analysis. Socio-Affective Computing*. Springer International Publishing, 2018.
25. Rajabi Z, Valavi MR, Hourali M. A context-based disambiguation model for sentiment concepts using a bag-of-concepts approach. *Cogn Comput*. 2020.
26. Justo R, Corcoran T, Lukin SM, Walker M, Torres MI. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowl-Based Syst*. 2014;69:124–33.
27. Poria S, Cambria E, Hazarika D, Vij P. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*. 2016. p. 1601–1612.
28. Zadeh LA. Fuzzy sets. *Inf Control*. 1965;8(3):338–53.
29. Chaturvedi I, Satapathy R, Cavallari S, Cambria E. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recogn Lett*. 2019;125:264–70.
30. Cambria E, Howard N, Hsu J, Hussain A. Sentic blending: scalable multimodal fusion for the continuous interpretation of semantics and sentics. In *Proceedings of the 2013 IEEE Symposium on Computational Intelligence for Human-like Intelligence*. 2013. p. 108–117.
31. Poria S, Cambria E, Howard N, Huang GB, Hussain A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*. 2016;174:50–9.
32. Tran HN, Cambria E. Ensemble application of ELM and GPU for real-time multimodal sentiment analysis. *Memetic Computing*. 2018;10(1):3–13.

33. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J Acoust Soc Am*. 1993;93(2):1097–1108.
34. Elshaer MEA, Wisdom S, Mishra T. Transfer learning from sound representations for anger detection in speech. *arXiv preprint* 2019. [arXiv:1902.02120](https://arxiv.org/abs/1902.02120).
35. Etienne C, Fidanza G, Petrovskii A, Devillers L, Schmauch B. Speech emotion recognition with data augmentation and layer-wise learning rate adjustment. *arXiv preprint* 2018. [arXiv:1802.05630](https://arxiv.org/abs/1802.05630).
36. Schuller BW, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association*. 2007. p. 2253–2256.
37. Tahon M, Devillers L. Towards a small set of robust acoustic features for emotion recognition: challenges. *IEEE Trans Audio Speech Lang Process*. 2016;24(1):16–28.
38. Tzirakis P, Zhang J, Schuller BW. End-to-end speech emotion recognition using deep neural networks. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2018. p. 5089–5093.
39. Aytar Y, Vondrick C, Torralba A. Soundnet: learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016. p. 892–900.
40. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, Zafeiriou S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2016. p. 5200–5204.
41. Gunawan T, Alghifari MF, Morshidi MA, Kartiwi M. A review on emotion recognition algorithms using speech analysis. *Indonesian Journal of Electrical Engineering and Informatics*. 2018;6:12–20.
42. Brady K, Gwon Y, Khorrami P, Godoy E, Campbell W, Dagli C, Huang TS. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016. p. 97–104.
43. AlHanai T, Ghassemi MM. Predicting latent narrative mood using audio and physiologic data. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. 2017. p. 948–954.
44. Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*. 2014. p. 223–227.
45. Lian Z, Li Y, Tao J, Huang J. Unsupervised representation learning with future observation prediction for speech emotion recognition. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*. 2019. p. 3840–3844.
46. Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*. 2017. p. 1089–1093.
47. Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS One*. 2018;13(5):1–35.
48. Zhang B, Essl G, Provost EM. Recognizing emotion from singing and speaking using shared models. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction*. 2015. p. 139–145.
49. Tomba K, Dumoulin J, Mugellini E, Abou Khaled O, Hawila S. Stress detection through speech analysis. In *Proceedings of the 2018 International Conference on Signal Processing and Multimedia Applications*. 2018. p. 394–398.
50. McFee B, McVicar M, Nieto O, Balke S, Thome C, Liang D, Battenberg E, Moore J, Bittner R, Yamamoto R, Ellis D, Stoter F-R, Repetto D, Waloschek S, Carr C, Kranzler S, Choi K, Viktorin P, Santos JF, Holovaty A, Pimenta W, Lee H. *librosa/ librosa: 0.8.0*, 2020.
51. Rao KS, Nandi D. *Language identification using excitation source features*. Incorporated: Springer Publishing Company; 2015.
52. Lech M, Stolar M, Bolia R, Skinner M. Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images. *Advances in Science, Technology and Engineering Systems Journal*. 2018;3:363–71.
53. Stolar MN, Lech M, Bolia RS, Skinner M. Real time speech emotion recognition using rgb image classification and transfer learning. In *Proceedings of the 2017 International Conference on Signal Processing and Communication Systems*. 2017. p. 1–8.
54. Koldijk S, Neerincx MA, Kraaij W. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Trans Affect Comput*. 2018;9(2):227–39.
55. Luo D, Zou Y, Huang D. Speech emotion recognition via ensemble neural networks. In *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. 2017. p. 1351–1355.
56. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600–612.
57. Haq S, Jackson PJB, Edge J. Audio-visual feature selection and reduction for emotion classification”audio-visual feature selection and reduction for emotion classification. In *Proceedings of the 2008 Conference on Auditory-Visual Speech Processing*. 2008. p. 185–190.
58. Cao H, Cooper D, Keutmann M, Gur R, Nenkova A, Verma R. Crema-d: crowd-sourced emotional multimodal actors dataset. *IEEE Trans Affect Comput*. 2014;5:377–390.
59. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan S. Iemocap: interactive emotional dyadic motion capture database. *Lang Resour Eval*. 2008;42(4):335–59.
60. Pichora-Fuller MK, Dupuis K. *Toronto emotional speech set (TESS)*, 2020.
61. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. A database of german emotional speech. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, vol. 5. 2005:1517–1520.
62. Chen M, Tian Y, Fortino G, Zhang J, Humar I. Cognitive internet of vehicles. *Comput Commun*. 2019;120:58–70.
63. Chen M, Herrera F, Hwang K. Cognitive computing: architecture, technologies and intelligent applications. *IEEE Access*. 2018;6:19774–83.
64. Budner P, Eirich J, Gloor PA. Making you happy makes me happy - measuring individual mood with smartwatches. *arXiv preprint* 2017. [arXiv:1711.06134](https://arxiv.org/abs/1711.06134).