applied sciences

MDPI

Article

Speech Emotion Recognition Using Transfer Learning: Integration of Advanced Speaker Embeddings and Image Recognition Models

Maros Jakubec [1,*], Eva Lieskovska [1], Roman Jarina [2], Michal Spisiak [2] and Peter Kasak [2]

[1] University Science Park UNIZA, University of Žilina, Univerzitna 8215/1, 010 26 Žilina, Slovakia; eva.lieskovska@uniza.sk
[2] Faculty of Electrical Engineering and Information Technology, University of Žilina, Univerzitna 8215/1, 010 26 Žilina, Slovakia; roman.jarina@uniza.sk (R.J.); michal.spisiak@feit.uniza.sk (M.S.); peter.kasak@uniza.sk (P.K.)
* Correspondence: maros.jakubec@uniza.sk

Abstract: Automatic Speech Emotion Recognition (SER) plays a vital role in making human–computer interactions more natural and effective. A significant challenge in SER development is the limited availability of diverse emotional speech datasets, which hinders the application of advanced deep learning models. Transfer learning is a machine learning technique that helps address this issue by utilizing knowledge from pre-trained models to improve performance on a new task in a target domain, even with limited data. This study investigates the use of transfer learning from various pre-trained networks, including speaker embedding models such as d-vector, x-vector, and r-vector, and image classification models like AlexNet, GoogLeNet, SqueezeNet, ResNet-18, and ResNet-50. We also propose enhanced versions of the x-vector and r-vector models incorporating Multi-Head Attention Pooling and Angular Margin Softmax, alongside other architectural improvements. Additionally, reverberation from the Room Impulse Response datasets was added to the speech utterances to diversify and augment the available data. Notably, the enhanced r-vector model achieved classification accuracies of 74.05% Unweighted Accuracy (UA) and 73.68% Weighted Accuracy (WA) on the IEMOCAP dataset, and 80.25% UA and 79.81% WA on the CREMA-D dataset, outperforming the existing state-of-the-art methods. This study shows that using cross-domain transfer learning is beneficial for low-resource emotion recognition. The enhanced models developed in other domains (for non-emotional tasks) can further improve the accuracy of SER.

Keywords: speech emotion recognition; IEMOCAP; CREMA-D; transfer learning; image classification; speaker embeddings; affective computing

check for updates

Citation: Jakubec, M.; Lieskovska, E.; Jarina, R.; Spisiak, M.; Kasak, P. Speech Emotion Recognition Using Transfer Learning: Integration of Advanced Speaker Embeddings and Image Recognition Models. *Appl. Sci.* **2024**, *14*, 9981. https://doi.org/10.3390/app14219981

Academic Editor: Douglas O'Shaughnessy

Received: 22 September 2024
Revised: 27 October 2024
Accepted: 28 October 2024
Published: 31 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In today's increasingly digital world, the pursuit of more intuitive and natural forms of human–computer interaction (HCI) is becoming more important than ever. This pursuit is not merely about making technology easier to use; it is about enabling devices to understand and respond to human emotions—a key goal of affective computing [1]. By incorporating emotional awareness into technology, we can enhance the quality of HCI, making interactions more adaptive and responsive, particularly in unpredictable situations.

Emotions are expressed through various channels, including body language, behaviour, speech, and physiological signals such as heart rate and skin temperature. Among these, speech is the most direct and natural mode of communication, rich with emotional nuances. Speech Emotion Recognition (SER) is less intrusive compared to physiological signal-based methods (e.g., ECG, EEG), which often require specialized sensors or wearable devices. Additionally, unlike facial recognition methods, SER is not affected by changes in lighting conditions, camera angles, or occlusions (e.g., masks or glasses). However, SER

can be negatively impacted by background acoustical noise or the quality of the audio recording. Furthermore, emotion expressed through speech might not always align with the true emotional state of speaker, leading to potential misinterpretations. Despite these challenges, SER has emerged as a critical component in the development of emotion-aware applications, with the potential to significantly enrich human–machine interaction.

While humans are naturally adept at discerning emotions through speech, replicating this capability in machines remains a challenging task. SER seeks to bridge this gap by utilizing speech signal processing and machine learning to decode emotional states. There is a vast range of SER applications, from personal virtual assistants and emergency call centres, where understanding the emotional state of a caller can be crucial [2,3], to educational tools, driver safety systems, therapeutic diagnostics, and immersive gaming experiences [4–6].

Despite the potential of SER, developing systems that can reliably capture subtle nuances in human speech remains a significant challenge. This difficulty arises from the diverse ways people express emotions, and the limited availability of large, diverse datasets needed to train state-of-the-art Deep Neural Network (DNN) models. Previous studies have explored various approaches to facilitate the development of the SER systems and improve their performance. Some have focused on learning from basic sound representations, such as raw waveforms or spectrograms, employing models based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [7–9]. Standard speech features like MFCCs and a range of spectral, temporal, and statistical features (e.g., the OpenSMILE set) have been utilized to improve DNN models [10,11]. A key strength of DNN-based approaches is their ability to learn generic representations that can be transferred across different tasks and domains [12].

A major challenge in developing SER systems is the lack of diverse and sufficiently large emotional speech datasets, which limits the use of advanced DNNs. Transfer learning (TL) helps overcome this issue by using knowledge from pre-trained models in a source domain to improve performance on new tasks in a target (downstream) domain. To address the challenge of cross-domain TL in SER, this study explores the use of models from two diverse source domains, including state-of-the-art voice-based speaker embedding models (d-vector, x-vector, and r-vector), originally developed for speaker recognition tasks [13–15], as well as image classification models (AlexNet, GoogLeNet, SqueezeNet, ResNet-18, and ResNet-50). Nowadays, image classification models, largely driven by deep learning, achieve impressive performance. Specifically, pre-trained architectures like ResNet are trained on large-scale image datasets, enabling them to extract rich and generalizable feature representations. These models are designed to identify patterns, edges, and textures, which can be analogous to the spectral features in audio data when transformed into spectrograms or other visual representations. The speaker embedding models simplify complex vocal features into compact manageable representations, in a way that captures unique characteristics of individual speakers. For instance, the x-vector model, developed by Snyder et al. [13], uses Time-delay Neural Networks (TDNNs) to create consistent representations of speech, while the r-vector, introduced by Zeinali et al. [15], employs a CNN-based ResNet architecture to further enhance the model's compactness and discriminative power. By adapting these models for SER, we aim to transfer the knowledge gained to improve SER performance, particularly in identifying and interpreting emotional nuances in speech [16].

In addition, we examined our own enhanced versions of the conventional x-/r-vectors, that were developed in our previous work [17]. These enhanced embeddings incorporate architectural improvements such as Multi-Head Attention Pooling and Angular Margin Softmax, which are expected to better exploit the potential of the models for SER. In this study, emotion-rich datasets such as IEMOCAP and CREMA-D were used for training and evaluation of the downstream SER task. To address the challenge of data scarcity, a data augmentation technique employing reverberation was applied. The availability of diverse training data plays a critical role in the development of robust SER systems.

The utilization of TL from image or speaker recognition models, along with the proposed enhancements and augmentation, is expected to lead to a more comprehensive modelling of emotional states in speech, promising improvements in the performance and robustness of SER systems.

The main contributions of this study, compared to other recently published TL approaches, are as follows:

- We applied TL from both speaker recognition and image classification to the SER task, utilizing d-vector, x-vector, and r-vector speaker embeddings, as well as image-based models like AlexNet, GoogLeNet, ResNet-18, and ResNet-50.
- We demonstrated that speaker models pre-trained on broad and diverse speech datasets are also effective for accurately modelling speech emotions.
- We refined the architectures of the x-vector and r-vector architectures with innovations such as Multi-Head Attention Pooling and Angular Margin Softmax to improve their ability to distinguish emotional states in speech. Additionally, incorporating data augmentation, particularly the addition of reverberation, could further enhance model generalization and reduce overfitting.
- We showed that the proper design and selection of TL models significantly impact the efficiency and accuracy of SER systems.

The structure of this paper is organized as follows: Section 2 presents a comprehensive review of related work in SER research. Section 3 describes the proposed methodology, detailing the architectures of the models utilized. Section 4 focuses on the databases and data augmentation techniques employed in the study, providing an explanation of the experimental setup, including the configuration of both image and speaker embedding models, as well as the evaluation metrics. Section 5 reports the experimental results and comparative analyses. Finally, Section 6 concludes the study, highlighting key findings and outlining potential directions for future research.

2. Emotion Recognition and Related Works

2.1. Understanding Emotions

The concept of emotions is multifaceted, with different perspectives offering unique definitions that highlight its complexity. Whether examined through the lens of psychology, sociology, or neurobiology, emotions are recognized as complex reaction patterns that encompass experiential, behavioural, and physiological elements. In the context of automatic emotion recognition, understanding how emotions are represented is crucial. They can be characterized using either a discrete or a spatial model. The discrete model adopts a categorical approach, breaking emotions down into fundamental categories, including sadness, joy, fear, anger, disgust, and surprise. The spatial model divides emotions into distinct spaces or dimensions, such as activation (arousal), valence, and dominance. Unlike the spatial model, which involves dimensions, the discrete model straightforwardly categorizes emotions, making it more accessible and intuitive for practical applications [18].

In general, the process of SER can be divided into two key phases: feature extraction and classification, as shown in Figure 1. In the feature extraction phase, the speech signal is converted into numerical values (feature vectors) that should ideally be concise, yet capture crucial information from the signal. In the back-end phase, a suitable classifier is chosen based on the specific task at hand. The next chapter outlines these phases and provides description of related works in the field.

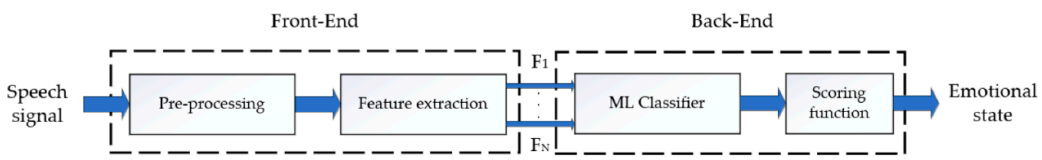


Figure 1. Block scheme of general SER system.

2.2. Related Works in SER

Speech is crucial for understanding human emotions, as it provides significant clues about emotional states. In addition to word choice and linguistic content, there are non-verbal components, such as vocal tone, pitch, and pace, that help listeners interpret the feelings and intentions behind spoken words. Prosodic features, such as intonation and tone, can be analyzed based on energy and pitch (F0), offering valuable insights. In the early stages of SER, the focus was on selecting speech characteristics or features that might reveal more about the emotional level of each utterance. Formants, pitch, duration, and energy contours, as well as statistical characteristics, are commonly accepted descriptors of emotional speech.

The next step, following the feature extraction, entails applying pattern recognition techniques or machine learning algorithms to classify the diverse array of emotional expressions embedded in speech. In the study of Dellaert et al. [19], the combination of k-Nearest Neighbours with the selection of the most prominent features outperformed other parametric and non-parametric approaches. In [20], decision trees effectively handled a variety of non-essential aspects of data. The Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Support Vector Machines (SVMs) are also frequently used methods. The GMM utilizes Gaussian distributions to characterize clusters of points for distinct categories. The Ergodic HMM is suitable for the acoustic modelling, capturing states like phonemes, pauses, and variations in F0 during speech [21]. The SVM allows for the classification of non-linearly separable data and is often used for regression tasks [22–24]. While these traditional methods are interpretable and require less data for training, they depend on handcrafted features and have a limited capacity to learn intricate hierarchical representations automatically.

Over the years, the utilization of neural networks for emotion recognition has seen continuous improvement. Multilayer Perceptron (MLP), a type of neural network consisting of multiple layers of interconnected nodes or neurons, has been employed in [25]. However, the shallow nature of MLPs lacks the capacity to effectively capture complex hierarchical patterns and handle sequential dependencies present in speech data. The emergence of deep learning signalled a notable change in the field of SER. CNNs, RNNs, and self-attention architectures, such as Transformer-based models, have demonstrated high performance in capturing intricate patterns and dependencies in speech data.

In their study, Lee and Tashev [26] employed a Bidirectional Long Short-Term Memory (BLSTM) to generate high-level feature representations of audio signals, which were represented by 32 acoustic features. Their model demonstrated an accuracy of 63.89% on four categorical emotions from the IEMOCAP dataset. Niu et al. [27] and Satt et al. [28] both used two-dimensional spectrogram features as input for a deep learning network. Niu et al. [27] proposed the application of a Deep Retinal CNN approach, which consists of two phases. The first phase involves a data augmentation, where retina and convex lens imaging principles are applied to spectrogram features. The second phase uses a Deep CNN to extract high-level representations from the spectrograms. This method achieved an overall accuracy of 48.8% across six categorical emotions from the IEMOCAP dataset. Satt et al. [28] first trained a CNN on the extracted spectrogram data, experimenting with various network topologies. The subsequent addition of temporal modelling using the BLSTM layer further improved overall accuracy, achieving 68.8%, compared to 66.1% with CNN layers alone, when classifying four emotions from the IEMOCAP dataset.

Integrating SER into multi-modal systems enhances the assessment of emotional states in complex environments, such as industrial, aviation, or healthcare scenarios. Combining SER with other modalities like facial expressions and physiological signals improves accuracy and provides a more comprehensive understanding, which is crucial for safety and productivity [29]. However, challenges such as background noise in real-world environments can affect SER performance. Additionally, integrating multiple modalities increases computational complexity and may impact real-time processing capabilities [30].

One well-known characteristic of DNNs is their requirement for a substantial amount of data to achieve satisfactory performance. In several studies [31–34], TL approaches using

pre-trained models have been applied to tasks such as speech recognition, acoustic event detection, music analysis, and emotion recognition, aiming to address the challenge posed by limited datasets.

Numerous studies have investigated the suitability of a TL approach from the image recognition domain for the SER task. Cummins et al. [35] employed deep spectrum features derived from the image classification of spectrograms using a pre-trained DCNN model called AlexNet [36]. The performance of deep spectrum features was compared to various acoustic feature representation sets specifically designed for emotion recognition and paralinguistic tasks, including eGeMAPS [37], COMPARE [38], and bag-of-audio word (BoAW) [39]. The results indicate that deep spectrum features either matched or outperformed traditional speech feature representations on the FAU-AIBO Emotion Corpus. Zhang et al. [40] proposed employing three channels of log Mel-spectrograms (static, delta, and delta–delta coefficients) as input for DCNN. The DCNN extracts segment-level features, which are then pooled into an utterance-level representation using the Discriminant Temporal Pyramid Matching strategy. The obtained results suggest that the pre-trained and fine-tuned AlexNet outperformed the simple DCNN, which was trained from scratch. Zhang et al. [41] combined DCNN pre-trained on the ImageNet dataset and BLSTM with an attention mechanism. The 3D Mel spectrograms were used as input to the DCNN model, where deep features were extracted at the utterance level and then fed into BLSTM for additional temporal modelling. The proposed pre-training improved accuracy by approximately 6% and 4% for the EmoDB and IEMOCAP datasets, respectively.

The utilization of TL from the speaker recognition domain aims to leverage the similarities between speaker recognition and emotion recognition. This approach could work both ways: integrating emotion and speaker embedding information enhanced the performance of speaker identification in [42]. Xi et al. [43] introduced a residual adapter architecture, in which the base trained ResNet model underwent pre-training for a speaker verification task using an extensive speaker database, VoxCeleb2. Modules containing domain-agnostic parameters (trained on a speaker database) were utilized alongside adapter modules containing domain-specific parameters for adaptation to the target emotion dataset. The proposed residual adapter outperformed the ResNet20 trained from scratch on emotional data. The fine-tuning of the ResNet20 approach did not yield clear results across various datasets, also resulting in a decrease in performance. This decline could be attributed to the potential forgetting of learned speaker information or the smaller size of the dataset used. In the study by Pappagari et al. [16], it was demonstrated that features of emotional speech extracted from pre-trained models on speaker recognition, such as the ResNet-based x-vector, outperformed randomly initialized ResNet models. Furthermore, fine-tuning for emotion classification led to a significant improvement in accuracy for datasets such as IEMOCAP, MSP-Podcast, and CREMA-D. Padi et al. [44] employed a ResNet model pre-trained on extensive datasets labelled with speaker information. Input representations in the form of spectrograms were utilized and augmented to mitigate overfitting. The proposed model outperformed SER systems employing handcrafted feature sets and demonstrated competitive results with other state-of-the-art systems.

The advantage of TL lies in its ability to improve generalization and the robustness of the model [45]. Despite these successes, fine-tuning approaches still exhibit some limitations [46]. The selection of a suitable source model is not always a trivial task, and prior knowledge of task similarity may be necessary. Transferring learned parameters between inconsistent architectures can pose challenges [47]. Furthermore, it is important to note that fine-tuning may not always be the most effective solution for every task, as its efficacy can depend on the specific characteristics of the target task or dataset.

3. Methodology

In this study, we implemented a TL strategy by adapting well-known image recognition models (AlexNet, GoogLeNet, SqueezeNet, ResNet-18, and ResNet-50) and speaker embedding models (d-vector, x-vector, and r-vector) for SER. Building on our previous

work [17], we further refined the architecture of the x-/r-vector models to enhance their emotion recognition capabilities. The original speaker embedding systems serve as baselines to evaluate the impact of proposed improvements on capturing and classifying emotional nuances in speech.

3.1. Transfer Learning

TL is a method that enhances the learning process for a new task by leveraging knowledge from a related, previously learned task. This is similar to how humans learn, where prior knowledge accelerates the acquisition of new skills. For example, when learning a new language, individuals use their understanding of the alphabet and grammar rules, which eliminates the need to start from scratch.

In our study, we employ TL to connect the domains of image recognition (computer vision) and speaker recognition with the downstream domain of SER. By repurposing pre-trained models originally developed for image and speaker recognition, we aim to adapt the rich features these models have learned to the task of SER. The fundamental difference between traditional machine learning and TL is shown in Figure 2.

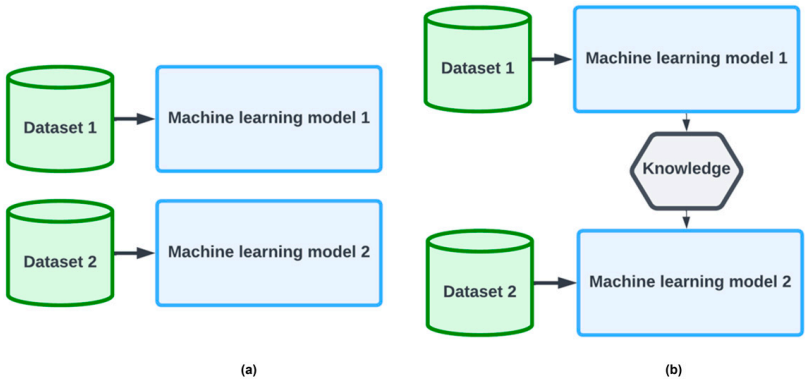


Figure 2. Comparison of traditional machine learning (**a**) and TL (**b**).

TL offers several key advantages in SER research. Primarily, it reduces the need for extensive labelled emotional speech datasets, which are often difficult to obtain. By utilizing pre-trained models, the training time and computational resources are significantly reduced, as models are not trained from scratch. Moreover, TL allows for the transfer of domain-specific knowledge, enabling the models to adapt learned features specifically to the SER task. This is achieved by fine-tuning pre-trained models on SER datasets, allowing them to learn the task-specific features.

3.1.1. Models Based on Image Recognition

Image classification is a fundamental task in computer vision, where the objective is to assign a specific label to an input image. CNNs have become the foundation of image recognition due to their unmatched ability to extract detailed features from images. They are composed of convolutional layers, pooling layers, and fully connected layers, arranged in a robust architecture to efficiently process and classify images. The convolutional layer extracts features from the input image by applying filters (kernels) through element-wise multiplication with local receptive fields of the image. These filters move across the image's width and height, generating a feature map that highlights important patterns. Afterward, a pooling layer reduces spatial dimensions of features, and fully connected layers perform the final classification. The architectural layout of CNN is shown in Figure 3.

Training CNN models for image recognition typically requires large datasets, such as ImageNet, which contains millions of labelled images across a wide variety of object classes. The ImageNet dataset [48] comprises 1,281,167 training images, 50,000 validation images, and 100,000 test images, and serves as a benchmark for evaluating the performance of different models.

Over the years, CNN architectures have undergone significant advancements, each introducing unique innovations that have contributed to the field of image recognition:

- AlexNet: Introduced in 2012, AlexNet [36] revolutionized the field by winning the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The architecture consists of 8 layers, including 5 convolutional layers and 3 fully connected layers. A key feature of AlexNet was the use of overlapping pooling layers, which efficiently extracted spatial features (Figure 4a).
- SqueezeNet: Known for its efficient use of parameters, SqueezeNet [49] achieved accuracy comparable to AlexNet while drastically reducing model size. It employs a Fire module, which combines squeeze and expand layers, frequently using 1×1 filters instead of 3×3 filters. SqueezeNet demonstrated that lightweight CNN architectures could be effective without compromising accuracy (Figure 4b).
- GoogLeNet: Also known as Inception, GoogLeNet [50] won the ILSVRC 2014 competition with its 22-layer architecture. The Inception modules within GoogLeNet use parallel convolutional filters to capture spatial patterns at multiple scales, allowing the network to process data efficiently across varying resolutions (Figure 4c).
- Residual Network (ResNet)**:** ResNet addressed the challenge of training deep networks by introducing skip connections, also known as residual blocks [51]. These blocks help maintain gradient flow during training, allowing for the successful training of very deep networks. ResNet-18 and ResNet-50 exemplify the power of these connections, with ResNet-18 using 8 iterations of a 2-layer residual block and ResNet-50 employing 16 iterations of a 3-layer bottleneck block (Figure 4d).

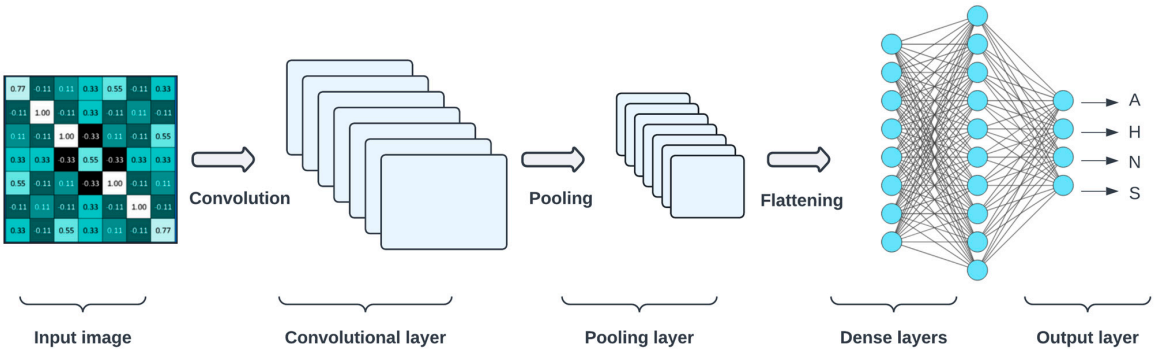


Figure 3. The basic principle of CNN.

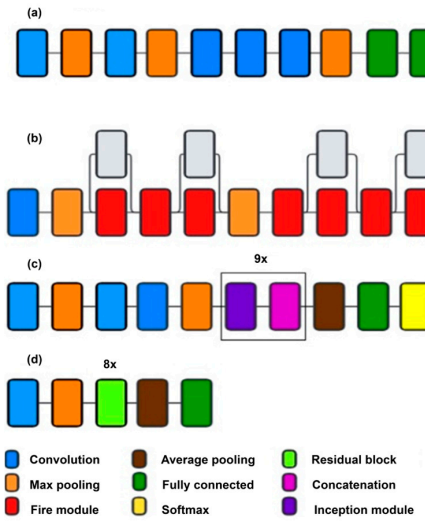


Figure 4. Pre-trained network architectures: (**a**) AlexNet (**b**) SqueezeNet, (**c**) GoogLeNet, and (**d**) ResNet-18. In ResNet-50, each 2-layer residual block of ResNet-18 is replaced by a 3-layer bottleneck block.

3.1.2. Models Based on Speaker Embeddings

Speaker recognition stands as a pivotal aspect of HCI, enabling the discernment of unique vocal signatures that define individuals. In the landscape of modern speaker recognition systems, the advent of DNNs has revolutionized the field, empowering the creation of vector representations of speakers' speech. These representations, commonly known as voiceprints or embeddings, encapsulate intricate information about a speaker's voice within a compact space.

One of the foundational approaches in speaker recognition is the i-vector, which models both speaker and channel variability within a unified low-dimensional space. However, recent advancements have led to the development of more sophisticated deep-learned embedding models, such as d-vector, x-vector, and r-vector, which have significantly improved the accuracy and robustness of speaker recognition systems.

- D-vector: Initially proposed by Variani et al. [52], the d-vector model uses DNNs to map acoustic features at the frame level, producing latent representations of the speech signal at the last hidden layer. Wang et al. [53] further enhanced this approach by incorporating LSTM layers, resulting in a more robust framework for speaker recognition. The d-vector is suitable for simpler tasks with limited computational resources and dataset size, making it ideal for small-scale speaker identification, where capturing temporal dynamics is important. This framework shows promise in emotion recognition, where it is crucial to capture temporal dependencies.
- X-vector: Unlike d-vector, x-vector is a robust solution for large-scale speaker recognition tasks, especially in challenging, noisy environments, though it comes with higher computational costs. Pioneered by Peddinti et al. [54], the x-vector approach uses TDNNs to extract speaker-specific features from speech. Snyder et al. [13] further refined this architecture, producing constant-length x-vectors that encapsulate detailed speaker information. The TDNN layers process speech sequences at the frame level, progressively expanding the temporal context to capture subtle variations in speech, making x-vectors highly effective for speaker recognition.
- R-vector: Drawing inspiration from the success of CNNs in image recognition, the r-vector model was introduced by Zeinali et al. [15]. This model integrates a ResNet-based architecture to create speaker embeddings, which offer a powerful combination of detailed feature representations, robustness to noise, and generalization to unseen speakers. The r-vector excels in modelling temporal dependencies and might be a good option for long-utterance speaker recognition, but its computational demands and complexity make it less practical for real-time applications.

In our study, we focused on enhancing the speaker embedding models, specifically the x-/r-vector, to improve their effectiveness in SER. Building on our previous work [17], we introduced several key modifications to the baseline x-/r-vector architectures as follows to optimize their performance.

For the x-vector model, we utilized a Factorized TDNN (F-TDNN) architecture [55]. This change involved reducing the dimension of feature maps of the original TDNN layers, which made the model more efficient by reducing the number of parameters and computational demands. Additionally, we added shortcut connections in specific layers (3, 5, 7, and 9) to help the model learn more effectively by improving the flow of gradients during training. We also expanded the temporal context in certain layers to capture a broader range of time-based information, which is crucial for detecting complex emotions. Specifically, the temporal context was modified to capture broader temporal information with contexts such as (t − 2, t + 2) in Layer 1, (t − 2, t, t + 2) in Layer 3, (t − 3, t, t + 3) in Layer 5, and (t − 4, t, t + 4) in Layer 7. Instead of using the traditional pooling method, we opted to replace it with Multi-Head Attention (MHA) Pooling. This modification enables the model to attend to multiple segments of the input data simultaneously, allowing it to focus on the most relevant and important parts of the speech signal. We believe MHA Pooling enhances the model's ability to capture nuanced emotional features that might be overlooked by traditional pooling techniques, which generally aggregate information

across the entire input uniformly. This approach enables the model to better handle the variability presented in speech, such as fluctuations in pitch, and prosody that are crucial for accurate emotion recognition. Additionally, we replaced the standard Softmax activation function with Angular Margin Softmax (AM-Softmax). This substitution can play a role in improving the model's ability to differentiate between subtle and similar emotional states. AM-Softmax introduces an angular margin between different class representations in the feature space, which effectively increases the decision boundaries. This can improve class separability, especially in the cases where emotional categories may overlap or present subtle differences in expression (similarly as we proved for speaker's voice modelling in [17]).

For the r-vector model, which is based on ResNet architecture, we also made several key enhancements. First, we replaced the standard ResNet blocks with Res2Net blocks [56]. This modification allows the model to capture features at multiple scales by dividing the feature maps within each block into smaller groups. Additionally, we incorporated a Squeeze-and-Excitation (SE) block, which adjusts the importance of different channels, helping the model focus on the most relevant features for emotion recognition. Similarly to the x-vector model, we replaced traditional pooling with MHA, enabling the model to pay attention to multiple important segments of speech simultaneously. Finally, we integrated AM-Softmax to improve its ability to separate different emotional categories, making it more precise and reliable in classifying emotions.

By building upon the original architectures and integrating advanced mechanisms for processing temporal information and channel-wise features, along with more sophisticated pooling and activation functions, this work aims to contribute to the advancement of SER. The architectural enhancements are summarized in Table 1.

Table 1. Comparison of the speaker-embedding architectures.

	Type Architecture			
Layer	**x-Vector (Baseline)**	**Enhanced x-Vector (Proposed)**	**r-Vector (Baseline)**	**Enhanced r-Vector (Proposed)**
1.	TDNN–ReLU	TDNN–ReLU	Conv2D–1	Conv2D–1
2.	TDNN–ReLU	F-TDNN–ReLU	ResNetBlock–1	SE-Res2NetBlock–1
3.	TDNN–ReLU	F-TDNN–ReLU	ResNetBlock–2	SE-Res2NetBlock–2
4.	Dense–ReLU	F-TDNN–ReLU	ResNetBlock–3	SE-Res2NetBlock–3
5.	Dense–ReLU	F-TDNN–ReLU	ResNetBlock–4	SE-Res2NetBlock–4
6.	Statistic Pooling	F-TDNN–ReLU	Pooling	MHA Pooling
7.	Dense–ReLU	F-TDNN–ReLU	Flatten	Flatten
8.	Dense–ReLU (x-vector)	F-TDNN–ReLU	Dense–ReLU (r-vector)	Dense–ReLU (r-vector)
9.	Dense–Softmax	F-TDNN–ReLU	Softmax	AM–Softmax
10.		Dense–ReLU		
11.		MHA Pooling		
12.		Dense–ReLU (x-vector)		
13.		AM–Softmax		

4. Experiments

This section outlines our approach to designing the SER system using TL with (a) pre-trained models from image recognition domain, and (b) speaker's voice embeddings. The experimental setup is described including the configurations of the networks, the databases, and the application of data augmentation technique.

It is important to mention that the original layers of the pre-trained models were retained, while only the final layers were replaced during the downstream task of TL. The experiments were conducted on workstations equipped with Intel Core i7-9750H CPUs (Intel Corporation, Santa Clara, CA, USA) and NVIDIA GeForce RTX 2070 Max-Q Design GPUs (NVIDIA Corporation, Santa Clara, CA, USA). Image recognition tasks were performed using Matlab 2020b with the Deep Learning Toolbox. Speaker recognition tasks were carried out using Python 3.8, utilizing the PyTorch version 2.3.1 and Librosa library

version 0.10.2 for model construction (including x-/r-vector models), deep learning, and audio processing.

4.1. Database

This section discusses the datasets used for the downstream task of SER in the context of TL, involving both computer vision and speaker recognition. The datasets examined—namely IEMOCAP and CREMA-D—are crucial in advancing research due to their comprehensive coverage and diverse emotional content.

Note, the study was exclusively carried out on English datasets, allowing for consistent benchmarking and comparison with existing state-of-the-art methods. We have not incorporated multilingual and cross-lingual SER issues that present unique challenges and would require further research and adaptation of models to handle diverse linguistic and cultural contexts effectively.

4.1.1. IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [57] is one of the most comprehensive multimodal datasets available for emotion recognition. It comprises approximately 12 h of audio-visual data, including speech, text transcriptions, and facial movement. The dataset features five sessions, with a total of ten professional actors (five males and five females) performing both scripted and improvised emotional scenarios. Each session was annotated by at least three people, providing categorical labels (neutral, happiness, sadness, anger) as well as continuous labels (valence, activation, dominance). The distribution of emotions across categories is imbalanced (Figure 5), which may lower the recognition accuracy for the under-represented samples.

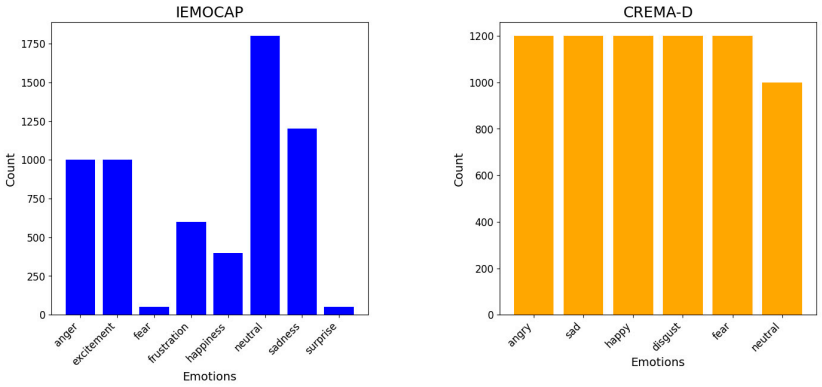


Figure 5. Comparison of emotion distribution between the IEMOCAP and CREMA-D datasets.

In our experiments, data from four sessions were used for training, while the remaining session was reserved for testing. The study focused on recognizing four basic emotion categories: anger, happiness, sadness, and neutral. To address data imbalance, happiness and excitement categories were combined into a single class. The number of samples in each emotional category is shown in Table 2, highlighting the uneven distribution of emotions, which could impact recognition accuracy.

Table 2. Distribution of utterances across different emotion classes in the IEMOCAP dataset.

Class	Number of Samples
Anger	1090
Happiness	1627
Neutral emotion	1704
Sadness	1077

4.1.2. CREMA-D

The crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [58] is a multimodal corpus designed for the study of emotional expression and perception. It includes 7442 audio-visual clips from 91 actors of varying ages, ethnicities, and genders. Each actor recorded 12 sentences in six emotional tones (anger, disgust, fear, happiness, neutral, and sadness) at four intensity levels. CREMA-D is notable for its detailed crowd-sourced evaluations, providing both categorical emotion labels and intensity ratings for each clip. The audio files are of high quality (.wav, mono, 48 kHz, 16-bit), ensuring consistency in data analysis. For our study, we used the standard 80/20 split for training and testing. Unlike IEMOCAP, the CREMA-D dataset offers a more balanced distribution of emotions across categories, as shown in Figure 5, which helps mitigate the issue of data imbalance often encountered in the SER tasks.

4.2. Settings of Image Models

Emotional recordings from the IEMOCAP and CREMA-D datasets were transformed into visual formats, such as scalograms and 3D Mel-spectrograms. These visual representations enabled the use of pre-trained image classification models like AlexNet, GoogLeNet, SqueezeNet, and ResNet for recognizing emotions from speech. This section describes the settings used to train these source image-based models for TL in SER.

Scalograms were generated using continuous wavelet transform filters with Morse and Morlet wavelets, selected for their ability to capture features across varying time and frequency scales [59]. This makes them particularly suitable for analyzing the complex temporal and spectral characteristics of emotional speech. Figure 6 provides examples of scalograms corresponding to different emotional states, demonstrating how these scalograms represent emotions like anger, happiness, neutrality, and sadness.

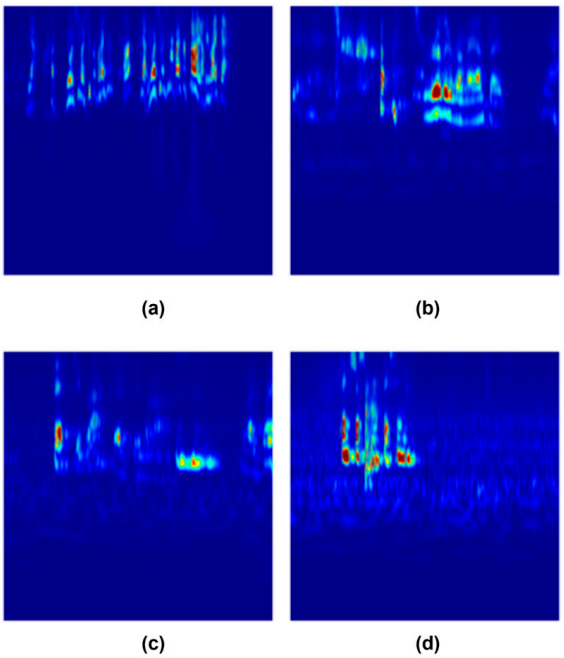


Figure 6. Examples of scalograms for different emotional states: (**a**) anger, (**b**) happiness, (**c**) neutral and (**d**) sadness. Color variations indicate energy distribution, with warmer colors (e.g., red and yellow) representing higher energy and cooler colors (e.g., blue) indicating lower energy.

As an alternative method for parameterizing the speech signal, we also utilized 3D Mel spectrograms, known for their strong performance in the SER tasks. These were generated by calculating static, delta (1), and delta–delta (2) coefficients of the Mel spectrogram. The

delta coefficients were computed along the time axis using a nine-frame section to capture temporal variations. The equations for calculating these coefficients are provided below:

$$m_i^d = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2}, \tag{1}$$

$$m_i^{dd} = \frac{\sum_{n=1}^{N} n\left(m_{i+n}^d - m_{i-n}^d\right)}{2\sum_{n=1}^{N} n^2}. \tag{2}$$

The Mel spectrograms were extracted using FFT window size of 512 samples, with 100-sample overlap. Hamming window was applied to each time frame. The Mel-filter banks covered a frequency range from 0 to 8000 Hz. Figure 7 shows the different representations of an emotional utterance labelled as anger, combined to form the resulting 3D Mel spectrogram.

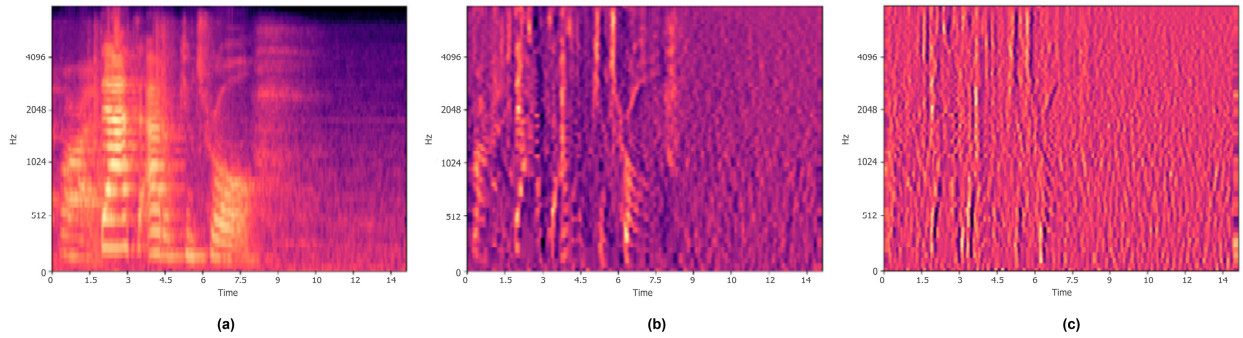


Figure 7. Components of a 3D Mel spectrogram: (**a**) static spectrum, (**b**) delta spectrum, (**c**) delta-delta spectrum. Color variations indicate energy levels, with warmer colors (e.g., yellow and orange) representing higher energy and cooler colors (e.g., purple) representing lower energy.

We tested the models by progressively transitioning from simpler to more complex CNN architectures, all pre-trained on over a million images from the ImageNet dataset. Table 3 summarizes the key characteristics of these pre-trained networks, including their depth, size, number of parameters, input image size, and accuracy metrics on the ImageNet dataset. This information serves as a basis for selecting the most appropriate model, depending on the specific requirements of the SER task.

Table 3. Characteristics and performance metrics of image classification models on the ImageNet.

Network	Depth	Size (MB)	Parameters (Millions)	Input Image Size	Top 1 Accuracy [%]	Top 5 Accuracy [%]
AlexNet [36]	8	227	61.0	227 × 227	63.30	84.60
GoogLeNet [60]	22	27	7.0	224 × 224	68.30	88.70
SqueezeNet [49]	18	4.6	1.24	224 × 224	60.40	82.50
ResNet-18 [51]	18	44	11.7	224 × 224	73.19	90.82
ResNet-50 [51]	50	96	25.6	224 × 224	80.67	95.09

To ensure a fair comparison across these architectures, we standardized the hyperparameters, as detailed in Table 4. Stochastic Gradient Descent (SGD) was selected as the optimization algorithm due to its balance between accuracy and efficiency. Particular attention was given to tuning the initial learning rate, which is critical for adjusting the model's weights during training.

Table 4. Hyperparameters used in the experiments for image classification models.

Hyperparameters	Value
Optimization algorithm	SGD
Momentum	0.9
Initial learning rate	1×10^{-3}
L2 Regularization	1×10^{-4}
Epochs	100
Batch size	32

4.3. Settings of Speaker Embedding Models

Speaker embeddings, which capture the unique vocal characteristics of individuals, can be highly effective for tasks such as SER. This section describes the settings used for TL in SER with both baseline and enhanced speaker embedding models.

In this work, we employed three primary speaker embedding architectures: d-vector, x-vector, and r-vector, along with our enhanced versions of the x-vector and r-vector models proposed in [17]. Following their success on speaker verification, we believe each model offers distinct advantages also in capturing speech features relevant to emotion recognition. Table 5 presents the core attributes of these models, including their depth, size, number of parameters, embedding size, and EER, reflecting their performance. This information serves as a basis for selecting the most appropriate model for the SER task, depending on the specific performance and resource constraints.

Table 5. Characteristics and performance metrics of speaker embedding models on the VoxCeleb1.

Network	Depth	Size (MB)	Parameters (Millions)	Embedding Size	EER [%]
d-vector	5	12.6	7.5	1024	6.44
x-vector (baseline)	9	25.8	8.5	512	3.98
Enhanced x-vector (proposed)	13	37.4	9.0	512	3.53
r-vector (baseline)	9	31.2	8.0	256	3.18
Enhanced r-vector (proposed)	9	36.9	8.5	256	2.71

Each speaker model was first trained on the VoxCeleb1 dataset [61], containing a comprehensive collection of real-world speech recordings from a large and diverse group of speakers. This benchmark dataset includes both training and evaluation subsets, ensuring robust model development. Key details about the dataset are summarized in Table 6.

Table 6. Summary of the VoxCeleb1 dataset.

Dataset	VoxCeleb1
# of speakers (Total)	1251
# of speakers (Training)	1211
# of speakers (Evaluation)	40
# of utterances (Total)	153,516
# of utterances (Training)	148,642
# of utterances (Evaluation)	4874
Average of utterances per speakers	116
Average length of utterances (s)	8.20

The symbol "#" denotes the count in each respective category.

All audio recordings were processed using a standardized pre-processing pipeline to ensure consistent data quality. This included removing the DC component, applying pre-emphasis filtering with a coefficient of 0.97, and using Hamming window on each speech frame. The speech utterances were segmented into 25 ms frames with a 10 ms overlap, and an 80-dimensional vector of filter bank (FB) features was computed for each frame,

capturing essential spectral information needed for accurate speech analysis. Additionally, energy-based voice activity detection was used to filter out silent frames, further enhancing the dataset's robustness. The training process for the speaker embedding models was configured using specific hyperparameters, as summarized in Table 7. These settings were applied for optimizing the performance of both the baseline and enhanced models.

Table 7. Hyperparameters used in the experiments for speaker embedding models.

Hyper-Parameters	Value
Optimization algorithm	SGD
Momentum	0.97
Initial learning rate	1×10^{-3}
L2 Regularization	2×10^{-5}
Epochs	100
Batch size	32

4.4. Data Augmentation

To address the problem of emotional speech data scarcity in SER, a data augmentation technique was employed, focusing on simulating reverberation effects using the Room Impulse Response (RIR) datasets [62]. RIR consist of recordings from various acoustic environments, such as rooms, halls, and auditoriums, each capturing the unique acoustic characteristics of the space.

In our approach, we augmented the original clean audio recordings by convolving them with randomly selected impulse responses from the RIR. This process introduced reverberation effects into the speech signals, effectively simulating the acoustics of different environments. Importantly, this convolution preserved the temporal alignment of the audio, ensuring that emotional expressions in the recordings were not distorted. The resulting augmented dataset provided a diverse range of acoustic conditions, which enriched the training data. This augmentation process not only increases the amount of training data but also introduces variability, helping the models better handle different acoustic environments during inference. Ultimately, this approach can lead to more reliable and accurate emotion recognition from speech.

4.5. Evaluation Criteria

In this study, we employed two common evaluation metrics in emotion recognition: Unweighted Accuracy (UA) and Weighted Accuracy (WA). These metrics provide a comprehensive assessment of model performance, particularly in cases of datasets with imbalanced class distributions, such as IEMOCAP.

UA (3) measures the overall accuracy of the model by calculating the proportion of correctly predicted instances across all classes, giving equal weight to each class regardless of its size. This metric is particularly useful for assessing the model's performance in recognizing less-represented emotions. WA (4), on the other hand, accounts for class imbalance by weighting the accuracy of each class according to the number of samples in that class (s_k). This provides a more balanced view of the model's performance, especially when some classes are more prevalent than others.

$$UA = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{K} \tag{3}$$

$$WA = \frac{\sum_{k=1}^{K} s_k \frac{TP_k}{TP_k + FN_k}}{\sum_{k=1}^{K} s_k} \tag{4}$$

In above formulas, TP_k is the number of true positive predictions for class k, where the model correctly identified the instances belonging to that class. FN_k is the number of

false negative predictions for class k, where the model failed to correctly identify instances belonging to that class, and K represents the total number of classes.

5. Experimental Results

The following sections detail the results obtained from image classification models, speaker embedding models, and our improved speaker embeddings. Additionally, the impact of data augmentation on these models is explored. Through these evaluations, we aim to identify the most effective approaches for improving the accuracy of the downstream SER task across the IEMOCAP and CREMA-D datasets.

5.1. Results on Image Models—TL from Computer Vision Domain

This section examines the effectiveness of the pre-trained image classification models when applied to the task of SER. Specifically, the models considered include GoogLeNet, AlexNet, SqueezeNet, ResNet-18, and ResNet-50, all of which were fine-tuned for SER using scalogram and spectrogram representations of downstream speech data. Table 8 presents the obtained results for the IEMOCAP and CREMA-D datasets, comparing the accuracy of models with and without the use of TL. According to the results, ResNet-50 consistently outperformed the other image recognition models, both with and without TL, making it the top performer across both datasets. Detailed conclusions can be drawn as follows:

- When comparing the different types of input features, scalogram (Morlet wavelet) consistently achieved the highest accuracy across all models and datasets. This feature type, which closely relates to human auditory perception, proved particularly effective in conjunction with ResNet-50, where it contributed to the highest UA and WA scores on both datasets. In contrast, the 3D Mel spectrogram generally resulted in lower accuracy compared to scalogram features.
- By applying TL, ResNet-50 achieved the highest accuracy on IEMOCAP, with a UA of 64.75% and a WA of 63.13%, showing improvements of up to 10% and 9%, respectively, compared to the results without TL. Similarly, on CREMA-D, ResNet-50 achieved a UA of 70.75% and a WA of 71.13%, with both metrics improving by slightly more than 6%. These results highlight ResNet-50's superior ability to adapt to SER when fine-tuned from image classification tasks, making it the most effective model in this comparison.
- ResNet-18 also demonstrated strong performance with TL, particularly on CREMA-D, achieving a UA of 70.18% and a WA of 70.45%, with both metrics improving by approximately 6.5%. While ResNet-18 does not achieve the same level of accuracy as ResNet-50, it offers a more lightweight architecture, making it suitable for applications with lower computational demands.
- It is worth noting that models such as AlexNet, GoogLeNet and SqueezeNet, when trained from scratch on the IEMOCAP dataset, have shown accuracies below 50%. For a multi-class problem with more than two classes, a 50% accuracy might be better than random guessing but still is not satisfactory. In the case of an imbalanced dataset, low accuracy might also indicate that the model is not effectively learning the minority class.
- Despite their low accuracy, GoogLeNet and AlexNet showed the highest improvements after the application of TL. For the IEMOCAP dataset, GoogLeNet achieved an improvement in UA of 9% and a WA of 14.7%, while AlexNet saw improvements of 8.6% in UA and 12.4% in WA. For the CREMA-D dataset, both models showed better accuracy with TL, but they still lagged behind the ResNet models, with GoogLeNet achieving a UA of 65.22% and a WA of 66.21%, and AlexNet reaching a UA of 65.71% and a WA of 67.63%.
- After applying TL, SqueezeNet is on par with AlexNet in terms of accuracy (a UA of 55.82% and a WA of 57.79%) for IEMOCAP, showing a very similar improvement range. For CREMA-D, SqueezeNet outperformed AlexNet, but it did not achieve

the TL-induced improvement as prominently. SqueezeNet's smaller model size and reduced complexity, compared to AlexNet, did not negatively impact the accuracy.

- Experiments on IEMOCAP showed a more prominent improvement in accuracy following the application of TL. Further experiments should be conducted to confirm high efficiency of TL in mitigating the impact of imbalanced datasets.

Table 8. Comparison of SER performance using image classification models with and without TL.

Models	Features	IEMOCAP				CREMA-D			
		Without TL		With TL		Without TL		With TL	
		UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]
GoogLeNet	Scalogram (Morse)	43.81	43.57	54.14	56.57	53.62	54.37	63.74	64.87
	Scalogram (Morlet)	46.15	44.52	55.22	59.21	54.15	56.52	65.22	66.21
	3D Mel spectrogram	43.32	41.64	54.55	54.55	52.32	53.64	63.55	64.54
AlexNet	Scalogram (Morse)	46.45	44.34	52.32	55.06	53.85	54.82	62.32	65.06
	Scalogram (Morlet)	47.08	45.24	55.71	57.63	55.48	57.04	65.71	67.63
	3D Mel spectrogram	45.82	43.65	52.40	53.59	53.37	53.65	62.40	63.59
SqueezeNet	Scalogram (Morse)	46.59	44.48	52.43	55.24	56.59	57.48	62.43	65.24
	Scalogram (Morlet)	47.34	45.36	55.82	57.79	57.34	58.76	65.82	67.79
	3D Mel spectrogram	45.90	43.71	52.61	53.68	55.90	56.71	62.61	63.68
ResNet-18	Scalogram (Morse)	51.08	52.38	61.74	60.87	62.08	62.38	69.74	69.87
	Scalogram (Morlet)	53.64	53.23	62.48	61.13	63.64	63.83	70.18	70.45
	3D Mel spectrogram	52.54	51.34	60.52	59.44	61.54	61.36	68.22	68.44
ResNet-50	Scalogram (Morse)	53.91	53.25	62.69	62.54	63.91	63.25	69.69	70.54
	Scalogram (Morlet)	54.37	54.12	64.75	63.13	64.37	64.52	70.75	71.13
	3D Mel spectrogram	52.41	51.95	61.98	60.46	62.41	62.95	69.38	69.46

5.2. Results on Speaker Models—TL from Speaker Recognition Task

This section examines the impact of pre-trained models for speaker recognition (or more specifically for the verification task), including the d-vector, x-vector, and r-vector systems, along with our enhanced model versions, on the performance of the downstream SER task. Initially, these models were also trained from scratch using emotional datasets only. Improvements in SER accuracy were examined after applying TL from the speaker recognition domain. Based on the results (see Table 9), the following conclusions can be drawn:

- When the baseline models were sorted according to accuracy, d-vector had the lowest accuracy, followed by x-vector, and then r-vector. Parallels can be drawn between the performance of the d-vector on the SER task and its performance in speaker recognition, as d-vector was generally outperformed by more advanced techniques like x-vectors. In contrast, the proposed modification to r-vector outperformed other models, across all settings, making it the top performer.
- TL notably improved the performance of the baseline x-/r-vector models across both datasets. For x-vector (baseline) using IEMOCAP, UA and WA increased by 12.1% and 12.3%, respectively. In contrast, using CREMA-D, the improvements were less pronounced, with a 7.5% increase in UA and a 7.8% increase in WA. For r-vector (baseline) using IEMOCAP, UA and WA increased by 11.8% and 12.1%, respectively. For CREMA-D, both metrics showed an increase of approximately 10%.
- The enhanced versions of the x-/r-vector systems consistently outperformed the baseline models on both datasets, both before and after applying TL, indicating that the proposed modifications were broadly beneficial. Specifically, using TL, the enhanced r-vector achieved 72.89% UA and 73.41% WA on IEMOCAP, and 79.34% UA and 79.72% WA on CREMA-D, suggesting that it is highly effective in capturing the nuances of speech emotion.

- The enhanced versions of the embeddings showed more substantial improvements after the application of TL. The r-vector exhibited the highest accuracy improvement across all models, with an increase of 12.8% in UA and 13.1% in WA on IEMOCAP. Similarly, on CREMA-D, it demonstrated improvements of about 10.5% for both metrics.
- The range of improvement varied between the datasets. TL led to more pronounced improvements on IEMOCAP compared to CREMA-D. The structured nature of CREMA-D, with consistent sentence use across actors (each emotion is expressed using the same 12 sentences) and well-balanced representation of emotions, may explain its higher pre-TL performance. In contrast, the mix of scripted and improvised utterances in IEMOCAP presented a more complex task, emphasizing the need for robust models and advanced techniques to achieve higher accuracy.

Table 9. Comparison of SER performance using speaker embedding models with/without TL [63].

Models	IEMOCAP				CREMA-D			
	Without TL		With TL		Without TL		With TL	
	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]
d-vector	51.36	52.06	63.18	63.92	61.84	62.18	70.25	70.16
x-vector (baseline)	57.37	58.14	69.49	70.44	66.85	67.08	74.36	74.87
Enhanced x-vector (proposed)	58.62	58.95	70.68	71.23	68.32	68.68	78.46	78.94
r-vector (baseline)	58.47	58.91	70.24	71.02	67.71	68.04	77.90	78.06
Enhanced r-vector (proposed)	60.08	60.36	72.89	73.41	68.96	69.13	79.34	79.72

5.3. Results on Augmented Data

This section examines the impact of data augmentation, specifically the addition of reverberation to emotional utterances, on the performance of both image-based and speaker-based models for SER. It focuses on how the introduction of augmented data affects accuracy, especially when combined with TL. Table 10 presents the accuracy results for the selected models: ResNet-18, ResNet-50, and enhanced versions of the x-/r-vector architectures.

Table 10. Impact of data augmentation on emotion recognition performance with and without TL.

Models	IEMOCAP				CREMA-D			
	Without TL		With TL		Without TL		With TL	
	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]
ResNet-18	53.64	53.23	62.48	61.13	63.64	63.83	70.18	70.45
ResNet-18 + augmentation	53.89	54.05	61.74	62.95	64.16	64.37	72.22	72.39
ResNet-50	54.37	54.12	64.75	63.13	64.37	64.52	70.75	71.13
ResNet-50 + augmentation	54.77	54.91	63.86	65.32	65.08	65.74	72.57	72.68
Enhanced x-vector	58.62	58.95	70.68	71.23	68.32	68.68	78.46	78.94
Enhanced x-vector + augmentation	59.24	59.63	71.12	71.84	68.75	68.98	78.86	79.33
Enhanced r-vector	60.08	60.36	72.89	73.41	68.96	69.13	79.34	79.72
Enhanced r-vector + augmentation	61.13	61.57	73.68	74.05	69.34	69.86	79.81	80.25

The accuracy (both UA and WA) of the ResNet image models is lower compared to the speaker-based models. When augmentation was applied to image-based models, the performance showed a slight improvement in most cases. For instance, for ResNet-18 using IEMOCAP, WA increased from 53.23% to 54.05% without TL. The biggest gains were observed when augmentation was combined with TL. For CREMA-D, ResNet-50 with augmentation achieved a significant boost (from 70.18% to 72.22% UA). However, after applying TL, augmentation caused a unique decrease in accuracy for both ResNet models using IEMOCAP, with a drop of 0.74% and 0.89% in UA.

Enhanced x-/r-vector models perform considerably better than the ResNet models. For example, the enhanced r-vector model with augmentation reaches a UA of 68.75% on IEMOCAP without TL, compared to ResNet-50's 65.08%. The introduction of augmentation for speaker-based models yielded an increase in accuracy across all settings and both datasets, with an average increase equal to 0.6%. The enhanced r-vector with augmentation and TL is the best-performing combination, achieving the highest UA (79.81%) and WA (80.25%) across both datasets.

The results suggest that speaker recognition models are more robust in handling the complexities of emotion recognition compared to image recognition models. The findings also highlight the subtle but positive impact of data augmentation on the performance of SER systems. The improvements observed across different models and datasets suggest that expanding the training data can lead to better accuracy and generalization in the SER tasks.

6. Discussion

In this study, we explore various methodologies to improve the performance of SER systems, focusing on the effects of data augmentation and TL using models from two different source domains: image classification (computer vision) and speaker recognition. The comparison of these approaches has provided valuable insights into their respective strengths and weaknesses.

The TL technique, when applied independently, has a significant positive impact on the performance of SER using both pre-trained image classification and speaker recognition models. When evaluating the overall impact of transferring knowledge from image-based models, an average increase of 9.1% in UA and 11.3% in WA was observed across models using the IEMOCAP dataset. For CREMA-D, both metrics showed an average increase of 8.5%. When assessing the impact of TL from speaker-based models, the improvements were higher, with an approximately 12% increase in both metrics for IEMOCAP and a 9% increase for CREMA-D. The results demonstrate that TL is effective for enhancing SER performance, especially when leveraging speaker-based models.

Data augmentation through the application of reverberation had a subtle but positive effect across all models. An exception was observed for the image-based models using IEMOCAP, where accuracy decreased when TL was applied in conjunction with augmentation. This raises the question of whether the imbalanced nature of the IEMOCAP dataset could lead to this instability in SER accuracy.

Figure 8 presents a comparison of model accuracy in the SER tasks using two best-performing models from both the image recognition domain (ResNet-18, ResNet-50) and speaker recognition domain (enhanced versions of x-/r-vector architectures). The performance metric used is UA to better understand how well the model is performing on under-represented classes, ensuring a balanced evaluation. A high UA score indicates that the model is performing well across all classes, not just the most frequent ones. The figure highlights individual increments in accuracy caused by the augmentation technique and TL. Following our previous work [17], the proposed enhancements of speaker-based models were also included.

In Figure 8, the baseline performance (green) represents the starting accuracy of each model without any enhancement, augmentation, or TL. The speaker-based models start with a higher performing baseline compared to image-based models. The enhancement

(blue) refers to the proposed modifications applied to the speaker-based models, which lead to noticeable improvements for the x-/r-vector. Augmentation (red) introduces additional slight improvements for models. TL (grey) shows the largest increase in accuracy, especially for the x-/r-vector models, where TL significantly boosts their performance.

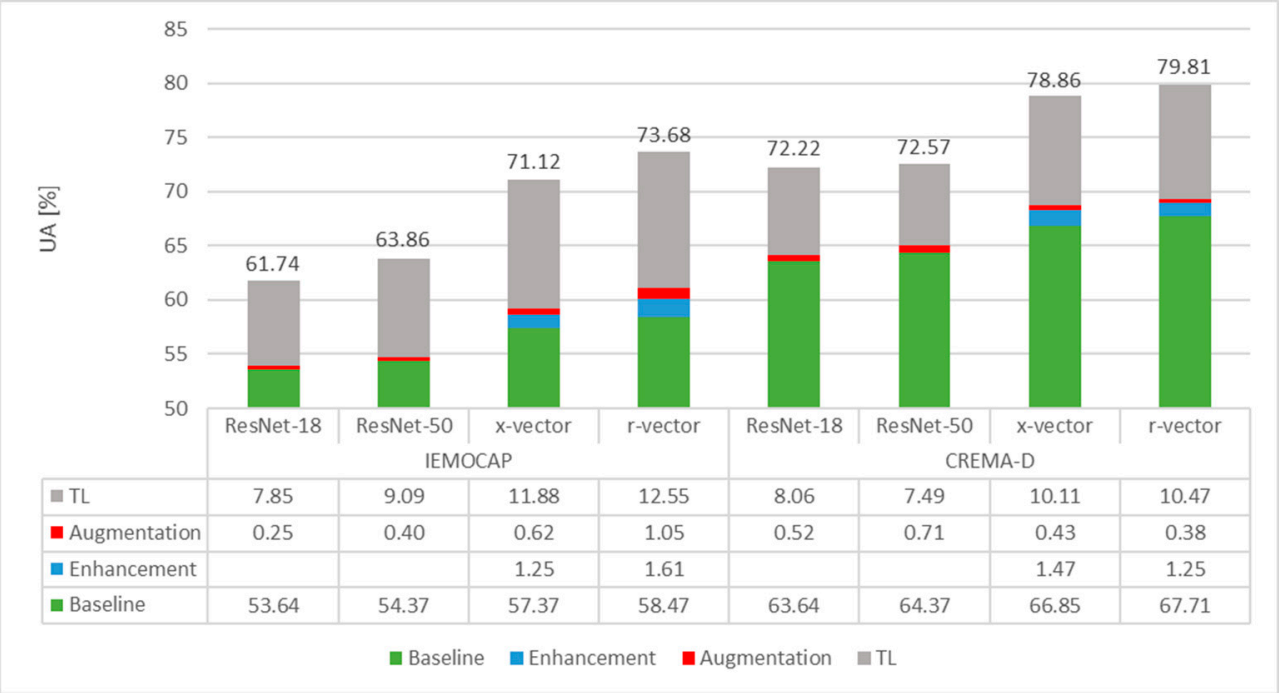


Figure 8. The results of the best models, showing baseline performance, individual increments in accuracy due to TL, augmentation, and the proposed enhancements.

Among the image-based models, ResNet-50, when used with TL and paired with scalogram input features, demonstrated the best accuracy. However, despite their competitive results, the overall performance of image-based models is lower compared to speaker recognition models. This difference is likely due to the domain similarity between speaker recognition and SER, which offers an advantage in capturing and classifying emotional nuances in speech. The combination of all techniques (enhancements, augmentation, and TL) resulted in highest accuracy. On IEMOCAP, the best-performing model is the r-vector with a UA of 73.68%, followed closely by x-vector (71.12%). On the CREMA-D dataset, x-vector achieves a UA of 78.86%, and r-vector reaches 79.81%, indicating higher performance on this dataset compared to IEMOCAP.

The field of SER has seen significant advancements, with numerous systems being evaluated using the IEMOCAP dataset as a standard benchmark. To ensure our results are comparable with existing research, we employed the Leave-One-Session-Out (LOSO) 5-fold cross-validation method, where four sessions were used for training and the remaining session for testing. This is a common approach for evaluating models on the IEMOCAP dataset.

Table 11 compares the obtained results on enhanced r-vector and x-vector models with existing studies. For instance, S. Padi et al. [44] used a pre-trained ResNet34 model with log-Mel spectrograms' input representations, achieving a UA of 64.14% and a WA of 63.61%. The approach by Noh and Jeong [42], which utilized acoustic low-level features and a fine-tuned ResNet, yielded a UA of 66.40% but a lower WA of 61.80%. Y. Xi et al.'s [43] method, incorporating a ResNet20 with a residual adapter, significantly improved performance, reaching a UA of 67.58% and a WA of 72.73% on improvised data. D. Hazarika et al. [64] fine-tuned ResNet34 on spectrograms resulted in a UA of 65.97% and a WA of 65.40%.

Table 11. Comparison of results with a 5-fold cross-validation approach (Source domain tasks: SR—speaker recognition; IC—image classification).

Author	Model	Input Representation (Task)	UA [%]	WA [%]
S. Padi et al. [44]	ResNet34	log-Mel spectrogram (SR)	64.14	63.61
D. Hazarika et al. [64]	ResNet34 fine-tuned	Spectrogram (SR)	65.97	65.40
Noh and Jeong [42]	ResNet fine-tuned	acoustic low-level descriptor (SR)	66.40	61.80
Y. Xi et al. [43]	ResNet20 with residual adapter	Spectrogram (SR)	67.58 *	72.73 *
H. Zhang et al. [41]	3D DCNN-BiLSTM s AM	3D Mel-spectrogram (IC)	68.50 *	-
Ours	Enhanced x-vector—proposed solution	FB features (SR)	69.45	69.24
Ours	Enhanced r-vector—proposed solution	FB features (SR)	71.08	70.97

* Improvised data.

In comparison, our r-vector model outperforms these existing approaches, achieving a UA of 71.08% and a WA of 70.97% on a mixed set of scripted and improvised data from IEMOCAP. These results demonstrate that the proposed enhancements, together with the TL approach, effectively improve model's abilities to classify emotional states from speech. The obtained results highlight the potential of the enhanced r-vector model as a promising method in the field of SER.

Limitations and Future Work

Despite the significant advancements in SER field, several limitations hinder its effectiveness in real-world applications:

Social desirability can influence how emotions are expressed in speech, as individuals may consciously or unconsciously adjust their tone or manner of speaking to align with socially acceptable norms or expectations. This masking of true emotions can reduce the accuracy of SER systems, as they rely on vocal cues to detect underlying emotions. Moreover, certain emotions may produce similar acoustic patterns, which can lead to difficulties in distinguishing between them based solely on speech features. For instance, emotions like frustration and anger might have overlapping vocal characteristics, complicating the task of accurate classification by SER systems.

SER systems that are developed under controlled conditions often struggle to maintain their accuracy in real-world settings. Background noise, overlapping voices, and suboptimal microphone quality are prevalent in practical applications, all of which can degrade system performance. Techniques such as noise reduction and speech enhancement methods could be explored to increase system reliability in diverse environments.

Many SER models are tested on a limited number of datasets, which can restrict their generalizability across various domains or speech environments. Emotional expression through speech varies across different cultures and languages, creating challenges for the development of universally applicable SER systems. What might be a typical vocal expression of anger in one culture could differ from another, which complicates the model's ability to accurately detect emotions across diverse populations.

Future work should aim to evaluate SER models on a broader range of datasets, encompassing different emotional contexts, languages, and cultural settings. Datasets such as RAVDESS, MSP-IMPROV, or MELD could be employed to assess model performance across more diverse populations and environments, helping to improve the generalizability and adaptability of SER systems. Efforts should be made to develop culturally and linguistically adaptive SER systems. This could involve training models on multilingual and multicultural datasets to enhance the system's ability to recognize emotions across diverse populations.

Future improvement strategies could enhance the robustness and generalizability of the SER system. One key challenge that remains unaddressed is the entanglement of speaker characteristics with emotional content, which may result in biased models that focus more on speaker identity than emotion. Developing techniques to disentangle these factors could significantly improve the system's performance [65]. Additionally, incorporating

knowledge distillation—transferring knowledge from larger models like ResNet-50 to smaller models—could improve the system's efficiency and reduce computational demands, making SER models more suitable for real-time applications [66].

7. Conclusions

SER remains a challenging endeavour due to the inherently subjective nature of emotions and the limited availability of comprehensive emotional speech datasets. In this study, we sought to address these challenges by leveraging successful models from different domains, specifically image and speaker recognition, and applying TL to enhance SER performance. The results demonstrate that the TL significantly improves the accuracy of the SER systems, particularly when models are fine-tuned from a domain with closer alignment to speech, such as speaker recognition.

Our experiments reveal that models utilizing TL outperformed those trained from scratch, validating the approach of transferring knowledge from other domains to SER. Among the models tested, speaker-based models, especially the enhanced r-vector model, demonstrate superior performance, achieving a UA of 71.08% and a WA of 70.97% after cross-validation on the IEMOCAP dataset. This performance not only highlights the effectiveness of TL in overcoming the limitations posed by small and imbalanced emotional speech datasets but also aligns with the outcomes of similar research in this area.

Although the augmentation technique is helpful, it plays a smaller role in improving the accuracy of SER compared to TL and other enhancement strategies. Overall, the improvements are more pronounced for the IEMOCAP dataset compared to CREMA-D. The difference is likely due to IEMOCAP's combination of scripted and improvised utterances, which presents a more complex task, whereas the structured nature and balanced emotion representation of CREMA-D led to pre-TL performance already being higher.

The methodologies and solutions presented in this research provide a promising foundation for further advancements in the SER field. The application of TL has proven to be an effective strategy, with significant potential for future innovation. Moreover, integrating multi-task learning (e.g., gender detection) could enhance the model's ability to disentangle emotion from other speech-related features. Additionally, exploring domain shifts, particularly when combining diverse emotional datasets, represents a valuable direction for future research.

Author Contributions: Conceptualization, M.J., E.L. and R.J.; methodology, M.J., E.L. and R.J.; writing—original draft preparation, E.L. and M.J.; writing—review and editing, M.J., R.J. and E.L.; supervision, R.J. and E.L.; data curation, M.S. and P.K.; project administration, E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Ekman, P. *Emotion in the Human Face, Annotated ed.*; Malor Books: Los Altos, CA, USA, 2015; ISBN 978-1-933779-82-9.
2. Chen, M.; Zhang, Y.; Qiu, M.; Guizani, N.; Hao, Y. SPHA: Smart Personal Health Advisor Based on Deep Analytics. *IEEE Commun. Mag.* **2018**, *56*, 164–169. [CrossRef]
3. Doctor, F.; Karyotis, C.; Iqbal, R.; James, A. An Intelligent Framework for Emotion Aware E-Healthcare Support Systems. In Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 6–9 December 2016; pp. 1–8.
4. Tawari, A.; Trivedi, M. Speech Based Emotion Classification Framework for Driver Assistance System. In Proceedings of the 2010 IEEE Intelligent Vehicles Symposium, La Jolla, CA, USA, 21–24 June 2010; pp. 174–178.

5. Ji, Q.; Zhu, Z.; Lan, P. Real-Time Nonintrusive Monitoring and Prediction of Driver Fatigue. *IEEE Trans. Veh. Technol.* **2004**, *53*, 1052–1068. [CrossRef]
6. Ando, A.; Masumura, R.; Kamiyama, H.; Kobashikawa, S.; Aono, Y.; Toda, T. Customer Satisfaction Estimation in Contact Center Calls Based on a Hierarchical Multi-Task Model. *IEEEACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 715–728. [CrossRef]
7. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093.
8. Sarma, M.; Ghahremani, P.; Povey, D.; Goel, N.K.; Sarma, K.K.; Dehak, N. Emotion Identification from Raw Speech Signals Using DNNs. In Proceedings of the Interspeech 2018, ISCA, Hyderabad, India, 2–6 September 2018; pp. 3097–3101.
9. Lieskovska, E.; Jakubec, M.; Jarina, R. Speech Emotion Recognition Overview and Experimental Results. In Proceedings of the 2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA), Kosice, Slovenia, 12–13 November 2020; pp. 388–393.
10. Yu, Y.; Kim, Y.-J. Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database. *Electronics* **2020**, *9*, 713. [CrossRef]
11. Zhang, C.; Xue, L. Autoencoder With Emotion Embedding for Speech Emotion Recognition. *IEEE Access* **2021**, *9*, 51231–51241. [CrossRef]
12. Triantafyllopoulos, A.; Schuller, B.W. The Role of Task and Acoustic Similarity in Audio Transfer Learning: Insights from the Speech Emotion Recognition Case. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, ON, Canada, 6–11 June 2021; pp. 7268–7272.
13. Snyder, D.; Ghahremani, P.; Povey, D.; Garcia-Romero, D.; Carmiel, Y.; Khudanpur, S. Deep Neural Network-Based Speaker Embeddings for End-to-End Speaker Verification. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; pp. 165–170.
14. Li, C.; Ma, X.; Jiang, B.; Li, X.; Zhang, X.; Liu, X.; Cao, Y.; Kannan, A.; Zhu, Z. Deep Speaker: An End-to-End Neural Speaker Embedding System. *arXiv* **2017**, arXiv:1705.02304.
15. Zeinali, H.; Wang, S.; Silnova, A.; Matějka, P.; Plchot, O. BUT System Description to VoxCeleb Speaker Recognition Challenge 2019. *arXiv* **2019**, arXiv:1910.12592.
16. Pappagari, R.; Wang, T.; Villalba, J.; Chen, N.; Dehak, N. X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–8 May 2020; pp. 7169–7173.
17. Jakubec, M.; Jarina, R.; Lieskovska, E.; Kasak, P. Deep Speaker Embeddings for Speaker Verification: Review and Experimental Comparison. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107232. [CrossRef]
18. Shah Fahad, M.; Ranjan, A.; Yadav, J.; Deepak, A. A Survey of Speech Emotion Recognition in Natural Environment. *Digit. Signal Process.* **2021**, *110*, 102951. [CrossRef]
19. Dellaert, F.; Polzin, T.; Waibel, A. Recognizing Emotion in Speech. In Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP '96), Philadelphia, PA, USA, 3–6 October 1996; Volume 3, pp. 1970–1973.
20. Pierre-Yves, O. The Production and Recognition of Emotions in Speech: Features and Algorithms. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 157–183. [CrossRef]
21. Nwe, T.L.; Foo, S.W.; De Silva, L.C. Classification of Stress in Speech Using Linear and Nonlinear Features. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03), Hong Kong, China, 6–10 April 2003; Volume 2, p. II-9.
22. Ke, X.; Zhu, Y.; Wen, L.; Zhang, W. Speech Emotion Recognition Based on SVM and ANN. *Int. J. Mach. Learn. Comput.* **2018**, *8*, 198–202. [CrossRef]
23. Fernandes, V.; Mascarehnas, L.; Mendonca, C.; Johnson, A.; Mishra, R. Speech Emotion Recognition Using Mel Frequency Cepstral Coefficient and SVM Classifier. In Proceedings of the 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 23–24 November 2018; pp. 200–204.
24. Kerkeni, L.; Serrestou, Y.; Raoof, K.; Mbarki, M.; Mahjoub, M.; Cléder, C. Speech Communication Automatic Speech Emotion Recognition Using an Optimal Combination of Features Based on EMD-TKEO. *Speech Commun.* **2019**, *114*, 22–35. [CrossRef]
25. Nicholson, J.; Takahashi, K.; Nakatsu, R. Emotion Recognition in Speech Using Neural Networks. *Neural Comput. Appl.* **2000**, *9*, 290–296. [CrossRef]
26. Lee, J.; Tashev, I. High-Level Feature Representation Using Recurrent Neural Network for Speech Emotion Recognition. In Proceedings of the Interspeech 2015, ISCA, Dresden, Germany, 6–10 September 2015.
27. Niu, Y.; Zou, D.; Niu, Y.; He, Z.; Tan, H. Improvement on Speech Emotion Recognition Based on Deep Convolutional Neural Networks. In Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, Chengdu, China, 12–14 March 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 13–18.
28. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech 2017, ISCA, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
29. Ahmed, N.; Aghbari, Z.A.; Girija, S. A Systematic Survey on Multimodal Emotion Recognition Using Learning Algorithms. *Intell. Syst. Appl.* **2023**, *17*, 200171. [CrossRef]

30. Antonaci, F.G.; Olivetti, E.C.; Marcolin, F.; Castiblanco Jimenez, I.A.; Eynard, B.; Vezzetti, E.; Moos, S. Workplace Well-Being in Industry 5.0: A Worker-Centered Systematic Review. *Sensors* **2024**, *24*, 5473. [CrossRef]
31. Gwardys, G.; Grzywczak, D. Deep Image Features in Music Information Retrieval. *Int. J. Electron. Telecommun.* **2014**, *60*, 321–326. [CrossRef]
32. Amiriparian, S.; Gerczuk, M.; Ottl, S.; Cummins, N.; Freitag, M.; Pugachevskiy, S.; Baird, A.; Schuller, B. Snore Sound Classification Using Image-Based Deep Spectrum Features. In Proceedings of the Interspeech 2017, ISCA, Stockholm, Sweden, 20–24 August 2017; pp. 3512–3516.
33. Baird, A.; Amiriparian, S.; Cummins, N.; Alcorn, A.M.; Batliner, A.; Pugachevskiy, S.; Freitag, M.; Gerczuk, M.; Schuller, B. Automatic Classification of Autistic Child Vocalisations: A Novel Database and Results. In Proceedings of the Interspeech 2017, ISCA, Stockholm, Sweden, 20–24 August 2017; pp. 849–853.
34. Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Tian, Q. Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 3030–3043. [CrossRef]
35. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An Image-Based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 478–484.
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
37. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
38. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 835–838. Available online: https://dl.acm.org/doi/10.1145/2502081.2502224 (accessed on 8 December 2022).
39. Schmitt, M.; Ringeval, F.; Schuller, B. At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In Proceedings of the Interspeech 2016, ISCA, San Francisco, CA, USA, 8–12 September 2016; pp. 495–499.
40. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2018**, *20*, 1576–1590. [CrossRef]
41. Zhang, H.; Gou, R.; Shang, J.; Shen, F.; Wu, Y.; Dai, G. Pre-Trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition. *Front. Physiol.* **2021**, *12*, 177. [CrossRef] [PubMed]
42. Noh, K.; Jeong, H. Emotion-Aware Speaker Identification With Transfer Learning. *IEEE Access* **2023**, *11*, 77292–77306. [CrossRef]
43. Xi, Y.; Li, P.; Song, Y.; Jiang, Y.; Dai, L. Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 513–518.
44. Padi, S.; Sadjadi, S.O.; Manocha, D.; Sriram, R.D. Improved Speech Emotion Recognition Using Transfer Learning and Spectrogram Augmentation. In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), Montréal, QC, Canada, 18–22 October 2021.
45. Feng, K.; Chaspari, T. A Review of Generalizable Transfer Learning in Automatic Emotion Recognition. *Front. Comput. Sci.* **2020**, *2*, 9. [CrossRef]
46. Rusu, A.A.; Rabinowitz, N.C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; Hadsell, R. Progressive Neural Networks. *arXiv* **2016**, arXiv:1606.04671.
47. Gideon, J.; Khorram, S.; Aldeneh, Z.; Dimitriadis, D.; Provost, E.M. Progressive Neural Networks for Transfer Learning in Emotion Recognition. *arXiv* **2017**, arXiv:1706.03256.
48. Recht, B.; Roelofs, R.; Schmidt, L.; Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 5389–5400.
49. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <0.5MB Model Size. *arXiv* **2022**, arXiv:1602.07360.
50. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
51. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.L.; Gonzalez-Dominguez, J. Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4052–4056.
53. Wang, Q.; Downey, C.; Wan, L.; Mansfield, P.A.; Moreno, I.L. Speaker Diarization with LSTM. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5239–5243.
54. Peddinti, V.; Povey, D.; Khudanpur, S. A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts. In Proceedings of the Interspeech 2015, ISCA, Dresden, Germany, 6–10 September 2015; pp. 3214–3218.

55. Povey, D.; Cheng, G.; Wang, Y.; Li, K.; Xu, H.; Yarmohammadi, M.; Khudanpur, S. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In Proceedings of the Interspeech 2018, ISCA, Hyderabad, India, 2–6 September 2018; pp. 3743–3747.
56. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
57. 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]
58. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [CrossRef]
59. Lilly, J.M. Element Analysis: A Wavelet-Based Method for Analysing Time-Localized Events in Noisy Time Series. *Proc. R. Soc. Math. Phys. Eng. Sci.* **2017**, *473*, 20160776. [CrossRef]
60. Iandola, F.N.; Ashraf, K.; Moskewicz, M.W.; Keutzer, K. FireCaffe: Near-Linear Acceleration of Deep Neural Network Training on Compute Clusters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
61. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. In Proceedings of the Interspeech 2017, ISCA, Stockholm, Sweden, 20–24 August 2017; pp. 2616–2620.
62. Szöke, I.; Skácel, M.; Mošner, L.; Paliesek, J.; Černocký, J. Building and Evaluation of a Real Room Impulse Response Dataset. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 863–876. [CrossRef]
63. Jakubec, M.; Jarina, R.; Lieskovská, E.; Kasák, P.; Spišiak, M. Enhancing Speech Emotion Recognition Using Transfer Learning from Speaker Embeddings. In *Text, Speech, and Dialogue*; Nöth, E., Horák, A., Sojka, P., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 184–195.
64. Hazarika, D.; Poria, S.; Zimmermann, R.; Mihalcea, R. Conversational Transfer Learning for Emotion Recognition. *Inf. Fusion* **2021**, *65*, 1–12. [CrossRef]
65. Zhao, X.; Sun, H.; Lei, Y.; Zhu, S.; Xiong, D. CCSRD: Content-Centric Speech Representation Disentanglement Learning for End-to-End Speech Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 5920–5932.
66. Lei, Y.; Xue, Z.; Zhao, X.; Sun, H.; Zhu, S.; Lin, X.; Xiong, D. CKDST: Comprehensively and Effectively Distill Knowledge from Machine Translation to End-to-End Speech Translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 3123–3137.

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.