Archives of Acoustics Vol. 49, No. 2, pp. 153–168 (2024), doi: 10.24425/aoa.2024.148784

Research Paper

Speech Emotion Recognition Using a Multi-Time-Scale Approach to Feature Aggregation and an Ensemble of SVM Classifiers

Antonina STEFANOWSKA, Sławomir K. ZIELIŃSKI*

Faculty of Computer Science, Białystok University of Technology
Białystok, Poland

*Corresponding Author e-mail: s.zielinski@pb.edu.pl

(*received January 31, 2023; accepted November 21, 2023; published online March 26, 2024*)

Due to its relevant real-life applications, the recognition of emotions from speech signals constitutes a popular research topic. In the traditional methods applied for speech emotion recognition, audio features are typically aggregated using a fixed-duration time window, potentially discarding information conveyed by speech at various signal durations. By contrast, in the proposed method, audio features are aggregated simultaneously using time windows of different lengths (a multi-time-scale approach), hence, potentially better utilizing information carried at phonemic, syllabic, and prosodic levels compared to the traditional approach. A genetic algorithm is employed to optimize the feature extraction procedure. The features aggregated at different time windows are subsequently classified by an ensemble of support vector machine (SVM) classifiers. To enhance the generalization property of the method, a data augmentation technique based on pitch shifting and time stretching is applied. According to the obtained results, the developed method outperforms the traditional one for the selected datasets, demonstrating the benefits of using a multi-time-scale approach to feature aggregation.

Keywords: speech emotion recognition; feature aggregation; ensemble classification.

1. Introduction

Since the publication of Picard's seminal report in 1995 (Picard, 1995), "affective computing", which concerns the identification, modelling, and reacting to human emotions by machines, has played an increasingly important role in the development of artificial intelligence algorithms. A growing interest of researchers in the area of affective computing is driven by the demands for emotion-aware applications. For example, the algorithms processing human emotions could be applied in health and safety systems, call centers, marketing recommenders, and forensic software. While human emotions could be recognized using a variety of methods, including facial recognition (Jain *et al.*, 2019), analysis of body movements (Zacharatos *et al.*, 2021), or through the exploration of physiological data (Yang *et al.*, 2023), the scope of this paper is limited to the identification of emotions based solely on speech signals.

The methods applied to computational emotions recognition can be divided into the following two groups. The first one consists of the algorithms using the audio feature extractors combined with the classical machine learning algorithms. The second one is based on modern deep learning algorithms such as the convolutional neural networks. The performance of the speech emotion recognition methods has recently greatly improved, primarily due to the incorporation of the aforementioned deep learning techniques (Khalil *et al.*, 2019; Pandey *et al.*, 2019). The main advantage of deep learning techniques, over the traditional methods, is that they normally do not require any feature extraction procedure, typically engineered manually by domain experts. The speech signals are either fed to the inputs of the deep learning algorithms directly (Tzirakis *et al.*, 2018) or indirectly through some form of intermediate transformations, most notably spectrograms (Eskimez *et al.*, 2018; Choi *et al.*, 2018; Zhao *et al.*, 2018; 2019; Zhang *et al.*, 2020; Guizzo *et al.*, 2020; Tang *et al.*, 2021). Nevertheless, the methods based on the deep learning approach still exhibit some limitations. For example, they require relatively large data sets for training. Moreover, they may suffer from over-learning during memorization of layer-wise information (Khalil *et al.*, 2019). Furthermore, due to their relatively high computational complexity, the optimization of the deep learning algorithms typically consumes more electric energy compared to the traditional techniques (the aspect of energy-efficiency of machine learning algorithms is often overlooked in the scientific literature (García-Martín *et al.*, 2019)). Hence, the traditional algorithms should not yet be considered as obsolete.

In this paper, we present an improved version of the traditional method applied to speech emotion recognition. In traditional speech emotion recognition algorithms, input signals are analyzed using a time window of a constant duration (Omman, Eldho, 2022; Seknedy, Fawzi, 2022; Shahin, 2020). Such an approach is based on an implicit assumption that the features analyzed using a fixed-duration time window capture sonic information equally well at a microscopic level (allophones, phones, syllables) and a macroscopic level (words, sentences). However, the way certain emotions affect the articulation of phonemes may be different from the way they influence the pronunciation of words or sentences (prosodic characteristics). The above assumption motivated these authors to design a method that explicitly takes into account information at multiple time scales. Such a strategy could be referred to as a multi-time-scale (MTS) approach to feature aggregation.

In machine audition, MTS methods are not new. For example, they proved to be effective in the area of respiratory sound classification (Monaco *et al.*, 2020). More recently, Guizzo *et al.* (2020) have redesigned a standard convolutional neural network to take into account multiple time scales, demonstrating the superiority of such an approach compared to the standard convolutional networks when applied to speech emotion recognition. However, to the best of the authors' knowledge, no one has attempted to introduce MTS techniques to the "traditional" classification algorithms in the field of speech emotion recognition.

The main contribution of this work is to demonstrate that the performance of the traditional methods can be improved by aggregating features concurrently using time windows of different lengths (MTS approach). Such an approach could be likened to taking pictures with a camera equipped with a set of different focal lenses, allowing a photographer to simultaneously acquire both micro- and macroscopic views of a photographed scene. The additional novelty of this work is the application of a genetic algorithm to optimize the parameters of the feature extractors. In machine learning, genetic algorithms are typically exploited for the purpose of feature "selection" (Sayed *et al.*, 2019; Jadhav *et al.*, 2018). Application of genetic algorithms for tuning feature extractors is very rare. In this study, a genetic algorithm was used to optimize the feature extractor responsible for the derivation of the Mel-frequency cepstral coefficients (MFCC). Although the research indicates that the parameters employed in the MFCC extraction algorithm should be optimized for a given task (Sahoo, Routray, 2016), undertaking a comprehensive optimization of MFCC extractors still constitutes an uncommon practice. Unlike most of the researchers, in this study, the authors decided to optimize 13 parameters of the MFCC extraction algorithm. Due to a relatively large number of parameters to be optimized, a popular grid-search optimization technique turned out to be impractical. While a genetic algorithm is commonly regarded as computationally demanding, in this study it proved to be more resourceful compared to the aforementioned grid-search technique.

To enhance the generalization property of the method, a data augmentation technique based on pitch shifting and time stretching was applied. In general, applying pitch shifting and time stretching effects to a speech signal may distort the overall prosody of the utterance, weakening its emotional expression. However, according to the research in the area of speech emotion recognition, the original emotional characteristics of speech signals may still be preserved if the above modulation processes are applied conservatively (Mohino-Herranz *et al.*, 2014; Tao *et al.*, 2023). Therefore, care was taken by the authors in employing pitch shifting and time stretching algorithms to maintain the original emotional characteristics of the speech recordings.

The proposed method was evaluated using five publicly available speech corpora, namely: CREMA-D (Cao *et al.*, 2014), eNTERFACE (Martin *et al.*, 2006), RAVDESS (Livingstone, Russo, 2018), SAVEE (Haq, Jackson, 2011), and TESS (Pichora-Fuller, Dupuis, 2020). The method was tested both under speaker-dependent and speaker-independent conditions. Moreover, its generalization property was also evaluated using cross-corpus tests. According to the obtained results, the developed method outperforms or it is comparable to the traditional ones for the selected datasets, demonstrating the benefits of using the MTS approach to feature aggregation.

The paper is organized as follows. In the next section we give an overview of the work of other researchers in the area of speech emotions recognition. In Sec. 3 we explain the methodology applied in our study. The obtained results are described in Sec. 4. The discussion of the obtained results and the conclusions are provided in Secs. 5 and 6, respectively.

2. Related work

Since the pioneering work of Picard (1995), the topic of the automatic speech emotion recognition has been investigated by many scientists, resulting in a considerable body of research. Table 1 overviews in chronological order the example studies in this area published over the past thirteen years. They were arbitrarily selected by these authors. The studies presented in the table are limited to the traditional algorithms as they are pertinent to the work presented in this paper. The methods based on deep learning techniques have been omitted from the table. An interested reader is referred to papers by Khalil *et al.* (2019) and Pandey *et al.* (2019), for comprehensive reviews of deep learning techniques and their applications to speech emotion recognition.

In the traditional methods used for speech emotion recognition, a classical two-stage machine-learning topology is used. It consists of an audio feature extractor followed by a classification algorithm. The features derived in the feature extractor typically include Mel-frequency cepstral coefficients (MFCC), linear predictive coding (LPC) coefficients, signal energy, fundamental frequency (F0), and zero-crossing rate (ZCR), as exemplified in the third column in Table 1. The classical machine learning algorithms are commonly utilized as classifiers, most notably support vector machines (SVM), random forests (RF), multilayer perceptrons (MLP), Gaussian mixture models (GMM), techniques employing linear discriminant analysis (LDA), hidden Markov models (HMM), dynamic time-warping (DT), and K-nearest neighbors (KNN) (cf. the second column in Table 1). The most recent studies in the area of speech emotion recognition have demonstrated that the performance of the traditional methods could be improved by the incorporation of the ensemble of classifiers (Seknedy, Fawzi, 2022; Omman, Eldho, 2022). Moreover, evolutionary algorithms, such as genetic algorithms, could be successfully used to further enhance their performance (Wang, Huo, 2019; Liu *et al.*, 2018). However, as already emphasized in Sec. 1, the genetic algorithms are

Table 1. Overview of the traditional methods used for speech emotions recognition since the year 2005 (in chronological order).

Reference	Model	Model input data	Corpus	Number of emotions	Reported accuracy [%]
Lin, Wei (2005)	HMM	F0, energy, F1-4, MFCC1-2, MBE1-5 with SFS selection	DES	5	99.5
	SVM	MEDC			88.9
Majkowski *et al.* (2016)	KNN	RMS, energy, MFCC1-12, delta features, ZCR, F0, SCG, SF, SRO with SFS selection	Polish radio broadcasts	6	75.6
	LDA				80.5
	SVM				79.2
Ghaleb *et al.* (2019)	SVM	low-level energy descriptors, spectral, vocal delta coefficients	CREMA-D	6	56.2
			eNTERFACE		55.9
Shahin (2020)	HMM (two-stage)	MFCC	in Arabic	6	72.8
	GMM				63.3
	SVM				64.5
	VQ				61.5
Abdel-Hamid (2020)	SVM	pitch, intensity, formants, MFCC, LTAS, wavelet features	EYASE	4	66.8
	KNN				61.7
Seknedy, Fawzi (2021)	MLP	RMS, MFCC1-12, ZCR, voicing probability, F0	RAVDESS	8	64.93
	SVM				70.56
	RF				59.31
	LR				62.64
Seknedy, Fawzi (2022)	MLP	MFCC1-40, Mel-spectrogram1-128, Chroma1-12, Tonnetz, Contrast1-8, RMS	EYASE	4	62.4
	SVM				50.6
	RF				62.4
	LR				62.9
	MLP + SVM + RT + LR (ensemble)				65.1
Omman, Eldho (2022)	SVM (ensemble)	MFCC, ΔMFCC, ΔΔMFCC, spectral subband centroids, logfbank	RAVDESS	8	80.07
Cao *et al.* (2022)	Hessian-based subspace learning + domain adaption	MFCC, ΔMFCC, ΔΔMFCC, LPC, LAFC, Philips fingerprint, spectral entropy	EMO-DB, NNIME, IEMOCAP, MSP-IMPROV, MSP-PODCAST	4	54.93

predominantly used for feature selection (Kanwal, Asghar, 2021; Yildirim et al., 2021; Sidorov et al., 2014), whereas in our study they were employed to optimize the parameters of the feature extractors.

Note that the emotion recognition accuracy reported by an early work of Lin and Wei (2005) (cf. top row of Table 1) exceeds the accuracy levels reported by many other authors, including the most recent work of Cao et al. (2022) (cf. the bottom row of the table). This observation highlights the difficulty in the direct comparison across the studies, caused by the differences in the number of investigated emotions, differences in speech corpora characteristics, or differences in testing procedures (e.g., dissimilar proportions between the train and test sets), just to mention a few factors. Therefore, caution has to be exercised when comparing the methods based on a single accuracy metric or a particular testing procedure.

The speech corpora used for evaluation of the methods can be divided into three groups according to the way the emotions were evoked, namely: acted, elicited, and natural. See the work of Basu et al. (2017) for the differentiation between these three groups. The speech corpora overviewed in Table 1 (fourth column) predominantly represent acted emotions (Lin, Wei, 2005; Ghaleb et al., 2019; Abdel-Hamid, 2020; Seknedy, Fawzi, 2021; 2022; Omman, Eldho, 2022; Cao et al., 2022). In the studies of Ghaleb et al. (2019) and Cao et al. (2022) in addition to the datasets incorporating acted emotions, the corpora employing elicited emotions were used as well. The remaining studies presented in Table 1 used either private corpora with an unknown type of emotions or corpora in which types of emotions are mixed or hard to verify (e.g., broadcasts). As mentioned earlier, the differences in the characteristics between the speech corpora could constitute a confounding factor when comparing the results. Therefore, it is imperative to employ several corpora when evaluating a given method. One of the most challenging evaluation scenarios involves testing new methods using corpora that were not "seen" during the training procedure (cross-corpus tests), including corpora representing different demographic, social, cultural, or language

Table 2. Overview of the speech corpora employed in this study.

Corpus	Reference	Number of speakers	Number of utterances	Duration of utterances [s]			Emotion categories	Emotion types
				Min.	Mean	Max		
CREMA-D	Cao et al. (2014)	92	7441	0.59	2.19	5.00	Happiness (1271)* Sadness (1270) Fear (1271) Anger (1271) Disgust (1271) Neutral (1087)	acted
RAVDESS	Livingstone, Russo (2018)	24	1248	1.00	1.74	4.21	Happiness (192) Sadness (192) Surprise (192) Fear (192) Anger (192) Disgust (192) Neutral (96)	acted
SAVEE	Haq, Jackson (2011)	4	480	0.86	3.22	7.14	Happiness (60) Sadness (60) Surprise (60) Fear (60) Anger (60) Disgust (60) Neutral (120)	acted
TESS	Pichora-Fuller, Dupuis (2020)	2	2800	1.13	1.90	2.86	Happiness (400) Sadness (400) Surprise (400) Fear (400) Anger (400) Disgust (400) Neutral (400)	acted
eNTERFACE	Martin et al. (2006)	10	1287	0.71	2.11	6.30	Happiness (212) Sadness (215) Surprise (215) Fear (215) Anger (215) Disgust (215)	elicited

* Number of recordings representing a given emotion category.

characteristics (Su, Lee, 2021; Seknedy, Fawzi, 2021; Tamulevičius *et al.*, 2020; Milner *et al.*, 2019; Kaya, Karpov, 2018; Cao *et al.*, 2022). In line with the abovementioned observations, in the present study, five following corpora were used, namely: CREMA-D (7441 utterances, 44 female and 48 male speakers), RAVDESS (1248 utterances, 12 female and 12 male speakers), SAVEE (480 utterances, 4 male speakers), TESS (2800 utterances, 2 female speakers), and eNTERFACE (1,287 utterances, 5 female and 5 male speakers). All of these datasets were recorded in English. Only the eNTERFACE dataset contained recordings of elicited emotions, as the other four corpora represented acted emotions obtained from amateur or professional voice actors. Table 2 provides a detailed overview of the five corpora used in this study. In addition to speaker-dependent and speaker-independent tests, a cross-corpus test was also included in the evaluation procedure.

3. Method

The conceptual topology of the proposed algorithm is shown in Fig. 1a. It consists of an ensemble of the feature extractors (FE) coupled with the individual SVM classifiers. The prediction of the emotion category is undertaken using the ensemble voting model. The distinct aspect of the proposed method is that the feature extraction procedure is concurrently undertaken using long-term, mid-term, and short-term time windows, as depicted in the figure. Their duration is adjusted adaptively, depending on the duration of the original excerpts, although it does not exceed 7 s for long-term windows, 2.33 s for mid-term windows, and 0.7 s for short-term windows.

The algorithm depicted in Fig. 1a is computationally inefficient since for the long-term, mid-term, and short-term windows, the same set of the "primary" features has to be calculated. The phrase "primary features" is used in this paper to denote the metrics calculated in the feature extractors such as the zero-crossing rate, whereas the expression "secondary features" represents the statistics derived from the primary features. A computationally optimized topology of the proposed method is illustrated in Fig. 1b. It consists of the single feature extractor (FE), providing a set of primary features, and the ensemble of the feature aggregators (FA) coupled with the individual SVM classifiers. The role of the feature aggregators (FA) is to convert specific parts of primary features into secondary statistical features.

In this study, a computationally optimized version of the algorithm has been implemented (Fig. 1b). Its constituent blocks are described in detail in the subsequent sections. More information on the MTS approach proposed in this study is provided in Subsec. 3.3.

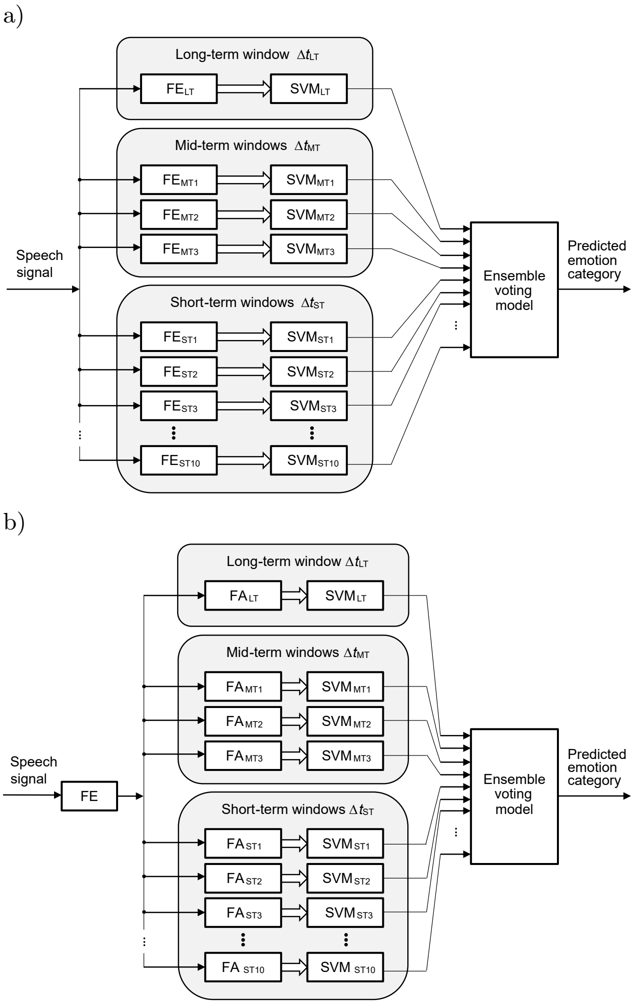

Fig. 1. Multi time-scale speech emotion classification method: a) conceptual algorithm; b) computationally optimized algorithm.

The number of recognizable emotion classes depended on the dataset used in the training stage. In this work, the focus has been put on the Ekman basic emotion set, comprising such emotion categories as anger, joy, disgust, sadness, fear, and surprise (Ekman, 1992), with the addition of the "neutral" emotion class representing utterances that were not emotionally charged, resulting in maximum of 7 classes in total. In some evaluation cases, a subset of this basic emotion set has been taken into consideration due to the dataset limitations (but never being smaller than 5 emotion classes).

3.1. Feature extraction

Prior to undertaking the feature extraction procedure, essential pre-processing tasks were carried out. Namely, the leading and the trailing silence of every recording was trimmed. The silence cutoff point in each recording was the first sample of which the absolute value exceeded 5% of the maximum absolute value of all samples in the recording. To equalize the audio signals' variance, each recording has been further z-standardized. Since the sample rate in the proposed algorithms was set to 44 100 Hz, all the audio recordings with a miss-matched sample rate were resampled accordingly.

For each recording, the signal was divided into time-frames of 1102 samples each, with an overlap of 827 samples. Such unusual values were adopted in this study, since the employed genetic algorithm (described in Subsec. 3.2) proved those values to be the best in the context of this experiment. Since the sample rate of the audio signals equaled 44 100 Hz, the duration of a single time-frame amounted to approximately 25 ms of which around 19 ms overlapped with the neighboring frame. The Hann window was applied to the signals in each frame. Similarly as before, the choice of the window-type was determined by the genetic algorithm. Finally, the features were extracted for each frame.

The following features have been taken into account: MFCC (20 coefficients), ZCR coefficient, fundamental frequency, and spectral flux, yielding 23 features in total. Additionally, the delta and delta-delta values were computed for ZCR, fundamental frequency, and spectral flux, respectively, as they provide information on abrupt changes and transitions of those features. Ultimately, for each audio frame, a primary feature vector of size 29 was computed. All of the features were calculated using the Essentia toolbox (Bogdanov *et al.*, 2013). For all the configuration parameters unspecified in this paper, default values provided by the toolbox were used.

3.2. Genetic algorithm

While most of the features were relatively straightforward to calculate, the estimation of MFCC turned out to be a more demanding task. The Essentia MFCC extractor takes 13 different parameters, including the number of Mel-frequency coefficients, the number of Mel-frequency bands, upper and lower bounds of the frequency range, discrete cosine transform type, type of spectrum, and the liftering coefficient. Hence, manual tuning proved to be a challenge and a need for an appropriate optimization method arose. The complete list of the optimized parameters is provided in Table 3.

While the popular greedy optimization algorithm Grid Search is usually very effective for parameter tuning, for this exact problem its computational complexity turned out to be impractical. Therefore, an alternative optimization method was utilized, namely the Genetic algorithm (Mitchel, 1996). In the context of this study, it is more computationally efficient than Grid Search as it avoids undertaking checks for every possible solution. According to the literature, genetic algorithms are most often utilized in the feature selection process and classifier hyperparameter optimization (Kanwal, Asghar, 2021; Wang, Huo, 2019). By contrast, in this study, a genetic algorithm has been deployed to tune the parameters of the MFCC extractor. The parameter values determined during this search were subsequently used in the Essentia extractor to calculate the MFCC coefficients. The genetic algorithm was implemented by the first author as a multithreaded Python script. For reproducibility of the research, the developed code is included in the publicly available repository at GitHub (Stefanowska, Zieliński, 2023).

The properties of the implemented genetic algorithm are overviewed in Table 4. A specimen in this problem is understood to be a specific parameter value combination from the set of considered values for each parameter. The fitness value for each specimen is calculated by extracting MFCC using its parameter values, training a single SVM classifier with those extracted coefficients, and checking its accuracy on a validation set. All the fitness values were calculated using the RAVDESS dataset (Livingstone, Russo, 2018)

Table 3. Parameters of the MFCC extraction algorithm optimized by the genetic algorithm.

Parameter	Considered values	Genetic algorithm results
Number of Mel coefficients	10, 13, 20, 40, 80, 120	20
Frame size (in samples)	512, 756, 1024, 1102	1102
Window type	Hamming, Hann	Hann
Mel scale implementation method	Auditory Toolbox (Slaney, 1998), HTK toolkit (Young *et al.*, 2006)	Auditory Toolbox
Logarithmic compression type	Natural, power, magnitudes, logarithmic	Magnitudes
Discrete cosine transform type	II, III	III
Normalization method	Unit sum, unit triangle, unit max	Unit triangle
The upper bound of the frequency range [Hz]	6000, 8000, 16 000, 20 000	16 000
The lower bound of the frequency range [Hz]	0, 50, 100, 200, 500	50
The number of Mel-bands in the filter	26, 128	128
Type of weighting function for determining triangle area	Warping, linear	Warping
Type of spectrum	Magnitude, power	Power
The liftering coefficient	0, 22, 10, 40, 100	40

Table 4. Properties of the genetic algorithm.

Property	Value
Maximum population size	10
Potential parent selection method	3-way tournament
Potential parent number	5
Crossover probability	0.7
Mutation probability	0.5

and then cached to save the computational power in case of reoccurring specimens. The basic properties of the genetic algorithm were picked based on how effectively they seemed to perform in the few initial iterations (Table 4). A relatively high mutation probability proved to help with reaching more effective specimens quicker.

For every parameter, a finite set of possible values was specified (Table 3). Certain parameters were numerical and their possible values were selected empirically, others were categorical (e.g., the Mel scale implementation method), and their possible values were already provided by the toolbox. The best set of final parameter values (see the last column in Table 3) was determined after 120 iterations of the genetic algorithm. A properly tuned MFCC extractor proved to significantly increase the accuracy of the trained model, as illustrated in Fig. 2.

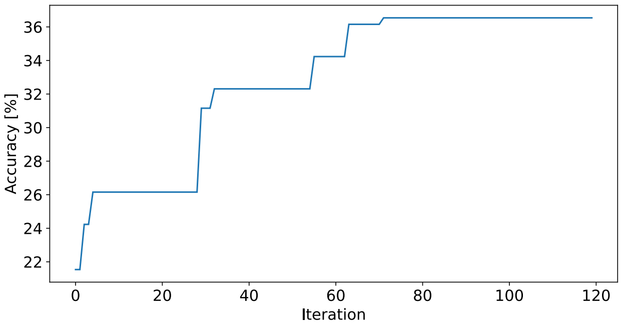

Fig. 2. Example learning curve of the genetic algorithm.

3.3. Multi-time-scale approach to feature aggregation

The primary features calculated by the FE, as described in the previous sections, are then processed using an ensemble of the FA. Each feature aggregator takes a specific "slice" of the primary features, according to the size of the corresponding time window. The following statistics are calculated in the process of the feature aggregation: mean values, standard deviations, minimum and maximum values, as well as lower and upper quartiles, yielding the set of 203 secondary features at the output of each feature aggregator. In accordance with the typical practice in machine audition, the secondary features are further z-standardized (Kreyszig, 1979).

According to the proposed topology, the algorithm consists of the three blocks signified by the shaded areas in Fig. 1b, each utilizing a different time window length. The top-most block comprises a single feature aggregator $\mathrm{FA}_{\mathrm{LT}}$ connected to its associated classifier ($\mathrm{SVM}_{\mathrm{LT}}$). A long-term window Δt_{LT} of the feature aggregation is used in this block. The duration of the time-window is in this case set to the duration of the whole utterance, constrained to 7 s maximum. In other words, the primary features are aggregated for the initial 7 s of each utterance. If the recording exceeds that limit, it is trimmed to the maximum permissible length of 7 s. It is presumed that the top-most block is responsible for capturing and processing the prosodic features from the whole speech utterance. Note that the way the signal is processed using the top-block (in isolation from the remaining two blocks) could be considered as the standard approach, commonly applied by the researchers in the field of speech emotion recognition (Omman, Eldho, 2022; Seknedy, Fawzi, 2022; Abdel-Hamid, 2020; Ghaleb *et al.*, 2019). Therefore, in this study this part of the algorithm is considered as the "baseline" method.

In the middle block depicted in Fig. 1b, the long-term window is divided into the three overlapping mid-term windows of maximum duration equal to $\Delta t_{\mathrm{MT}} = 2.33$ s each, with an overlap of approximately 0.1 s. These windows are responsible for dividing all the primary features into the ones representing the initial, middle, and ending part of each utterance, respectively. The primary features from these three mid-term time windows are then processed individually by the three feature aggregators. The statistics calculated by the feature aggregators are the same as the ones described above in the case of the long-term window. In the next step, the secondary features derived by the feature aggregators are fed to the three SVM classifiers. Due to the shorter length of the window of analysis, it could be supposed that the middle block would better utilize information conveyed by individual words.

The finest temporal resolution is exhibited by the bottom-block shown in Fig. 1b. In this case the long-term window of analysis is divided into ten overlapping short-term windows. Consequently, the window of analysis is further reduced down to $\Delta t_{\mathrm{ST}} = 0.7$ s at most, with an overlap of 0.05 s. The primary features encompassed by each of the ten short-term windows are processed independently by the ten feature aggregators, and then the ten classifiers. Out of the three blocks included in the algorithm, the bottom one is the most complex, as it consists of the ten feature aggregators combined with the ten associated SVM classifiers. It could be hypothesized that the bottom-block would be particularly efficient in capturing and processing information represented by short words or syllables. In total, each speech utterance is concurrently analyzed and classified using 14 time windows (one long-term, three mid-term, and ten short-term windows).

3.4. Classification algorithm

The support vector machine (SVM) classifier was selected as the base model for the proposed method. SVM is one of the most commonly used traditional machine learning techniques, which despite being potentially less effective than modern deep learning models, still prove advantageous in certain cases – especially when available datasets are sparse or too small to train effective deep models (Omman, Eldho, 2022; Seknedy, Fawzi, 2022; Abdel-Hamid, 2020; Ghaleb *et al.*, 2019; Shahin, 2020). In the proposed method, the SVM's hyperparameters are optimized using the Grid Search algorithm. The parameters chosen for tuning include the SVM's C coefficient with possible values of 0.1, 1, and 10; gamma coefficient with possible values of $\frac{1}{\text{feature number}} \times 0.1$, $\frac{1}{\text{feature number}}$, $\frac{1}{\text{feature number}} \times 10$; and the Kernel function that might be chosen to be linear, polynomial of the 3rd degree, or radial basis function (RBF).

Another parameter that gets optimized by the Grid Search algorithm, yet does not belong to SVM's hyperparameters, is the number of selected features that are used as the final input vector. The list of possibilities include: X = 100, 90, 50 or 25% of all the original features. This optimized value is used at the feature selection stage, which consists of filtering out all the constants and then utilizing the selection method based on the ANOVAF statistic from the scikit-learn library (Pedregosa *et al.*, 2011). A rank of features is created, of which only top X features with the best score get selected (as mentioned previously, the X value is determined by the optimization algorithm). The sequential process of parameter tuning, feature selection and classifier training were managed with the use of the pipeline tool from the scikit-learn toolbox (Pedregosa *et al.*, 2011).

3.5. Ensemble voting model

To make use of all the micro and macro information contained in each of the statistical feature vectors obtained as described in the previous sections, they were used as inputs for separate SVM classifiers which were then combined into an ensemble voting classifier (cf. Fig. 1). The final assembling stage involved building the voting classifier. The soft voting method was utilized. The winning class is the one with the greatest total sum of probability of occurring in each component classifier. Additionally, every probability was weighted based on how well the classifier performed on the validation dataset during the tuning phase. In summary, the score for each class was calculated using the equation:

$$s_c = \sum_{i=0}^{N} w_i * p_{c,i}, \tag{1}$$

where s_c – score of the emotion class c; w_i – voting weight of the i-th classifier (its accuracy on the validation dataset during the tuning stage); $p_{c,i}$ – probability of the emotion class c in the i-th classifier; N – number of classifiers.

The emotion class with the maximum score is considered to be the final 'decision' of the ensemble voting classifier.

3.6. Data augmentation

In order to enhance the generalization property of the classification model, all the speech recordings went through the data augmentation process. Simple pitch shifting and time stretching operations available in the librosa toolbox (McFee *et al.*, 2015) were applied to enrich the existing datasets. Introducing pitch shifting and time stretching effects to speech signals influences the overall prosody of the utterance. Consequently, such processes may modify emotional expressions. However, the authors assumed that the original emotional characteristics of the speech recordings would be preserved if these effects were applied cautiously, that is using conservative pitch shifting and time stretching limits. This assumption is in accordance with the research in the area of speech emotion recognition (Mohino-Herranz *et al.*, 2014; Tao *et al.*, 2023). In line with the above considerations, the pitch has been shifted up and then down by three semitones whereas the audio signals have been sped up and slowed down by 25%, respectively, resulting in four new audio files for each existing audio file. All the augmented recordings were further used only in the training sets (the test sets comprised solely the original recordings).

The developed method was implemented in Python. The code was made publicly available at GitHub repository (Stefanowska, Zieliński, 2023).

4. Results

The performance of the developed method was evaluated in five experiments. The comparisons were made both against the traditional algorithms as well as the deep learning techniques, published recently in the literature. Three different experimental methodologies have been considered, including speaker-dependent tests, speaker-independent tests, and cross-corpus tests.

4.1. Speaker-dependent tests

In this approach, recordings coming from the same speakers can appear in validation, training, and testing sets. The speaker-dependent tests were conducted for a single speaker using the TESS dataset. In this case, only samples belonging to the younger actress were utilized. In total, 1400 audio recordings were employed, representing seven emotion categories (200 recordings per emotion). Moreover, additional 5 600 augmented excerpts were utilized in this experiment. The tests followed the methodology from the work of Chatterjee *et al.* (2021). The dataset was split with a 65/15/20 percentage ratio in order to obtain the training, validation, and test subsets, respectively. For the sake of comparison, aside from testing only the main proposed method utilizing an ensemble of SVM classifiers with the MTS approach and data augmentation (MTS + Aug), a variant without data augmentation (MTS), as well as variants based on a single SVM classifier with augmentation (SVM + Aug) and without it (SVM), were tested too, which resulted in four test cases. All the experiments were repeated 30 times with different randomization seeds and the final result was the mean value of all the individual accuracy values obtained in the repeated trials. Those results were presented on the mean accuracy chart along with the corresponding standard deviations (Fig. 3). It can be seen that the proposed ensemble classifier with the MTS approach, labeled as MTS in the figure, outperformed the method proposed by Chatterjee *et al.* (2021). This outcome was statistically significant, based on the one sample t-test (p-value was less than 10^{-4} for the dependent t-test with a 0.05 alpha level). It also performed better than the single SVM with the use of augmented data ($p < 10^{-4}$). Moreover, it outperformed the standard SVM algorithm without data augmentation (p-value was approximately equal to 10^{-4}). Hence, the addition of the augmented data for the training stages did not improve the accuracy for this case. Figure 4 shows the accuracy of recognition of the individual emotions for the TESS dataset. It can be seen that all the emotions were identified with almost 100% accuracy using the proposed MTS approach, except for the “surprise” category, which was recognized with 98.67% (SD 1.7%) accuracy.

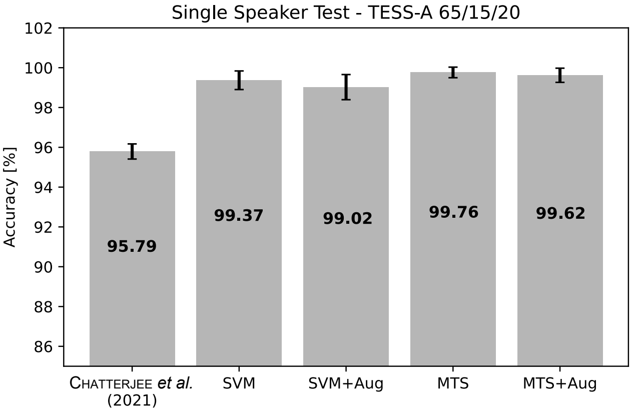

Fig. 3. Accuracy chart for the TESS 65/15/20 split ratio experiment with the speaker-dependent testing approach – only younger actress' samples were used. The results represent the mean accuracy values and associated standard deviations.

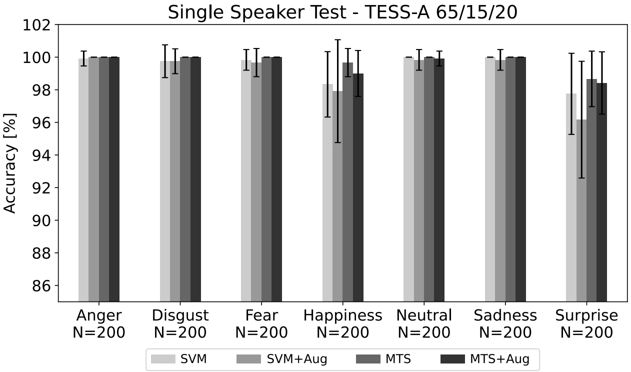

Fig. 4. Accuracy of recognition of individual emotions for the TESS 65/15/20 split ratio experiment with the speaker-dependent testing approach – only younger actress' samples were used. The results represent the mean accuracy values and associated standard deviations.

A separate experiment involving a speaker-dependent test of the proposed MTS method was performed. It was based on the 10-fold cross-validation procedure, conducted using the RAVDESS dataset. It contained 192 recordings per emotion, apart from the neutral state, which was represented by 96 excerpts, giving 1248 audio files in total (plus the addition of 4992 augmented samples). The results showed that the class that was relatively the hardest one to classify was the neutral emotional state (Fig. 5a). It was often mistaken with sadness. Another class often mistaken with sadness was fear. Classes that seemed to be the most recognizable by the proposed method were anger and disgust – they also tended to be mistaken with each other more than with any other emotion.

4.2. Speaker-independent tests

In this experiment, the first test with the speaker-independent constraint was a 10-fold cross-validation and it was conducted using the eNTERFACE dataset with 6 emotion classes. For this repository, each emotion was represented by 215 recordings except happiness which was exemplified by 212 audio excerpts. The reported results are the average values of the accuracies from all the folds. Corresponding standard deviations were also calculated. The literature reference was a method based on a SVM classifier, utilizing multimodal inputs (Ghaleb *et al.*, 2019). For comparison purposes, solely audio-only average accuracy was taken into account. The obtained results are presented in Fig. 6. In this case, a single SVM classifier trained with the aid of the augmented data (SVM + Aug) performed better than a single SVM with no augmentation ($p = 0.0045$). Similarly, the proposed MTS model

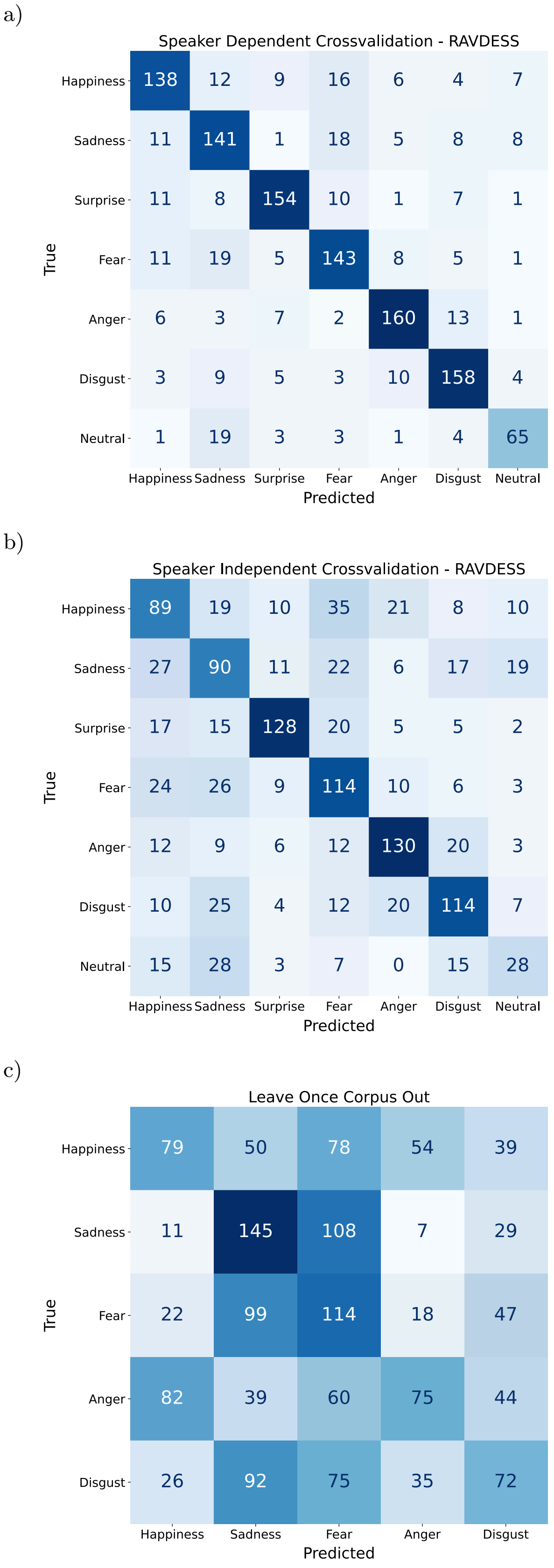

Fig. 5. Classification accuracy tests results for the developed model presented as confusion matrices. Three different testing approaches: a) speaker-dependent; b) speaker-independent; c) corpus-independent.

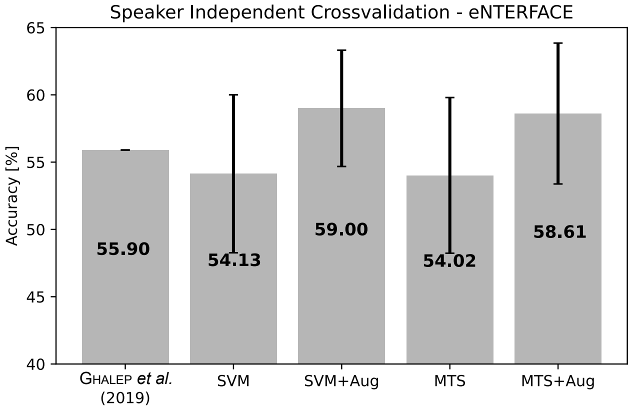

Fig. 6. Accuracy chart for the eNTERFACE cross-validation experiment with the speaker-independent testing approach. The results represent the mean accuracy values and associated standard deviations.

using the augmented data (MTS + Aug) performed better than the same model not utilizing it (MTS) ($p = 0.0036$). The advantage of using MTS model with the augmented data over the literature example could not be verified as the p-value for the one sample t-test was equal to 0.0773, unlike the advantage of exploiting a single SVM classifier with the use of the augmented data (SVM + Aug) ($p = 0.03$). The reasons for such a high accuracy of a single SVM classifier trained using the augmented data could be attributed to properly tuned MFCC extractor and model's hyperparameters. For this amount of data, a properly tuned single classifier proved to be sufficient. As the advantage of the proposed MTS method over the literature example could not be verified ($p = 0.9227$), the results were statistically comparable. According to both the results obtained for the single SVM classification algorithm and the MTS ensemble method, the chosen augmentation procedure substantially improved the accuracy using the selected dataset. Figure 7 shows the accuracy of

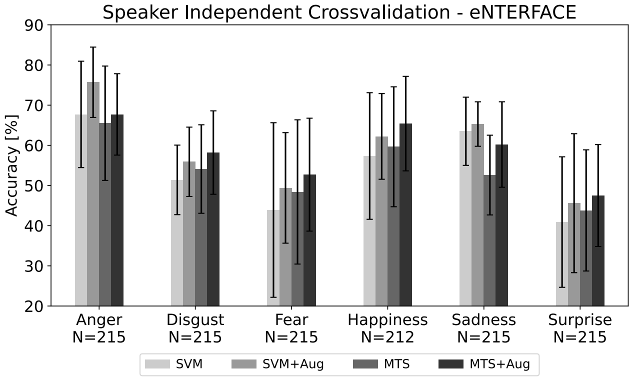

Fig. 7. Accuracy of recognition of individual emotions for the eNTERFACE cross-validation experiment with the speaker-independent testing approach. The results represent the mean accuracy values and associated standard deviations.

recognition of the individual emotions for the eNTERFACE dataset. It can be seen that the anger category is recognized with the highest accuracy, reaching almost 76%, whereas the fear and surprise categories are identified with the lowest accuracy at a level ranging from 41% to 53%.

The second test set was based on the same literature reference. In this case, it was conducted on the CREMA-D dataset with 6 emotion classes. Due to time constraints, in this experiment 996 samples were randomly chosen from the complete dataset while maintaining the original distribution of speakers and emotions, which resulted in the repository of 166 utterances per emotion class. The results are presented in Fig. 8. It can be seen that a single SVM classifier performed better with the aid of augmented data ($p = 0.0072$) and that the MTS ensemble model with the augmented data performed better than the same model without it ($p = 0.0002$). The advantage of the literature example over the proposed method turned out not to be statistically significant ($p = 0.2385$), therefore two methods seem to be comparable. Figure 9 shows the accuracy of recognition of the individual emotions for the CREMA-D dataset. Similar to the previous experiment employing the eNTERFACE dataset, the anger category is recognized with the highest accuracy, reaching 76%. The disgust and fear categories are identified with the lowest accuracy at a level ranging from 38 to 45%.

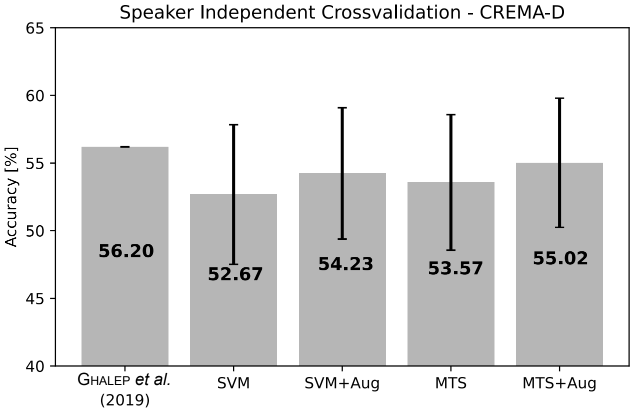

Fig. 8. Accuracy chart for the CREMA-D cross-validation experiment with the speaker-independent testing approach. The results represent the mean accuracy values and associated standard deviations.

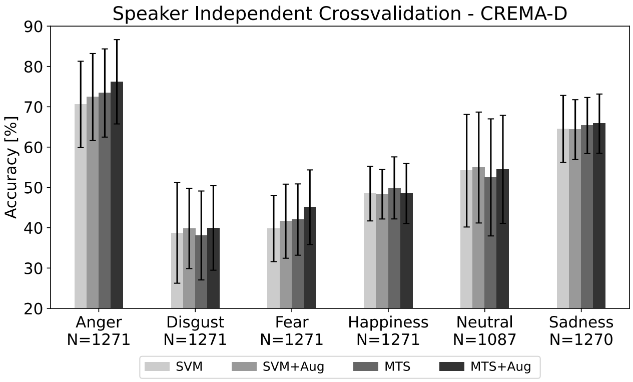

Fig. 9. Accuracy of recognition of individual emotions for the CREMA-D cross-validation experiment with the speaker-independent testing approach. The results represent the mean accuracy values and associated standard deviations.

The third test set was based on the work of Guizzo *et al.* (2020), who developed an advanced model employing convolutional neural networks. Their model was trained on the RAVDESS dataset, comprising 192 recordings for each emotion (96 for neutral state), as mentioned before. The results reported by the quoted authors constitute the average accuracy values obtained from a 4-fold cross-validation test. The cited work also utilized the MTS approach by introducing multiple convolution kernels and obtaining differently scaled feature maps that were all used as the model input. The dataset was split with an approximate ratio of 70/20/10 into training, validation, and testing subsets, respectively. In this study, the test was repeated 30 times, and the final result was estimated as the mean value of all the individual accuracy values. According to the obtained results (see Fig. 10), the MTS ensemble model performed better than the single SVM classifier for the case without the use of the augmented data ($p = 8 \times 10^{-4}$). The advantage of using the augmentation process in this case could not be statistically verified. Importantly, the difference between the result reached by Guizzo *et al.* (2020) and that obtained using the proposed MTS method (MTS + Aug) was not statistically significant ($p = 0.2875$). Consequently, it could be concluded that the accuracy of the proposed MTS method proved to be comparable to the one reached using a state-of-the-art deep learning technique applied to the RAVDESS data set. Figure 11 shows the accuracy of recognition of the individual emotions for the RAVDESS dataset. Similar to the results obtained in the previous two experiments,

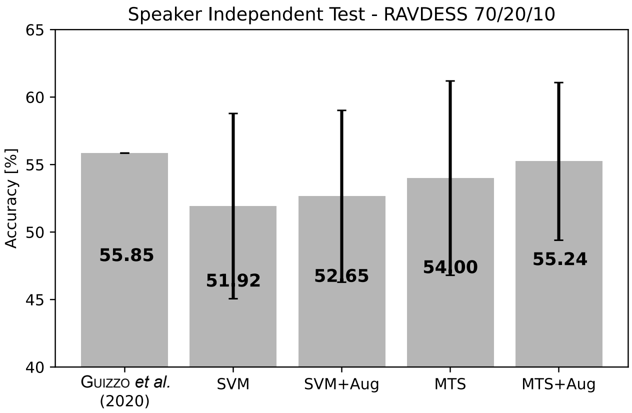

Fig. 10. Accuracy chart for the RAVDESS 70/20/10 split ratio experiment with the speaker-independent testing approach. The results represent the mean accuracy values and associated standard deviations.

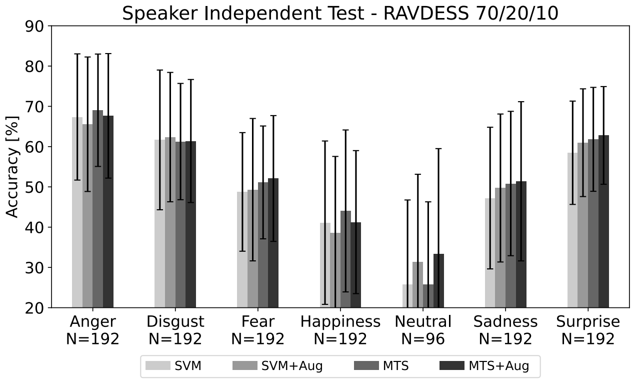

Fig. 11. Accuracy of recognition of individual emotions for the RAVDESS 70/20/10 split ratio experiment with the speaker-independent testing approach. The results represent the mean accuracy values and associated standard deviations.

the anger category was recognized with the highest accuracy, reaching approximately 69%. The neutral category was identified with the lowest accuracy at a level ranging from 26 to 33%.

For the 10-fold speaker-independent cross-validation test using the RAVDESS dataset, similarly to the previously discussed speaker-dependent evaluation (see Subsec. 4.1), a neutral emotional state turned out to be the class with the lowest accuracy (Fig. 5b). Likewise, it was often mistaken with sadness, although this time even more frequently. As the overall accuracy is substantially lower than that obtained for the speaker-dependent case, the system mixes up emotional states much more often. However, the anger emotion still exhibits a relatively high recognition rate.

4.3. Cross-corpus tests

For the cross-corpus test of the proposed MTS method, a leave-one-corpus-out cross-validation experiment was conducted. For this purpose, the following English datasets were used: CREMA-D, RAVDESS, SAVEE, TESS, and eNTERFACE. The selected corpora represented different but overlapping sets of emotions and contained vastly different number of recordings (see Table 2). Due to the above factors, the datasets were unified. Namely, the files with emotions that were not present in every corpus were discarded (neutral and surprise), which resulted in five common emotional classes – happiness, sadness, fear, anger, and disgust. The remaining recordings were sampled (while maintaining the original distribution of emotions and sexes) so that all the datasets were of the same size – 300 recordings each, which was the size of the smallest dataset (SAVEE). Considering that each repository was balanced in terms of the number of recordings representing each emotion, the number of recordings taken from every dataset representing a given emotion equaled 60. With the addition of the augmented files (1200 for each corpus), the total number of files was equal to 7500. In each iteration one corpus became a testing set, another one was selected as a validation dataset, whereas the remaining three datasets were used for training.

Even with the reduced number of emotional classes, the corpus-independent test turned out to give the lowest overall accuracy so far, being equal to 32.33%. Nevertheless, this value was still statistically greater than the chance level, which in the experiment amounted to 20% (p = 0.0035). Despite this outcome, sadness turned out to be recognized comparatively often, with the classification accuracy reaching as much as 48.33%. As shown in Fig. 12, its recognition rate turned out to be comparable to the recognition rate of sadness in the speaker-independent test on a single corpus. It was, however, often mistaken with fear and disgust (Fig. 5c). Unlike in previous tests, for the cross-corpus test, anger and disgust were the hardest emotions to classify.

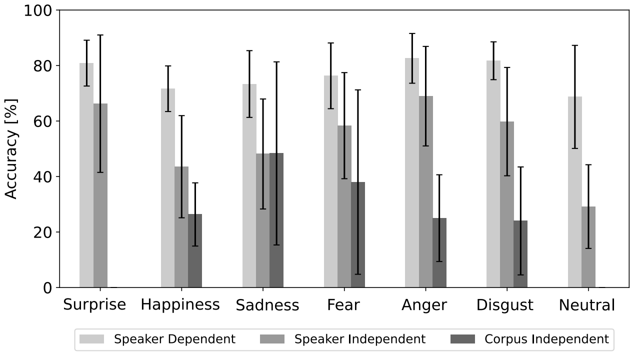

Fig. 12. Classification accuracy tests results obtained using the proposed MTS model under the three testing conditions: speaker-dependent, speaker-independent, and corpus-independent. Note that the corpus-independent approach applies only to five out of seven emotions considered in the remaining tests (see the text for details).

5. Discussion

Out of the four experiments conducted in this study, the two experiments proved that the multi-time scale approach to feature aggregation yields better classification results compared to the conventional techniques. These two experiments were based on the TESS and RAVDESS speech corpora, respectively. Moreover, in one of the experiments, involving the RAVDESS corpus, the MTS method achieved a very similar classification accuracy level compared to the one obtained using the state-of-the-art deep learning technique (Guizzo *et al.*, 2020), as illustrated in Fig. 10. Furthermore, the acquired results highlight the benefits of using a data augmentation technique (Figs. 6 and 8).

There are two recent studies resembling our work as they also employed the ensembles of classifiers.

In the work by Seknedy and Fawzi (2022), an ensemble of four different classification algorithms was used, reaching a maximum classification accuracy of 65.1%. In their work, they utilized an Arabic speech corpus. Therefore, due to the linguistic mismatch, our method cannot be directly (and fairly) compared to the aforementioned one. In their recent work, Omman and Eldho (2022) have employed an ensemble of 20 SVM classifiers. They used a bootstrap aggregating technique to train their ensemble model, reaching an accuracy of 80.07% when tested on the RAVDESS corpus, outperforming our method by 3 percentage points. While this outcome may indicate an inferiority of our method, the cited authors did not provide sufficient details regarding their testing methodology (e.g., whether the tests were speaker-independent), preventing other researchers from a rigorous comparison of the results.

According to the results of the fifth experiment, involving the cross-corpus test (Subsec. 4.3), the classification accuracy of the proposed method was substantially lower than the accuracy levels obtained within the selected corpora using speaker-dependent and speaker-independent tests, respectively. This outcome does not invalidate the proposed method but indicates that its generalization property needs to be improved. Note, that the cross-corpus tests are still very rare in the literature as they constitute the most rigorous way of testing the speech emotions recognition systems (Tamulevičius *et al.*, 2020). The recent study by Cao *et al.* (2022) confirms that the average accuracy for this testing approach remains relatively low, especially for the simpler methods.

The presented results indicate that the proposed MTS method has an advantage over the baseline technique employing a single classifier with a fixed time-window applied during the feature aggregation. It is, however, more computationally complex, as it utilizes multiple classifiers instead of one. For example, it took 2.73 ms for a single SVM model to classify one recording. Compared to that, using the MTS model for classification took about 5.86 times longer (16 ms). There was also a significant difference between the total duration time of tuning and training. Namely, tuning and training a single SVM classifier on 4160 training files and 260 validation files took in total 8.37 s. In contrast, tuning and training an MTS model on the same dataset took 8 min 33 s and, consequently, it requires more resources. The training of a genetic algorithm itself to tune the feature extractor took 12 h 11 min 47 s, which is the reason why it was not used as a part of a training pipeline but constituted a separate procedure conducted once. All the calculations were carried out using parallel processing on 8 threads of the Intel Core i5 1.6G Hz processor.

There are some limitations of this study that need to be acknowledged. Firstly, the undertaken experiments were based on only five datasets. Broader conclusions could be reached if more corpora were taken into consideration. Secondly, due to the data storage and computation constraints of the hardware used, a subset of the CREMA-D corpus was employed, as described in Subsec. 4.2. Thirdly, the feature extractor tuning procedure was performed using a single speech corpus (RAVDESS), potentially biasing the model towards the selected data set. Fourthly, the duration of the long-term window applied for the feature aggregation was limited to 7 s. In retrospect, the above constraint could be too short for some applications, potentially causing the method to discard important information conveyed by the prosodic characteristics at the ending parts of the sentences. In the present study, this issue affected only one recording belonging to the SAVEE repository (the audio excerpt was trimmed as its duration exceeded the limit). Optimization of the long-term window applied for the feature aggregation as well as the exploration of different optimization strategies for the feature extraction may constitute the subject of future work.

6. Conclusions

This study presents an improved method of speech emotions recognition using an ensemble of SVM classification algorithms. The novelty of the proposed method consists in using a MTS approach to the feature aggregation procedure. Another distinct aspect of the proposed technique is the application of the genetic algorithm to optimize the feature extraction process. Out of the four experiments conducted in this study, the two experiments support the hypothesis that the MTS approach to feature aggregation yields better classification results compared to the conventional way of feature aggregation based on a fixed-duration time window. This implies that the proposed MTS approach is not always superior compared to the conventional technique. Nevertheless, it exhibits satisfactory performance for the selected datasets, matching or outperforming the recently published methods. Interestingly, in one of the experiments conducted within this study, the results obtained using the proposed MTS method proved to be comparable to the ones achieved by means of the state-of-the-art deep learning technique. This outcome indicates that a properly developed traditional classification method could be competitive to a deep learning algorithm. As a side observation, the obtained results exemplified the benefits of data augmentation. The technique of data augmentation is commonly used for the training of deep learning models (Milner *et al.*, 2019). However, this study demonstrated the advantages of applying such a technique during the development of the traditional model. Future work may be focused on testing the MTS method using a broader scope of speech corpora, with the aim of gaining knowledge as to how to further optimize the technique within individual data sets while still maintaining a satisfactory cross-corpus generalization property.

Acknowledgments

The work was supported by the grants from the Białystok University of Technology (WZ/WI-IIT/5/2023 and WI/WI-IIT/6/2023) and funded with resources for research by the Ministry of Science and Higher Education in Poland.

References

1. Abdel-Hamid L. (2020), Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features, *Speech Communication*, **122**: 19–30, doi: 10.1016/j.specom.2020.04.005.

2. Basu S., Chakraborty J., Bag A., Aftabuddin M. (2017), A review on emotion recognition using speech, [in:] *International Conference on Inventive Communication and Computational Technologies* (*ICICCT*), doi: 10.1109/ICICCT.2017.7975169.

3. Bogdanov D. *et al.* (2013), ESSENTIA: An audio analysis library for music information retrieval, [in:] *International Society for Music Information Retrieval Conference* (*ISMIR'13*), pp. 493–498.

4. Cao H., Cooper D.G., Keutmann M.K., Gur R.C., Nenkova A., Verma R. (2014), CREMA-D: Crowdsourced emotional multimodal actors dataset, *IEEE Transactions on Affective Computing*, **5**(4): 377–390, doi: 10.1109/TAFFC.2014.2336244.

5. Cao X., Jia M., Ru J., Pai T. (2022), Cross-corpus speech emotion recognition using subspace learning and domain adaption, *EURASIP Journal on Audio, Speech, and Music Processing*, **2022**: 32, doi: 10.1186/s13636-022-00264-5.

6. Chatterjee R., Mazumdar S., Sheratt R.S., Halder R., Maitra T., Giri D. (2021), Real-time speech emotion analysis for smart home assistants, *IEEE Transactions on Consumer Electronics*, **67**(1): 68–76, doi: 10.1109/TCE.2021.3056421.

7. Choi W.Y., Song K.Y., Lee C.W. (2018), Convolutional attention networks for multimodal emotion recognition from speech and text data, [in:] *Proceedings of Grand Challenge and Workshop on Human Multimodal Language* (*Challenge-HML*), pp. 28–34, doi: 10.18653/v1/W18-3304.

8. Ekman P. (1992), An argument for basic emotions, [in:] *Cognition and Emotion*, **6**(3–4): 169–200.

9. Eskimez S.E., Duan Z., Heinzelman W. (2018), Unsupervised learning approach to feature analysis for automatic speech emotion recognition, [in:] *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 5099–5103, doi: 10.1109/ICASSP.2018.8462685.

10. García-Martín E., Rodrigues C.F., Riley G., Grahn H. (2019), Estimation of energy consumption in machine learning, *Journal of Parallel and Distributed Computing*, **134**: 75–88, doi: 10.1016/j.jpdc.2019.07.007.

11. Ghaleb E., Popa M., Asteriadis S. (2019), Metric learning-based multimodal audio-visual emotion recognition, *IEEE MultiMedia*, **27**(1): 37–48, doi: 10.1109/MMUL.2019.2960219.

12. Guizzo E., Weyde T., Leveson J.B. (2020), Multi-time-scale convolution for emotion recognition from speech audio signals, [in:] *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), doi: 10.1109/ICASSP40776.2020.9053727.

13. Haq S., Jackson P.J.B. (2011), Multimodal emotion recognition, [in:] *Machine Audition: Principles, Algorithms and Systems*, Wang W. [Ed.], pp. 398–423, IGI Global Press, Hershey, doi: 10.4018/978-1-61520-919-4.

14. Jadhav S., He H., Jenkins K. (2018), Information gain directed genetic algorithm wrapper feature selection for credit rating, *Applied Soft Computing*, **69**: 541–553, doi: 10.1016/j.asoc.2018.04.033.

15. Jain D.K., Shamsolmoali P., Sehdev P. (2019), Extended deep neural network for facial emotion recognition, *Pattern Recognition Letters*, **120**: 69–74, doi: 10.1016/j.patrec.2019.01.008.

16. Kanwal S., Asghar S. (2021), Speech emotion recognition using clustering based GA-optimized feature set, *IEEE Access*, **9**: 125830–125842, doi: 10.1109/ACCESS.2021.3111659.

17. Kaya H., Karpov A.A. (2018), Efficient and effective strategies for cross-corpus acoustic emotion, *Neurocomputing*, **275**: 1028–1034, doi: 10.1016/j.neucom.2017.09.049.

18. Khalil R.A., Jones E., Babar M.I., Jan T., Zafar M.H., Alhussain T. (2019), Speech emotion recognition using deep learning techniques: A review, *IEEE Access*, **7**: 117327–117345, doi: 10.1109/ACCESS.2019.2936124.

19. Kreyszig E. (1979), *Advanced Engineering Mathematics*, 4th ed., Wiley.

20. Lin Y.-L., Wei G. (2005), Speech emotion recognition based on HMM and SVM, [in:] *Fourth International Conference on Machine Learning and Cybernetics*, doi: 10.1109/ICMLC.2005.1527805.

21. Liu Z.-T., Xie Q., Wu M., Cao W.-H., Mei Y., Mao J.-W. (2018), Speech emotion recognition based on an improved brain emotion learning model, *Neurocomputing*, **309**: 145–156, doi: 10.1016/j.neucom.2018.05.005.

22. Livingstone S.R., Russo F.A. (2018), 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*, **13**(5): e0196391, doi: 10.1371/journal.pone.0196391.

23. Majkowski A., Kołodziej M., Rak R.J., Korczyński R. (2016), Classification of emotions from speech signal, [in:] *Signal Processing Algorithms, Architectures, Arrangements and Applications* (*SPA*), doi: 10.1109/SPA.2016.7763627.

24. Martin O., Kotsia I., Macq B., Pitas I. (2006), The eNTERFACE'05 audio-visual emotion database, [in:] *22nd International Conference on Data Engineering Workshops* (*ICDEW'06*), doi: 10.1109/ICDEW.2006.145.

25. McFee B. *et al.* (2015), librosa: Audio and music signal analysis in Python, [in:] *14th Python in Science Conference*, pp. 18–25, doi: 10.25080/Majora-7b98e3ed-003.

26. Milner R., Jalal M.A., Ng R.W.N.M., Hain T. (2019), A cross-corpus study on speech emotion recognition, [in:] *IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*), doi: 10.1109/ASRU46091.2019.9003838.

27. Mitchel M. (1996), *An Introduction to Genetic Algorithms*, MIT Press, Cambridge.

28. Mohino-Herranz I., Gil-Pita R., Alonso-Diaz S., Rosa-Zurera M. (2014), MFCC based enlargement of the training set for emotion recognition in speech, *Signal & Image Processing: An International Journal* (*SIPIJ*), **5**(1), doi: 10.48550/arXiv.1403.4777.

29. Monaco A., Amoroso N., Bellantuono L., Pantaleo E., Tangaro S., Bellotti R. (2020), Multi-time-scale features for accurate respiratory sound classification, *Applied Sciences*, **10**(23): 8606, doi: 10.3390/app10238606.

30. Omman B., Eldho S.M. (2022), Speech emotion recognition using bagged support vector machines, [in:] *International Conference on Computing, Communication, Security and Intelligent Systems* (*IC3SIS*), doi: 10.1109/IC3SIS54991.2022.9885578.

31. Pandey S.K., Shekhawat H.S., Prasanna M.S. (2019), Deep learning techniques for speech emotion recognition: A review, [in:] *29th International Conference Radioelektronika* (*RADIOELEKTRONIKA*), doi: 10.1109/RADIOELEK.2019.8733432.

32. Pedregosa F. *et al.* (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12**(85): 2825–2830.

33. Picard R.W. (1995), *Affective computing*, M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321.

34. Pichora-Fuller M.K., Dupuis K. (2020), Toronto emotional speech set (TESS) (V1), *University of Toronto Dataverse*, doi: 10.5683/SP2/E8H2MF.

35. Sahoo S., Routray A. (2016), MFCC feature with optimized frequency range: An essential step for emotion recognition, [in:] *2016 International Conference on Systems in Medicine and Biology* (*ICSMB*), doi: 10.1109/ICSMB.2016.7915112.

36. Sayed S., Nassef M., Badr A., Farag I. (2019), A Nested Genetic Algorithm for feature selection in high-dimensional cancer microarray datasets, *Expert Systems with Applications*, **121**: 233–243, doi: 10.1016/j.eswa.2018.12.022.

37. Seknedy M.E., Fawzi S. (2021), Speech emotion recognition system for human interaction applications, [in:] *Tenth International Conference on Intelligent Computing and Information Systems* (*ICICIS*), doi: 10.1109/ICICIS52592.2021.9694246.

38. Seknedy M.E., Fawzi S. (2022), Speech emotion recognition system for Arabic speakers, [in:] *4th Novel Intelligent and Leading Emerging Sciences Conference* (*NILES*), doi: 10.1109/NILES56402.2022.9942431.

39. Shahin I. (2020), Emotion recognition using speaker cues, [in:] *Advances in Science and Engineering Technology International Conferences* (*ASET*), pp. 1–5, doi: 10.48550/arXiv.2002.03566.

40. Sidorov M., Brester C., Minker W., Semenkin E. (2014), Speech-based emotion recognition: Feature selection by self-adaptive multi-criteria genetic algorithm, [in:] *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (*LREC'14*), pp. 3481–3485, http://www.lrec-conf.org/proceedings/lrec2014/pdf/341_Paper.pdf (access: 10.11.2023).

41. Slaney M. (1998), *Auditory Toolbox: A MATLAB Toolbox for Audtiory Modeling Work*, version 2, Interval Research Corporation.

42. Stefanowska A., Zieliński S.K. (2023), Software repository for speech emotion recognition using a multi-time-scale approach to feature aggregation and an ensemble of SVM classifiers, GitHub, https://github.com/antoninastefanowska/MTS-SVM-EmotionRecognition (access: 27.10.2023).

43. Su B.-H., Lee C.-C. (2021), A conditional cycle emotion gan for cross corpus speech emotion recognition, [in:] *IEEE Spoken Language Technology Workshop* (*SLT*), doi: 10.1109/SLT48900.2021.9383512.

44. Tamulevičius G., Korvel G., Yayak A.B., Treigys P., Bernatavičienė J., Kostek B. (2020), A study of cross-linguistic speech emotion recognition based on 2D feature spaces, *Electronics*, **9**(10): 1725, doi: 10.3390/electronics9101725.

45. Tang D., Kuppens P., Geurts L., van Waterschoot T. (2021), End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network, *EURASIP Journal on Audio, Speech and Music Processing*, **18**(2021), doi: 10.1186/s13636-021-00208-5.

46. Tao H., Shan S., Hu Z., Zhu C., Ge H. (2023), Strong generalized speech emotion recognition based on ef-

fective data augmentation, *Entropy*, **25**(1): 68, doi: 10.3390/e25010068.

47. Tzirakis P., Zhang J., Schuller B.W. (2018), End-to-end speech emotion recognition using deep neural networks, [in:] *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), doi: 10.1109/ICASSP.2018.8462677.

48. Wang Y., Huo H. (2019), Speech recognition based on genetic algorithm optimized support vector machine, [in:] *6th International Conference on Systems and Informatics* (*ICSAI*), doi: 10.1109/ICSAI48974.2019.9010502.

49. Yang K. *et al.* (2023), Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition, [in:] *IEEE Transactions on Affective Computing*, **14**(2): 1082–1097, doi: 10.1109/TAFFC.2021.3100868.

50. Yildirim S., Kaya Y., Kiliç F. (2021), A modified feature selection method based on metaheuristic algorithms for speech emotion recognition, *Applied Acoustics*, **173**: 107721, doi: 10.1016/j.apacoust.2020.107721.

51. Young S. *et al.* (2006), *The HTK Book*, Cambridge University Engineering Department.

52. Zacharatos H., Gatzoulis C., Charalambous P., Chrysanthou Y. (2021), Emotion recognition from 3D motion capture data using deep CNNs, [in:] *IEEE Conference on Games* (*CoG*), doi: 10.1109/CoG52621.2021.9619065.

53. Zhang S., Chen A., Guo W., Cui Y., Zhao X., Liu L. (2020), Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition, *IEEE Access*, **8**: 23496–23505, doi: 10.1109/ACCESS.2020.2969032.

54. Zhao J., Mao X., Chen L. (2018), Learning deep features to recognise speech emotion using merged deep CNN, *IET Signal Process*, **12**(6): 713–721, doi: 10.1049/iet-spr.2017.0320.

55. Zhao J., Mao X., Chen L. (2019), Speech emotion recognition using deep 1D & 2D CNN LSTM networks, *Biomedical Signal Processing and Control*, **47**: 312–323, doi: 10.1016/j.bspc.2018.08.035.