Emotional Deep Learning; A Cross-lingual Speech Emotion Recognition Thesis on English and Dutch

Sietse Schröder

s.schroder@students.uu.nl

7.5 ECTS. BSc. Thesis

M.F. Pessanha (First Supervisor)

Dr. R. van Lambalgen (Second Supervisor)

Bachelor Artificial Intelligence, Utrecht University

July 2, 2021

ABSTRACT

Research on speech emotion recognition encounters the problem that the availability of well-annotated corpora is scarce for many languages. In this research, a cross-lingual deep learning approach is presented, in which a deep learning model is trained on English corpora annotated with seven different emotional labels and tested on a Dutch corpus annotated similarly, originating from a Dutch oral history archive. A one-dimensional multilayered convolutional neural network architecture is used and tested during a mono-lingual speech emotion recognition experiment using the English corpora. Results show that the architecture used is capable of approximating the state-of-the-art performance for mono-lingual speech emotion recognition, retrieving an average accuracy of 0.585 during 5-fold cross-validation. Results on the cross-lingual experiment show that cross-lingual speech emotion recognition is feasible across English and Dutch by retrieving a well above chance accuracy of 0.311 on the Dutch corpus. These results enable future work to further explore speech emotion recognition for Dutch by validating and enlarging the Dutch corpus and to implement techniques reported to significantly improve performance on the cross-lingual speech emotion recognition task from English to Dutch.

Keywords: CNN, Speech Emotion Recognition, cross-corpus.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

1.1 Introduction

The field of advanced human-computer interaction has been receiving increasing amounts of attention as the demand for high-level implementations grows. Speech Emotion Recognition (SER) is characterized as the process of recognizing emotion from audio signals, an important aspect of human-computer interaction [2]. Machine learning for SER was originally approached using traditional classification methods such as Hidden Markov Models and Support Vector Machines [15, 21]. This focus shifted after significant improvements were made using deep learning techniques [2]. Over the last years, several studies have been aimed at exploring deep learning approaches to improve the performance on SER. However, in contrast to more exploited area's of machine learning, such as computer vision [33], little consensus exists among these studies, each proposing fundamentally different deep learning architectures, such as recurrent neural networks [34], text-based convolutional neural networks [31] and long short-term memory networks [2]. The performance of these individual architectures is heavily dependent on multiple factors, such as the availability and complexity of well-annotated data and the number of emotions to be distinguished.

Generally, SER approaches attempt to classify emotional speech that is annotated by multiple human annotators, due to the subjective nature of emotion perception [7]. Speech perceived as emotionally ambiguous across annotators is usually discarded in the process. The number of emotions to be recognized varies among research, ranging from positive/negative in early research [19], to six or more emotions, for which the standard set by Ekman is generally adopted [10, 25], which consists of the emotions anger, happiness, fear, sadness, disgust and surprise. Recent work often complements Ekman's emotions with the label calm or neutral [4, 24].

Current development in SER is heavily thwarted by limitations in the transcriptions and annotations of datasets containing emotional speech, especially for languages other than English. These limitations include the subjectivity of annotators and the general availability of datasets in languages less explored with machine learning techniques. Because emotional speech is a univer-

sally more recognized phenomenon than language, the paralinguistic features of emotional speech might enable machine learning methods to generalize better across languages [20]. Although significantly lower than mono-lingual SER results, research concerning cross-lingual SER methods reports feasible results for several language combinations [12, 20, 27].

Although quite some research has been aimed at cross-lingual machine learning across a wide variety of languages, little attention has been paid to any combination containing the Dutch language, as can be concluded from the cross-lingual machine learning overview provided by Feng and Chaspari [11]. The primary purpose of this study is to investigate whether cross-lingual SER is feasible from English to Dutch, by training on English corpora and testing on a Dutch corpus. Obtaining feasible results on this task will open possibilities for analyzing Dutch oral history archives using machine learning methods, a process that is significantly thwarted by the limited availability of Dutch annotated corpora. Analyzing whether cross-lingual SER is applicable for analyzing Dutch oral history archives is a secondary goal of this study. To summarize, this study will primarily attempt to show that cross-lingual emotion recognition is feasible across English and Dutch emotional speech, without adaptation to Dutch, and secondarily that cross-lingual SER is applicable for the analysis of Dutch oral history archives.

The English emotional speech is extracted from four corpora, built by actors uttering phrases with one of seven basic emotions, which will be described in detail in the section on English Emotional Speech. The Dutch emotional speech will originate from a Dutch oral history archive, the Dutch Veterans Institute's database, consisting of interviews with Dutch veterans about their military experiences [32]. The Dutch emotional speech data will be described in detail in the corresponding section. An attempt will be made to produce feasible results on annotated data from Dutch veteran interviews with a deep learning architecture trained on English SER datasets. Both the primary and secondary purpose of this study will be served in this process, attempting to show the feasibility of SER from English to Dutch as well as the applicability of cross-lingual SER to help analyze oral history archives, such as the Dutch Veterans Institute's database. The sidenote should be made that, as is the case with the majority of English annotated datasets for the SER task, the English data is gained by actors uttering phrases with a certain emotion, whereas the annotated interviews in Dutch represent a more natural situation for data retrieval. Earlier research proved

that a positive information transfer between acted and natural datasets exists [25].

Ideally, This research will contribute to the relatively unexplored field of cross-lingual deep learning for SER between English and Dutch and will therefore be relevant for the fields of deep learning and SER within the field of Artificial Intelligence. More specifically, this research might contribute to the possibilities of deep learning to Dutch speech, which can be of assistance to future attempts to apply advanced machine learning methods to problems concerning Dutch speech. Exploring and annotating Dutch datasets with emotional speech will expand the availability of Dutch SER datasets, therefore being possibly valuable for future research as well. A feasible SER model will be able to highlight emotional sections of the interviews and might therefore be useful for a wide variety of applications, such as an aid to summarize emotional moments in therapeutic conversations or oral history archives.

1.2 Related Work

Cross-lingual machine learning for SER is relatively well-explored within the field of machine learning. Several languages have been the topic of cross-lingual SER research in combination with English. However, due to the limitations in available resources for many languages, research for specific language combinations or combinations of some specific languages with English is sparse. Despite the limited resources, several studies have aimed to explore cross-lingual machine learning for SER on language combinations, for example, Japanese and English [20], French and English [27], and any pair of German, Danish, Spanish, English, Romanian, Turkish and Mandarin [12]. These cross-lingual studies operate on valence and arousal rather than seven basic emotions, but each report positive cross-lingual information transfer leading to above-chance performance. The recent state-of-the-art cross-lingual research aimed at predicting Ekman's seven emotions, training on English corpora, and testing on corpora containing other languages, reports macro F1-scores between 0.39 and 0.46 and accuracies between 0.40 and 0.50 [35]. Related work in other cross-corpus research shows the possibility of information transfer between acted datasets and natural datasets for the SER task [25]. Some effort has been made to produce Dutch corpora for emotion recognition tasks, among which SER [6]. However, these resources appear to have been insufficient to properly be used in a mono-lingual or cross-lingual study towards the SER task on Dutch spoken data, of which very little to no related work exists.

Mono-lingual work on SER is correspondingly well-explored, although even more dependent on the availability of well-annotated corpora. Mono-lingual studies concerning Dutch speech are scarce, in accordance with the availability of Dutch annotated corpora containing emotional speech. However, due to numerous corpora containing emotional speech annotated in English, several studies have aimed to reach state-of-the-art results, presenting the current state-of-the-art performance and allowing comparison with these results. Such results are, among others, 0.887 on the RAVDESS dataset [8], 0.658 on the SAVEE dataset, 0.557 on the TESS dataset, and 0.658 on the CREMA-D dataset [23]. These results are retrieved from an overview of SER methods applied to different datasets [2]. The different datasets mentioned will be further examined in the section on English emotional speech. Besides deep learning, some techniques tend to have a positive effect on performance on the mono-lingual SER task, such as gender differentiation [26] and attention mechanisms [25].

CHAPTER 2. METHODOLOGY

2.1 Model Architecture

For this study, a one-dimensional convolutional neural network (1D CNN) is trained on English spoken emotional speech. The model architecture is largely adopted from Dmity Babko's work on SER [3], and follows the convolutional approach suggested by Neumann and Vu [27]. The model consists of five one-dimensional convolutional layers using the rectifying linear unit (ReLU) activation function, respectively containing 512, 512, 256, 256, and 128 kernels, followed by a flatten-layer and two dense layers. The final layer uses softmax as activation. Max pooling, with a kernel size of five for the first four convolutional layers and a kernel size of three for the last convolutional layer, is applied to simplify the output of each convolutional layer and reduce the required computational time. Batch normalization is used after each layer to standardize the layer's output.

As input to the 1D CNN, the zero-crossing rate (ZCR) and root mean square energy (RMSE) were extracted as features. Additionally, the Mel-frequency cepstral coefficient (MFCC) was extracted to represent the audio signal itself, following the recommendations reported independently by Ganchev et al. and by Dolka et al. on the appliance of MFCC for speech recognition purposes [9, 13]. These features were concatenated time-wise to form the input to the 1D CNN. The features were extracted from audio fragments of a fixed length of 2.5 seconds. The audio fragments were loaded using a sample rate of 22,050 frames per second. The first 0.6 seconds of each audio fragment was ignored, due to silence at the beginning of most fragments. Each audio fragment of 2.5 seconds resulted in a time-wise concatenation of ZCR (108 features), RMSE (108 features), and MFCC (2,160 features). Shorter audio fragments were padded at the beginning to reach a length of 2.5 seconds. Each fragment contained 2,376 features upon entering the model.

2.2 English Emotional Speech

Four corpora of English emotional speech were used to prevent overfitting on a specific data format. The two main criteria used in the selection process were the free availability of the datasets and the annotation with Ekman's six emotions, which are happiness, sadness, fear, disgust, anger, and surprise, complemented with a neutral or calm label. The four selected corpora meeting those

criteria were TESS [28], CREMA-D [5], RAVDESS [22], and SAVEE [16].

The number of utterances each corpus contains that are annotated to be applicable to SER can be referenced in Table 2.1. All utterances are concatenated to create an English corpus containing 12,162 labeled audio fragments.

Dataset	**Number of utterances**
TESS	2,800
SAVEE	480
RAVDESS	1,440
CREMA-D	7,442
TOTAL	12,162

Table 2.1: *Number of utterances for each corpus.*

The Toronto Emotional Speech Set (TESS) contains 2,800 utterances spoken by two actresses, aged 26 and 64. In each recording, one of the actresses speaks the phrase "Say the word", followed by a random sample of a set of 200 target words. Each recording was made portraying one of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).

The Surrey Audio-Visual Expressed Emotion dataset (SAVEE) contains 480 utterances, recorded by four actors in a neutral British accent, portraying seven emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). The actors are postgraduate students and researchers at the University of Surrey at the time of recording.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a gender-balanced multimodal database of emotional speech and song, consisting of 24 professional actors who vocalize lexically matched statements in a neutral North American accent. The recordings are labeled using seven emotions (calm, happy, sad, angry, fearful, surprise, and disgust). Each recording in the resulting set was rated ten times on emotional validity, intensity, and genuineness by 247 participants, resulting in 1,440 valid utterances labeled with emotional speech.

The Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) is constructed using 91 actors with diverse ethnic backgrounds. The seven categorical labels (happy, sad, anger, fear, disgust, surprise, and neutral) used were annotated using crowd-sourcing, produced by 2,443 raters. The dataset contains 7,442 recordings.

To estimate the performance of state-of-the-art techniques on the corpus defined above, the weighted average of the performance of state-of-the-art models on the individual dataset is calculated based on the number of utterances each dataset contributed. The individual performances used are accuracies of 0.887 on RAVDESS, 0.658 on SAVEE, 0.557 on TESS, and 0.658 on CREMA-D [2]. Calculating the weighted average for these results leads to an estimation of state-of-the-art performance on the corpus defined above, resulting in an accuracy of 0.661.

2.3 Dutch Emotional Speech

The corpus created to test the cross-lingual predictive ability of the model consists of audio fragments with a duration of 2.5 seconds. These fragments were obtained from seven interviews with Dutch veterans about their military experiences and missions, made freely available and summarized by the Dutch Veteran's Institute [32]. Only interviews concerning relatively recent missions (1970-present) were selected, preventing the possible corruption of the results due to old fashioned use of language. Additionally, interviews were selected based upon expected emotionality judged from the summary provided. Emotional parts of the interview were annotated by a single annotator, the author of this study. Perceived ambiguousness in the emotional content of a fragment led to exclusion of this fragment from the corpus. Both fragments spoken by the interviewer and interviewee were annotated, due to the primary purpose of this study to attempt to show the possibility of cross-lingual SER, which is likely to benefit from the increasing amount of variance in the test data, resulting from annotating multiple speakers.

Finally, an attempt was made to estimate the emotion of the fragment, based on pitch and voice, preventing linguistic features such as naming the emotion in the fragment from corrupting the corpus (the case might present itself where a veteran explains his disgust using a humorous or happy tone). Analyzing and annotating seven interviews resulted in 164 fragments with a duration of 2.5 seconds and relatively unambiguous emotional content, that were then stored using wav audio format. Since the interviews were mainly directed at war experiences of veterans, containing conversations about traumatic experiences, some emotions (anger and fear) were overrepresented compared to others (surprise and disgust). Anecdotally, fragments containing surprise were almost exclusively spoken by the person conducting the interview, reacting to a surprising statement of

the veteran. Table 2.2 shows an overview of the number of fragments labeled with each emotion for the Dutch veteran's corpus.

Neutral	**Fear**	**Angry**	**Happy**	**Sad**	**Disgust**	**Surprise**
8	34	38	48	26	5	5

Table 2.2: *Number of utterances annotated with each emotion for Dutch corpus.*

2.4 Experimental Setup

A mono-lingual experiment will be conducted to provide a baseline. In this experiment, the English corpus will be used for both training and testing. To optimize hyperparameters, the model will be trained ten times on a random sample of 72% of the English corpus. Further information concerning these runs will be discussed in the hyperparameter section. A small random sample of 8% is used as validation. The model will be tested against the remaining 20% of the corpus. Since there is no predetermined distribution of train and test datasets, cross-validation (CV) will be applied to evaluate the model performing best on the randomly sampled test dataset, to validate results and prevent overfitting on a small portion of the corpus. Due to limited computational resources, the CV will not be used to select the best-performing model. To serve the CV, the corpus will be divided into five segments, resulting in 5-fold cross-validation. Each model will be evaluated using the reported macro F1-score and accuracy. The best-performing model will be the model with the highest F1-score. The F-measure, resulting in an F1-score, is defined as the harmonic mean of the precision (true predicted positives divided by all predicted positives) and recall (true predicted positives divided by all true positives), as shown in Formula 2.1. By calculating the F1-score for the performance on each class, and averaging the result, the macro F1-score can be obtained. The macro F1-score will be the measure of performance used in this study.

$$F = \frac{(2PR)}{(P+R)} [29]. \tag{2.1}$$

Previous research in machine learning methods applied to speech shows the importance of data augmentation, introducing perturbations into the speech signal [18]. In this study, the following three augmentation techniques are applied to each fragment in the English corpora: a pitch shift with a random rate [36], the addition of white noise [14], and a combination of these augmentations. Due to these augmentations, the dataset contains four times more fragments than the corpus

itself, resulting in 9,730 test fragments, 38,918 train fragments, and 3,892 validation fragments, randomly divided each run. To validate whether data augmentation is of importance for the mono-lingual SER task, the best-performing model is tested against the original corpus, without data augmentation, and the results on the test dataset are compared.

A cross-lingual experiment will be conducted to test the ability of the model to transfer information between English and Dutch corpora. During this experiment, the model will be trained several times using the full English corpus, on which the data augmentation described above is applied. The different runs for this experiment will be described in detail in the hyperparameter section. Each model will be tested against the Dutch corpus and evaluated using the reported macro F1-score and accuracy. A pipeline of the cross-lingual experiment is illustrated in Figure 2.1.

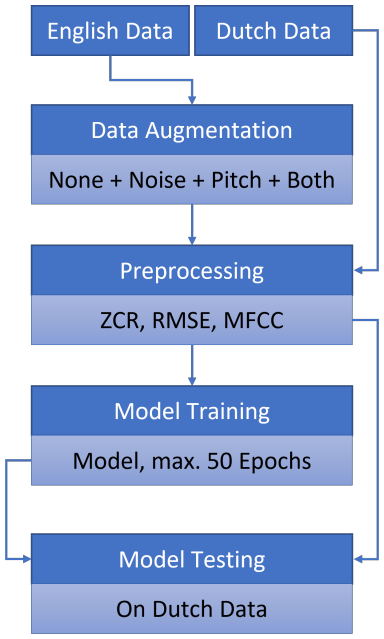

Figure 2.1: *Pipeline of cross-lingual experiment.*

2.5 Hyperparameters

The following static hyperparameters are implemented. Each 1D CNN model is implemented using Keras, the high-level TensorFlow API [1]. The 1D CNN model is trained using stochastic gradient descent for a maximum of 50 epochs. An early stopping metric is implemented monitoring the validation accuracy with a patience value of 5 epochs, meaning the model will stop training if no improvement is measured in the accuracy on the validation data for 5 epochs. The learning rate is initially set to 0.001 (the default value for both optimizers that will be used in the 1D CNN model) and is reduced using the ReduceLROnPlateau callback monitoring the validation accuracy with a patience of 3 epochs and a factor of 2, meaning that if the validation accuracy does not improve for 3 epochs, the learning rate is divided by a factor of 2. Both the early stopping and the ReduceLROnPlateau method are implemented to prevent the phenomenon of overfitting that is possibly occurring when the accuracy retrieved on the validation data during training stagnates or decreases for multiple epochs.

The best-performing optimizer and batch size for the mono-lingual experiment on the English corpus were selected experimentally by performing ten runs. As documented above, each run

consists of training on 72% of the corpus while being validated on 8% and tested on 20%, which are all sampled randomly from the corpus. Each run has a unique combination of the optimizers Adam [17] and RMSProp [30], and the batch sizes 16, 32, 64, 128, and 256. Each run is evaluated using the F1-score, selecting the model with the highest F1-score, after which this model is trained again on unaugmented data to test the positivity of the influence of the augmentation methods used. The best-performing model is trained and evaluated using 5-fold cross-validation to determine the final performance of the model.

The ten parameter combinations described above are applied to the cross-lingual experiment as well, resulting in ten runs. The 1D CNN model is trained on the entire augmented English corpus and tested using the F1-score and accuracy on the Dutch corpus. Since a clear separation between the train and test dataset exists, cross-validation is redundant and an evaluation on the test dataset suffices. The model with the highest reported F1-score on the Dutch corpus will be marked as best-performing and the results of this model will be compared to the results of the best-performing model resulting from the mono-lingual experiment. The difference between the performance of the two best-performing models will be compared to the reported results of other cross-lingual experiments mentioned in the introduction.

CHAPTER 3. RESULTS

As mentioned above, the performance measure that will be used throughout all experiments is the macro F1-score. A secondary performance measure used will be the accuracy (correct predictions divided by all predictions), to be able to conclude more about the applicability of the reported methods in real-world situations. The reported performance of the experiments will be compared to a baseline defined by the performance of a model predicting emotions randomly for each fragment. This corresponds to a macro F1-score and accuracy of respectively 0.143 and 0.143. These values were found experimentally by performing five runs, consisting of one million randomly predicted labels each, and calculating the corresponding F1-score and accuracy for each run. The results from these runs were averaged to estimate the baseline performance.

3.1 Mono-lingual experiment

The results for the mono-lingual experiment are shown in Table 3.1. The model with the highest performance is trained using an Adam optimizer and batch size of 256, scoring 0.6034 and 0.6030 F1-score and accuracy respectively. The decreasing loss and progressing accuracy while training this model is visualized in Figure 3.1. To validate the positive influence of the augmentation techniques used, this model is trained and tested again using the same parameters and corpora, without using the data augmentation techniques. The resulting F1-score and accuracy for this experiment are respectively 0.5986 and 0.6071, indicating that the data augmentation techniques used, adding noise and changing pitch, have very little influence on the performance when applied to the mono-lingual SER task on the English corpora.

To validate the performance of the highest-scoring model, a second experiment using 5-fold CV is held, using the same parameters and data augmentation techniques as in the original mono-lingual experiment. This experiment prevents the model from accidentally overfitting on the randomly sampled train and test set. The 5-fold CV resulted in an average F1-score of 0.5852 and an average accuracy of 0.5846, proving the validity of the results reported in the mono-lingual experiment for the model trained with an Adam optimizer and a batch size of 256. The average results of the cross-validation exceed any single-run results reported in Table 3.1 except those from the similarly trained model, therefore validating the choice for an Adam optimizer and batch size of 256. The

Optimizer	Batch Size	F1-score	Accuracy
Adam	16	0.5832	0.5835
RMSProp	16	0.5755	0.5739
Adam	32	0.5804	0.5794
RMSProp	32	0.5620	0.5640
Adam	64	0.5790	0.5765
RMSProp	64	0.5690	0.5679
Adam	128	0.5748	0.5771
RMSProp	128	0.5791	0.5767
Adam	**256**	**0.6034**	**0.6030**
RMSProp	256	0.5834	0.5837

Table 3.1: *Results of mono-lingual experiment on English corpora.*

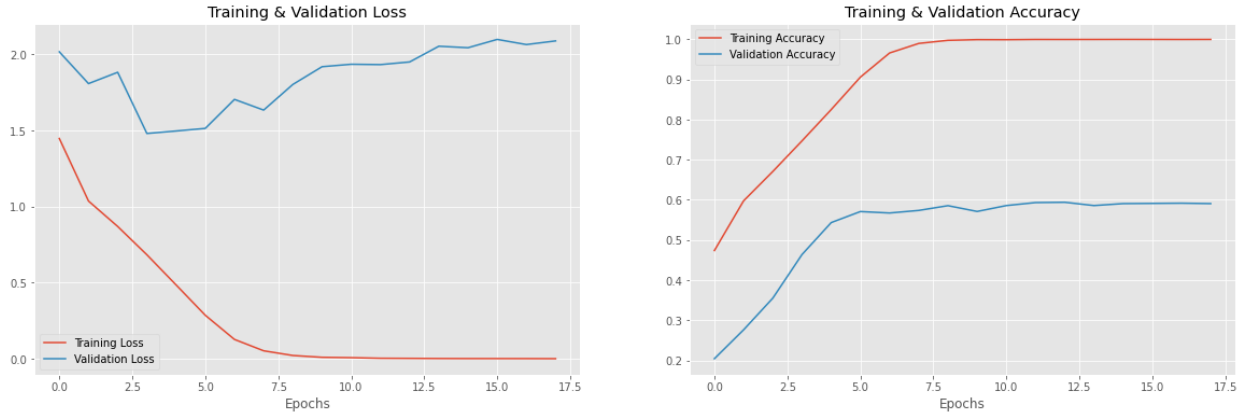

Figure 3.1: *Loss and Accuracy during training Adam optimizer with batch size 256.*

results originating from the mono-lingual experiment and the cross-validation are congruent with findings reported in related mono-lingual studies, which lead to an estimation of state-of-the-art performance on SAVEE, TESS, RAVDESS, and CREMA-D of a weighted average accuracy of 0.661. A retrieved average accuracy resulting from the mono-lingual experiment of 0.5846 shows a performance loss of approximately 0.076 when compared to an estimation of the state-of-the-art techniques on SER for the specific corpora used (state-of-the-art reports 0.661, cross-validation reports 0.585). A side note should be made that the state-of-the-art performance is estimated based on the performance of state-of-the-art baselines on the individual datasets. Concatenating the datasets might require the model to generalize better, therefore scoring slightly worse on the individual datasets. This phenomenon might explain a portion of the gap between state-of-the-art performance and that of the model created in the mono-lingual experiment. However, generalization might in this case be beneficial for the cross-lingual SER task, due to the large difference between the train corpora and test corpus. Other possible explanations for the performance loss

in the mono-lingual experiment can be sought in the architecture of the model used. Whereas the mono-lingual experiment is limited to a relatively straightforward 1D CNN, state-of-the-art implementations use several additional performance-boosting techniques such as gender differentiation and attention mechanisms. Those techniques were excluded from this research due to limitations in time and resources, their absence possibly contributing to the performance loss between the state-of-the-art techniques and the mono-lingual experiment. Finally, the results show the surprisingly minimalistic effect of data augmentation techniques, causing insignificant changes in F1-score and accuracy of the resulting model. Related work reports a positive influence of augmentation techniques, among which the techniques used in this study, indicating a change should be made in either the selection of the augmentation techniques used, or the implementation of them.

3.2 Cross-lingual experiment

The results for the cross-lingual experiment, in which the model is trained on the English corpora and evaluated on the Dutch corpus containing annotated fragments with emotional speech in Dutch, are shown in Table 3.2. The Dutch corpus is annotated using seven basic emotions (fear, anger, happiness, surprise, disgust, sadness, and neutral), correlating with an accuracy of around 0.143 for random performance. The results in Table 3.2 indicate that most models trained with a batch size lower than 128 perform above chance on the cross-lingual SER task, suggesting a positive information transfer from English to Dutch corpora on the SER task. The model trained using the RMSProp optimizer and 32 batch size is evaluated with an F1-score and accuracy of respectively 0.252 and 0.311, suggesting cross-lingual training works to some extent, performing approximately twice as high as chance. These results are congruent with state-of-the-art cross-lingual research (F1-score 0.39-0.46 and accuracy 0.40-0.50) for the SER task on seven emotions, although reporting significantly lower results. The model performs relatively well on anger and fear, two of the well-represented classes in the Dutch corpus, performing significantly above chance for both of them, therefore predicting significantly more than half of the fragments labeled with anger or fear correctly. The model performed worse than chance on sadness and surprise, predicting very little fragments labeled with them correctly.

Optimizer	**Batch Size**	**F1-score**	**Accuracy**
Adam	16	0.1838	0.1646
RMSProp	16	0.1174	0.1585
Adam	32	0.1738	0.1585
RMSProp	**32**	**0.2517**	**0.3110**
Adam	64	0.1945	0.1768
RMSProp	64	0.1720	0.1585
Adam	128	0.1561	0.1463
RMSProp	128	0.0990	0.1159
Adam	256	0.1126	0.1341
RMSProp	256	0.1265	0.1463

Table 3.2: *Results of cross-lingual experiment, training on English and testing on Dutch emotional speech.*

The results of the cross-lingual experiment indicate a larger deviation from state-of-the-art performance, which reports F1-scores between 0.39 and 0.46 and accuracies between 0.40 and 0.50, whereas the cross-lingual experiment reports an F1-score of 0.252 and an accuracy of 0.311. Several explanations for this performance gap can be forwarded. First of all, annotating emotional speech is a highly subjective task, as mentioned in the introduction. Annotating emotional speech is commonly done by multiple annotators, discarding any speech perceived as ambiguous among them. Due to limitations in resources, the Dutch corpus in this study is annotated by a single person, possibly resulting in subjectively labeled data that would otherwise have been discarded, enlarging the difference between the English and Dutch corpora and complicating the cross-lingual SER task for the model.

Secondly, the origin of the Dutch corpus, namely interviews with Dutch veterans, can be perceived as imbalanced in the context of the SER task with Ekman's seven emotions, due to the emotionally imbalanced subjects discussed in them. Anger and fear appeared much more often than surprise and disgust, resulting in a corpus containing much more fragments labeled with some emotions than others, limiting the possibility to conclude the predictive ability of the model for the primary purpose of this study, which is the feasibility of the cross-lingual SER task between English and Dutch. However, emotionally imbalanced datasets will often occur in the analysis of oral history archives, due to specific contexts these entail. Therefore, the ability to conclude about the predictive performance of the model for the secondary purpose of this study might be enlarged by the imbalance of the Dutch corpus.

Finally, an information gap is caused by the origin of the corpora. The English corpora are constructed by actors uttering phrases with certain emotional content, whereas the Dutch corpus is constructed under more natural circumstances. As mentioned before, positive information transfer between acted and natural datasets is possible. However, this difference may enlarge the information gap, and the assumption that this effect complicates the cross-lingual SER task between the English and Dutch corpora is therefore justified.

However, despite these complications that might explain the difference in state-of-the-art performance and the results of the cross-lingual experiment, the results indicate a positive information transfer between the English and Dutch corpora. Whereas on chance performance would result in respectively 0.143 and 0.143 F1-score and accuracy, the results show respectively 0.252 and 0.311 F1-score and accuracy, indicating the ability of the model to adapt to Dutch emotional speech using only English emotional speech with a performance exceeding chance significantly.

CHAPTER 4. CONCLUSION

The results on the mono-lingual experiment are concordant with state-of-the-art performance and show the feasibility of the model to recognize English emotional speech, thus showing the ability of machine learning techniques to recognize English emotional speech. The results of the cross-lingual experiment show performance significantly worse than state-of-the-art and mono-lingual performance. However, the performance of the model is significantly better than chance, showing that cross-lingual SER is to some extend practicable between English and Dutch, therefore confirming the feasibility concerning the primary goal of this research. These results indicate numerous possibilities for analyzing Dutch emotional speech, despite the absence of properly annotated Dutch resources applicable to SER. The application of cross-lingual SER to analyze oral history archives such as the database containing interviews with Dutch veterans should be treated with more caution, as the performance of the model on the cross-lingual SER task might be insufficient to be of proper assistance since most sections of emotional speech predicted by the model might not correspond to the observations made by the human observer. However, the model is capable of distinguishing emotions like anger and fear relatively accurately, and could therefore be of assistance in recognizing sections of speech in which these emotions are numerous. Combined with the existence of techniques reported to increase accuracy for the cross-lingual SER task that were not implemented in this research, cross-lingual SER can certainly be an auxiliary tool for analyzing Dutch oral history archives. In conclusion, this research reports positive results for cross-lingual SER between English and Dutch and cautionary suggests the feasibility of cross-lingual SER for analyzing Dutch oral history archives.

This outcome suggests multiple possibilities for future work on cross-lingual SER between English and Dutch. First of all, results would be more conclusive if the Dutch corpus would be enlarged and validated by multiple human annotators, preventing individual subjectivity to influence the predictive ability of the model. Research could be aimed at building, validating, and testing an expanded Dutch corpus for SER. Second of all, multiple techniques such as gender differentiation and attention mechanisms report improvement on the SER task. Future work could be aimed at exploring the effect of these methods on the cross-lingual SER task. Last of all, the lack of improvement the data augmentation techniques yielded is contradictory to the findings of

related work and could be thoroughly examined in future work, to construct an augmentation set improving performance on the SER task.

REFERENCES

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[2] Abbaschian, B. J., Sierra-Sosa, D., and Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249.

[3] Babko, D. (2021). Speech emotion recognition. Kaggle.

[4] Bashirpour, M. and Geravanchizadeh, M. (2018). Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1):1–13.

[5] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

[6] Chiţu, A. G., Van Vulpen, M., Takapoui, P., and Rothkrantz, L. J. (2008). Building a dutch multimodal corpus for emotion recognition. In *Programme of the Workshop on Corpora for Research on Emotion and Affect*, page 53.

[7] Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019). Deep neural networks for emotion recognition combining audio and transcripts. *arXiv preprint arXiv:1911.00432*.

[8] Darekar, R. V. and Dhande, A. P. (2018). Emotion recognition from marathi speech database using adaptive artificial neural network. *Biologically Inspired Cognitive Architectures*, 23:35–42.

[9] Dolka, H. and Juliet, S. (2021). Speech emotion recognition using ann on mfcc features. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pages 431–435. IEEE.

[10] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

[11] Feng, K. and Chaspari, T. (2020). A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science*, 2:9.

[12] Feraru, S. M., Schuller, D., et al. (2015). Cross-language acoustic emotion recognition: An overview and some tendencies. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 125–131. IEEE.

[13] Ganchev, T., Fakotakis, N., and Kokkinakis, G. (2005). Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194.

[14] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[15] Ingale, A. B. and Chaudhari, D. (2012). Speech emotion recognition using hidden markov model and support vector machine. *International Journal of Advanced Engineering Research and Studies*, 1(3):316–318.

[16] Jackson, P. and Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

[17] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[18] Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., and Wermter, S. (2018). On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 854–860. IEEE.

[19] Lee, C. M. and Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303.

[20] Lee, S.-w. (2019). The generalization effect for multilingual speech emotion recognition across heterogeneous languages. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5881–5885. IEEE.

[21] Lin, Y.-L. and Wei, G. (2005). Speech emotion recognition based on hmm and svm. In *2005 international conference on machine learning and cybernetics*, volume 8, pages 4898–4901. IEEE.

[22] Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

[23] Mekruksavanich, S., Jitpattanakul, A., and Hnoohom, N. (2020). Negative emotion recognition using deep learning for thai language. In *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, pages 71–74. IEEE.

[24] Meng, H., Yan, T., Yuan, F., and Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE access*, 7:125868–125881.

[25] Milner, R., Jalal, M. A., Ng, R. W., and Hain, T. (2019). A cross-corpus study on speech emotion recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 304–311. IEEE.

[26] Mishra, P. and Sharma, R. (2020). Gender differentiated convolutional neural networks for speech emotion recognition. In *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 142–148. IEEE.

[27] Neumann, M. et al. (2018). Cross-lingual and multilingual speech emotion recognition on english and french. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5769–5773. IEEE.

[28] Pichora-Fuller, M. K. and Dupuis, K. (2020). Toronto emotional speech set (tess). *Scholars Portal Dataverse*.

[29] Sasaki, Y. et al. (2007). The truth of the f-measure. 2007. *URL: https://www. cs. odu. edu/˜ mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07. pdf [accessed 2021-05-26]*.

[30] Tieleman, T., Hinton, G., et al. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

[31] Tripathi, S., Kumar, A., Ramesh, A., Singh, C., and Yenigalla, P. (2019). Deep learning based emotion recognition system using speech features and transcriptions. *arXiv preprint arXiv:1906.05681*.

[32] Veteraneninstituut, N. (2021). Interviewcollectie nederlandse veteranen.

[33] Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018.

[34] Weninger, F., Ringeval, F., Marchi, E., and Schuller, B. W. (2016). Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In *IJCAI*, volume 2016, pages 2196–2202.

[35] Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U., and Gadekallu, T. R. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, pages 1–10.

[36] Zhou, Y., Xiong, C., and Socher, R. (2017). Improved regularization techniques for end-to-end speech recognition. *arXiv preprint arXiv:1712.07108*.