01010
01010
01010 *information*

MDPI

Article

Novel Task-Based Unification and Adaptation (TUA) Transfer Learning Approach for Bilingual Emotional Speech Data

Ismail Shahin [1,*,†], Ali Bou Nassif [2,†], Rameena Thomas [1,†] and Shibani Hamsa [3,†]

[1] Department of Electrical Engineering, University of Sharjah, Abu Dhabi P.O. Box 127788, United Arab Emirates
[2] Department of Computer Engineering, University of Sharjah, Abu Dhabi P.O. Box 127788, United Arab Emirates
[3] Center for Cyber Physical Systems, Department of Electrical and Computer Engineering, Khalifa University, Abu Dhabi P.O. Box 127788, United Arab Emirates
* Correspondence: ismail@sharjah.ac.ae
† These authors contributed equally to this work.

Abstract: Modern developments in machine learning methodology have produced effective approaches to speech emotion recognition. The field of data mining is widely employed in numerous situations where it is possible to predict future outcomes by using the input sequence from previous training data. Since the input feature space and data distribution are the same for both training and testing data in conventional machine learning approaches, they are drawn from the same pool. However, because so many applications require a difference in the distribution of training and testing data, the gathering of training data is becoming more and more expensive. High performance learners that have been trained using similar, already-existing data are needed in these situations. To increase a model's capacity for learning, transfer learning involves transferring knowledge from one domain to another related domain. To address this scenario, we have extracted ten multi-dimensional features from speech signals using OpenSmile and a transfer learning method to classify the features of various datasets. In this paper, we emphasize the importance of a novel transfer learning system called Task-based Unification and Adaptation (TUA), which bridges the disparity between extensive upstream training and downstream customization. We take advantage of the two components of the TUA, task-challenging unification and task-specific adaptation. Our algorithm is studied using the following speech datasets: the Arabic Emirati-accented speech dataset (ESD), the English Speech Under Simulated and Actual Stress (SUSAS) dataset and the Ryerson Audio-Visual Database of Emotional Speech and Song dataset (RAVDESS). Using the multidimensional features and transfer learning method on the given datasets, we were able to achieve an average speech emotion recognition rate of 91.2% on the ESD, 84.7% on the RAVDESS and 88.5% on the SUSAS datasets, respectively.

Keywords: deep learning; emotion recognition; speech processing; transfer learning

Citation: Shahin, I.; Nassif, A.B.; Thomas, R.; Hamsa, S. Novel Task-Based Unification and Adaptation (TUA) Transfer Learning Approach for Bilingual Emotional Speech Data. *Information* **2023**, *14*, 236. https://doi.org/10.3390/info14040236

Academic Editor: Diego Reforgiato Recupero

Received: 24 January 2023
Revised: 2 April 2023
Accepted: 4 April 2023
Published: 12 April 2023

1. Introduction

There is a considerable difference between robots and humans in this age of rising artificial intelligence. Machines are incapable of understanding or expressing emotion, unlike humans. Speech emotion recognition, which examines emotions from spoken utterances, is given much attention. As more human–machine interactions use voice as an input, the importance of emotion identification from speech grows. Automated call centers, human–robotic interfaces and onboard computer systems in cars that use virtual assistants analyze the speaker's emotional state to increase safety and deliver a prompt answer. The speaker could exhibit any of the following emotions: neutral, happy, sad, anger, disgust, excitement and boredom. Emotion Recognition can be employed in criminal investigations by police and military to observe the mental and emotional state of a suspect, to tackle telephone extortion and personal attacks, in civil cases involving recorded conversations, in calls to insurance companies, by media to identify prank calls, in blended classroom training and in many other situations [1]. Speaker recognition from emotion becomes crucial in affective computing, where machines interact with humans effectively by recognizing, interpreting and expressing human emotions. Apart from Speaker Identification (SI), there is another wing called Speaker Verification (SV), which aims to accept or reject a speaker for their claimed activity. Telephone banking, credit card transactions, access to confidential government facilities and services, server access, biometric authentication and many more services rely on SV technology [2]. Traditionally, emotion recognition was accomplished using different kinds of emotional features such as keywords, facial expressions, speech signals, etc. Conventional methods that use keywords from spoken sentences suffer from uncertainty in emotional keyword interpretation and the lack of ability to understand sentences with no emotion-based keywords [3].

Speech signals are the most favoured and intelligible feature. Many research works suggest acoustic or prosodic features, namely, pitch, intensity, frequency, energy and speaking rate, be used for emotion recognition. Recent speech emotion recognition techniques have three fundamental phases: signal pre-processing followed by feature extraction and emotion classification. The primary stage of signal processing involves the denoising of a speech signal to remove corrupted noise from the signal. This is a critical step to reinforce the input data and elevate chances of more accurate results in further stages. The second phase has two parts, namely, feature extraction and feature selection. Extraction and selection of relevant features from the segmented signal takes place in the first stage. Mapping of these features to the right emotions using classifiers is the final stage. Classifiers, also called pattern recognizers, are broadly of two types: linear classifiers and non-linear classifiers. A few of the linear classifiers used are the Bayes Classifier, Linear Support Vector Machine, and discriminative classifiers such as Logistic regression, Least square methods and Perceptron classifiers. Common non-linear classifiers include the Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Decision trees, Polynomial classifiers and Artificial Neural Networks. Deep learning has picked up steam as a research field in machine learning. Deep learning methods compute on a parallel basis, with deeper layers of architecture constructed in order to overcome the limitations of existing methods. The use of advanced technologies such as Deep Boltzmann Machine (DBM), Deep Belief Network (DBN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) can enhance the comprehensive performance of designed system.

We have devised an advanced transfer learning system for speech emotion recognition called the Task based Unification and Adaptation (TUA) system. It primarily intends to bridge the gap between two of the crucial techniques, such as downstream customization and large-scale upstream training. The system could be used for a variety of specialized solutions in places of diversified demands such as well-prepared pre-trained weights and task-specific architectures. For datasets with very few datapoints, TUA can work alongside them to collect relevant data. Through this method, we aim to bring forth an intergrated system to scientists in the sphere of emotion recognition. Task-challenging unification and task-specific adaptation are the two major elements of TUA. Data are used in the task-agnostic unification process, followed by the system architecture. Inputs from numerous sources are gathered to create a data bank with a unified label space. This information is then used to determine the extent to which a given transfer learning method can be employed in speech emotion recognition. Incorporating a large amount of categories enables data to enhance the performance of a system and ensures task-specific adaptation. We employ the weight-sharing scheme [4–7] in our system to coherently train models with different widths and depths on a huge amount of upstream data. In methods such as task-specific adaptation, TUA is required to find the right model architectures for its specified tasks. [*R*2,1] The rationale for designing a transfer learning system for bilingual emotion recognition from audio signals lies in the potential benefits of using pre-trained models and knowledge transfer between languages to refine the accuracy and efficiency of the emotion recognition system. Transfer learning is a machine learning approach that involves leveraging knowledge gained from training on one task to boost performance on a related but different task. In connection with emotion recognition, transfer learning can be applied to improve the performance of the system by leveraging pre-existing knowledge and resources from other languages. The complex and varied nature of emotional expression across multiple languages and cultures makes bilingual emotion recognition a much strenuous task at hand. However, by using transfer learning techniques, the system can learn from the emotional characteristics of one language and apply that knowledge to another language, improving its accuracy and generalization capabilities. By designing a transfer learning system for bilingual emotion recognition, the researchers can capitalize on pre-existing knowledge and resources from one language to enhance the performance of the system in another language, while also lowering the amount of training data required. This can be particularly useful in scenarios where the amount of training data available is limited, or where it is difficult to obtain labeled data for a particular language. Overall, the rationale for designing a transfer learning system for bilingual emotion recognition from audio signals is to capitalize on pre-existing knowledge and resources to enhance the accuracy and performance of the system, while also reducing the amount of training data required. This approach has the potential to revamp the performance of emotion recognition systems in multilingual and cross-cultural contexts, with wide-ranging applications in areas such as speech therapy, affective computing and human-computer interaction. This research work contributes towards the following criteria:

- Our work explicitly exhibits how a weight-sharing scheme and transfer learning can be integrated into a frame powerful pre-trained prototype over a diverse set of architectures at once.
- According to the information we have, this is the first work utilizing a task-specific adaptation-based transfer learning approach for emotion recognition from speech.
- A huge data pool is compiled with a unified label for emotion classes available in the Arabic and English languages.

This paper is designed as follows: First, a literature review is put forth in Section 2. Then, the system description is provided in Section 3. This section presents the Feature Extraction and Selection in the first part followed by the Proposed TUA Model. The results along with supporting experiments are illustrated in Section 4. Lastly, the conclusion is presented in Section 5.

2. Literature Review

Communication through speech is the foremost medium of interaction among humans. Body language, heart rate, voice modulation, facial mien and blood pressure can divulge more information on the emotional state of a person. Comprehending these factors can help detect the emotional intensity behind the words spoken by a person. Speech signals are the most constructive signals, with linguistic and acoustic features such as intensity, pitch, vowel and tonal factors embedded in the signals, which make comprehension more feasible. Training a machine to establish the link between spoken sentences and the sentiments behind those spoken sentences is still a challenge, especially with the amount of training data being relatively small. Shahin [8] implemented and tested a new method using HMM classifiers for speaker identification in his paper. The experiment was completed using a speech database created by 20 male and 20 female adults uttering four sentences with an American accent. Each sentence was uttered while articulating emotions of anger, sadness, disgust, happiness, fear and neutral state. Speech signals were converted and sampled to undergo the Hamming window every 5 ms to extract Linear Prediction Coefficients (LPCs). These LPCs were then reconstructed to Linear Prediction Cepstral Coefficients (LPCC) to equate for vectors in HMMs. The average speaker identification rate was 78.8%, which is evidently a high performance rate at that time. The best speaker identification was obtained from the neutral state, whereas the worst speaker identification was obtained from the angry emotional state. Rong et al. [9] exploited the basic acoustic features from raw speech signals by applying Discrete Fourier Transform (DFT) and Mel-Frequency

Cepstral Coefficients (MFCC) to focus on a novel algorithm named Ensemble Random Forest to Trees (ERFTrees) to effectively select features from the available small dataset [9]. With the undesirable data removed, training sets turn more capable and intelligible. This work used two speech corpora from common sources: acted speech corpora and natural speech corpora in Chinese (Mandarin). The K-NN algorithm processed data to recognize emotions, and an emotion recognition rate of 72.3% was observed on the given speech datasets using Random Forest classifier. The emotions in the dataset were angry, happy, sad, fear and neutral.

Shahin [10] employed a two-stage recognizer approach that combines both HMM and Suprasegmental Hidden Markov Models (SPHMMs) as classifiers. MFCCs were extracted to give higher estimates of human auditory perception. Eight sentences were uttered nine times by 25 male and female speakers each, with a native American accent used for all six emotions: sad, disgust, happy, neutral, angry and fear. Here, the total number of utterances amount to 21,600. The average speaker identification rate obtained by a one-stage recognizer was 71.6%, while the performance improved to 79.9% by a two-stage recognizer using SPHMMs. In his research [11] on emotion recognition under stressful and emotional talking environments, Shahin experimented with three different classifiers: HMM, Second-order Circular Hidden Markov Model (CHMM2) and SPHMM classifiers. The speech database was gathered from 30 speakers under both talking environments. Observation vectors in both talking environments for the classifiers were obtained from MFCCs. SPHMM achieved the highest emotion recognition rates, 72.0% and 69.7%, in stressful and emotional environments, respectively. The results of SPHMMs surpassed both HMMs and CHMM2s under similar talking conditions. Additionally, the outcomes of stressful talking recognition were more true than the emotional talking recognition rate. The emotional talking environment had neutral, angry, sad, happy, fear and disgust states, whereas the stressful talking environment had loud, soft, neutral, shouted, slow and fast talking conditions. In a similar work [12], Shahin illustrated how the proposed system of combining both gender and emotion cues offer higher results than emotion or gender cues independently. Gender identification was computed by HMM on the dataset to categorize into male and female emotion groups. The second stage, called gender-specific emotion identification, used SPHMM to deduce the unknown emotion. The third stage was speaker identification. The first stage furnished a gender identification performance rate of 96.9% and the gender-dependent emotion identification stage on SPHMMs gave an average rate of 89.3%, which was higher than the rate in previous studies. The emotions used by the Collected Speech Database (CSD) are angry, neutral, sad, happy, fear and disgust.

Sun et al. [13] suggested that the distribution of energy on a spectogram would differ considerably for each emotion type. Weighted spectral features with Local Hu moments (HuWSF) have the potential to distinguish between energy concentrations of different emotion types on a spectogram. The results of speaker-independent emotion recognition experiments indicated 74.7% classification rates with the Berlin German Emotional Voice Library (EmoDB) [14], 45.4% with the Surrey Audio-Visual Expressed Emotion (SAVEE) [15] database and 41.9% with the speech emotion database of Institute of Automation, Chinese Academy of Sciences (CASIA) [16] using HuWSF. Speaker-dependent emotion recognition demonstration produced rates of 84.7%, 70.6% and 76.1% with EmoDB, SAVEE and CASIA, respectively, while using HuWSF. Wang et al. [17] proposed a Fourier parameter model using the permanent content of voice quality and speaker-independent speech emotion recognition. They enhanced the recognition rates using MFCC features and Support Vector Machine (SVM) [18] classifiers on the EmoDB of 73.3%. Shahin and Ba-Hutair [19] drew our attention to the results obtained using Second-Order Circular Suprasegmental Hidden Markov Model (CSPHMM2) classifiers to identify emotions in emotional and stressful talking environments. They substantiated that the performance of CSPHMM2s using MFCCs surpasses all other contemporary models. Average emotion identification in stressful talking environments using Speech Under Simulated and Actual Stress (SUSAS) [20] datasets was evaluated as 64.4%, 68.5%, 72.4% and 76.3% with HMMs, CHMM2, SPHMMs and CSPHMM2s, respectively. When using the Emotional Prosody Speech and Transcripts (EPST) database, the average performance was evaluated as 63.0%, 67.4%, 70.5% and 73.6% with HMMs, CHMM2s, SPHMMs and CSPHMM2s, respectively. In [21], Deng et al. used Semi-supervised autoencoders (SS-AE) to enhance emotion recognition. A revised form of SS-AE known as SS-AE-Skip, which presented skip connections from lower layer to upper layer, was introduced in this technique. In this model [21], SS-AE and SS-AE-Skip methods demonstrated a higher margin compared to other supervised and semi-supervised methods. Huang and Bao [22] proposed a Convolutional Neural Network (CNN) classifier-based system for the Ryerson Audio-Visual Database of Emotional Speech and Song dataset (RAVDESS) [23] to recognize four emotions. They used a 10-fold validation method, which calculates the average recognition rate as 72.2%. The emotions considered were angry, happy, neutral and sad. Zhao et al. [24] achieved emotion recognition accuracies of 95.3% and 95.9% for EmoDB for speaker-dependent and speaker-independent experiments, respectively, using a two-dimensional CNN "Long shortterm memory" (LSTM) [25] network. Happiness, sadness, neutral, surprise, disgust, fear and anger were the six emotions recognized in this model. Bhavan et al. [26] used MFCC, spectral centroid and MFCC derivative features on RAVDESS dataset to obtain an accuracy of 75.69%. Bagged ensemble of SVM was used as a classifier. Meng et al. [27] introduced a new architecture, ADRNN, expanded with CNN and Bidirectional Long Short-Term Memory (BiLSTM) with the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [28] dataset and EmoDB datasets. They attained 74.9% in speaker-dependent and 69.3% in speaker-independent environments with the IEMOCAP dataset. Additionally, accuracy rates were 90.4% and 84.9% with EmoDB for speaker-dependent and speaker-independent emotion recognition ex-periments. The Parallelized Convolutional Recurent Neural Network (PCRN) [29] proposed by Jiang et al. [29] extracts the frame-level features from each pronouncement. LSTM portrays these features by each frame. The SoftMax classifier categorizes the emotions. The performance of the system was higher than state-of-the-art works. The system attained a recognition rate of 58.3% for CASIA, 86.4% with EmoDB, 61.6% with Airplane Behaviour Corpus (ABC) [30] and 62.5% with SAVEE. The highest classification rates accomplished were 75.0% for "anger" with CASIA, 95.3% for "anger" with EmoDB, 86.3% for "aggressive" on ABC and 84.2% for "neutral" on SAVEE datasets.

Hamsa et al. [31] proposed a system for emotion recognition. They used Wavelet Packet Transform (WPT)-based cochlear filtering to extract MFCC features and Random Forest Classifier to classify emotions. The system achieved an average emotion recognition rate of 93.8% with RAVDESS when using five-fold validation and 98.0% when using ten-fold validation. The emotions observed were angry, happy, neutral, sad, calm and fearful. An average emotion recognition rate of 89.6% was recognized in the Arabic Emirati-emphasized speech dataset (ESD) [32] for six emotions. They are neutral, angry, sad, disgust, happy and fearful. In [32], Shahin et al. executed text-independent speaker identification under emotional conditions using a cascaded Gaussian Mixture Model and Deep Neural Network (GMM-DNN) as a classifier on the ESD and SUSAS datasets. The average speaker identification rate for ESD using GMM-DNN was 81.7%, which is superior to other existing methods, which were 77.2% for Deep Neural Network Bottleneck (DNN-BN), 78.6% for Single DNN and 77.4% for DBN methods. In another work [33], Shahin et al. validated the improved emotion recognition rate in normal and noisy talking conditions using the GMM-DNN classifier. MFCC features were extracted during training phase as it represents vectors closer to human voice. ESD dataset with six emotions such as neutral, fearful, angry, happy, sad and disgusted emotions verified an average emotion recognition of 83.97% using the GMM-DNN classifier as opposed to 80.3% with SVM and 69.8% with Multi-Layer Perceptron (MLP) classifiers. The hybrid classifier proved escalated performance in a turbulent environment. Hamsa et al. [34] proposed a novel deep sparse matrix representation (DSMR) approach for emotion recognition and reported an accuracy of 89.75% using the RAVDESS dataset.

Chen et al. [35] put forward a novel framework addressing the issue of unbalanced data distribution in the existing datasets. The system employed a unified first-order attention network with data balance to increase and stabilize the training data. The accuracies of emotion recognition obtained were 48.8%, 37.6% and 43.9% with the Bahçes¸ehir University Multimodal Affective Database-1 (BAUM-1s) [36], Acted Facial Expressions in the Wild (AFEW5.0) [37] and CASIA Chinese Natural Emotional Audio-Visual Database (CHEAVD2.0) [38], respectively. The results achieved by data balance were found to be more accurate, as it refines the stability of trained deep models. Peng et al. [39] explored the temporal modulations from the speech signals to design a model using Attention-based Sliding Recurrent Neural Networks (ASRNNs) and 3D convolutions. Experiments conducted with the IEMOCAP and Multimodal Signal Processing Improvisations databases (MSP-IMPROV) [40] achieved emotion recognition rates of 62.6% and 55.7%, respectively, as compared to earlier state-of-the-art methods. Zhong et al. [41] advanced with a framework wherein the performance of the system is calculated from empirically learned features (ELFs) and automatically learned features (ALFs). The comparison methods involved fused and independent training on each of the datasets used. Accuracy rates for the classification of the whole test set and each emotion were attained. In this experiment, independent training methods demonstrated a higher percentage as compared to fused training on all of the three datasets used. The presented approach achieved emotion recognition rates of 74.9% and 68.8% on the whole test set and each emotion sets, respectively, for the IEMOCAP database. Similarly, the rates amounted to 85.8% and 86.1% for EmoDB. On CASIA, this was observed to be 98.2% and 98.2% for the whole test set and each emotion set, respectively. Zhang et al. [42] computed a strategy using binaural representations and deep convolutional neural networks where a block-based temporal feature pooling method is used to form fixed-length utterance-level features and SVM is adopted for emotion classification. The system achieved 36.3% and 44.3% emotion recognition rates for the AFEW5.0 and BAUM-1s databases, respectively, which happens to be the highest for similar works with the same datasets. Shahin et al. [43] made advances in speech emotion recognition by using MFCC's spectogram features with a dual-channel long short-term memory compressed-CapsNet (DC-LSTM COMP-CapsNet) algorithm employed as the classifier. The average emotion recognition accuracy attained using this model on the Arabic Emirati dataset is 89.3%. The accuracies using the same dataset with various systems, such as Capsule Network (CapsNet) [44], CNN, SVM and KNN, were 84.7%, 82.2%, 69.8% and 53.8%, respectively.

Transfer learning has been used as one of the most potent techniques for deep learning. However, though it shows significant success with pre-training data, transfer learning is stringent with choosing its model architectures. Different datasets need individual model architectures, and different applications may demand new models of varying scales. As suggested in the "no free lunch" theorem [45] by David and William, there is no algorithm that is suited for distinguished application scenarios and datasets. In order to employ the full functionality of transfer learning, the models need to be custom-made and must provide training from scratch on the upstream datasets, which could be unreasonably overpriced. Therefore, the demand for task-specific architecture adaptation is much stronger in emotion recognition in various languages.

3. System Description

The basic schematic block of the proposed framework is shown in Figure 1. There are three stages executed in speech emotion recognition in this work: speech processing, features extraction and selection and, lastly, classification using a classifier. First, the speech signals are processed to remove unwanted noise. Our system uses OpenSMILE [46] software to extract desired features from the raw speech signal. SMILE stands for Speech and Music Interpretation by Large-space Extraction. Scientists can benefit from both the domains of music and speech signals when using OpenSMILE. It delivers an easy application with which different components can be computed. The results from one feature separator can be internally exported as an input to other units. We have the advantage of extracting Low-Level Descriptors (LLD) using OpenSMILE and can administer various filters, functional and transformation equations to it. Transfer learning is used to categorize the extracted signal components to predict emotion. Applications of transfer learning include pattern recognition and cross-corpus problems in image and audio processing systems [47]. People tend to exhibit similar attributes during speaking. Fear induces reduced loudness while an angry person has powerful facial expressions and increased loudness in their voice. Such robust traits of emotions are common among various available emotional datasets. The idea behind transfer learning in speech emotion recognition is to exploit these common characteristics. The source domain refers to a dataset with innumerable high-quality labeled data, while a target domain may contain a limited amount of data, either labeled or unlabeled [48]. Transfer learning is essentially the learning of a target domain using the classification knowledge obtained from the source domain.

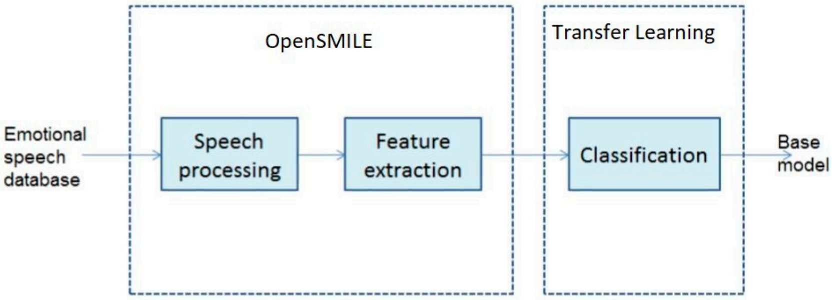

Figure 1. Proposed framework block diagram.

3.1. Feature Extraction and Selection

The acoustic features used in this experiment are of two kinds: statistical functions, also called functionals, and Low Level Descriptors (LLD). The most significant features of the signal are determined by various feature-selection algorithms. This reduces the calculation complexity of the high-dimensional feature sets. The commonly used feature selection method is the greedy algorithm, which is also known by the name forward selection algorithm. Initially it has an empty model, which adds features to the model by gradual regression until the termination condition.

The features used in this project for speech emotion recognition comprise the 384 extracted features from the INTERSPEECH2009 Emotion Challenge feature set [49] and 988 features from Emo-DB analysis. The Naïve Bayes classifier and Sequential Minimal Optimization algorithm (SMO) on SVM are the base models for training. The Naïve Bayes classifier exhibits high speed, better accuracy, reliability, less complexity and is also easier to work with in any domain. For an n dimensional feature space with random variables Y and $X_1, X_2, \ldots, X_n$; the feature vector components are given by $x_1, x_2, \ldots, x_n$.

For any variable y coupled with the feature vector components $x_1, x_2, \ldots, x_n$, the probability is constructed as:

$$P\,(Y = y, X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \quad (1)$$

By using the Naïve Bayes theorem, we come to the conclusion [50]:

$$P\,(Y = y,\ X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n) = P\,(Y = y) \prod_{i\,=\,1}^{n} P\ \ (X = x_1\ Y = y) \quad (2)$$

Naïve Bayes classifiers are effectively used in many applications, such as text classification, also known as text tagging or text categorization in Natural Language Processing. SVMs are usually used as the classifiers in problems dealing with pattern recognition and matching. A multiclass classification problem such as speech emotion recognition can be dealt with using the highly efficient SVM SMO classification method with linear kernel type.

3.2. Proposed TUA Model

This project explores emotion recognition using English and Arabic databases. The base model for speech emotion recognition is built from a huge data pool of English and Arabic datasets. The Arabic data used in this work is a standard Emirati-accented Arabic dataset [32]. The Machine Learning and Arabic Language Processing Research group at the University of Sharjah collected the ESD database. Fifty actors provided emotion expressions for the voiced communication in the Arabic language.

In this section, we present a modern transfer learning framework by the abbreviation TUA and its implementation details. TUA primarily consists of two modules of operations: task-challenging unification and task-specific adaptation. In the initial taskchallenging unification module, data is stored from numerous references to construct a data pool with a unified label space. Later, the models of different architectures are optimized together by implementing a weight-sharing learning scheme. The task-specific adaptation module selects the most suited architectures for any defined tasks, preps the network with the weights yielded from the first module and refines it on the downstream data. Due to the methods involved, this process is called task-specific architecture selection (TUA-AS). Moreover, sometimes the tasks may have very scant data points. TUA is designed to determine relevant data points from the data pool that are in tune with the specified tasks. This potential of the TUA is called the task-specific data selection (TUA-DS) shown in Figure 2.

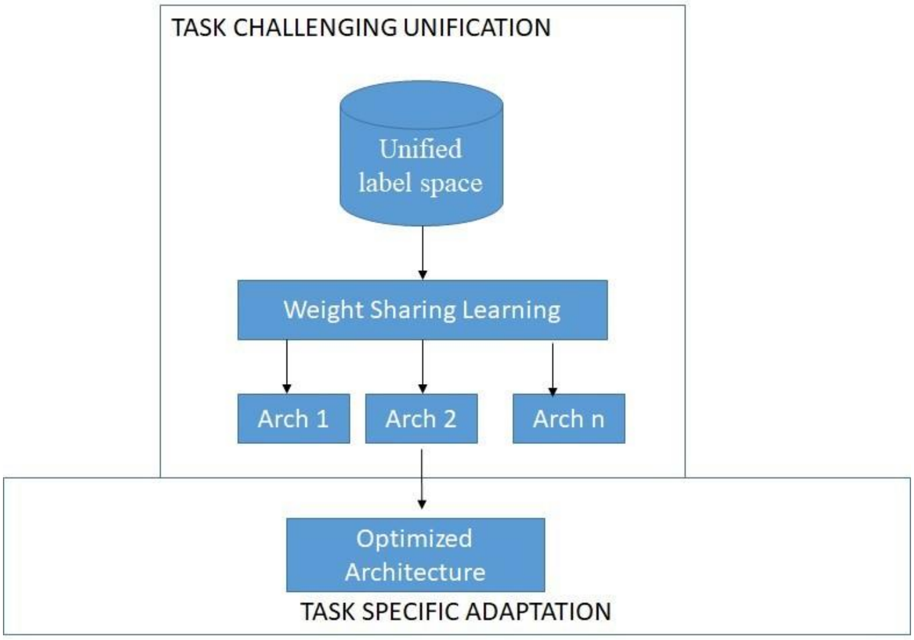

Figure 2. Proposed transfer learning approach.

3.2.1. Unified Data and Label Space

Different datasets are restricted to specific domains and are independent of one another. To advance the usefulness of an emotion identification system, we have combined datasets of the English and Arabic languages into a sizable data pool with a common label space. In our system, datasets $D = d_1, d_2, \ldots, d_N$ and label spaces $L = l_1, l_2, \ldots, l_N$ are represented. Six emotional categories $c_1, c_2, \ldots, c_6$ related to dataset d_i make up each label space. The initial unified label space $UL = c_1, c_2, \ldots, c_6$. was selected since it was the largest among the available label spaces, L. Then, we map the rest of the label spaces to the superset UL. While mapping, if there is no similarity found among the given categories, then we label it as a new category and add it to the *UL*. The mappings are then verified for credibility. The unified label space is always distinctive and variable, since the TUA can intake more datasets. The above-mentioned mapping process is repeated again whenever a new dataset is incorporated into the unified label space. When a pre-trained network needs to be appended into the label space, the terminal layers bestow new ways to connect while the rest of the layers remain the same. By using this unified label space, we can reduce the cost of integrating datasets, which lessens possible conflicts between redundant categories. This permits TUA to support down-stream tasks more efficiently. Sometimes, unified label space can bring in long-tail and partial annotation problems, which barely affects the fine tuning procedures in our study.

3.2.2. Anchor Based Gradual Down-Sizing

We initiate a training strategy called anchor-based gradual down-sizing (ABGD) that reduces multiple search dimensions steadily. We decide on a model anchor and create a search space encompassing it, while the depth and input scales remain intact. We curtail the model anchor after a few trainings and focus on refining the search space around the anchor. The same process is repeated until the complete search space has a vast latency range.

3.2.3. Model Adaptation

We randomly pick a model anchor, divide the subnets into groups and study the ranking to work efficiently. From each group, a set of models is evaluated in order to obtain the best-performing model. This model is then fine tuned using a 1 scheduler followed by a 0.2 "fast-fine tuning" to compare the ranking references. As the application of fast-fine tune to such a huge set of models is quite an extravagant process, we scale down the search space to a small but informative one. We then sample different models having varying depths and input scales around a defined Floating-point operations per second (FLOPs) and then execute fast-fine tuning to obtain the desired aspects. From the experiments, it is observed that models having similar input scales and depths have near precisions, while those with different inputs or depths have mixed precisions. Hence, in TUA, we implement a two-step search scheme. First, we sample n models in each sub search space. Then, the most efficient model from each group is selected and the top 50% of these models are fast-fine tuned to choose the apt architecture.

3.3. Transfer Learning Classification for Emotion Recognition

Transfer learning is a method of transferring the knowledge obtained in one model to process another model with a comparatively smaller set of data. This process is randomly sorted into two groups on the basis of (i) number of source datasets and, (ii) utilization of data in the target domain. Transfer learning methods may be single source transfer learning with one dataset in the source domain, or multi-source transfer learning having many datasets in the source domain. The second outlook on transfer learning results in supervised transfer learning and unsupervised transfer learning. Supervised transfer learning takes in labeled data from both source and target domains during training, while unsupervised transfer learning takes in data from the source domain only. Statistical-based transfer learning and deep transfer learning methods are the common learning practices of classification in speech emotion recognition. Domain adaptation by deep learning techniques deploy the layout of pre-trained models for transfer learning. In most cases, the last layers are replaced by new layers by fine-tuning some of the parameters of the models. By doing this, the attributes of the source task are forgotten. Progressive Neural Network (PNN) overcomes this by forming a layer onto the source network during the training phase. Most recent studies use the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) feature set, which contains frequency, spectral and dynamic information. In [51], Ghifary et al. proposed a conventional deep learning model with adaptive layers where Maximum Mean Discrepancy (MMD) could be utilized to tackle domain distribution difference.

In [52], Sawada et al. adopted a transfer learning method using Multi-Prediction Deep Boltzmann Machine (MPDBM). This method overwhelms the conventional methods of Deep Boltzmann Machine (DBM). Latif et al. [53] also used DBNs for recognition of emotion from speech. Since DBNs contain stacked Restricted Boltzmann Machines (RBMs), they work in a greedy manner to create a probabilistic model. RBM consists of three layers: visible layer, hidden layer and a bias unit. The bias unit is linked to each unit in both visible and hidden layers, whereas each visible unit is linked to units in the hidden layer. DBNs can be successfully trained to learn features and classify emotions, which can be exploited in cross-corpus and cross-language emotion recognition.

The Faster Region-based Convolutional Neural Network (Faster R-CNN) [54] is a deep convolutional network that is trained end-to end to form a single unified matrix. By using simple Faster R-CNN with Feature Pyramid Network (FPN) [55] as the base framework in our system, we can see a significant development in our efforts to design a state of the art model in speech emotion recognition. In the work [55] by Lin et al., they used a similar proposition in object detection. A Region of Interest (RoI) contains the extracted features for the Faster R-CNN. For a resolution level k in an image pyramid P_k, an RoI is set with width b and height h. Here, k is given by,

$$k = [k_0 + \log_2(\sqrt{bh})/224] \tag{3}$$

In our work, the task-specific adaptation and task-agnostic unification are applied to the Residual Network (ResNet) [56] prototype implemented in TUA. ResNet is chosen since it is much closer to the real-world applications and is the most realistic backbone in a similar field such as object detection. We selected three models during our training for the anchor-based gradual down-sizing (ABGD) method. In each round of training, we picked out a model at random that follows at least one of the stated guidelines of search space. The guidelines include: models having a minimum prescribed width, a maximum defined depth and other likeness.

4. Results and Discussion

[$R_{3,3}$] We augmented the data by mixing it with noise in a 2:1 and 3:1 ratio to scale the dataset and to enhance the noise susceptibility of the system. We used Adam optimizer to schedule the learning rate. We prepared the model for 150 epochs with an initial learning rate of 0.0005; after the 10th epoch, the learning rate is reduced by half every ten epochs. Our work presents a novel method using Transfer Learning Classification for speech emotion recognition using ESD, RAVDESS and SUSAS datasets. Figure 3 illustrates a high recognition rate for almost all the emotions in the Arabic dataset. The below chart demonstrates the highest average emotion recognition rate of 90.2% using the novel classifier on the ESD using the six emotions. The highest emotion recognition rate is observed for neutral and happy emotions; 97.2% and 96.1%, respectively. A comparative drop in the recognition rate is observed for the disgust emotion, with a rate of 79%. The proposed method is compared with the earlier works using GMM-DNN, MLP and SVM classifiers. The GMM-DNN is a hybrid classifier consisting of Gaussian mixture model and deep neural network. The GMM classifier evaluates the vectors and assigns a binary digit for each emotion, known as a GMM tag. The GMM tag is loaded into the DNN, which yields a probability distribution for each emotion. A performance analysis of the proposed systems, GMM-DNN, MLP and SVM, are graphically represented in Figure 4. The average emotion recognition accuracy rate using ESD is 83.9%, 69.7% and 80.3% for GMM-DNN, MLP and SVM, respectively. Compared to the existing methods, our proposed system exhibits an inflated average accuracy rate of 90.2% using the ESD dataset.

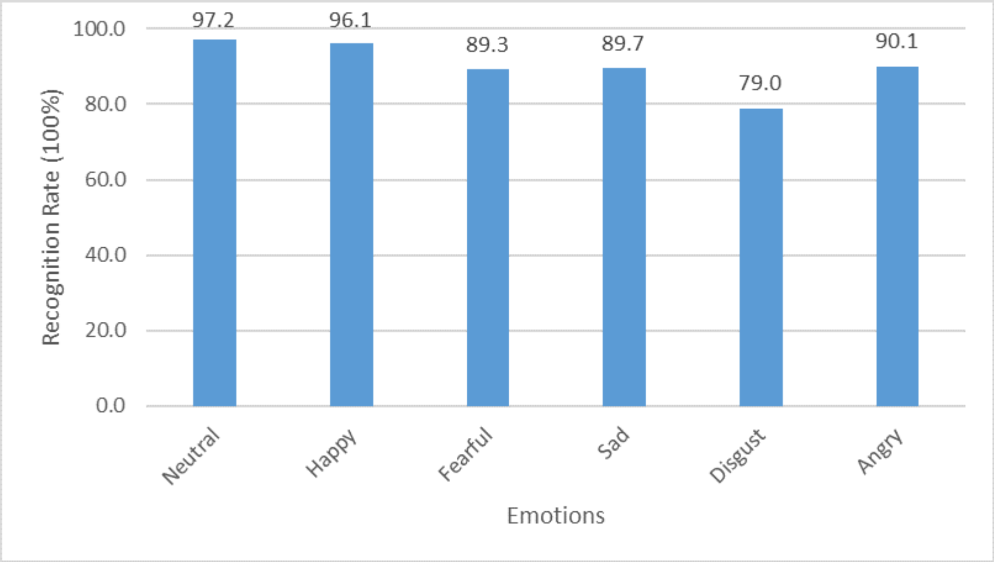

Figure 3. Emotion identification accuracy analysis of the presented system using ESD.

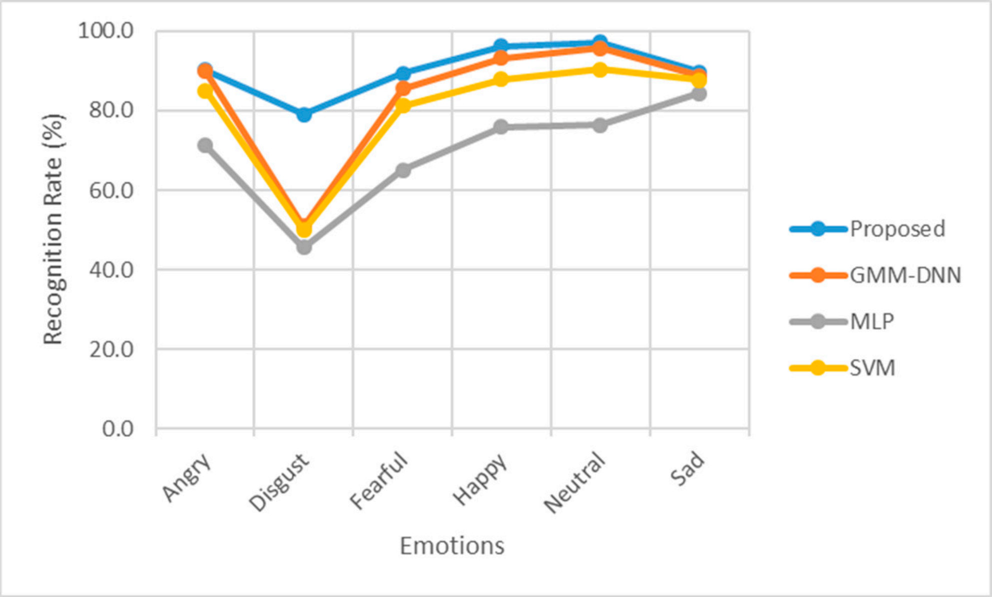

Figure 4. Emotion recognition rate obtained based on the proposed system, GMM-DNN, MLP and SVM using ESD.

The statistical significance test is an analysis test to measure the existence of a relationship between two variables. This indicates whether the deviation emotion recognition accuracy between the proposed system and each of the GMM-DNN, SVM and MLP are likely to occur due to a random statistical variation or to an actual factor. The analysis is prepared by stating the Null Hypothesis, wherein Student's *t* Distribution test is employed. The statistic *t* value to calculate the difference is given by [57],

$$t_{1,2} = \frac{\bar{X}_1 - \bar{X}_2}{SD_{Pooled}} \quad (4)$$

where,

$$SD_{Pooled} = \sqrt{\frac{{SD_1}^2 + {SD_2}^2}{2}} \quad (5)$$

with two samples of equal size n.

In the equation, $\bar{X}_1$ and $\bar{X}_2$ stand for the mean of the first and second sample, respectively, with size n each. SD_{Pooled} is the pooled standard deviation for size n. The standard error between the two means is calculated by (X_1, X_2). SD_1 is the standard deviation of the first sample, SD_2 demonstrates the standard deviation of the second sample, each having equal sample of size n. In this research, the assessed t values between the proposed transfer learning system and each of GMM-DNN, SVM and MLP are shown in Table 1. It is evident from Table 1 that the t values deliberated are higher than the critical threshold value $t_{0.05}$ = 1.654 at 0.05 level of significance. The measured t value with the proposed system and GMM-DNN model is 1.70, which is higher than the threshold value $t_{0.05}$. Hence, the calculation is statistically stable. When analyzing the MLP model, the t value between the proposed transfer learning system and the MLP model is 1.89, which is the highest among the three models.

Table 1. Computed t values of the presented transfer learning method and GMM-DNN, SVMs and MLP using ESD.

$t_{1,2}$	**Calculated *t* Value**
t_{proposed}, GMM-DNN	1.70
t_{proposed}, SVM	1.81
t_{proposed}, MLP	1.89

Two additional experiments were discreetly conducted to evaluate the performance of the proposed system. The two experiments are:

Experiment 1: In this experiment, the RAVDEES dataset is used to test the classifier to support the proposed system for emotion recognition. The "RAVDESS" is a well-substantiated dynamic multimodal gender balanced database of emotional speech and song. It consists of 24 players uttering lexically matched assertions in an unbiased North American articulation [23]. This speech corpora includes happy, surprised, calm, sad, angry, fearful and disgust expressions, and the song corpora contains happy, sad, angry, calm and fearful emotions. All emotions are created at two levels of intensity: normal and strong. Each actor accounts for 60 utterances and 44 song utterances, amounting to 104 utterances. There are three modalities designed using this data: audio-video, audio and video. This produces 312 records per person and a total of 3036 song recordings and 4320 speech utterances. The illustration in Figure 5 demonstrates the evaluation based on the proposed system using the RAVDESS dataset. The results indicate that the emotion recognition rate is steady across all the sets of emotions when using the RAVDESS dataset. The mean emotion recognition rate of the proposed system using the RAVDESS dataset is 84.7%, which is closer to the results obtained using Random Forest Classifier [31]. The results clearly specify that the highest emotion recognition rate of 96.3% is attained for neutral utterance condition, while all the other emotion classes display a lucid recognition rate higher than 80%.

Figure 6 illustrates a comparison of the proposed work and the state-of-the-art Random Forest Classifier model. The figure shows that the recognition rate is steady over the emotion sets when using the RAVDESS dataset. The graph demonstrates that there is an enhanced recognition rate for talking conditions with angry, calm and surprised emotions.

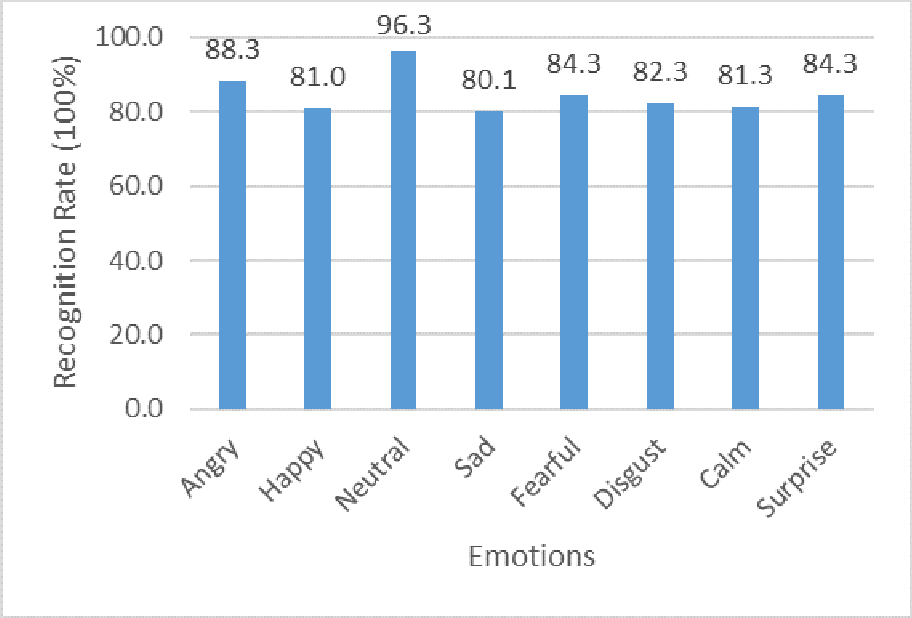

Figure 5. Emotion identification accuracy analysis of the proposed system using the RAVDESS dataset.

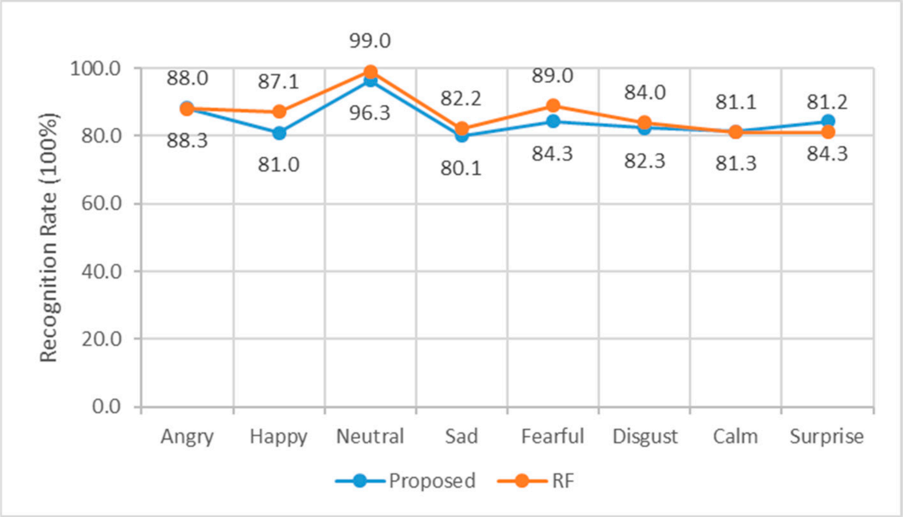

Figure 6. Emotion recognition rate obtained based on the proposed system and Random Forest Classifier using the RAVDESS dataset.

Tables 2 and 3 represent the comparison of the performance analysis of some of the earlier research works. The efficacy of the proposed system is compared with existing works using two different datasets: ESD and RAVDESS. The results of Table 2 show that our present work provides a boost in the mean emotion recognition rate of 6.3%, 2.0% and 0.6% over the outcome obtained by Shahin et al. [33], Hamsa et al. [31] using Gradient Boosting Classifier and Hamsa et al. [31] using Random Forest Classifier, respectively. The Arabic ESD dataset is used in these experiments, where the emotion classes considered are happy, neutral, sad, disgust, fearful and angry. Shahin et al. [33] reinforced his work with MFCC feature extraction model and GMM-DNN classifier. Hamsa et al. [31] used ten-fold cross validation to analyze the datasets while using both gradient boosting classifier and random forest classifier. Table 3 outlines the performance of recent experiments using the RAVDESS dataset. The results show that our proposed system attains increases in emotion recognition rate of 12.5%, 5.4%, 5.0%, 3.0% and 1.1% over the research approach put forward by Huang et al. [22], Gao et al. [58], Biqiao et al. (St Hier) [59], Biqiao et al. (Mt Hier) [59] and Shahin et al. [33], respectively.

Table 2. Performance analysis of various advanced methods using the ESD dataset.

Method	Features	Classifier	Validation	Accuracy (%)
I. Shahin [33]	MFCC	GMM-DNN hybrid classification	1:2 ratio	83.9
S. Hamsa [31]	MFCC	Gradient Boosting	K-fold	88.2
S. Hamsa [31]	MFCC	Random Forest Classifier	K-fold	89.6
S. Hamsa [34]	DSMR	Random Forest Classifier	K-fold	90.9
Proposed	Learned	Transfer Learning	K-fold	91.2

Table 3. Performance study of various methods employing the RAVDESS dataset.

Method	Features	Classifier	Validation	Accuracy
Y. Gao [58]	MFCC, LSP, Pitch, ZCR	SVM	10-fold	79.3
A. Huang [22]	MFCC, STFT	CNN	10-fold	72.2
I. Shahin [33]	MFCC	GMM-DNN	1:2 ratio	83.6
S. Hamsa [31]	MFCC	Random Forest	5-fold	86.4
Z. Biqiao [59]	LLD (Mt Hier)	SVM	5-fold	81.7
S. Hamsa [34]	DSMR	Random Forest	5-fold	93.8
Z. Biqiao [59]	LLD (St Hier)	SVM	5-fold	79.7
Proposed	Learned	Transfer Learning	Cross validation	84.7

Experiment 2: The second experiment has been conducted to support the proposed model, using the SUSAS dataset to accomplish the emotion recognition rate. The SUSAS dataset encloses utterances of two varieties: actual and simulated speech under stress. We use utterances from thirty-two speakers in six speaking-style classes. The six speaking styles are angry, neutral, slow, loud, soft and fast. Figure 7 illustrates the emotion recognition rate of the proposed algorithm using the SUSAS dataset. The results indicate a stable accuracy rate, with neutral having the highest accuracy rate of 93.2%, whereas the accuracy rate of angry utterances is relatively less than other utterances. The average emotion recognition rate attained for this algorithm using the SUSAS dataset is 88.5%, which is on the higher end compared to the existing state-of the-art algorithms.

A comparison between the proposed and recent algorithms using the SUSAS dataset is depicted in Figure 8. The graph exhibits an emotion recognition rate that is fairly stable, using the proposed algorithm as differentiated by the previous research works. The average emotion recognition rate is 88.5%. An accuracy rate of 93.2% is obtained by the neutral talking condition, whereas the lowest accuracy rate in the emotion classes used is as high as 84.2% which is the highest of all the recognition rates for the angry talking condition in recent research works. The differences in the recognition rate of the proposed system and the earlier works are mentioned in the given graph. An advancement of 32.2%, 26.6%, 28.8%, 23.4% and 30.5% is observed when using the proposed algorithm over the HMMs model in angry, slow, loud, soft and fast talking conditions, respectively. Similarly, compared with the CHMM2s model, our system yields a 27.7%, 20.6%, 23.3%, 18.9% and 26.5% rise in the emotion recognition rate for angry, slow, loud, soft and fast talking conditions, respectively. Comparisons with the SPHMMs model bring forth a surge in the percentages. The proposed system gives a 25.7%, 15.1%, 19.3%, 14.9% and 21% increase in recognition rates in angry, slow, loud, soft and fast utterances when compared with the SPHMMs model. Table 4 exemplifies the performance evaluation of the proposed feature extraction and classifier methods with respect to the numerous feature extraction approaches employed in earlier works. The results from the table indicate that the proposed algorithm offers better performance when compared with other commonly used techniques.

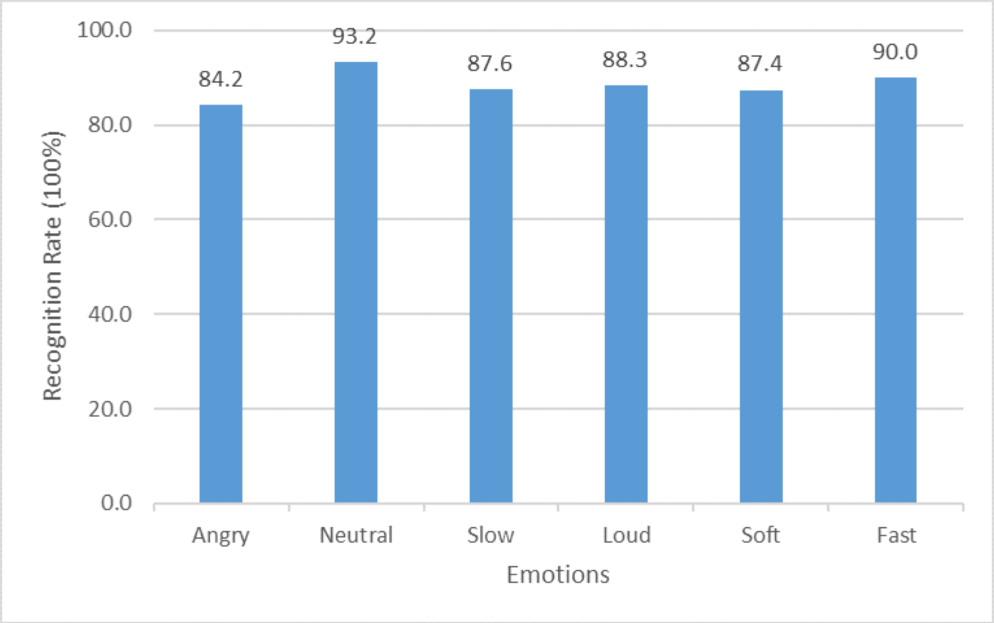

Figure 7. Emotion recognition accuracy assessment of the proposed system using SUSAS dataset.

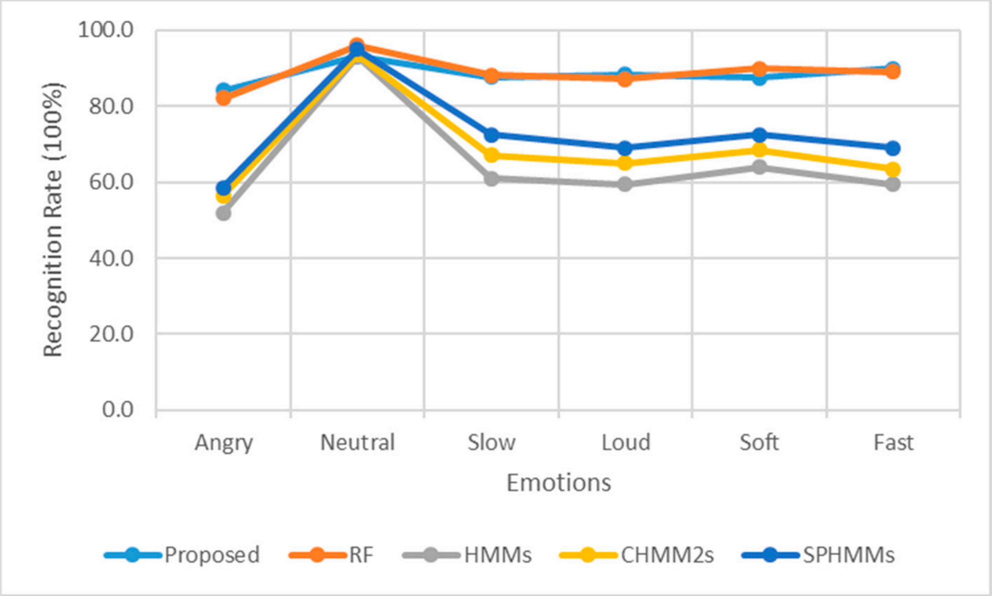

Figure 8. Comparison of the Emotion Recognition rate attained by the proposed algorithm, Random Forest Classifier model, HMMs, CHMM2s and SPHMMs using the SUSAS dataset.

As mentioned in Table 2, our proposed work offers an almost identical or even better performance when using the Arabic ESD dataset as compared to other models. This is because the proposed work was trained using the Arabic ESD dataset and is tested using thhe RAVDESS and SUSAS datasets, as mentioned in Tables 3 and 4, respectively. While the algorithms using the RAVDESS and SUSAS datasets yield superior performance than the proposed algorithm, since these works are trained and tested on the same dataset for emotion recognition.

Table 4. Performance study of various methods utilizing the SUSAS dataset.

Method	Features	Classifier	Validation	Accuracy
Q.Y. Hong [60]	MFCC	GA	-	68.7
I. Shahin [33]	MFCC	GMM-DNN	10-fold	86.6
S. Hamsa [31]	MFCC	Random Forest	10-fold	88.6
T. Kinnunen [61]	MFCC	VQ	Not mentioned	68.4
S. Hamsa [62]	DSMR	Random Forest	10-fold	90.9
W.M. Campbell [63]	MFCC	SVM	Not mentioned	72.8
Proposed	Learned	Proposed TUA	Cross validation	88.5

Table 5 demonstrates the complexity of the recently proposed emotion recognition models. The results show that the proposed transfer learning approach considerably reduces the complexity of the SER algorithm. It required 69,331.1 s for training, versus 84,128.13 s for Random Forest and 95,921.45 for GMM-DNN hybrid classifier models.

Table 5. Analysis of the model in terms of computational complexity.

Model	Training	Testing
Random Forest	84,128.13	2.46
GMM-DNN	95,921.45	4.12
Proposed	69,331.1	2.01

Table 6 reports the accuracy, precision, recall and F1-score. The proposed framework offers superior performance over GMM-DNN in terms of accuracy. However, the proposed framework performance is less than the random forest classifier model in terms of accuracy, though it outperforms the random forest classifier based on the F1-score. In terms of F1-score and computational complexity, the proposed framework offers superior performance over the random-forest-based SER model [34].

Table 6. Analysis in terms of performance evaluation matrices using the RAVDESS dataset.

Model	Accuracy	Precision	Recall	F1-Score
Proposed	0.84	0.83	0.84	0.84
Random Forest [31]	0.86	0.85	0.85	0.82
GMM-DNN [1]	0.83	0.83	0.82	0.83

5. Conclusions

[$R_{6,3}$] In this work, we designed, implemented and evaluated a model for emotion recognition. The proposed framework utilizes the benefits of a transfer learning approach on a bilingual platform and obtained remarkable results in terms of performance without repeating complex and time-consuming training procedures. Task-based unification and adaptation is an approach that involves unifying and adapting multiple related tasks to improve performance on each individual task. This approach can be applied to other feature recognition problems in other domains where high performance transfer learning has become an attractive solution. One instance is in computer vision, where transfer learning is commonly used to improve performance on specific tasks such as object detection or image classification. By unifying and adapting multiple related tasks, such as pedestrian detection and vehicle detection, transfer learning can be used to improve performance on each individual task. This can be particularly useful in domains such as autonomous driving, where accurately detecting and classifying objects in the environment is critical for safety. Task-based unification and adaptation can help improve performance on individual tasks while reducing the need for large amounts of labeled data. By leveraging related tasks and adapting existing models, transfer learning can help reduce the cost and time required for training new algorithms. This approach can be particularly valuable in domains where the cost of producing new data for training is high, or where labeled data is scarce.

Our system uses a pragmatic key for the research and development of feature pyramids where projections are made on each level. The obtained results in different experimentations ensure the model's performance in both English and Arabic languages. Our model achieves competitive results compared to previously published DSMR speech emotion recognition models and state of-the-art hybrid models, with lower latency, higher performance and fewer parameters. The limitations of this system are the amount of time and heavy resources required to build an operable TUA. This leaves room for future work in creating TUA frameworks much more efficiently. Our future work aims to incorporate multiple languages in the data pool to facilitate a unified transfer-learning model suitable for affective computing-based applications.

Author Contributions: All authors have equal participation in Conceptualization, methodology, software, validation, formal analysis, writing, original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University of Sharjah in the United Arab Emirates through the internal competitive research project entitled Investigation and Analysis of Emirati-Accented Corpus in Neutral and Abnormal Talking Environments for Engineering Applications using Shallow and Deep Classifiers, 20020403159.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions.

Acknowledgments: The authors of this research would like to express their gratitude to the "University of Sharjah in the United Arab Emirates for providing financial support to this study through the competitive research project entitled Investigation and Analysis of Emirati-Accented Corpus in Neutral and Abnormal Talking Environments for Engineering Applications using Shallow and Deep Classifiers, 20020403159".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shahin, I. Emotion Recognition Using Speaker Cues. In Proceedings of the 2020 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 4–6 February 2020; pp. 1–5.
2. Shahin, I. Speaker verification in emotional talking environments based on three-stage frame-work. In Proceedings of the 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 21–23 November 2017; pp. 1–5.
3. Wu, C.H.; Liang, W.B. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Trans. Affect. Comput.* **2010**, *2*, 10–21.
4. Liu, H.; Simonyan, K.; Yang, Y. Darts: Differentiable architecture search. *arXiv* **2018**, arXiv:1806.09055.
5. Pham, H.; Guan, M.; Zoph, B.; Le, Q.; Dean, J. Efficient neural architecture search via parameters sharing. In Proceedings of the International Conference on Machine Learning, Macau, China, 26–28 February 2018; pp. 4095–4104.
6. Cai, H.; Gan, C.; Wang, T.; Zhang, Z.; Han, S. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv* **2019**, arXiv:1908.09791.
7. Yu, J.; Jin, P.; Liu, H.; Bender, G.; Kindermans, P.J.; Tan, M.; Huang, T.; Song, X.; Pang, R.; Le, Q. Bignas: Scaling up neural architecture search with big single-stage models. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2020; pp. 702–717.
8. Shahin, I. Using emotions to identify speakers. In Proceedings of the The 5th International Workshop on Signal Processing and its Applications (WoSPA 2008), Sharjah, United Arab Emirates, 18–20 March 2008.
9. Rong, J.; Li, G.; Chen, Y.P.P. Acoustic feature selection for automatic emotion recognition from speech. *Inf. Process. Manag.* **2009**, *45*, 315–328. [CrossRef]
10. Shahin, I. Identifying speakers using their emotion cues. *Int. J. Speech Technol.* **2011**, *14*, 89–98. [CrossRef]
11. Shahin, I. Studying and enhancing talking condition recognition in stressful and emotional talking environments based on HMMs, CHMM2s and SPHMMs. *J. Multimodal User Interfaces* **2012**, *6*, 59–71. [CrossRef]

12. Shahin, I. Employing both gender and emotion cues to enhance speaker identification performance in emotional talking environments. *Int. J. Speech Technol.* **2013**, *16*, 341–351. [CrossRef]
13. Sun, Y.; Wen, G.; Wang, J. Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomed. Signal Process. Control* **2015**, *18*, 80–90. [CrossRef]
14. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
15. Haq, S.; Jackson, P.J.; Edge, J. Speaker-dependent audio-visual emotion recognition. In Proceedings of the AVSP, Norwich, UK, 10–13 September 2009; pp. 53–58.
16. Li, Y.; Tao, J.; Chao, L.; Bao, W.; Liu, Y. CHEAVD: A Chinese natural emotional audio–visual database. *J. Ambient. Intell. Humaniz. Comput.* **2017**, *8*, 913–924. [CrossRef]
17. Wang, K.; An, N.; Li, B.N.; Zhang, Y.; Li, L. Speech emotion recognition using Fourier parameters. *IEEE Trans. Affect. Comput.* **2015**, *6*, 69–75. [CrossRef]
18. Schuller, B.; Rigoll, G.; Lang, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; Volume 1, pp. 1–577.
19. Shahin, I.; Ba-Hutair, M.N. Talking condition recognition in stressful and emotional talking environments based on CSPHMM2s. *Int. J. Speech Technol.* **2015**, *18*, 77–90. [CrossRef]
20. Hansen, J.H.; Bou-Ghazale, S.E. Getting started with SUSAS: A speech under simulated and actual stress database. In Proceedings of the Fifth European Conference on Speech Communication and Technology, Rhodes, Greece, 22–25 September 1997.
21. Deng, J.; Xu, X.; Zhang, Z.; Frühholz, S.; Schuller, B. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *26*, 31–43. [CrossRef]
22. Huang, A.; Bao, P. Human vocal sentiment analysis. *arXiv* **2019**, arXiv:1905.08632.
23. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
24. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
26. Bhavan, A.; Chauhan, P.; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst.* **2019**, *184*, 104886. [CrossRef]
27. Meng, H.; Yan, T.; Yuan, F.; Wei, H. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]
28. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
29. Jiang, P.; Fu, H.; Tao, H.; Lei, P.; Zhao, L. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access* **2019**, *7*, 90368–90377. [CrossRef]
30. Schuller, B.; Arsic, D.; Rigoll, G.; Wimmer, M.; Radig, B. Audiovisual behavior modeling by combined feature spaces. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; Volume 2, pp. 2–733.
31. Hamsa, S.; Shahin, I.; Iraqi, Y.; Werghi, N. Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier. *IEEE Access* **2020**, *8*, 96994–97006. [CrossRef]
32. Shahin, I.; Nassif, A.B.; Hamsa, S. Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments. *Neural Comput. Appl.* **2020**, *32*, 2575–2587. [CrossRef]
33. Shahin, I.; Nassif, A.B.; Hamsa, S. Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE Access* **2019**, *7*, 26777–26787. [CrossRef]
34. Hamsa, S.; Shahin, I.; Iraqi, Y.; Damiani, E.; Nassif, A.B.; Werghi, N. Speaker identification from emotional and noisy speech using learned voice segregation and speech VGG. *Expert Syst. Appl.* **2023**, *224*, 119871. [CrossRef]
35. Chen, G.; Zhang, S.; Tao, X.; Zhao, X. Speech Emotion Recognition by Combining a Unified First-Order Attention Network with Data Balance. *IEEE Access* **2020**, *8*, 215851–215862. [CrossRef]
36. Zhalehpour, S.; Onder, O.; Akhtar, Z.; Erdem, C.E. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Trans. Affect. Comput.* **2016**, *8*, 300–313. [CrossRef]
37. Dhall, A.; Ramana Murthy, O.; Goecke, R.; Joshi, J.; Gedeon, T. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 423–426.
38. Li, Y.; Tao, J.; Schuller, B.; Shan, S.; Jiang, D.; Jia, J. Mec 2017: Multimodal emotion recognition challenge. In Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 20–22 May 2018; pp. 1–5.
39. Peng, Z.; Li, X.; Zhu, Z.; Unoki, M.; Dang, J.; Akagi, M. Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access* **2020**, *8*, 16560–16572. [CrossRef]

40. Busso, C.; Parthasarathy, S.; Burmania, A.; AbdelWahab, M.; Sadoughi, N.; Provost, E.M. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Trans. Affect. Comput.* **2016**, *8*, 67–80. [CrossRef]
41. Zhong, S.; Yu, B.; Zhang, H. Exploration of an Independent Training Framework for Speech Emotion Recognition. *IEEE Access* **2020**, *8*, 222533–222543. [CrossRef]
42. Zhang, S.; Chen, A.; Guo, W.; Cui, Y.; Zhao, X.; Liu, L. Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition. *IEEE Access* **2020**, *8*, 23496–23505. [CrossRef]
43. Shahin, I.; Hindawi, N.; Nassif, A.B.; Alhudhaif, A.; Polat, K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Syst. Appl.* **2022**, *188*, 116080. [CrossRef]
44. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
45. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
46. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
47. Latif, S.; Rana, R.; Younis, S.; Qadir, J.; Epps, J. Transfer learning for improving speech emotion classification accuracy. *arXiv* **2018**, arXiv:1801.06353.
48. Feng, K.; Chaspari, T. A review of generalizable transfer learning in automatic emotion recognition. *Front. Comput. Sci.* **2020**, *2*, 9. [CrossRef]
49. Schuller, B.; Steidl, S.; Batliner, A. The interspeech 2009 emotion challenge. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009.
50. Yao, Q. Multi-Sensory Emotion Recognition with Speech and Facial Expression. Ph.D. Thesis, The University of Southern Mississippi, Hattiesburg, MS, USA, 2016.
51. Ghifary, M.; Kleijn, W.B.; Zhang, M. Domain adaptive neural networks for object recognition. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, 1–5 December 2014; pp. 898–904.
52. Sawada, Y.; Kozuka, K. Transfer learning method using multi-prediction deep Boltzmann machines for a small scale dataset. In Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May 2015; pp. 110–113.
53. Latif, S.; Rana, R.; Younis, S.; Qadir, J.; Epps, J. Cross corpus speech emotion classification-an effective transfer learning technique. *arXiv* **2018**, arXiv:1801.06353v2.
54. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef] [PubMed]
55. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
57. Hogg, R.V.; McKean, J.; Craig, A.T. *Introduction to Mathematical Statistics*; Pearson Education: London, UK, 2005.
58. Gao, Y.; Li, B.; Wang, N.; Zhu, T. Speech emotion recognition using local and global features. In Proceedings of the International Conference on Brain Informatics, Beijing, China, 16–18 November 2017; pp. 3–13.
59. Zhang, B.; Essl, G.; Provost, E.M. Recognizing emotion from singing and speaking using shared models. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 139–145.
60. Hong, Q.; Kwong, S. A genetic classification method for speaker recognition. *Eng. Appl. Artif. Intell.* **2005**, *18*, 13–19. [CrossRef]
61. Kinnunen, T.; Karpov, E.; Franti, P. Real-time speaker identification and verification. *IEEE Trans. Audio Speech Lang. Process.* **2005**, *14*, 277–288. [CrossRef]
62. Hamsa, S.; Iraqi, Y.; Shahin, I.; Werghi, N. An Enhanced Emotion Recognition Algorithm Using Pitch Correlogram, Deep Sparse Matrix Representation and Random Forest Classifier. *IEEE Access* **2021**, *9*, 87995–88010. [CrossRef]
63. Campbell, W.M.; Campbell, J.P.; Reynolds, D.A.; Singer, E.; Torres-Carrasquillo, P.A. Support vector machines for speaker and language recognition. *Comput. Speech Lang.* **2006**, *20*, 210–229. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.