IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 1, February, 2021, pp. 221-230
Journal homepage: http://iieta.org/journals/ts

Classification of Pitch and Gender of Speakers for Forensic Speaker Recognition from Disguised Voices Using Novel Features Learned by Deep Convolutional Neural Networks

Athulya M. Swamidasan Unni Nair*, Sathidevi P. Savithri

Department of Electronics and Communication Engineering, National Institute of Technology Calicut, Kozhikode 673601, India

Corresponding Author Email: athulya_p150030ec@nitc.ac.in

https://doi.org/10.18280/ts.380124

Received: 28 November 2020
Accepted: 20 January 2021

Keywords:
deep convolutional neural network, FASR, Mel-spectrogram, MFCC, pitch disguise

ABSTRACT

Voice disguise is a major concern in forensic automatic speaker recognition (FASR). Classifying the type of disguise is very important for speaker recognition. Pitch disguise is a very common type of disguise that criminals try to attempt. Among the different types of disguises, high pitch and low pitch voices show more distortion. The features that are robust for high pitch and low pitch voices are different. Moreover, the effect of disguise on male and female voices are also different. In this work, we classified high pitch and low pitch disguised voices for male and female voices using a novel set of features. We arranged Mel frequency cepstral coefficients (MFCC), ΔMFCC, and ΔΔMFCC features as three-dimensional features, and these are given as the RGB equivalent spectrogram inputs to pretrained AlexNet deep convolutional neural network (DCNN). We fused the AlexNet output features with corresponding MFCC correlation features. These fused features are the proposed novel features for disguise classification. Classification using neural network (NN) and support vector machine (SVM) classifiers are performed. Simulation results show that classification with SVM classifier using these novel features gives improved accuracy of 98.89% compared to 95.99% accuracy obtained by using DCNN output features using traditional spectrogram inputs.

1. INTRODUCTION

Disguised speech samples are often found in forensic scenarios like anonymous calls, threatening calls, blackmailing, kidnapping, police calls, bribery, terrorist activities etc. [1-5]. Automatic speaker recognition (ASR) from disguised speech is a highly challenging task and relevant concern in forensic applications [2, 5-9] since the speakers intentionally modify their voice by several means. ASR in forensic applications is called forensic automatic speaker recognition (FASR). A very elaborate study of voice disguises is presented in the studies [8, 10, 11]. The study [8] reveals that voice disguise is a highly complex problem in forensics. Voice disguise changes the vocal source and vocal tract properties of voice. The extent of variation in these properties depends on the type of disguise applied to a voice, the gender of the speakers and to a minute level on the individual speakers also. Perrot et al. [11] observed that using different disguises at the same time is a serious issue, but in forensics mostly the impostor uses a specific disguise only. In forensic scenarios, they also try to detect the disguise and type of disguise by using specific features. Speaker recognition problem from disguised voice mainly involves three steps:

1. Identifying the disguised voices
2. Classifying the type of disguise
3. Speaker recognition by performing preprocessing and robust feature extraction particular to that type of disguise.

Comparatively more works are done in the first and third stages of the problem. But very few works are concentrated on the second stage. Classifying the type of disguise can also include identifying the gender, age etc. which will aid in improving the performance of the third stage [8]. After identifying the type of disguise, speaker recognition can be done by extracting robust features for that type of disguise or by training using speech with the same type of disguise. This will improve the speaker recognition accuracy.

Speaker recognition from disguised voices significantly reduces the performance of the ASR system [4, 12, 13]. It is shown [3, 12] that some feature extraction methods and classification methods improve the recognition performance of ASR systems for disguised voices. The effect of disguise on ASR was evaluated by asking the subjects to do voice modification of their choice [14]. Majority of the subjects had chosen pitch disguise. The performance of the ASR system was reduced significantly with disguised test voices. The authors remarked that if the training is done with disguised voices, then the error can be reduced significantly.

Identifying whether a given test speech is disguised or original is the first step in ASR from disguised voices. In some works, deep features and neural network classifiers are used for this classification [15-18]. This classification is done in literature using both prosodic and cepstral features [16, 18-21]. Specific types of disguises are considered in most of the works like pitch disguised voices [16, 18-20], creaky voices [9, 17], mimicked voices [15, 21] etc.

Some of the related works available in literature analyzed various voice features affected by disguise and also robust to disguise. Glottal plosive is identified as robust to deliberate disguise of voice [22]. Mathur et al. [2] observed voice disguise as a serious threat in forensic scenarios and examined

the variation of fundamental frequency, F0, under different disguise conditions. They observed the variation of F0 to be different in different types of disguises. Singh et al. [23] analyzed the formant variations in phonemes due to expert mimicking and observed that expert mimicking has complex variation of formants for some phonemes and none for some other phonemes. They suggest the study of spoofing invariant-features to improve the biometric analysis of disguised voices. Hautamäki et al. [8] observed changes in average formant values when speakers tried to sound older or younger than their actual age. Leemann and Kolly [1] analyzed the impact of imitation of foreign dialect on suprasegmental temporal features like speaking rate and concluded high between-speaker variation and low within-speaker variation of suprasegmental features are required for improving speaker recognition performance in forensic context. The authors point out that pitch and formants show high within-speaker variation. In case of mimicking which is an extreme case of disguise, imitators usually try to imitate F0 and formants and succeed in doing so [24, 25]. By analyzing speech spectrograms, Endres et al. [25] found that there is a strong variation in formant structure between normal and disguised voices. Even though the imitators were able to change their formant structure and fundamental frequency, they were not able to adapt it to match or even be similar to the imitated person. Vestman et al. [26] analyzed the use of one automatic speaker verification (ASV) system to find the closest person who could be imitated by an impersonator and using this they checked how other ASV systems could be attacked. In their study, they observed that attackers were able to considerably change their speaking rate compared to F0 and formants. So from all these studies, we can see that the type of features which change and those which remain robust depends on the type of disguise and many other factors.

Voice disguise can be classified into deliberate and non-deliberate and also into electronic and non-electronic [11, 20, 27, 28]. Deliberate voice disguises are purposefully made voice disguises; non-deliberate disguises are those which happen unintentionally. Electronic disguises are made with some software and non-electronic ones are those which are made manually. Based on this, there are four broad types of voice disguises:

1. Deliberate and electronic
2. Deliberate and non-electronic
3. Non-deliberate and electronic
4. Non-deliberate and non-electronic

In this paper, we address the problem of disguise classification for deliberately electronic pitch disguised voices. We classify them into high pitch or low pitch voice. Pitch disguise classification can aid in speaker recognition performance in many ways like robust feature extraction, feature compensation, best feature selection, matched training etc. Apart from this, we also classify high pitch and low pitched voices into male and female voices. Classification of gender along with disguise classification can help significantly in the performance improvement of ASR from disguised voices. This is because, in many previous studies in literature, it is found that disguises affect male and female voices differently. Cross gender conversion increases the error rates in ASR systems with respect to the forensic scenario [10]. When a male tries to change his speech to female, some of the maleness remains [5]. Cross gender voice conversion causes more identification error rates compared to intragender conversions [10]. Gender classification can also be helpful in other speaker recognition aiding methods like using gender dependent background model set for normalizing the scores as done in the study [29]. So, in summary, disguises affect male and female voices in different ways. Knowing the gender will be helpful in the following stages of speaker recognition for many purposes, some of which are: extracting robust features specific to each gender, using the same gender speech for developing models for speaker recognition, minimizing the recognition error rates by removing opposite gender training models, removing features which undergo more distortion due to disguise which is specific to each gender, etc. So if we can determine whether the disguised voice is a male or female voice, then in the following stages of speaker recognition, this information can be used to improve speaker recognition performance.

The remaining part of this paper is organized as follows. Section 2 discusses the related work available in literature. Proposed method is described in section 3. Section 4 explains the experiments and results and section 5 concludes the paper.

2. RELATED WORK

A background study of the related methods and literature used in this work is presented here.

2.1 Pitch disguise

Changing pitch is a common disguise method done by criminals in forensic scenario [30]. Pitch disguised voices affect ASR performance significantly [4, 30]. Acoustic properties of disguised voices of 11 Chinese male speakers with raised and lowered pitch were investigated [30]. The authors indicate different abilities of speakers to adjust pitch. Very poor recognition rate was obtained for pitch disguised voices. High pitched voice has shown more degradation (only 10% recognition accuracy) compared to low pitched voice (55%). The effect on parameters like intensity, vowel formant frequency, speaking rate, syllable duration, long term average spectrum (LTAS) etc. were found to be different for raised pitch and lowered pitch.

As mentioned earlier, pitch disguise can be done by both electronic and manual (non-electronic) ways. Nowadays, due to technological advancements, there are lots of available software which can be used for this purpose. Voice changing software are widely used in audio forensics [18]. In the coming years, electronic means of disguise will be utilized to a large extent. Wu et al. [20] considered electronic means of pitch disguise and classified the speech into disguised or original speech using MFCC static and correlation features and a novel classification algorithm using support vector machines (SVM). Identification of weakly pitch shifted voice is performed in view of the forensic scenario [16]. We have also considered electronic pitch disguise in our work and extended the problem addressed in the study [20] for the classification of disguised voice into high pitch or low pitch voices. This is important because, in some previous works, it is shown that understanding the type of disguise can be utilized for improving the speaker recognition accuracy. Farrus [28] presents a very elaborate survey of existing works discussing the effect of deliberate/non-deliberate and electronic/non-electronic types of disguise on ASR systems and their role in modifying voice features. They suggest that the understanding of disguise based altered features will assist in the design of

voice recognition systems. So, considering the need for disguise and gender classification for improving the performance of the speaker recognition stage, we propose a new set of features for the classification of disguise type and gender. The features are extracted from pretrained AlexNet DCNN model. Normally, for speech processing applications like speech emotion recognition, spectrograms are given as inputs to the pretrained models. But in our work, we give MFCC features, their delta values and delta values as equivalent to the spectrogram images. The pretrained DCNN output features are appended with the correlation features. These novel features give good classification accuracy as compared to the state-of-the-art methods.

2.2 Gender and disguise type impact

The features affected, extent of feature distortion, robust features etc. vary, depending on the type of disguise, gender of the speakers, age of the speakers and the individual speakers [4, 8, 13]. Hautamäki et al. [8] find that the extent to which F0 and formants varied depends upon the speakers. Zhang and Tan [4] conclude that the effect of disguise on performance of ASR depends on the type of disguise. They considered 10 types of disguises common in forensic case works in their study. San Segundo et al. [9] put forward a finding that speaker recognition is easier under falsetto than under creaky condition at least in female voices in forensic phonetics application to speaker identification. Cross gender conversion increases the error rates in ASR systems with respect to the forensic scenario [10].

Some works analyze the difference in feature variations due to disguise for different genders [5, 7]. Tavi et al. [17] suggests that the effect of speaker's sex on creakiness should be treated carefully. González Hautamäki et al. [31] did an extensive study of how certain features are affected in male and female speakers differently in three voice conditions; modal, intended old and intended child. The features considered are formants F1-F4, bandwidths B1-B4, F0 and speaking rate. Prior to this, they analyzed the performance of ASR systems for males and females separately and found significant differences in the error rates. Identification rates were shown to be different in pitch lowered and pitch raised voices [32]. 10 types of disguises were considered and they analyzed the effect of disguise type on FASR performance [4]. They observed that depending on the type of disguise, the FASR performance varies. They also pointed out that some speakers can perform some types of voice disguise well and some speakers are well identified when performing certain types of disguise. Also, it is noted that raised pitch is more capable of defeating the FASR system.

Matched training conditions of gender and disguise have been addressed in previous works and have shown improvement in speaker recognition accuracy under matched training conditions. Training with matched gender [33] and disguise type [34] was found to help the speaker recognition stage. FSR system performance was affected only marginally when the same type of disguise was used for training [34]. The authors state that the effects are severe if training data is assembled with normal speech only. In their work, they considered high pitch, low pitch and pinched-nose disguises. The degradation in performance was more in high and low pitched voices. Domain mismatch between training and evaluation data reduces the scores [35]. Prasad, S. and Prasad, R. [3] addressed the mismatch problem in forensic caseworks by using multistyle training method and obtained improved performance. This was done by mixing speech of different speaking styles in different ways and using these for training. Above observations point to the necessity of classifying the gender of the speaker along with the type of disguise present in the voice.

2.3 Discriminative features

Extracting secondary features from speech is mainly done with spectrograms where a second set of features are extracted from a set of features derived earlier. Many speech applications like emotion classification [36-42], speech classification [43], sound event classification [44, 45], speaker recognition [46, 47], acoustic scene classification [48] use spectrograms as inputs to derive secondary features. SIFT features extracted from spectrograms are used for speech classification to perform speech classification [43]. Ren et al. [44] extracted local binary pattern (LBP) from the logarithm of the Gammatone like spectrogram to do sound event classification. Ajmera et al. [46] extracted Radon transforms from speech spectrograms as features and obtained superior performance for speaker identification. Features are derived by dividing the spectrogram into 4 by 4 matrices for emotion classification [38]. Hyder et al. [48] performed acoustic scene classification using CNN super vector derived from spectrogram images. Dennis et al. [45] carried out sound event classification by quantizing the grey scale range of spectrogram thereby limiting the effect of noise to certain quantization regions. This was found to be effective for sound event classification under mismatched conditions.

Discriminative features can be extracted from speech spectrograms by training DCNN with speech spectrograms [36, 37, 39-41, 47, 49]. Stolar et al. [40] used pretrained AlexNet to extract features from spectrogram for emotion classification. Discriminative features are learned from speech spectrograms for emotion recognition using CNN with rectangular kernels of varying shapes and sizes [37]. This method showed better performance compared to state-of-the art methods. A deep ResNet-based architecture is proposed for extracting features for speaker recognition [47]. Zheng et al. [41] also uses DCNN to learn features from log spectrograms. The log spectrograms are split into non-overlapping segments and given as input. The method showed superior performance compared to using standard hand-crafted acoustic features for emotion recognition.

A different approach is employed in the study [39] for emotion recognition using spectrograms. Here, instead of giving spectrogram images as inputs to DCNN, the authors extract static, delta and delta delta channels of log Mel-spectrograms similar to red, green, blue (RGB) image representation to train the DCNN. The AlexNet DCNN model pretrained on the large ImageNet dataset is employed to learn high-level feature representations and authors observe that DCNN model pretrained for image applications performs reasonably well in affective speech feature extraction also. We extend this idea of high level feature extraction to the application of disguise classification in our work. In our work, instead of using spectrogram static and dynamic features, we train the DCNN using MFCC static and dynamic features. We show that the proposed features perform better than those secondary features obtained from spectrogram static and dynamic features.

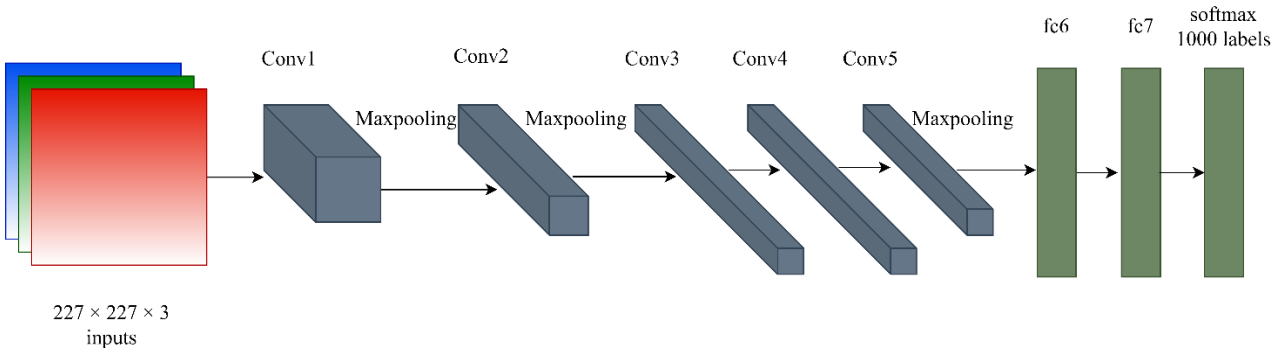

Figure 1. AlexNet architecture

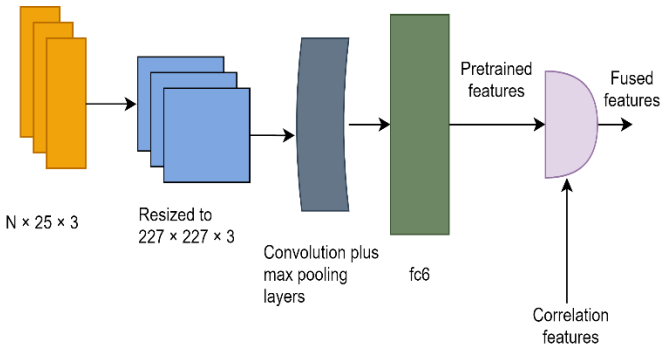

Figure 2. Feature extraction

2.4 Pretrained deep convolutional neural network

We have used AlexNet pretrained DCNN network for extracting the secondary features from MFCC inputs. AlexNet is a convolutional neural network trained on ImageNet dataset. ImageNet dataset consists of 15 million images. The AlexNet network architecture is shown in Figure 1. It consists of 5 convolutional layers followed by 3 fully connected layers. The last fully connected layer is connected to softmax layer, which does classification to 1000 classes. Basically, AlexNet is pretrained to classify images to 1000 classes. But it has shown very good performance in the case of speech classification applications also. Usually for classification of speech into various classes, AlexNet network is given speech spectrograms as inputs, which are equivalent to image representation. Spectrograms depict the intensity of speech for different time and frequency frames. Zhang et al. [39] used a different approach to give speech spectrogram input to AlexNet. They derived the delta and delta delta features of log Mel- spectrograms and represented them as 3-D equivalent for RGB representation of image. AlexNet then extracted the high-level feature representations from these inputs which resulted in good classification accuracy. In this paper, we employed well known MFCC features and their delta and delta delta representations as RGB equivalent which gave better performance than the corresponding spectrogram features. We took 25 dimension MFCC features. The number of frames varies for each speech utterance. So the size of MFCC features will be $N \times 25$ where N is the number of frames in each speech signal. Now the size of delta and delta delta features will also be $N \times 25$. Now these three set of features; i.e., static and dynamic features; are arranged to get the 3-D RGB equivalent of an image. So the size of the 3D feature is $N \times 25 \times 3$. The inputs are resized to 227×227×3 to match the AlexNet input dimensions. We took the output features from the fc6 layer, which is of dimension 4096. The correlation features of all static and dynamic sets of features are derived. Hence, the size of each of static and dynamic features becomes 1×300. These are appended to the fc6 output features resulting in the proposed feature set of dimension 4996. The feature extraction procedure is shown in Figure 2.

3. PROPOSED METHOD

3.1 Multilevel feature extraction

We proposed a novel set of features for high pitch and low pitch classification derived using pretrained AlexNet DCNN. Usually, for speech, the input data to DCNN are spectrogram images. We derived a new set of features by inputting a three dimensional feature matrix to the pretrained DCNN. The three dimensional RGB equivalent data is formed from the MFCC, their delta and delta delta feature coefficients. MFCC features are one of the most commonly used speech features. MFCC and their correlation features are used for disguise classification of electronic disguised voices [20]. Here, the authors classified the voice as original or disguised which is the first stage in speaker recognition from disguised speech. In this work, we classified the voice as high pitch or low pitch disguised voice. For this classification, the proposed novel features are shown to give better results compared to the features used in the study [20]. The features obtained from pretrained AlexNet are appended with the MFCC correlation features to get the final feature set.

24 dimension MFCC features and the energy feature are used in this work. Their delta and delta delta features and their energy features are also derived. So the three dimensional matrix used as input to AlexNet is initially of size $N \times 25 \times 3$. This is resized to 227 × 227 × 3, which is the required input image size for AlexNet. The output features of AlexNet are taken from the fc6 layer, which is of dimension 4096 as shown in Figure 2. This is appended with the correlation features of MFCC which are derived as explained in the study [20]. The general procedure of correlation feature extraction with reference to the study [20] is explained here.

Let L be the MFCC feature dimension and N be the number of frames in the speech signal. Let V_j be the set of j^{th} feature vector for all frames. We take it as the j^{th} column of the feature matrix. Correlation coefficient $CR_{jj'}$ between j^{th} and j'^{th} feature vector is calculated as,

$$CR_{jj'} = \frac{cov(V_j, V_{j'})}{\sqrt{VAR(V_j)}\sqrt{VAR(V_{j'})}}, i \le j < j' \le L \qquad (1)$$

Let set of all such correlation coefficients be denoted as C_{MFCC}. Now, the same calculation is done for delta MFCC matrix and delta delta MFCC matrix. Let them be denoted as$C_{\Delta MFCC}$ and $C_{\Delta\Delta MFCC}$. Now the total correlation features are:

$$C = [C_{MFCC}C_{\Delta MFCC}C_{\Delta\Delta MFCC}] \quad (2)$$

The dimension of correlation coefficients for MFCC, delta and delta delta MFCC is $L(L-1)/2$; each accounting for a total dimension of $3L(L-1)/2$ dimension correlation features.

Now, in our work, MFCC feature matrix of dimension 25 is taken and correlation coefficient of each column of this matrix with other columns are found according to Eq. (1). i.e., first column with second column, first column with third column... first column with last column. This gives 24 correlation coefficients. Next correlation of second column with third column, second column with fourth column... second column with last column is found. This gives 23 correlation features. Likewise we get 22, 21, ..., 2, 1 correlation features for the third, fourth, ... twenty three and twenty fourth columns of the matrix. So total dimension of correlation features for the MFCC matrix alone is 1+2+...+24 which is 300. The same procedure is repeated for delta and delta delta feature matrix also of same size. So the dimension of total correlation features becomes 3 times 300 which is 900. The final feature set is obtained by appending the pretrained AlexNet features of dimension 4096 with the MFCC correlation features to get a final feature dimension of 4996. The features extracted for comparison with Ref. [20] will be the 25 dimension MFCC mean features appended with 25 dimension mean delta delta features, 25 dimension mean delta delta features and 900 dimension correlation features counting to a dimension of 975.

3.2 Disguise classification framework

The disguise classification algorithm is shown in Figure 3. In phonetics, voice pitch is measured in 12-semitone division i.e. -1 to -12 and +1 to +11 [20]. We have considered semitones from -4 to -8 and +4 to +8. So there are total 10 levels or classes of pitch each for male and female speakers resulting in 20 classes. We performed the classification of test speech into these base classes initially from which they are again classified into male or female high pitch or low pitch classes. First we extracted MFCC features, their delta and delta delta features from all speech samples of all speakers from each of the pitch disguise classes. The static and dynamic features of each utterance are arranged as a three dimensional matrix which is equivalent to RGB image representation. Pretrained AlexNet DCNN is used for deriving the secondary features from the RGB equivalent three dimensional matrix representing MFCC and delta delta features. AlexNet, which is finely tuned by pretraining with more than a million images from the ImageNet database, is very efficient in classifying into 1000 image classes. We assume that using this model, we can efficiently extract more discriminative features of speech. The motivation behind this is the work [39] which shows better performance of secondary features in improving classification accuracy. Also, since secondary features derived from spectrograms are well performing, we hope secondary features derived from more acceptable MFCC features can perform much better.

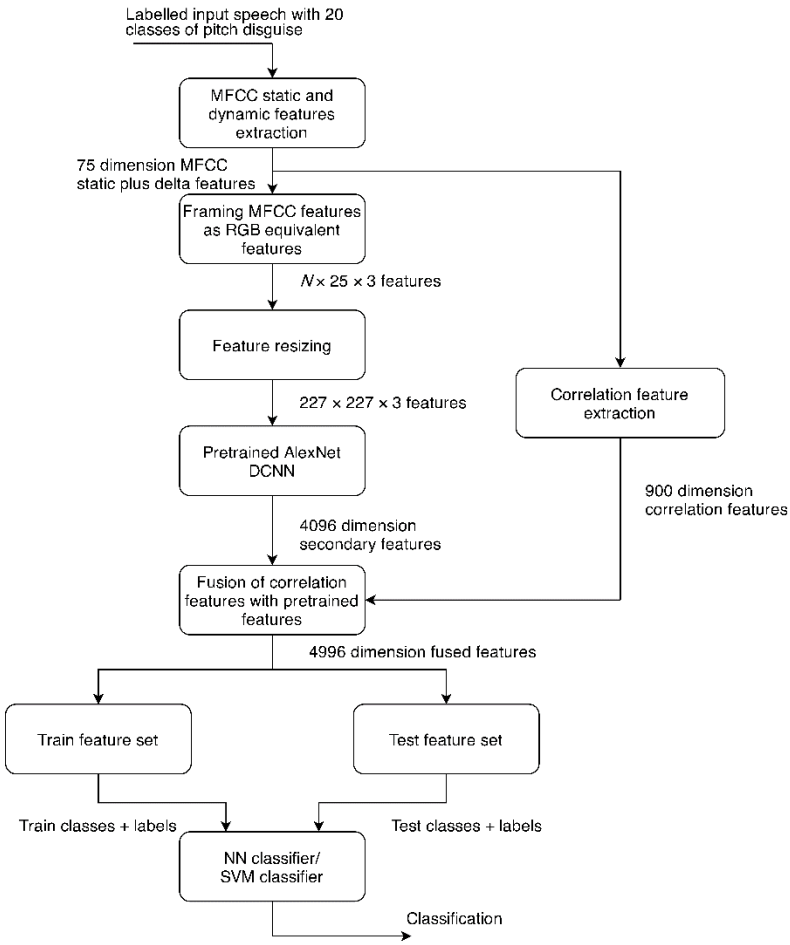

Figure 3. Disguise classification framework block diagram

The discriminative features derived from MFCC by training AlexNet DCNN are of dimension 4096, which is taken from the fc6 layer. Next, the pretrained AlexNet features are appended with the correlation features derived directly from the MFCC static and dynamic features resulting in a total feature dimension of 4996. We partitioned these features into train and test feature set. Classification is done and compared using a neural network (NN) classifier and a support vector machine (SVM) classifier in order to generalize the improved performance of the proposed feature set. The 20 classes of train and test speeches with corresponding labels are inputted to the classifier and the test speeches are classified into high pitch and low pitch male and female voices.

3.3 Classifiers

As mentioned in the previous section, classification of the fused DCNN extracted features is performed using the SVM classifier and NN classifier and performances are compared. Initial classification is done to different levels of pitch disguise and based on this classification the final low pitch and high pitch male and female classification is performed.

3.3.1 NN classifier

A fully connected feedforward neural network classifier [50] is shown in Figure 4. The inputs are the features and corresponding labels of ten levels of pitch disguise. Gradient descent back propagation method is used for minimizing the error between the output matrix and the target matrix/labels by adjusting the weights. After training the network, we get the final updated weight matrix which represents the classifier for classifying different levels of pitch disguise. Using this weight matrix, the test samples are classified. The output is thresholded by using the step activation function with threshold value of 0.5 and depending on the maximum element in the output matrix of the classifier, the level of disguise is determined.

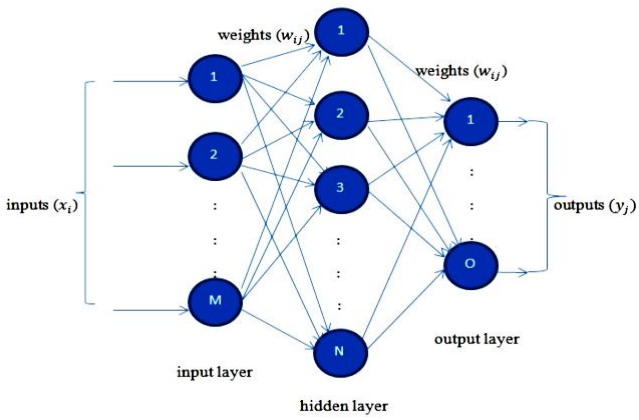

Figure 4. Neural network classifier

3.3.2 SVM classifier

The SVM classification algorithm is implemented by utilizing LIBSVM [51] through the MATLAB interface. SVM is a supervised type binary classifier based on structural risk minimization [52, 53]. Extension to multiclass classification is accomplished using the OAO (one against one) strategy. For an n class problem, the OAO formulation will involve ${}^{n}C_{2}$ binary SVMs; the decisions being fused using a majority voting scheme [54]. In this work, we utilize an OAO SVM scheme using the radial basis function (RBF) kernel. Equation for the RBF kernel transformation is given below.

$$K(v, w) = e^{-\gamma |v-w|^2} \quad (3)$$

where, γ is the RBF kernel parameter and v, w are the data samples. RBF kernel is the typically employed transformation for dealing with non-linear data classification problems using SVM [55]. The training phase involves finding the optimal hyperplane separating the classes. In this phase, the optimal values for the variable parameters in the RBF kernel SVM - penalty parameter, c and RBF kernel parameter, γ (since RBF kernel based SVM is used) are identified. 3 fold cross validation is performed on the training data for finding the parameters. Grid search is done in the ranges- $[2^{-15}, 2^{15}]$ and $[2^{-15}, 2^{15}]$ for c and γ parameters respectively. Parameter values associated with the best cross validation accuracy are identified through this process and training is performed for the SVM using these optimal values. The trained SVM model obtained is used to classify the test data samples [55]. Samples are generated by imposing both speaker independent and speech independent conditions. The utterances and speakers are selected randomly. Test samples are selected so that there is no overlap in the speech utterances also, aside from them originating from different speakers as to those considered in the training set.

4. EXPERIMENTAL SETUP AND RESULTS

In pitch and gender classification, 20 base classes are considered - 10 levels of male classes and 10 levels of female classes. The aim of this work is to finally classify the speech into one of the four classes namely;

(i) male high pitch
(ii) male low pitch
(iii) female high pitch
(iv) female low pitch

Among the 10 levels, levels corresponding to -4, -5,..., -8 semitones are considered as low pitch and +4, +5,..., +8 semitones are considered as high pitch.

4.1 Experiments

The database used for the experiments is TIMIT (Texas Instruments Massachusetts Institute of Technology) database. TIMIT database consists of 630 speakers which includes 438 males and 192 females with 10 different utterances for each speaker. For performing pitch disguise classification, the utterances are pitch disguised by ten disguise levels using Audacity software. The ten levels of disguise are from −4 to −8 semitones and +4 to +8 semitones. Each pitch level is further divided gender wise thus generating 10 male and 10 female pitch classes. In order to have speech independent testing, 5 speech utterances are used for testing and the remaining 5 speech utterances for training.

Initially, pretrained AlexNet features are extracted from all train speech utterances and correlation features are appended. Same features are extracted from the test speech. We simulated and compared classification accuracy using two classifiers; SVM and neural network classifier. For SVM classifier, we input the train data features for 20 classes with labels and train the classifier. When test input comes, it is classified to one of the 20 classes. The NN classification also

follows a similar procedure. Here, we consider only a single hidden layer since adding more hidden layers did not show performance improvement. Using one hidden layer is generally found to be sufficient for most classification problems [50]. If large dataset is used, then more hidden layers may give improved performance. The activation function used for the output layer is step activation function. The final classification for the test speech is done as follows:

(i) male classes corresponding to +4, +5,..., +8 semitones - male high pitch
(ii) male classes corresponding to -4, -5,..., -8 semitones - male low pitch
(iii) female classes corresponding to +4, +5,..., +8 semitones-female high pitch
(iv) female classes corresponding to -4, -5,..., -8 semitones -female low pitch

4.2 Results and discussion

Table 1 shows the percentage classification accuracy obtained for low pitch and high pitch classification for male and female speakers with NN classifier. We have compared the results with other features and the proposed novel features are shown to give better accuracy for individual classes as well as overall accuracy as shown in Table 1. The best accuracies are shown in bold numbers. One of the features with which we compared the proposed feature accuracy is MFCC + correlation features which have given improved results for performing pitch disguised or original speech [20]. But as shown in the table, these features perform comparatively poor when we use them for further classification into low pitch and high pitch classes for male and female speakers. There is an overall accuracy improvement of 2% for the proposed features with NN classifier. The second feature set with which comparison is done is the log Mel-spectrogram features. These features were shown to give good performance for emotion classification which is also a type of voice modification [39]. But compared to the proposed feature set they give poor classification accuracy. The proposed features give an overall accuracy improvement of about 3.7% with NN classifier. SVM classifier also identifies better classification for the proposed feature set with an improvement of 2.45% compared to MFCC + correlation features and 2.9% compared to log Mel-spectrogram features. This is tabulated in Table 2. We performed classification with two well-known classifiers to validate and generalize the effectiveness of the proposed features in classification compared to the most related features.

Table 1. Classification accuracy (in %) with different feature sets for NN classification

Class	Features		
	MFCC + Correlation	log Mel-spectrogram	Pretrained + Correlation (Proposed features)
Female low pitch	94.50	95.10	**98.10**
Female high pitch	92.73	89.20	**96.17**
Male low pitch	99.03	96.73	**99.63**
Male high pitch	99.03	97.43	**99.43**
Average	96.33	94.62	**98.33**

Table 2. Classification accuracy (in %) with different feature sets for SVM classification

Class	Features		
	MFCC + Correlation	log Mel-spectrogram	Pretrained + Correlation (Proposed features)
Female low pitch	94.37	96.57	**98.20**
Female high pitch	93.07	91.90	**98.13**
Male low pitch	99.07	97.90	**99.73**
Male high pitch	99.27	97.60	**99.50**
Average	96.44	95.99	**98.89**

Table 3. Confusion matrix for classification using pretrained AlexNet fc6 layer+correlation features with neural network

	C1	C2	C3	C4
C1	98.1	0	1.67	0.23
C2	0.03	96.17	0.03	3.77
C3	0.30	0	99.63	0.07
C4	0.13	0.40	0.03	99.43

Table 4. Confusion matrix for classification using pretrained AlexNet fc6 layer+correlation features with SVM

	C1	C2	C3	C4
C1	98.2	0	1.80	0
C2	0	98.13	0	1.87
C3	0.27	0	99.73	0
C4	0	0.50	0	99.5

Tables 3 and 4 show the confusion matrices for classification with the proposed feature set with NN and SVM classifiers respectively. C1, C2, C3 and C4 are the four classes: female low pitch, female high pitch, male low pitch and male high pitch respectively. In all the features using both classifiers, we can see that C1 is misclassified mostly into C3 and C2 to C4; i.e.; male low pitch to female low pitch and male high pitch to female high pitch and vice versa. This may be due to some common features added or removed in both genders when pitch is raised to or lowered from original pitch.

5. CONCLUSION AND FUTURE SCOPE

High pitch and low pitch classification system for both males and females are implemented in this work. A novel set of secondary features is derived for this classification and is shown to be more effective when compared to the state-of-the-art features. The secondary features are derived by inputting primary set of most commonly used speech features to a pretrained DCNN. Usually, DCNN are given image inputs. Our major contribution here is deriving the secondary features from DCNN by training the DCNN using non-image input data and appending them with correlation features which gives the proposed novel feature set. Extraction of secondary features was done by giving an RGB equivalent representation for the primary features.

The simulation results proved that we succeeded in deriving more robust secondary features as compared to the secondary features usually derived by giving spectrogram image inputs.

We used MFCC features for deriving the secondary features. We performed comparisons of results by using MFCC features as such and also with secondary features derived from log Mel-spectrogram inputs to DCNN. Also, classification using two classifiers; SVM classifier and NN classifier; was performed. For both classifiers, the proposed novel set of features outperformed the pitch classification done with other feature sets.

Disguise classification is very important as far as speaker recognition from disguised speech is concerned; especially in certain critical applications like forensic automatic speaker recognition. We can enhance the speaker recognition performance with prior knowledge of disguise and gender type through several means; either by doing matched training or by extracting disguise and gender specific robust features or by analyzing the disguise and gender specific varying features. We can further extend our work for disguise classification by including more types of disguises and also for different types of distortions. Our future work also aims to classify disguise type for speech utterances degraded by other types of distortions since such combined distortions often occur in forensic automatic speaker recognition scenarios.

REFERENCES

[1] Leemann, A., Kolly, M.J. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech. Speech Communication, 75: 97-122. http://dx.doi.org/10.1016/j.specom.2015.10.002

[2] Mathur, S., Choudhary, B., Vyas, C. (2016). Effect of disguise on fundamental frequency of voice. Journal of Forensic Research, 7(327): 2. http://dx.doi.org/10.4172/2157-7145.1000327

[3] Prasad, S., Prasad, R. (2019). Fusion multistyle training for speaker identification of disguised speech. Wireless Personal Communications, 104(3): 895-905. http://dx.doi.org/10.1007/s11277-018-6057-y

[4] Zhang, C., Tan, T. (2008). Voice disguise and automatic speaker recognition. Forensic Science International, 175(2-3): 118-122. http://dx.doi.org/10.1016/j.forsciint.2007.05.019

[5] Kumar, M., et al. (2019). Forensic Speaker Identification: A Review of Literature and Reflection on Future. Language in India, 19(7).

[6] Hove, I., Dellwo, V. (2014). The effects of voice disguise on f0 and on the formants. Proceedings of IAFPA 2014.

[7] Zhang, X., Zheng, L. (2019). Features extraction and analysis of disguised speech formant based on SoundTouch. 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, pp. 502-508. http://dx.doi.org/10.1109/IMCEC46724.2019.8983831

[8] Hautamäki, R.G., Sahidullah, M., Hautamäki, V., Kinnunen, T. (2017). Acoustical and perceptual study of voice disguise by age modification in speaker verification. Speech Communication, 95: 1-15.

[9] San Segundo, E., Alves, H., Trinidad, M.F. (2013). CIVIL corpus: Voice quality for speaker forensic comparison. Procedia-Social and Behavioral Sciences, 95: 587-593. http://dx.doi.org/10.1016/j.sbspro.2013.10.686

[10] Farrús Cabeceran, M., Wagner, M., Erro Eslava, D., Hernando Pericás, F.J. (2010). Automatic speaker recognition as a measurement of voice imitation and conversion. The International Journal of Speech. Language and the Law, 1(17): 119-142. http://dx.doi.org/10.1558/ijsll.v17i1.119

[11] Perrot, P., Aversano, G., Chollet, G. (2005). Voice disguise and automatic detection. In: Workshop on Nonlinear Speech Processing, WNSP 2005, pp. 101-117.

[12] Prasad, S., Tan, Z.H., Prasad, R. (2016). Multiple frame rates for feature extraction and reliable frame selection at the decision for speaker identification under voice disguise. Journal of Communication, Navigation, Sensing and Services (CONASENSE), 2016(1): 29-44. http://dx.doi.org/10.13052/jconasense2246-2120.2016.003

[13] Tan, T. (2010). The effect of voice disguise on automatic speaker recognition. 2010 3rd International Congress on Image and Signal Processing, Yantai, China, pp. 3538-3541. http://dx.doi.org/10.1109/CISP.2010.5647131

[14] Kajarekar, S.S., Bratt, H., Shriberg, E., De Leon, R. (2006). A study of intentional voice modifications for evading automatic speaker recognition. 2006 IEEE Odyssey - The Speaker and Language Recognition Workshop, San Juan, PR, USA, pp. 1-6. http://dx.doi.org/10.1109/ODYSSEY.2006.248123

[15] Chen, N., Qian, Y., Dinkel, H., Chen, B., Yu, K. (2015). Robust deep feature for spoofing detection—The SJTU system for ASV spoof 2015 challenge. In: Sixteenth Annual Conference of the International Speech Communication Association.

[16] Ye, Y., Lao, L., Yan, D., Wang, R. (2020). Identification of weakly pitch-shifted voice based on convolutional neural network. International Journal of Digital Multimedia Broadcasting, 2020: 1-10. http://dx.doi.org/10.1155/2020/8927031

[17] Tavi, L., Alumäe, T., Werner, S. (2019). Recognition of creaky voice from emergency calls. Proc. Interspeech 2019: pp. 1990-1994. http://dx.doi.org/10.21437/Interspeech.2019-1253

[18] Liang, H., Lin, X., Zhang, Q., Kang, X. (2017). Recognition of spoofed voice using convolutional neural networks. 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, pp. 293-297. http://dx.doi.org/10.1109/GlobalSIP.2017.8308651

[19] Singh, M.K., Singh, A., Singh, N. (2019). Multimedia analysis for disguised voice and classification efficiency. Multimedia Tools and Applications, 78(20): 29395-29411. http://dx.doi.org/10.1007/s11042-018-6718-6

[20] Wu, H., Wang, Y., Huang, J. (2014). Identification of electronic disguised voices. IEEE Transactions on Information Forensics and Security, 9(3): 489-500. http://dx.doi.org/10.1109/TIFS.2014.2301912

[21] Mary, L., Babu, K.A., Joseph, A. (2012). Analysis and detection of mimicked speech based on prosodic features. International Journal of Speech Technology, 15(3): 407-417. http://dx.doi.org/10.1007/s10772-012-9163-3

[22] Taseer, S.K. (2005). Speaker identification for speakers with deliberately disguised voices using glottal pulse information. 2005 Pakistan Section Multitopic Conference, Karachi, Pakistan, pp. 1-5. http://dx.doi.org/10.1109/INMIC.2005.334384

[23] Singh, R., Gencaga, D., Raj, B. (2016). Formant manipulations in voice disguise by mimicry. 2016 4th International Conference on Biometrics and Forensics (IWBF), Limassol, Cyprus, pp. 1-6. http://dx.doi.org/10.1109/IWBF.2016.7449675

[24] Eriksson, A., Wretling, P. (1997). How flexible is the human voice? A case study of mimicry. Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece.

[25] Endres, W., Bambach, W., Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. The Journal of the Acoustical Society of America, 49(6B): 1842-1848. http://dx.doi.org/10.1121/1.1912589

[26] Vestman, V., Kinnunen, T., Hautamäki, R.G., Sahidullah, M. (2020). Voice mimicry attacks assisted by automatic speaker verification. Computer Speech & Language, 59: 36-54. http://dx.doi.org/10.1016/j.csl.2019.05.005

[27] Rodman, R. (1998). Speaker recognition of disguised voices: A program for research. In: Proceedings of the 8th COST 250 Workshop, Ankara: «Speaker Identification by Man and by Machine: Directions for Forensic Applications, pp. 9-22.

[28] Farrus, M. (2018). Voice disguise in automatic speaker recognition. ACM Computing Surveys (CSUR), 51(4): 1-22. http://dx.doi.org/10.1145/3195832

[29] Mary, L., Yegnanarayana, B. (2006). Prosodic features for speaker verification. In: Ninth International Conference on Spoken Language Processing.

[30] Zhang, C. (2012). Acoustic analysis of disguised voices with raised and lowered pitch. 2012 8th International Symposium on Chinese Spoken Language Processing, Hong Kong, China, pp. 353-357. http://dx.doi.org/10.1109/ISCSLP.2012.6423510

[31] González Hautamäki, R., Hautamäki, V., Kinnunen, T. (2019). On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. The Journal of the Acoustical Society of America, 146(1): 693-704. http://dx.doi.org/10.1121/1.5119240

[32] Clark, J., Foulkes, P. (2007). Identification of voices in disguised speech. International Journal of Speech, Language and the Law, 14: 195-221.

[33] Snyder, D., Garcia-Romero, D., Povey, D. (2015). Time delay deep neural network-based universal background models for speaker recognition. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, pp. 92-97. http://dx.doi.org/10.1109/ASRU.2015.7404779

[34] Künzel, H.J., Gonzalez-Rodriguez, J., Ortega-García, J. (2004). Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In: ODYSSEY04-The Speaker and Language Recognition Workshop.

[35] Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L.P., Richardson, F., Dehak, R., Torres-Carrasquillo, P.A., Dehak, N. (2020). State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. Computer Speech & Language, 60: 101026. http://dx.doi.org/10.1016/j.csl.2019.101026

[36] Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., Cai, L. (2018). Emotion recognition from variable-length speech segments using deep learning on spectrograms. In: Interspeech, pp. 3683-3687. http://dx.doi.org/10.21437/Interspeech.2018-2228

[37] Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M.Y., Kwon, S., Baik, S.W. (2019). Deep features-based speech emotion recognition for smart affective services. Multimedia Tools and Applications, 78(5): 5571-5589. http://dx.doi.org/10.1007/s11042-017-5292-7

[38] Prasomphan, S. (2015). Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), London, UK, pp. 73-76. http://dx.doi.org/10.1109/IWSSIP.2015.7314180

[39] Zhang, S., Zhang, S., Huang, T., Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Transactions on Multimedia, 20(6): 1576-1590. http://dx.doi.org/10.1109/TMM.2017.2766843

[40] Stolar, M.N., Lech, M., Bolia, R.S., Skinner, M. (2017). Real time speech emotion recognition using RGB image classification and transfer learning. 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS), Surfers Paradise, Australia, pp. 1-8. http://dx.doi.org/10.1109/ICSPCS.2017.8270472

[41] Zheng, W., Yu, J., Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, pp. 827-831. http://dx.doi.org/10.1109/ACII.2015.7344669

[42] Meng, H., Yan, T., Yuan, F., Wei, H. (2019). Speech emotion recognition from 3D log-Mel spectrograms with deep learning network. IEEE Access, 7: 125868-125881. http://dx.doi.org/10.1109/ACCESS.2019.2938007

[43] Nguyen, Q.T., Bui, T.D. (2016). Speech classification using SIFT features on spectrogram images. Vietnam Journal of Computer Science, 3(4): 247-257. http://dx.doi.org/10.1007/s40595-016-0071-3

[44] Ren, J., Jiang, X., Yuan, J., Magnenat-Thalmann, N. (2016). Sound-event classification using robust texture features for robot hearing. IEEE Transactions on Multimedia, 19(3): 447-458. http://dx.doi.org/10.1109/TMM.2016.2618218

[45] Dennis, J., Tran, H.D., Li, H. (2010). Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Processing Letters, 18(2): 130-133. http://dx.doi.org/10.1109/LSP.2010.2100380

[46] Ajmera, P.K., Jadhav, D.V., Holambe, R.S. (2011). Text independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recognition, 44(10-11): 2749-2759. http://dx.doi.org/10.1016/j.patcog.2011.04.009

[47] Chung, J.S., Nagrani, A., Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622. http://dx.doi.org/10.21437/Interspeech.2018-1929

[48] Hyder, R., Ghaffarzadegan, S., Feng, Z., Hansen, J.H., Hasan, T. (2017). Acoustic scene classification using a CNN-SuperVector System trained with auditory and spectrogram image features. In: Interspeech, pp. 3073-3077. http://dx.doi.org/10.21437/Interspeech.2017-431

[49] Satt, A., Rozenberg, S., Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In: INTERSPEECH, pp. 1089-1093. http://dx.doi.org/10.21437/Interspeech.2017-200
[50] Heaton, Jeff. (2008). Introduction to Neural Networks with Java. Heaton Research, Inc.
[51] Chang, C.C., Lin, C.J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(3): 1-27. http://dx.doi.org/10.1145/1961189.1961199
[52] Vapnik, V. (2013). The Nature of Statistical Learning Theory. Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4757-3264-1
[53] Burges, C.J. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121-167.
[54] Milgram, J., Cheriet, M., Sabourin, R. (2006). "One against one" or "one against all": Which one is better for handwriting recognition with SVMs?
[55] Hsu, C.W., Chang, C.C., Lin, C.J. (2003). A practical guide to support vector classification.

NOMENCLATURE

N	dimensionless number of frames of each speech signal
L	dimensionless number of MFCC features
V_j	dimensionless set of j^{th} feature vector for all frames
CR	dimensionless correlation coefficient
C_{MFCC}	dimensionless set of all correlation coefficients of MFCC feature matrix
$C_{\Delta MFCC}$	dimensionless set of all correlation coefficients of delta MFCC feature matrix
$C_{\Delta\Delta MFCC}$	dimensionless set of all correlation coefficients of delta delta MFCC feature matrix
n	dimensionless number of classes in SVM

Greek symbols

ν	dimensionless data samples
ω	dimensionless data samples

Subscripts

j	dimensionless j^{th} feature vector
j'	dimensionless j'^{th} feature vector