applied sciences

MDPI

Article

Cross-Corpus Training Strategy for Speech Emotion Recognition Using Self-Supervised Representations

Miguel A. Pastor *, Dayana Ribas, Alfonso Ortega, Antonio Miguel and Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, 50009 Zaragoza, Spain; dribas@unizar.es (D.R.); ortega@unizar.es (A.O.); amiguel@unizar.es (A.M.); lleida@unizar.es (E.L.)
* Correspondence: mapastor@unizar.es

Abstract: Speech Emotion Recognition (SER) plays a crucial role in applications involving human-machine interaction. However, the scarcity of suitable emotional speech datasets presents a major challenge for accurate SER systems. Deep Neural Network (DNN)-based solutions currently in use require substantial labelled data for successful training. Previous studies have proposed strategies to expand the training set in this framework by leveraging available emotion speech corpora. This paper assesses the impact of a cross-corpus training extension for a SER system using self-supervised (SS) representations, namely HuBERT and WavLM. The feasibility of training systems with just a few minutes of in-domain audio is also analyzed. The experimental results demonstrate that augmenting the training set with EmoDB (German), RAVDESS, and CREMA-D (English) datasets leads to improved SER accuracy on the IEMOCAP dataset. By combining a cross-corpus training extension and SS representations, state-of-the-art performance is achieved. These findings suggest that the cross-corpus strategy effectively addresses the scarcity of labelled data and enhances the performance of SER systems.

Keywords: speech emotion recognition; cross-corpus; data augmentation; self-supervised representation

Citation: Pastor, M.A.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Cross-Corpus Training Strategy for Speech Emotion Recognition Using Self-Supervised Representations. *Appl. Sci.* **2023**, *13*, 9062. https://doi.org/10.3390/app13169062

Academic Editor: Douglas O'Shaughnessy

Received: 21 June 2023
Revised: 28 July 2023
Accepted: 1 August 2023
Published: 8 August 2023

1. Introduction

With the advancement of human-machine interaction systems, we have witnessed the integration of related technologies into our daily lives. These technologies have found applications in various industries and households [1], such as automated customer support in call centers and popular virtual home assistants like Alexa and Google Home. However, despite these advancements, these systems still struggle to accurately discern the emotional state of their users, which is a fundamental aspect of human communication. The current SER systems are not yet proficient enough to be implemented in commercial systems [2,3]. The challenge lies in the complexity of accurately identifying the speaker's emotions from a single speech component. Recognizing and interpreting emotions require robust SER systems, which would enhance the fluidity of human-machine communication and unlock new business opportunities.

One of the primary challenges encountered in the development of SER systems is the limited availability of emotional audio data for training, especially when compared to other speech classification tasks. Obtaining authentic emotional audio is a highly complex endeavor due to the intricate nature of emotions. Moreover, researchers face the challenge of subjective labelling once the audio is recorded. As a result, most datasets include artificially simulated emotions and feature a small number of audio samples and speakers. To enhance the overall performance and generalization capabilities of subsequent systems, researchers have employed a technique known as cross-corpus or joint training. This approach involves amalgamating multiple datasets during the training phase. By combining datasets, SER systems aim to address the previously mentioned limitations associated with the scarcity of real emotional audio and subjective labelling. Cross-corpus or joint training

has been utilized in certain SER systems as a means to overcome the challenges posed by the aforementioned dataset limitations [4,5].

This paper explores the cross-corpus strategy as an extension of the training set, building upon the study conducted by Pastor et al. [6]. In the previous work, the authors combined three datasets (EmoDb, RAVDESS, and IEMOCAP) to train a DNN-based SER system, utilizing the HuBERT SS representation. The results demonstrated an improvement in SER accuracy, even in the presence of language mismatch between training and testing datasets. This finding highlighted the potential of the cross-corpus strategy in enhancing the robustness of SER systems. Expanding on these findings, this paper investigates the cross-corpus strategy in greater detail. It explores additional language variability by incorporating a dataset in a different language, and it evaluates the system's performance by gradually increasing the amount of matched data in the training set. Furthermore, another SS representation, WavLM, is assessed, considering its favorable performance in previous studies.

In summary, this paper makes the following contributions:

1. It assesses the effectiveness of the cross-corpus strategy for SER systems.
2. It evaluates the incorporation of multiple languages within the cross-corpus strategy.
3. It investigates the feasibility of training SER systems using predominantly out-domain data with a limited amount of in-domain data.

In the following, Section 2 provides a review of prior research on the cross-corpus strategy for training extension within the context of SER. Subsequently, Section 3 introduces the databases and performance metrics utilized in the experimental setup. Section 4 offers a comprehensive description of the SER system's architecture, while Section 5 outlines the experiments conducted in this study. Finally, Section 6 presents and discusses the obtained results, and Section 7 concludes the paper.

2. Previous Work

The scarcity of emotional audio data for training SER systems has led to the application of the cross-corpus strategy in this field. However, the results obtained with this strategy have been inconclusive, with some studies reporting positive outcomes. For example, Schuller et al. [7] and Zehra et al. [8] report improvements of around 5% when a cross-corpus strategy is applied. On the other hand, Braunschweiler et al. [5] report no difference or a reduction of 1–2% in system accuracy when using a cross-corpus strategy.

Early research focused on determining the most effective approach for applying the cross-corpus strategy. Schuller et al. [7] compared two different methods: voting and pooling. The voting method involved training individual systems for each database and combining the final scores, while the pooling method trained a single system using data from all databases. The results demonstrated that the pooling method consistently yielded significantly better performance than voting. In a different context, Zehra et al. [8] applied the cross-corpus strategy to address data scarcity in non-English languages. They achieved notable improvements in accuracy by combining databases in English, German, Italian, and Urdu.

However, not all studies have observed accuracy improvements when employing the cross-corpus strategy. Braun et al. [5] did not find significant enhancements in their work, which could be attributed to the larger size of the databases used compared to previous studies.

Various approaches have been explored to implement the cross-corpus strategy. Hongchao [9] and Lian [10] utilized Domain Adversarial Neural Networks to learn invariant feature sets across databases, aiming to create a model that is independent of the specific database. Liu [11] proposed a domain-adaptive subspace learning technique to derive a common subspace for all databases.

In the field of SER, most studies have utilized spectral [12] and cepstral [5] parameters, with a few incorporating handcrafted [7] features, such as GeMAPS [13] and eGeMAPS. Additionally, only a limited number of works have explored the use of SS representa-

tions for SER tasks. For example, Pepino et al. [14] compared the performance of the eGeMAPS parameter set [13] with that of Wav2Vec2 [15] and found that the SS representation outperformed eGeMAPS, although their study did not employ the cross-corpus strategy. The results obtained in this study are a recall of 66.3% for the IEMOCAP database. In this work, we utilize two SS representations, HuBERT [16] and WavLM [17], which have demonstrated superior performance to Wav2Vec2 in speech processing tasks.

3. Materials: Experimental Setup

3.1. Databases

The experiments in this study utilize several datasets, namely EmoDb [18], RAVDESS [19], CREMA-D [20], and IEMOCAP [21]. All of these datasets are freely available for research purposes. EmoDb, RAVDESS, and CREMA-D are acted datasets, meaning that the audio recordings involve actors intentionally portraying different emotions. All of them were recorded in a studio, so the audio samples have high quality.

The selection of appropriate datasets is of paramount importance in Speech Emotion Recognition (SER) research to ensure the validity and generalizability of the proposed models. In this study, we carefully curated four specific databases, namely EmoDb, RAVDESS, IEMOCAP, and CREMA-D, as they represent the most extensively utilized datasets in the SER community, providing a rich array of references for result comparison. Notably, EmoDb, RAVDESS, and IEMOCAP exhibit a comparable number of speakers, contributing to a controlled evaluation environment. In contrast, the CREMA-D dataset stands out by encompassing a significantly larger number of speakers compared to the cumulative speaker count of the other three datasets. The inclusion of CREMA-D enables a comprehensive analysis of the impact of heightened variability in the cross-corpus strategy.

Table 1 shows further details about the datasets, where *Time* is the duration in minutes of each database, *Lang.* is the language spoken in each database, and *Spk* is the number of speakers. Then, *#Neutral*, *#Happy*, *#Anger*, and *#Sad* indicate the number of utterances with the corresponding label of each emotion. *Text* indicates how the database was recorded: *Read* means the text has been read when acting an emotion, while *Improv.* indicates the actors improvise what they are saying based on general instructions provided by the creator of the database.

Table 1. Information about speech emotion databases included in the study.

Database	Time (m)	Lang.	Spk	#Neutral	#Happy	#Anger	#Sad	Text
EmoDb	16	Ger.	10	79	71	127	62	Read
RAVDESS	42	Eng.	24	96	192	192	192	Read
CREMA-D	203	Eng.	91	1087	1271	1271	1271	Read
IEMOCAP	420	Eng.	10	1708	1636	1103	1084	Improv.

3.1.1. Speech Emotion Collection

The process of collecting this type of audio typically involves engaging actors and actresses to read a text with specific emotions. Acted datasets offer certain advantages compared to real datasets. Firstly, researchers have complete control over the recording conditions, ensuring optimal audio quality. Additionally, it is easier to create label-balanced datasets as opposed to real datasets where the distribution of labels heavily relies on the recording conditions specific to each dataset. Furthermore, in acted datasets, the same sentences can be recorded with different emotions, allowing for greater flexibility in experimentation without relying solely on textual information.

However, acted datasets also come with some disadvantages. Firstly, the expressed emotions may be less realistic compared to natural conversations where emotions tend to be more subtle and often appear in combination with one another. Moreover, the number of individuals recorded in acted datasets is typically limited, which can pose challenges in terms of generalizability to a broader population.

3.1.2. Special Case: IEMOCAP Database

IEMOCAP was specifically designed to address some of the challenges associated with acted datasets and to provide a more realistic representation of emotions in speech. While it is still an acted dataset, the audio recordings in IEMOCAP were obtained through improvised conversations between two actors, aiming to capture more authentic and natural emotions compared to scripted readings. To ensure objectivity and a wider perspective, the audio data in IEMOCAP was labelled by three different individuals who were independent of the researchers. This reduces biases and provides diverse emotional annotations. IEMOCAP is highly used in previous work due to its authentic nature and rigorous labelling process. It is considered a valuable dataset for studying SER, enabling researchers to develop advanced algorithms and models for emotion analysis and human-computer interaction applications.

3.2. Performance Metrics

The classification performance is evaluated using Unweighted Average Recall (UAR) (Equation (1)). The values in the confusion matrix are utilized to compute the score, including true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). UAR is commonly employed in emotion classification as it considers each class individually, making it suitable for handling the typical imbalance in the number of audio samples for each class. In balanced datasets, UAR results are similar to those obtained with accuracy.

$$UAR = 0.5 \cdot \frac{TP}{TP + FN} + 0.5 \cdot \frac{TN}{FP + TN} \quad (1)$$

4. Methods: Speech Emotion Recognition System

A complete flowchart of the training scheme is depicted in Figure 1.

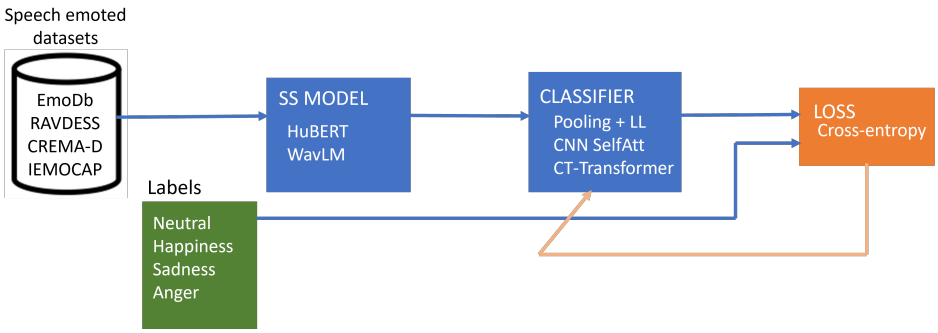

Figure 1. Flowchart of the SER system.

4.1. Feature Extraction

Self-supervised learning is a machine learning approach that involves training a model to extract meaningful representations from unlabelled data by creating surrogate tasks. Unlike supervised learning, which relies on labelled data, self-supervised learning utilizes the inherent structure or patterns in the data to generate labels. This method has the potential to expand the training data available for networks, addressing the challenge of data labelling. In various speech-related tasks like speaker identification or automatic speech recognition, self-supervised speech representation networks have achieved superior results, as reported in previous studies [22].

4.1.1. HuBERT

This paper uses the base HuBERT [16] model, which calculates a 768-dimensional representation of the audio. The model architecture consists of a local encoder followed by a transformer. During training, the contiguous time steps of the local encoder representations are randomly masked. To generate labels for the initial pre-training iteration, a k-means clustering mechanism is applied to 39-dimensional Mel Frequency Cepstral Coefficient

(MFCC) features. In subsequent iterations, the k-means clustering utilizes the latent features from the previous iterations to improve the quality of the targets. To predict cluster labels, a projection layer is added over the transformer blocks.

To pre-train the model, the Librispeech dataset consisting of 960 h of speech is utilized. The pre-training process involves clustering the output of the 6th transformer layer from the first iteration of the HuBERT base model to generate labels. These labels are then used to train the model on the Librispeech dataset, enabling it to learn meaningful representations from the unlabelled data.

4.1.2. WavLM

In this paper, WavLM [17] representation is employed, a more recent speech recognition system that builds upon the HuBERT network. The WavLM architecture follows a similar structure to HuBERT, comprising a local encoder and a transformer.

During the training process, the WavLM model learns to predict hidden segments of speech signals by masking certain parts of the input audio. The architecture incorporates a convolutional representation encoding step. The output of the convolutional encoder is then passed through a transformer encoder, generating hidden states. To create discrete target sequences for training, the k-means algorithm is applied to the training data. Initially, Mel Frequency Cepstral Coefficients (MFCCs) are used for clustering and, in subsequent iterations, the learned latent representations are employed.

The specific base model employed in this paper consists of 12 transformer encoder layers, 768-dimensional hidden states, and 8 attention heads. For pre-training the model, a dataset of 960 speech hours from Librispeech is used. The labels for pre-training are generated by clustering the output of the 6th transformer layer from the first iteration of the HuBERT base model.

4.2. Classification

Three classification models based on neural networks are used for processing the aforementioned features of speech emotions.

4.2.1. Pooling + Linear Layer

The first network employed in the classification task is a pooling operation followed by a linear layer. This approach aims to assess the performance of the raw self-supervised (SS) representation without any additional complex architecture.

4.2.2. CNN Self Attention

Convolutional Networks with self-attention mechanisms have demonstrated excellent performance in various speech- and image-related tasks. This network architecture consists of a convolutional neural network (CNN) followed by a self-attention pooling layer [23].

In this network, a single hidden layer computes a representation C using the following formula:

$$C = Softmax(W_c H^T) H \tag{2}$$

where $W \in \mathbb{R}^{d_m}$ is a trainable parameter and $H = [h_1, h_2, \ldots, h_T]^T \in \mathbb{R}^{T \times d_m}$ represents the output of the previous layer in the network. The resulting representation C can be viewed as a weighted average of the previous sequence of features. This attention-pooling method provides an efficient self-attention mechanism with reduced computational cost.

4.2.3. Class Token Transformer

Finally, a classifier based on a Class Token Transformer (CT-Transformer) is employed. This is a sequence classification architecture inspired by the Vision Transformer [24]. It processes the temporal sequence of embeddings derived from the self-supervised (SS) representations [25]. The CT-Transformer incorporates the concept of a Class Token (CT) through multiple layers of self-attention mechanisms to encode temporal information.

During training, a configurable number of heads and layers in the multihead self-attention block are employed to process the entire sequence of embeddings. In our implementation, we used two layers and six heads. The attention mechanism learns the weights to combine these embeddings at each layer, resulting in a concatenated vector that represents the CT. This CT serves as a global representation of the utterance, with the multiple attention heads acting as individual slots. Compared to the pooling approach used in previous works, the CT-Transformer with multiple attention heads effectively captures underlying information in the sequence.

5. Experiments

5.1. Experiments Description

This study comprises a series of experiments aimed at investigating the influence of dataset selection on the performance of an emotion recognition system. The following subsections describe each experiment and present a table with the organization of train and test sets for carrying them out.

5.1.1. Experiment 1: Matched vs. Cross-Corpus Training with EmoDb, RAVDESS, and IEMOCAP Databases

The first experiment studies the behavior of the SER system using the cross-corpus strategy for training. Initially, a matched train–test dataset approach was employed as a baseline, utilizing the same dataset for both training and testing. The row entitled MATCHED TRAIN in Table 2 corresponds to the training set of this experiment. Subsequently, the cross-corpus strategy was applied by expanding the training set including additional datasets, corresponding to the row entitled EXTENDED TRAIN in Table 2. Both experiments employ the same test set indicated in the row TEST of Table 2.

Table 2. Train and test databases used in experiment 1: cross-corpus strategy for training. Audio and speakers in the train set are not included in the test set.

MATCHED TRAIN	EXTENDED TRAIN	TEST
EmoDb	EmoDb RAVDESS IEMOCAP	EmoDb
RAVDESS	EmoDb RAVDESS IEMOCAP	RAVDESS
IEMOCAP	EmoDb RAVDESS IEMOCAP	IEMOCAP

5.1.2. Experiment 2: Matched vs. Cross-Corpus Training with EmoDb, RAVDESS, IEMOCAP, and CREMA-D Databases

The second experiment further studies the effect of dataset size and diversity on the system's performance. So, following a similar strategy to the previous MATCHED/EXTENDED TRAIN experiments, the training set was expanded also using CREMA-D dataset. This dataset includes many more speakers than the previously used datasets (see Table 1, which allows challenging the cross-corpus training in the face of plenty of variability). Table 3 presents the organization of datasets in this experiment.

Table 3. Train and test databases used in experiment 2: effect of dataset size and diversity in cross-corpus strategy for training. Audio and speakers in the train set are not included in the test set.

MATCHED TRAIN	EXTENDED TRAIN	TEST
IEMOCAP	EmoDb RAVDESS CREMA-D IEMOCAP	IEMOCAP
CREMA-D	EmoDb RAVDESS CREMA-D IEMOCAP	CREMA-D

5.1.3. Validity of Out-Domain Training

The third experiment studies the role of in-domain training data in a cross-corpus framework. In this case, the system baseline was trained and tested on different datasets, namely training on IEMOCAP and evaluating with EmoDb and RAVDESS. Then, this baseline was progressively expanding the training set with audio from the test set: EmoDb and RAVDESS. The objective of this experiment is to assess the impact of including additional matched audio samples in the training data. For a detailed organization of the data in this experiment, please refer to Table 4.

Table 4. Train and test databases used in experiment 3: role of in-domain data in the training set. Train and test sets are always gender-balanced; audio and speakers in the train set are not included in the test set.

STEP	EmoDb		RAVDESS	
	TRAIN	TEST	TRAIN	TEST
0	IEMOCAP	2 speakers of EmoDb	IEMOCAP	4 speakers of RAVDESS
1	IEMOCAP + 3 min EmoDb (2 speakers)	2 speakers of EmoDb	IEMOCAP + 7 min RAVDESS (4 speakers)	4 speakers of RAVDESS
2	IEMOCAP + 6 min EmoDb (4 speakers)	2 speakers of EmoDb	IEMOCAP + 14 min RAVDESS (8 speakers)	4 speakers of RAVDESS
3	IEMOCAP + 9 min EmoDb (8 speakers)	2 speakers of EmoDb	IEMOCAP + 21 min RAVDESS (12 speakers)	4 speakers of RAVDESS
4	IEMOCAP + 12 min EmoDb (10 speakers)	2 speakers of EmoDb	IEMOCAP + 28 min RAVDESS (16 speakers)	4 speakers of RAVDESS
5			IEMOCAP + 35 min RAVDESS (20 speakers)	4 speakers of RAVDESS

5.2. Experimental Framework Description

Experiments were conducted in a five-fold cross-validation scheme, for a multi-class classification framework with four classes consisting of the emotions: neutral, happiness, sadness, and anger. To ensure a robust and reliable evaluation, 10% of the training partitions were randomly set aside as a development set at each epoch. The final results were obtained from the test set of the iteration that yielded the best development outcome. The test set was carefully constructed to ensure a balanced representation of gender. It comprised one-fifth of the speakers, and the distribution of male and female speakers was equal.

For the IEMOCAP dataset, the test set consisted of one recording session involving a spontaneous conversation between a male and a female speaker. This experimental design aimed to provide a comprehensive evaluation of the model's performance across multiple emotion categories while maintaining a fair and equitable representation of speakers and gender in the test set.

6. Results and Discussion

6.1. Experiment 1: Cross-Corpus Strategy for Training

The first experiment examines the viability of cross-corpus strategies for improving the performance of SER systems. The impact of utilizing different combinations of classifiers and feature extractors from multiple datasets was investigated. The experiment aimed to evaluate the effectiveness of both matched training, where only the test dataset was used for training, and extended training, where all available datasets were used for training. The results of the cross-corpus experiments are presented in Table 5. The performance metrics for each combination of datasets, classifiers, and representations were computed and analyzed.

Table 5. Accuracy in terms of UAR for the SER system using SS representations (HuBERT and WavLM) with DNN-based classifiers. Matched train: database for training and test is the same; extended training: database for training is IEMOCAP + RAVDESS + EmoDb. Bold text indicates the best performance for each database.

Classifier	MATCHED TRAIN		EXTENDED TRAIN	
	HuBERT	**WavLM**	**HuBERT**	**WavLM**
Evaluation dataset: EmoDb				
MLP	87.86%	85.64%	84.64%	84.64%
CNNSelfAtt	87.34%	89.70%	87.26%	87.02%
Transformer	90.60%	**90.64%**	81.16%	79.44%
Evaluation dataset: RAVDESS				
MLP	68.34%	70.82%	64.14%	65.70%
CNNSelfAtt	62.44%	67.04%	68.88%	**70.92%**
Transformer	64.78%	66.58%	69.26%	69.14%
Evaluation dataset: IEMOCAP				
MLP	63.52%	63.38%	63.38%	63.21%
CNNSelfAtt	64.94%	65.90%	65.75%	**66.90%**
Transformer	60.82%	62.08%	61.90%	62.98%

The obtained results reveal a discrete improvement of accuracy in terms of UAR score for the RAVDESS and IEMOCAP datasets when cross-corpus training is applied. Despite this being a discrete improvement, the positive tendency aligns with the conclusions reported in the previous works of Schuller et al. [7] about the positive impact of a cross-corpus strategy in emotion recognition. Note that no significant improvement was observed for the EmoDb database. This could be attributed to the limited amount of audio available in the EmoDb dataset.

Nevertheless, the little difference among the results with and without cross-corpus suggests that further research is needed to clarify the advantage of extending the train set with audio out-domain.

Furthermore, these results support the previous findings reported in the work by Zehra et al. [8], which showed that training DNN-based systems with datasets containing multiple languages leads to performance improvements. This suggests that, despite the linguistic differences, augmenting the training data can help the model generalize better across languages.

6.1.1. Analysis of the Performance among Representations and Classifiers

The findings of our experiments seem to indicate that WavLM outperforms HuBERT in terms of results, although the performance difference between them is not big enough to give a definite answer. Throughout most experimental configurations, the performance of WavLM representation either matches or surpasses that of HuBERT. The observed difference between them amounts to approximately 1.5%. These findings are consistent with the results presented by Chen et al. [17], who showed how WavLM outperforms HuBERT in several alternative speech-processing tasks.

However, it is noteworthy that, in the experiments conducted on the EmoDb database, the advantage of WavLM over HuBERT is less pronounced. In fact, in this dataset, most combinations demonstrate superior performance with HuBERT compared to WavLM. This particular outcome may be attributed to the limited size of the EmoDb database, rendering the performance disparities between the two models irrelevant.

Regarding the classifiers, there is not a clear outperformance among them due to their results being very similar. However, analyzing further, the CNNSelfAtt gets better results no matter the representation employed or the kind of training for the largest database (IEMOCAP). On the other side, despite the CT-Transformer being the largest and most complex model, its results are not correspondingly improved, indicating that this model is not using its full potential, which may be related to the scarcity of audio data for fulfilling the complexity of the model to be properly trained.

The following experiments are performed with the WavLM representation and CNN with self-attention to classification.

Analysis of the System Errors among Emotions

The confusion matrix in Table 6 shows that the most common failure is confusing neutrality with happiness and vice versa. This can be due to the lack of contextual information the classifier has. Another common source of confusion is anger and happiness. A possible reason for this is both are high-valence emotions according to Russell's emotional model [26].

Table 6. Confusion matrix for IEMOCAP with the model based on WavLM-CNNSelfAtt on the EXTENDED TRAIN configuration.

		Predicted Value			
		Neutral	**Happiness**	**Anger**	**Sadness**
Real Value	Neutral	976	345	136	251
	Happiness	338	981	167	150
	Anger	142	155	771	35
	Sadness	220	116	9	709

The recall per label for the optimal combination for IEMOCAP is shown in Table 7. The "Neutral" label gets considerably worse results than the average (7% below). On the other side, the accuracy of "anger" is 6% better than the average. This is consistent with the findings of Petrushin [27], who found that anger is easier to detect in speech for humans than other emotions of the spectrum, with "normal state" (Neutral) being the most difficult to identify among the ones used in this work.

Table 7. Recall by class for IEMOCAP with the model based on WavLM-CNNSelfAtt on the EXTENDED TRAIN configuration.

Emotion	**Neutral**	**Happiness**	**Anger**	**Sadness**
Recall	59.37%	65.06%	71.81%	66.72%

6.2. Experiment 2: Effect of Dataset Size and Diversity in Cross-Corpus Strategy for Training

The second experiment studies the impact of increasing the variability of the training set by adding the CREMA-D database. This consists of emotional speech utterances from 91 speakers, which is a considerable increase in population compared to IEMOCAP (10 speakers), EmoDb (10 speakers), and RAVDESS (24 speakers). Then, the methodology followed in the previous experiment is repeated, adding this database to the extended training set.

The results in Table 8 show a deterioration of performance when the cross-corpus strategy is applied, both for IEMOCAP and CREMA-D, indicating that too much variability can be detrimental to the system performance. The creators of the CREMA-D dataset aimed to include speakers with diverse ethnicities and accents, resulting in considerable variability among the samples. This forecast was not taken into account in the other databases analyzed in the previous experiment.

In the case of the CREMA-D database, the best classifier is not CNN with self-attention, as in the previous databases, but the transformer. This is possible because of the bigger diversity of the database, which gives an advantage to the most complex classifier.

Table 8. Accuracy in terms of UAR for the SER system using SS representations (HuBERT and WavLM) with DNN-based classifiers. Matched train: the database for training and test is the same; extended training: the database for training is IEMOCAP + RAVDESS + EmoDb + CREMA-D. Bold text indicates the best performance for each database.

Classifier	MATCHED TRAIN		EXTENDED TRAIN	
	HuBERT	**WavLM**	**HuBERT**	**WavLM**
Evaluation dataset: IEMOCAP				
MLP	63.52%	63.38%	62.56%	63.28%
CNNSelfAtt	64.94%	**65.90%**	64.72%	65.06%
Transformer	60.82%	62.08%	62.56%	62.85%
Evaluation dataset: CREMA-D				
MLP	81.64%	80.85%	80.20%	79.19%
CNNSelfAtt	80.75%	78.51%	79.80%	78.09%
Transformer	**83.04%**	79.57%	72.99%	78.83%

6.3. Experiment 3: Role of In-Domain Data in the Training Set

The third experiment conducted in this study aimed to evaluate the results of training a model using data from a different database and then fine-tuning it with increasing amounts of audio in-domain. The objective was to assess the feasibility of using a trained model as a baseline for a specific task, which can reduce the amount of speech required for training the model and the development time. This can be useful in circumstances in which obtaining labelled data is difficult or expensive. One example of this is a call-center service where the satisfaction of the customer is required to be measured.

In this experiment, a model was initially trained with the IEMOCAP database and subsequently retrained with progressively increasing amounts of EmoDb and RAVDESS audio data. Each step involved adding a specific number of speakers to the database, namely, two in EmoDb and four in RAVDESS maintaining the gender balance. This corresponds to approximately three minutes of audio per iteration for EmoDb and seven minutes for RAVDESS.

Figures 2 and 3 depicted the accuracy in terms of UAR of the sequence of experiments. The obtained results reveal a notable improvement when incorporating in-domain audio into the train set, corresponding to approximately three minutes of speech in each step. This addition leads to a significant enhancement in the model's performance (see the percentage of improvement with respect to the baseline trained with out-of-domain data in Table 9), demonstrating the value of including in-domain data.

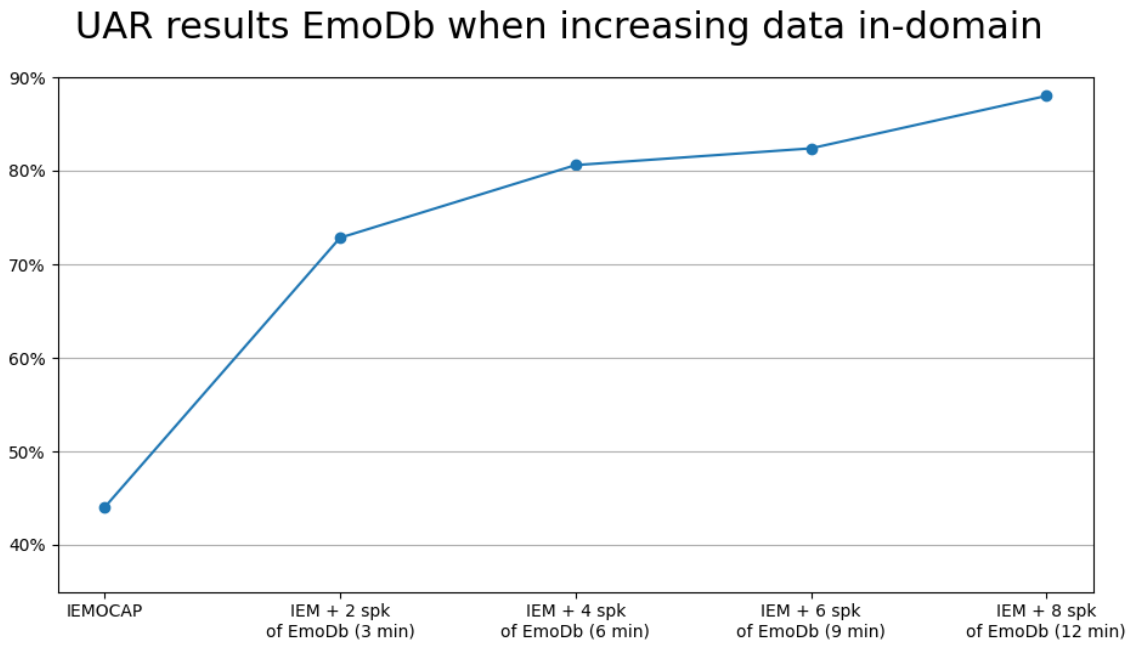

Figure 2. Accuracy in terms of UAR for EmoDb database with a progressive increment of audio in-domain in the training set.

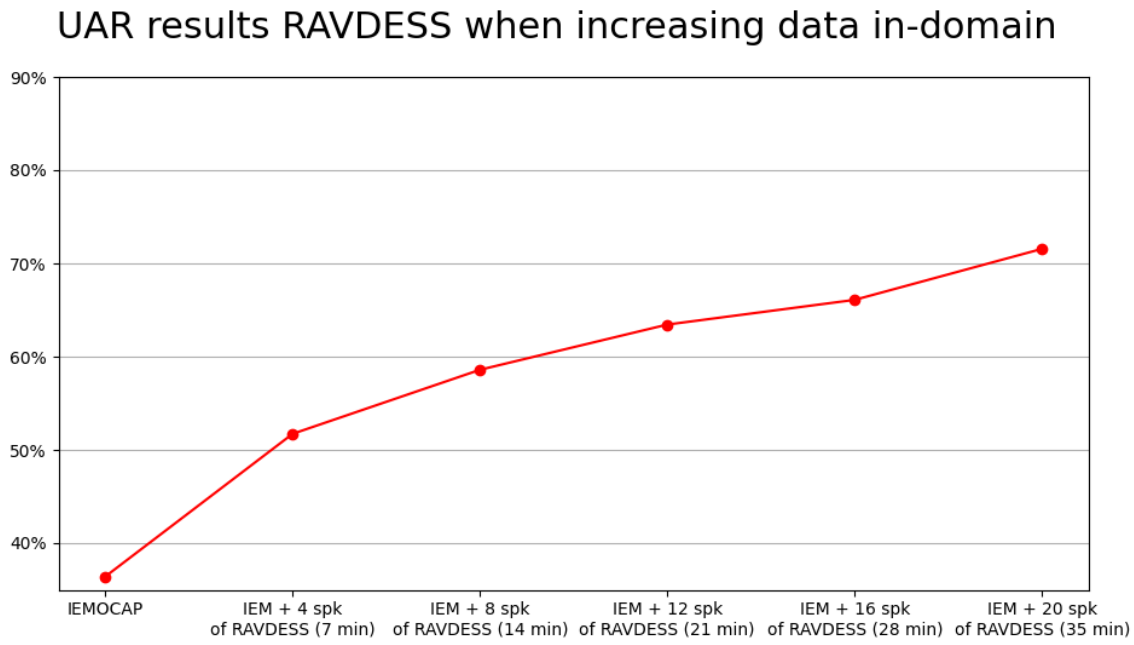

Figure 3. Accuracy in terms of UAR for RAVDESS database with a progressive increment of audio in-domain in the training set.

Table 9. Improvement of the sequence of experiments progressively adding in-domain data over the baseline trained only with out-of-domain data.

	Step 1	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**
EmoDb	Baseline	28.82%	36.59%	38.38%	43.99%	–
RAVDESS	Baseline	15.31%	22.18%	27.03%	29.68%	35.15%

These findings highlight the potential of leveraging out-domain training with a different database as a starting point for specific emotion recognition tasks. The ability to

achieve substantial performance gains with minimal additional in-domain audio data showcases the efficacy of out-domain training in reducing the data requirements and development time of emotion recognition models. This approach offers a promising avenue for leveraging existing resources and leveraging knowledge from related tasks to enhance performance in targeted domains.

Analysis of the System Errors among Emotions

Figure 4 illustrates the two-dimensional projection of HuBERT representation vectors using TSNE [28] in the series of experiments conducted for evaluating the EmoDb and RAVDESS databases. The figure demonstrates the improved clustering of emotions as the amount of in-domain data increases. The relative positions of the emotion clusters are also noteworthy. The neutral emotion is situated between emotions of low valence, such as sadness, and those of high valence, namely anger and happiness.

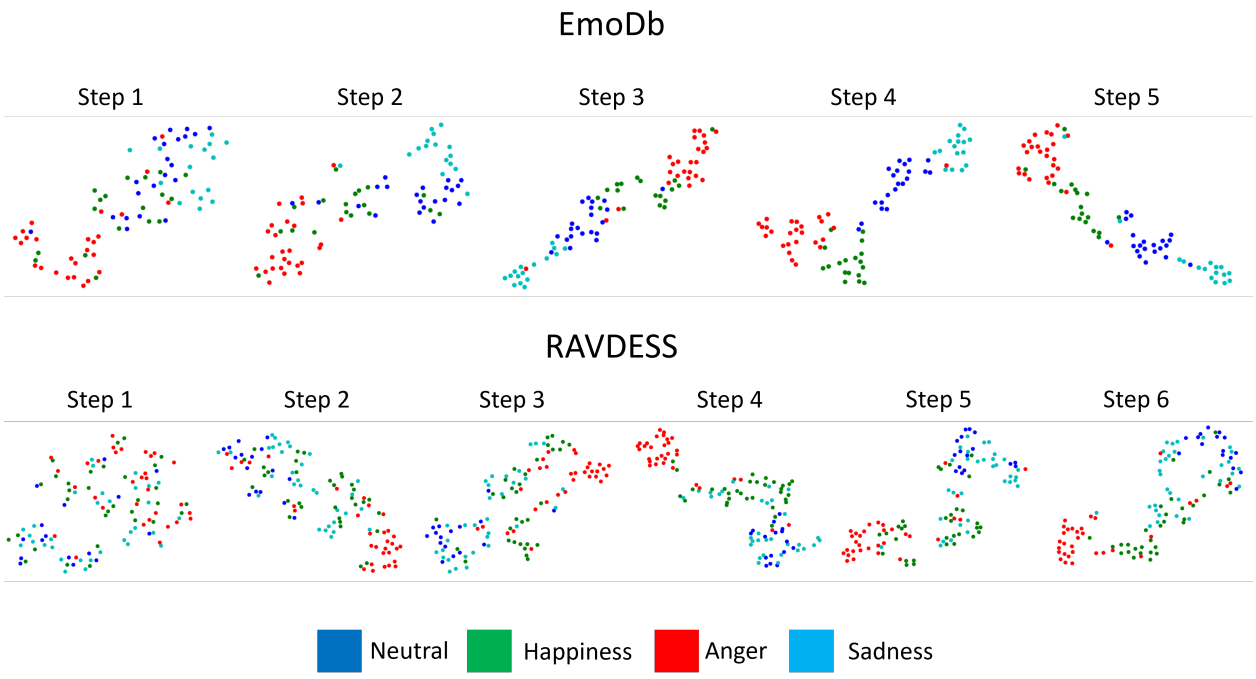

Figure 4. TSNE representations of HuBERT representation vectors in the sequence of experiments evaluating with EmoDb and RAVDESS.

In Tables 10 and 11 the confusion matrices for the systems trained with all the IEMOCAP database, and all audio recordings from EmoDb and RAVDESS, excluding the test set, are shown. For EmoDb, the most common confusion is happiness with anger and vice versa, due to the high valence of both emotions. In the RAVDESS test, the most accurate label is anger, which is consistent with the previously mentioned research of Petrushin [27].

Table 10. Confusion matrix for EmoDb with the model based on WavLM-CNNSelfAtt on the last step with all the IEMOCAP dataset and 15 min of data in-domain.

		Predicted Value			
		Neutral	**Happiness**	**Anger**	**Sadness**
Real Value	Neutral	77	0	0	2
	Happiness	2	51	18	0
	Anger	0	20	107	0
	Sadness	4	0	0	58

Table 11. Confusion matrix for RAVDESS with the model based on WavLM-CNNSelfAtt on the last step with all the IEMOCAP dataset and 35 min of data in-domain.

		Predicted Value			
		Neutral	**Happiness**	**Anger**	**Sadness**
Real Value	Neutral	52	7	1	20
	Happiness	10	102	15	33
	Anger	3	20	124	13
	Sadness	13	24	7	116

7. Conclusions

This paper presents an evaluation of the cross-corpus strategy for data augmentation in DNN-based SER systems. The study reveals that the WavLM representation outperforms the HuBERT representation in the SER task, aligning with other signal-processing tasks' findings. The key advantage of the WavLM representation lies in its additional training approach, where, apart from masking segments of the audio, it also involves denoising segments that remain unmasked.

The results also show that combining databases, even when they encompass different languages, can enhance the system's performance. This finding opens up possibilities for developing SER systems in non-English languages. Further investigation is needed to identify the database properties that significantly impact the system's performance. Additionally, employing broader SS representations may yield improved results in SER systems.

The experiments conducted in this study also explored the impact of progressively incorporating in-domain data into the training set of an initially out-domain trained system. The results revealed that, by adding just a few minutes of in-domain audio, the system achieved performance levels comparable to those of state-of-the-art models in the field.

Author Contributions: M.A.P. designed and performed the set of experiments, analyzed the results, and wrote the manuscript. D.R. helped to analyze the results and to write the manuscript. A.O., A.M. and E.L. helped to revise the manuscript and approved it for publication. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant 101007666; in part by MCIN/AEI/10.13039/5011-00011033 and by the European Union "NextGenerationEU"/PRTR under Grants PDC2021-120846-C41 & PID2021-126061OB-C44, and in part by the Government of Aragon (Grant Group T36_20R).

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

Abbreviations

The following abbreviations are used in this manuscript:

SER	Speech Emotion Recognition
DNN	Deep Neural Network
SS	Self-Supervised
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives
UAR	Unweighted Average Recall
MFCC	Mel Frequency Cepstral Coefficient
CNN	Convolutional Neural Network
CT-Transformer	Class Token Transformer
CE	Cross Entropy

References

1. Gupta, N. *Human-Machine Interaction and IoT Applications for a Smarter World*; Taylor & Francis Group: Milton, UK, 2022.
2. Castellano, G.; Kessous, L.; Caridakis, G. Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*; Peter, C., Beale, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 92–103. [CrossRef]
3. Thakur, A.; Dhull, S. Speech Emotion Recognition: A Review. In *Proceedings of the Advances in Communication and Computational Technology*; Hura, G.S., Singh, A.K., Siong Hoe, L., Eds.; Springer: Singapore, 2021; pp. 815–827.
4. Zong, Y.; Zheng, W.; Zhang, T.; Huang, X. Cross-Corpus Speech Emotion Recognition Based on Domain-Adaptive Least-Squares Regression. *IEEE Signal Process. Lett.* **2016**, *23*, 585–589. [CrossRef]
5. Braunschweiler, N.; Doddipatla, R.; Keizer, S.; Stoyanchev, S. A Study on Cross-Corpus Speech Emotion Recognition and Data Augmentation. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 24–30. [CrossRef]
6. Pastor, M.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Cross-Corpus Speech Emotion Recognition with HuBERT Self-Supervised Representation. In Proceedings of the IberSPEECH 2022, Granada, Spain, 14–16 November 2022; pp. 76–80. [CrossRef]
7. Schuller, B.; Zhang, Z.; Weninger, F.; Rigoll, G. Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? In Proceedings of the Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011; pp. 1553–1556.
8. Zehra, W.; Javed, A.R.; Jalil, Z.; Gadekallu, T.; Kahn, H. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex Intell. Syst.* **2021**, *7*, 1845–1854. [CrossRef]
9. Ma, H.; Zhang, C.; Zhou, X.; Chen, J.; Zhou, Q. Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation. *Appl. Sci.* **2022**, *12*, 5436. [CrossRef]
10. Lian, Z.; Tao, J.; Liu, B.; Huang, J. Domain Adversarial Learning for Emotion Recognition. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019.
11. Liu, N.; Zong, Y.; Zhang, B.; Liu, L.; Chen, J.; Zhao, G.; Zhu, J. Unsupervised Cross-Corpus Speech Emotion Recognition Using Domain-Adaptive Subspace Learning. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5144–5148. [CrossRef]
12. Etienne, C.; Fidanza, G.; Petrovskii, A.; Devillers, L.; Schmauch, B. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. *arXiv* **2018**, arXiv:1802.05630.
13. 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]
14. Pepino, L.; Riera, P.E.; Ferrer, L. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021.
15. Baevski, A.; Zhou, H.; rahman Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv* **2020**, arXiv:abs/2006.11477.
16. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
17. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
18. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the INTERSPEECH, Lisbon, Portugal, 4–8 September 2005; pp. 1517–1520.
19. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak. *Zenodo* **2018**. [CrossRef]
20. 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] [PubMed]
21. 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]
22. Yang, S.W.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. Superb: Speech processing universal performance benchmark. *arXiv* **2021**, arXiv:2105.01051.
23. Safari, P.; India, M.; Hernando, J. Self-attention encoding and pooling for speaker recognition. *arXiv* **2020**, arXiv:2008.01077.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 3–5 June 2019; pp. 4171–4186.
26. Russell, J. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]

27. Petrushin, V. Emotion in Speech: Recognition and Application to Call Centers. In Proceedings of the Artificial Neural Networks in Engineering, St. Louis, MO, USA, 7–10 November 1999.
28. van der Maaten, L.; Hinton, G.E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.