INVESTIGATING EMOTION EMBEDDING BASED TEXT-TO-SPEECH MODELS UNDER LIMITED TRAINING DATA

HAROLD PIJPELINK

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

235038

COMMITTEE

dr. Dimitar Shterionov
dr. Chris Emmery

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

July 15, 2022

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Dimitar Shterionov for his continued support of my thesis, and his wide variety of suggestions and ideas.

"Work on this thesis (did/did not) involve collecting data from human participants or animals. The author of this thesis acknowledges that they do not have any legal claim to this data. The code used in this thesis is publicly available: https://github.com/haza97/Thesis-Project-Emotion-Embeddings"

INVESTIGATING EMOTION EMBEDDING BASED TEXT-TO-SPEECH MODELS UNDER LIMITED TRAINING DATA

HAROLD PIJPELINK

Abstract

This thesis studies the effect of limiting training data on emotion embedding models for Text-to-Speech (TTS) systems. In order to reproduce natural human prosody, TTS models use emotion embedding layers. The datasets used to TTS models are large and require parallel samples, which makes them costly to obtain. This study investigated the effect of using smaller datasets and lower quality datasets when training these models. In this case, lexical diversity was taken as a proxy for quality. Two model architectures were used: one SVM-based machine learning model (Zhu, Yang, Yang, & Xie, 2019), and a deep learning model (Liu, Sisman, & Li, 2021b), both of which are trained of pairs of <emotional, neutral> speech utterances and its transcriptions. To determine the effect of limited lexical diversity, two models with high lexical diveristy and two models with low lexical diversity were trained using the SVM model architecture.

The SVM models were evaluated on two tasks: the emotional vs neutral text classification task that they were trained on, and an emotion classification task with four emotion categories. The deep learning model was evaluated using only the latter method. The SVM models show high accuracy on the emotional vs. neutral task, but only reach chance performance on the other emotion classification task. The level of lexical diversity of the training data does not affect the accuracy. The deep learning model performs slightly better than chance level as the amount of training data increases, but there appears to be a lower bound where the model does not learn to generalize anything.

Word count: 8026

1 INTRODUCTION

Text to speech (TTS) models are systems that produce an audio signal from a text input. The aim of these models is to synthesize speech that is indistinguishable from human speech, replicating every aspect of the prosody of human speech. TTS models as we know them now have existed since the 1960s in the form of Linear Predictive Coding (Ding & O'Shaugnhnessy, 2003), with a large variety in models following it until present day. The current state of the art models are based on deep learning, and are able to imitate human speech to large degree. However, one of the aspects of TTS models that has not been perfected yet is the emotionality of the output speech. The audio signal has dimensions such as pitch, prosody and tempo which are not reflected in the text, but are to be replicated in order for the output to sound natural. The challenge of emotional TTS is to devise a methodology that allows a model to transform the low information text signal into a high information speech signal (Wang et al., 2018). Another issue with large TTS models is that suitable training data is limited. Successful TTS models such as TacoTron2 (Shen et al., 2018) can only be trained on large speech corpus datasets and their transcriptions. However, such datasets are costly to obtain and the publicly available ones are limited in number and in context. This is why recent trends in deep learning models have looked at the effect of lower volumes of training data. This is known as a form of domain adaptation (Farahani, Voghoei, Rasheed, & Arabnia, 2021), which aims to obtain a small, high quality source dataset for training that will generalize well to a different dataset. Moreover, emotion data is generally not labeled, so it cannot be used in a supervised learning problem as is. One solution to both the problem of poor emotionality and unlabeled data is the addition of so-called emotion embeddings. These are extra layers in deep learning models that capture the emotion from unlabeled-training data and use them to improve the output speech of the model.

1.1 *Social Relevance*

TTS models in general have a large number of applications, ranging from assistance for the visually impaired to the generation of automatic reading of websites, to social media such as TikTok. High quality emotional TTS models have been suggested to be particularly important in applications that require interactivity between the user and the system, or when involving interactions that take place over longer periods of time, such as virtual assistants, call centers, online education (Liu, Sisman, & Li, 2021a). Another example is assistance for people with visual impairments. Contrary to

other systems, in which speech is an addition to what happens on a screen, these systems are required to operate using only audio. A TTS model with better emotional expression will be able to convey the its message to the impaired user more clearly Edward (2018). Emotional TTS is also found in the field of speech-to-speech translation. Current translation algorithms can provide high quality translations quickly, but TTS models using only text as input are not able to express all the information in the input speech (Akagi, Han, Elbarougy, Hamada, & Li, 2014). Szekely, Steiner, Ahmed, and Carson-Berndsen (2013) argues that speech-to-speech translation systems would benefit from better emotional TTS, as it captures the intention of the source speaker better than a non-emotional equivalent. Examples that brings several of these fields together can be found in robotics. One example is ORLANoid, which is a robot alter-ego for the French artist ORLAN. ORLANoid consists of a torso, two movable arms and a head. It is also able show some facial emotions and to speak using the artist's voice. One part of this robot is an end-to-end machine translation system that translates French speech input directly to English speech output using a TTS module. This makes ORLANoid a system with two separate requirements for emotions modeling: the facial emotions and vocal emotions must be in line with each other, so the system should have an overall emotion module.

1.2 *Problem statement*

Currently, for many TTS models the quality of the output speech is lacking in both intonation and prosody; the artificial speech is unable to replicate variations in human speech accurately and sounds unnatural. According to Wang et al. (2018), this is due to speech being a rich signal that contains information about style an prosody, while text includes only the literal meaning. Recent literature suggests that training a TTS model using inputs that quantify the level and nature of emotion in the training audio can lead to improvements in emotional expression (Wang et al., 2018; Zhu et al., 2019). Another challenge challenge with deep learning models is their low sample efficiency. For a model to perform well, it requires a large training set of labeled data. If no such values are available, developing such a data set would require manual grading by native speakers.

This thesis will address both problems by training two emotional embedding models using limited training data, similar to what a real world situation could look like. The two models that will be compared both include emotion labels into the training of a TTS model and will be compared on their accuracy. In the context of this thesis, limited data refers to both quantity as well as quality. The amount of data is sometimes

referred to as its volume, note that this has nothing to do with any of the sound characteristics or loudness of the data.

When trained on smaller subsets of the original data, the performance for both models is likely to go down, as the models have fewer examples to learn from. However, the magnitude of this effect is not clear in advance, and may differ between the two models. The first methodology is based on Zhu et al. (2019), who use a Support Vector Machine (SVM) to train a ranking function which quantifies the degree of emotionality for different audio inputs. The second methodology is StrengthNet (Liu et al., 2021b), which is a neural network based on a Convolutional Neural Network (CNN) and a Bidirectional LSTM (BiLSTM) architecture to quantify a probability distribution over the emotion categories, but also an overall emotional intensity score.

The ultimate goal for these models is to be included in a full TTS model as part of the emotional embedding system. Evaluating a TTS model is generally done through human scoring on the speech output. However as the aim of this thesis is to replicate a real life situation with restricted data, the speech output of the TTS system would be affected by the limited data not only in its emotionality, but also in the overall performance. This would make it difficult to evaluate only the effect of limited data on the emotional models. As a result, the trained will be evaluated objectively on their accuracy of emotion predictions without human judgement.

1.3 *Research Questions*

How well do emotion embedding models learn when training data is limited? To ding out, Strengthnet (Liu et al., 2021b) and Zhu et al. (2019) based models will be trained using varying amounts of input data. Moreover, the effect of data quality will be investigated using lexical diversity as a proxy, by training models based on (Zhu et al., 2019).

To what extent can emotional embeddings be trained accurately with limited training data?

RQ 1 *To what extent does the accuracy for both models decrease as the volume of input data decreases?*

RQ 2 *To what extent does the accuracy for the Zhu et al. (2019) model decrease as the lexical diversity of input data decreases?*

2 LITERATURE REVIEW

This section starts with a general overview of neural speech synthesis models to provide a general background of work in the field. This is followed by a more detailed explanation of the methodologies that we will implement: the Zhu et al. (2019) methodology and the StrengthNet methodology.

2.1 *TTS models*

Historically, TTS models consist of three parts: text analysis, an acoustic model and a vocoder. The text analysis models extract linguistic features from a given text. The acoustic model transforms the linguistic features into acoustic features, which are then turned into audio wave forms by the vocoder (Tan, Qin, Soong, & Liu, 2021). Prior to the advent of neural networks and deep learning, TTS models made use of concatenation based models.These models produced speech by concatenating combinations of phones into new sentences. Variations of concatenative models include formant synthesis, which used artificially generated phones (Burkhardt & Sendlmeier, 2000), and diphone and triphone speech synthesis , which used combinations of phones from existing speech corpora (O'Shaughnessy, Barbeau, Bernardi, & Archambault, 1988). Other models took this further and used entire words from existing corpora (Campbell & Black, 1997). The main drawback of these models were that the phones themselves were fixed, and could not represent the dynamic changes often seen in emotional speech. These were partly solved by increasing the number of emotional utterances, such as in Turk, Schröder, Bozkurt, and Arslan (2005), but the parametrization of emotional speech was limited to a small number of discrete states. This was solved better through Hidden Markov Models (HMM) (Tokuda et al., 2013), also known as statistical parametric speech synthesis (SPSS). These models consist of a series of subsequent internal states that are connected by way of probability distributions. For a given input sequence of acoustic vectors, the HMM will try to find the most likely output sequence of based on maximum likelihood estimation. As a result, the level of expression became parameterized; prosody generated by the model could be tweaked by changing the parameters of the probability distributions.

Contrary to of data science problems, the evaluation of TTS systems is not straightforward. The TTS system is only as good as it is perceived to be by its users. The most used evaluation method for TTS systems is the Mean Opinion Score. MOS is a human evaluated metric used to determine the subjective quality of audiovisual systems, including TTS systems. It asks

participants to rate TTS generated speech a 5-point scale which is usually expressed as bad, poor, fair, good and excellent. These scores are then averaged over the number of participants to give the MOS (**?**). Separate from the subjective MOS, there are also objective measures that can be calculated without getting costly human judgement. Commonly used ones are statistics based models such as peak-signal-to-noise ratio, and there is an entire field of perceptual modeling (Brunnstrom, Hands, Speranza, & Webster, 2009). However, since MOS represents the user experience directly, it has become the most used quality metric (Streijl, Winkler, & Hands, 2016).

In neural TTS, at least one of the main three modules is a neural network. Common applications include models that replace the acoustic model with a CNN which can directly interpret the audio input as a spectrogram, replacing the need for a separate text analysis module (Tan et al., 2021). Generally, the input and output speech in TTS models are shown as mel spectrograms. These are audio spectrograms that are logarithmically-scaled using measurements of human hearing and to emphasize lower frequencies. These frequencies are essential to human comprehension of speech, making mel spectrograms a more realistic representation of human hearing (Davis & Mermelstein, 1980). The alternative to the mel spectrogram would be a linear-scale spectrogram, such as the Short-Term Fourier Transform (STFT). The aim of TTS models is to generate speech that sounds natural to humans, which is the mel spectrogram. For deep learning, there is the additional advantage of mel spectrograms discarding more information than STFTs, but retaining most of the important features. This effectively makes the mel spectrogram a lower feature representation of the sound, lowering the variance of the signal. This makes them the preferred choice over the STFT when used in deep learning (Choi, Fazekas, Cho, & Sandler, 2018).

2.2 *Defining Style*

Before describing methods for style modeling, we first look at what kind of information is actually learned by the models. Tan et al. (2021) define style as "any aspect of the speech utterance that is not determined by its linguistic content and the speaker's inherent voice characteristics". Wang et al. (2018) emphasizes that style contains "rich information, such as intention and emotion". This is quite vague. In order to make the distinction between and information more clear, there is a framework by Fujisaki (2004). This framework which divides the information found in speech into three separate categories: linguistic, para-linguistic and non-linguistic information. Linguistic information is the meaning of the sentence uttered, similar to how meaning would be conveyed in writing. He defines para-

linguistic information as the information that is added by the speaker to change or supplement the linguistic information, such as emotion. Lastly, non-linguistic information contains all the other information that is contained by the sentence, such as information about the gender, age or emotional state of the speaker. Clearly, emotional TTS is concerned with the connection between linguistic and para-linguistic information.Emotional speech is one way to express the connection between these two kinds of information, which makes it necessary for a good TTS model to replicate this. Other models that use this framework explicitly are (Kano et al., 2012), who modeled phoneme duration and power in their TTS module, and (Székely, Henter, & Gustafson, 2019), who explicitly model the breaths taken between utterances to make a TTS model sound more natural.

2.3 *Modeling Emotion*

In order for a TTS model to generate emotional speech, the model must be able to make generalizations between the emotionality of its input data and the test data. This requires a classification model for emotions. The field of emotion classification contains a variety of models, which can generally be subdivided in dimensional models and discrete models. Examples are the circumplex model, which considers emotion as a two dimensional system of valence and arousal (Russell, 1980). Discrete models include Plutchik's model, which describes emotions as a 3-dimensional cone which places emotion categories on the base circle of the cone and the intensity on the vertical axis (Plutchik, 1984). These types of simple models are used for a variety of other fields, such as in modeling the effect of emotions on the economy (Tilly, Ebner, & Livan, 2021), or the . There are also more complicated models, such as Velsquez (1997), who developed an emotion model for artificial agents. This model, which is called Cathexis, provides artificial agents with an internalized state of emotions and motivations which affect their interaction with their environment. More recent developments in the field of affective computing include the use of deep learning methods such as such as CNNs (Khatua, Khatua, & Cambria, 2020) and RNNs (J, Trueman, & Cambria, 2021). Another set of models that may be interesting in the context of TTS are multi-modal models, that take into account several types of input data. Neural TTS models are by definition trained on both audio and speech and, although there are a variety of TTS models that take multi-modal input (Effendi, Tjandra, Sakti, & Nakamura, 2020; Shen et al., 2018), none of these seem to make any explicit references to existing multi-modal emotion models.

2.4 *Emotional TTS Models*

Despite this large variety of emotion classification models, modeling emotionality in a TTS setting is done mostly through trained embeddings. This works by adding extra layers to the deep learning model that can capture the relationship and express them as an embedding. The emotion modeling in these embeddings is generally kept simple, and consist of n-dimensional vectors for each discrete emotion. TacoTron2 (Shen et al., 2018) is a popular model to train using (external) emotion embeddings. It consists of a combination of a sequence-to-sequence model (seq2seq) combined with a vocoder model which are then trained simultaneously. TacoTron2 maps a text input sequence to a MEL spectrum output sequence using an encoder-decoder structure. These mel spectra are then transformed into an audio signal using the associated Wavenet (Oord et al., 2016) vocoder model. The prosody and emotion in TacoTron2 are found in the latent space between the encoder and decoder. By inserting a (pretrained) embedding vector in this space, the emotionality of the output speech can be altered. At inference, the model either uses its predicted embedding for the input sentence, or the user is able to specify an embedding and influence the emotion of the speech output directly. This process is known as conditioning the TacoTron2 model on the emotional embedding. It is also possible to use external embeddings that were trained by a separate model, such as in Hayashi et al. (2019). The authors develop an encoder-decoder methodology based on (Shen et al., 2018), but rather than only using text as training data, word embeddings are extracted using a BERT model (Devlin, Chang, Lee, & Toutanova, 2018).

The advantage of using an end-to-end model is that does not require data with explicit emotion labels. Training TTS models requires fairly large speech corpora. For reference, existing TacoTron2 implementations have been trained on the LJ speech dataset (24 hours of speech) (Ito & Johnson, 2017) and the JSUT corpus (Sonobe, Takamichi, & Saruwatari, 2017) (10 hours of speech). Generating emotion scores for such volumes of data using human judgement would be very costly. As we saw in the section on emotion classification models, it would also be unclear what methodology to use to label the data. End-to-end modeling takes these problems away by learning emotion embeddings in the latent space of the model.

The idea to use an embedding to represent differences between input speech was first used in multi-speaker modeling.

2.5 *Data Volume and Data Quality*

One requirement for neural TTS models is a high volume of input data; Prajwal and Jawahar (2021) estimates this to be 20 hours per speaker per language. Getting such data is costly, and is not available in many languages. To overcome this problem, there are a variety of data augmentation applications specific to TTS, such as models that diversify their training set using data generated by other TTS models. These models are all trained on the same input data, but generate new speech themselves which is then used to further train the models (Hwang, Yamamoto, Song, & Kim, 2021; Laptev et al., 2020). The same idea has been applied using voice conversion models, which take audio input from a target speaker and aim to transfer it to output that resembles a source speaker (Huybrechts et al., 2021). (Latorre et al., 2019) investigated the effect of restricting training data by comparing TacoTron2 models trained on a large one-speaker dataset to a TacoTron2 model trained using smaller samples from multiple speakers. They find that, despite using less overall data, the latter model outperforms the first one, possibly due to the extra diversity in its input data.

This diversity is an aspect of overall data quality. It is also a known metric in linguistics, where it is known as lexical diversity. The goal is to capture the overall diversity of a text corpus through a variety of metrics, most of which are based on frequency analysis of words in the total vocabulary (Tweedie & Baayen, 1998). In the field of Automatic Speech Recognition, lexical diversity has been used as a measure of quality for input data. An example is Rosenberg et al. (2019), who apply data augmentation using the lexical diversity metric of maximum entropy to sample its training data from a corpus. By training on this more diverse data, the quality of the output speech of the model is improved.

2.6 *Two Embedding Methodologies*

To answer the research question, this thesis will train two emotional embedding models using various volumes of training. These two models are StrengthNet (Liu et al., 2021b), and the model described in (Zhu et al., 2019), which will hereafter be referred to as "the Zhu model".

Both of these methodologies are extensions of the Global Style Token (GST) (Wang et al., 2018) methodology. GST generates emotion embeddings by comparing mel spectrum input of audio input to randomized embeddings in the latent space. At training, the model receives mel spectrograms of unlabeled audio data as input. Using a combination of convolutional layers and an RNN, the mel spectrograms of the input audio is turned into a 3 dimensional embedding. This embedding is then fed through

an attention layer and into the latent space, where the model learns the similarity between the embedding of the input and a set of randomly initialized embeddings known style tokens. As a result, the model learns which style tokens contribute to each training utterance, and the style tokens can be used to fine-tune a TacoTron2 model on. At inference, the model can be conditioned on a particular style token, allowing the user to directly change the prosody of the output. In the GST methodology, the entire embedding of the input sequence is compared against the style tokens in the latent space. This leaves out differences in emotions between words in the input sentence and results in an output with averaged emotions across the output sequence. Zhu et al. (2019) use a machine learning methodology that works as an extension of GST, but includes a more fine-grained control over the emotion in the output sentence. The novelty in this model is the application of the relative attributes methodology (Parikh & Grauman, 2011). This replaces binary feature vectors with a learned ranking function, which allows the binary representation to be replaced with a continuous one. The original model applies this methodology to the relative attributes found in images, but (Zhu et al., 2019) extend this to emotions. The ranking function is based on a constrained optimization problem that tries to fit a hyperplane between two categories in the form of a weight matrix. This weight matrix can then be used as a ranking function for that category. The mathematics behind this optimization problem are similar to the rank support vector machine (Joachims, 2002). In practice, the Zhu model trains one SVM model for each emotion category on pairs of <neutral, emotional> utterances. Figure 2 shows an overview of the model pipeline. Rather than comparing the different emotional utterances to each other, the model compares them all against a neutral utterance and gives a score for each category, effectively solving the problem using transfer learning (Pan & Yang, 2010). The mathematical details will be discussed further the methods section. By training a ranking function for four discrete emotion categories, this allows each utterance to be represented by 4-dimensional vector with scores for each emotion. Unlike GST, which provides an emotion embedding that is averaged over the output sequence, every time step in the sequence has an associated emotion score vector, which leads to the increase in fine-grainedness.

The other methodology is StrengthNet (Liu et al., 2021b). This two-part neural TTS model is based on a combination of convolutional neural layers and a bidirectional LSTM layer. It takes the same <neutral, emotional> pairs found in Zhu et al. (2019) but transforms them into mel-spectrograms before using them as input. It uses the output from an external ranking function, which could be taken from the Zhu model, as its ground-truth and outputs a probability distribution over four emotion categories, as well

as a scalar that represents emotional intensity. These outputs can then be used as an emotion embedding in a TTS model.

2.7 *Other Embedding Models*

The success of GST has led to a large variety in TacoTron2 based model architectures that are conditioned on some implementation of emotion scores. To provide a comparison between the two methodologies chosen in this thesis, this section will describe several alternative models, and why they were not included in the model comparison. (Cai et al., 2020) note that using a supervised learning model to improve emotional TTS requires large amounts of labeled data, which is not always available. To solve this, they developed a TTS model that contains an explicit Speech Emotion Recognition module and train both of these together. Although this idea seems like a better form of StrengthNet, it still explicitly uses a GST based embedding, which does not allow for fine-grained emotionality.. (Kwon, Jang, Ahn, & Kang, 2019) developed a different methodology to achieve fine-grained emotionality. This is done by comparing the distribution of different emotion vectors in the latent space of a Tacotron2 model, and taking the center of this distribution to be a weighting value for that emotion, leading to a better MOS than the original TacoTron2. Another is Chen et al. (2019) who take inspiration from multi-speaker models and create a specific embedding for each language. They use a CNN based encoder which creates an embedding for each speaker in the training set. Using a triplet loss function, the distance between speaker embeddings is then maximized. The main strength of this model is its ability to synthesize the text with a new voice using only three minutes of audio. The reasons for choosing StrengthNet and the Zhu et al. (2019) methodology over these other models is the fine-grainedness of the emotion. Human evaluators found that the emotions and emotional strength in the audio generated by Zhu et al. (2019) were more clear than those of the Tacotron2 model for all categories of emotions. The reason to choose the StrengthNet methodology is that it is similar to Zhu et al. (2019), but provides and end-to-end model, which will make it a good comparison. One last note on the embedding model literature is that many do not elaborate on the methodology of how the model is to be inserted in to a TTS model. Neither the (Zhu et al., 2019) or (Liu et al., 2021b) provide a detailed methodology for conditioning the TacoTron2 model on the externally learned embeddings. (Zhu et al., 2019) discusses that the embedding is appended to the latent space between the encoder and the attention layer. However, there are a variety of ways that the embedding could be implemented into the TTS model. Elyasi and Bharaj (2021) explore this by inserting the embedding in TacoTron2

at different locations, and find that conditioning the model on multiple embeddings can lead to better pitch accent and syllable stress.

3 METHODOLOGY

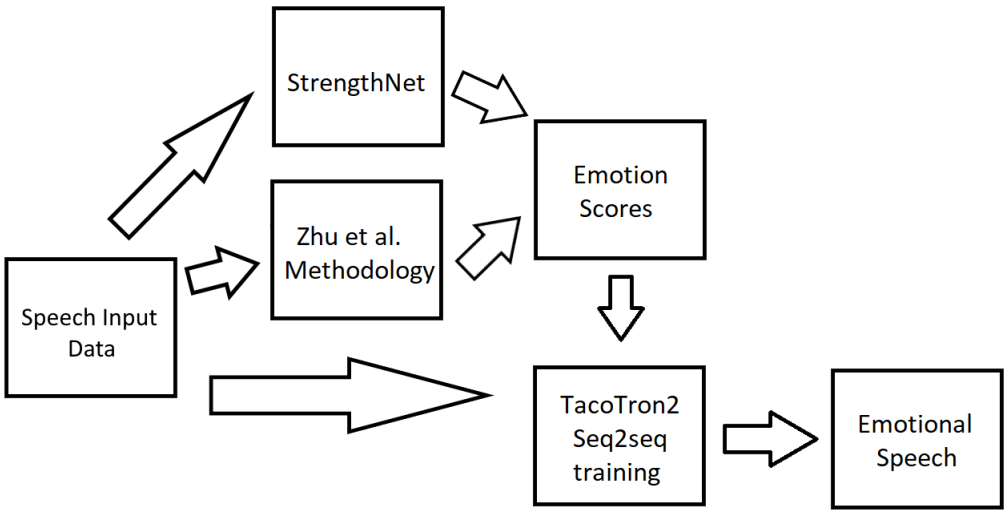


Figure 1: Overview of a completed emotional TTS model pipeline.

To answer the research question, this thesis will compare the Zhu model and Strengthnet in two experiments. The first experiment investigates the effect of low volumes of training on accuracy for both models by training both models on increasingly small subsets of input data. The second experiment is inspired by Rosenberg et al. (2019), and investigates the effect of different qualities of training data, which is operationalized as the difference in lexical diversity. The first section describes the two models in detail, followed by a description of the experiment and the data involved. In the last subsection, some of the implementation details are elaborated on, including some practical adaptations to the Zhu methodology.

3.1 *Architectures of Embedding Methodologies*

As the Zhu model is trained using a method that uses pairwise comparison, it has to be trained on a dataset with parallel utterances. To extract the feature vectors for the input data, this corpus is put through the emotional feature extraction toolkit OpenSMILE to extract the statistical properties of the speech pairs into 384-dimensional feature vectors (Eyben, Wöllmer, & Schuller, 2010). This toolkit takes the mel spectrograms of

the audio input and calculates a range of statistics such as fundamental frequency or number of zero-crossing. OpenSMILE provides a large variety of configurations that determine the exact composition of these statistical properties. The configuration used in this thesis is the Interspeech 2009 Emotion Challenge configuration (Schuller, Steidl, & Batliner, 2009), an exact overview of its statistical properties can be found in Appendix A (page 30).

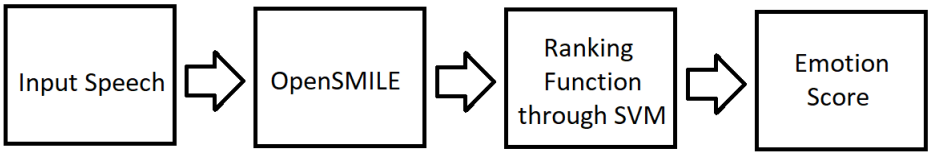


Figure 2: Overview of a the model trained in Zhu et al. (2019).

The model then learns a ranking function for each emotion based on two separate training sets containing the feature-vectors. The set $\boldsymbol{O}$ contains pairs consisting of a neutral utterance and an emotional utterance. The set $\boldsymbol{S}$ contains pairs of neutral utterances. The underlying reasoning is that for each pair in $\boldsymbol{O}$, the emotional strength of the emotional speech has to be higher than the emotional strength of the neutral speech. Mathematically, learning the ranking function corresponds to minimizing the following optimization problem in a pairwise fashion, taking the assumptions as constraints that the model cannot violate. This optimization problem can be shown as follows, taking into account a quadratic loss function given here by $\boldsymbol{C}$:

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w}_m^T||_2^2 + C(\sum \xi_{ij}^2 \ + \sum \gamma_{ij}^2)\right) \tag{1}$$

$$\text{s.t.} \quad \boldsymbol{w_m^T x_i} \geq \boldsymbol{w_m^T x_j} + 1 - \xi_{ij}; \forall (i,j) \in O_m \tag{2}$$

$$|\boldsymbol{w_m^T x_i} - \boldsymbol{w_m^T x_j}| \leq \gamma_{ij}; \forall (i,j) \in S_m \tag{3}$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0 \tag{4}$$

With $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ representing the pairs of input data, $\boldsymbol{w}$ representing the learned weights for the ranking function and $\boldsymbol{m}$ representing the different emotion categories. ξ and γ are slack variables, and $\boldsymbol{C}$ is a constant determining the trade-off between the minimizing the objective function and keeping the two constraints. $\boldsymbol{O}$ and $\boldsymbol{S}$ are the two training sets described in the previous paragraph. $||w^{T_m}||_2^2$ stands for the Euclidean norm of the

weights matrix squared, the intuition behind this, along with an example can be found in section 4 of Joachims (2002)

In practice, it may be time consuming to find the exact solution to the optimization problem, so to make this problem more computationally efficient, the solution is approximated. This is done by adding two slack variables that represent the extent to which the model is allowed to violate the constraints while training. This approximate solution will give a function that explicitly ranks the speech pairs based on the level of emotionality for each emotion.

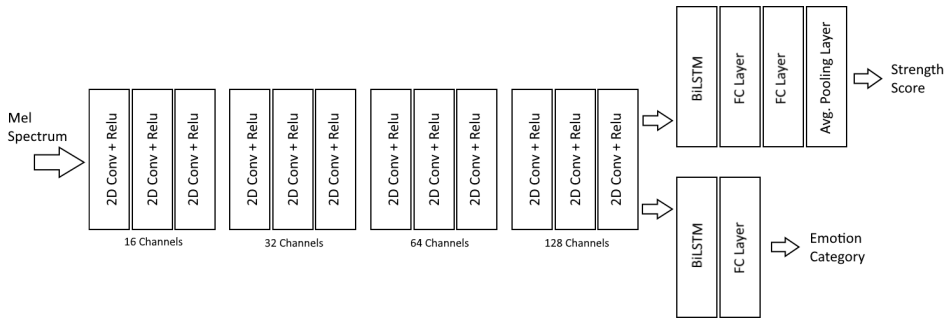


Figure 3: StrengthNet Model Architecture, adapted from Fig. 1 in Liu et al. (2021b).

StrengthNet is a deep learning model that takes pairs <neutral, emotional> of mel spectrum frames as input. Its architecture can be seen in Figure 3. Given that the sampling rate of native audio signal is unnecessarily high and training on individual samples would be computationally expensive, 256 audio samples are concatenated into one frame. These frames are input into an acoustic encoder which consists of twelve ReLu activated Convolutional CNN layers with an increasing number of channels. The idea of using a large number of convolutional layers was Mosnet (Lo et al., 2019). As the input to StrengthNet is in frames of mel spectra, the higher number of filters per convolutional layer allows the model to keep increasingly long periods of time into account per neuron. The output of the acoustic encoder is used to simultaneously train two output heads: the Strength Predictor and the Emotion Predictor. The purpose of the Strength Predictor is to use the so called "high level features" taken from the Acoustic Encoder to predict how strong the emotion presented in the utterances is. The first layer of the strength predictor is a Bidirectional LSTM (BiLSTM). This which extends the LSTM architecture to have both forward and backward dependencies in time. Each unit consists of a forward and a backwards layer, who take the sequential data in its regular order and re-

verse order, respectively. The outputs for both layers are then concatenated at each time step, providing every time step with information on both the future as well as the past (Cui, Ke, Pu, & Wang, 2018), making it possible for StrengthNet to predict emotional strength more accurately.The strength predictor is followed by two fully connected layers and an average pooling layer, after which it outputs a scalar that indicates the emotional strength of the input utterance. The Emotion Predictor classifies the emotion into one of four categories: happy, sad, angry and surprised. This output head also starts with a BiLSTM layer, followed by a single fully connected layer with four output units.

The model is trained using three different loss functions. In the Strength Predictor, there is an Mean Average Error (MAE) loss function that minimizes the loss before the second fully connected layer and the average pooling layer, and a second MAE loss function that minimizes the loss of the strength score against the strength of the ground-truth. The authors suggest that this extra loss function will make the model converge faster. The third loss function is a categorical cross-entropy loss function in the emotion category predictor, between the category found in the training sample and the category of the ground-truth.

The original StrengthNet was trained on three datasets, ESD (Zhou, Sisman, Liu, & Li, 2022), RAVDESS (Livingstone & Russo, 2018), and SAVEE (Ul-haq & Jackson, 2010). All three datasets contain parallel samples of the utterances spoken in several emotion categories. These samples are then used to create the <neutral, emotional> pairs for the training data. RAVDESS and SAVEE were only used for data augmentation purposes, and as this is outside the research niche of this thesis, they were not used. The ground truth for StrengthNet is determined by a separately trained ranking function, similar to (Zhu et al., 2019). The final output of StrengthNet predicts both the kind of emotion expressed in the utterance, as well as the strength of the emotion in a 5-dimensional vector. This is the difference in the embedding size between StrengthNet and the Zhu methodology: StrengthNet provides a score for each emotion for categorization and a separate overall emotion strength score. The Zhu model provides individual strength scores for each emotion, but no overall score.

3.2 *Data*

The dataset used in this thesis is the ESD dataset used in the original StrengthNet methodology. The EDS is a bilingual speech dataset that includes 10 native English speakers and 10 native Chinese speakers providing utterance samples. It consists of 350 parallel utterances that are each

Training Set Size	Number of Sample Sentences
4000	100
2000	50
1000	25
200	5
120	3
40	1

Table 1: Description of training set sizes for the first experiment

spoken in a neutral, happy, angry, sad and surprised tone, along with a text transcription which is labeled with the emotion of the utterance. Since every participant reads every sentence in all 5 emotions, the labels of the dataset are balanced exactly, with no missing data. This thesis uses the English data for the all five emotions, which adds up to a total of 17500 utterances for a length of approximately 14.5 hours.

3.3 *Experimental setup*

To answer the first research question, both models were trained on increasingly smaller data sets. The models were trained on the four emotion categories found in the ESD: happy, angry, sad and surprised. A random sample of 2500 sentences was taken from the 10 English speakers in the ESD dataset, and the associated audio files were taken as separate training sets all four emotion categories and the neutral categories. Sampling the sentences instead of individual utterances ensures that the sample is balanced across all emotion categories. As the Zhu model is trained on pairs of <neutral, emotional> data, this four separate training sets, each with feature vector representations of 2500 neutral sentences and 2500 sentences of the respective emotion category. Six models were trained, with the volume of training data decreasing as is visible in Figure 1. The data restriction was done by keeping the first N in the sample and removing the unrelated utterances. By training all models on the same subset of data , the results can be seen as a function of only the model. The same sentence sampling was used to create test and validation sets for the four emotional speech categories, but these were kept the same for all models and not restricted in size. There are no ground truth for the distribution of the emotion scores, but there are binary ground truths for the emotion label. Both models were evaluated on an emotion classification task, while the Zhu model was also evaluated on the emotional vs. neutral speech task that it was trained on. More on this in the Evaluation subsection.

As StrengthNet is a supervised deep learning model, it requires a ground truth for emotion scores. ESD does not provide this, so it was taken by using the ranking function trained using the Zhu et al. (2019) methodology, just like in the original paper. This could be seen as a methodological weakness: the ground truth for the second model depends on the accuracy of the first model, so if the latter performs poorly, the former may perform poorly as well. However, this is akin to what a real-life situation would look like when training these models for new or limited data.

The second experiment consists of a methodology similar to Rosenberg et al. (2019), and investigate the performance of models of trained on datasets with different levels of lexical diversity. To obtain text samples with varying degrees of lexical diversity, 100,000 random samples of 20 sentences out were taken from the ESD corpus. The measure of lexical diversity chosen is the Token Type Ratio (TTR) which is defined as the number of words in the vocabulary divided by the total number of words in the text (Tweedie & Baayen, 1998). The TTR of these samples was calculated, and the top 2 and bottom 2 samples were selected as training data for a total of four Zhu models. The sample with the average TTR score was taken as test data to evaluate the models. The choice for TTR was a practical one, as it is easier to calculate than the maximum entropy described in Rosenberg et al. (2019).

3.4 *Evaluation*

Evaluation for the Zhu model is not straightforward, since the task that they are trained on is not the task that they would perform in a TTS model. Regarding the Zhu model, the difficulty is that the model consists of four separate SVMs that have been trained on a neutral vs. emotion task, but the task that the model performs as the embedding in a TTS model is an emotion vs. emotion task. Arguably, the second task is the most important one, but to capture the overall performance, the models were evaluated for both tasks. The output of the models is a 4-dimensional vector with scores for each emotion, but the test data available only has binary emotion categories. To compare them, each model prediction was assigned the emotion category for which its emotion score was highest. As per the original methodology, the scores for each emotion were normalized on the [0,1] interval, eliminating any scale difference between the four emotion SVMs. As the StrengthNet model is not trained using neutral data, it will only be evaluated on the emotion category prediction task.

Evaluation for the second experiment was done on a separate test dataset, which consisted of ta , The evaluation was done using the same

methodology as in the first experiment, so both evaluating the emotion vs neutral task, as well as overall emotion prediction. The measure chosen for lexical diversity was TTR, as it is simple to calculate, but this experiment could be repeated with any measure for lexical diversity.

3.5 *Implementations*

The implementation of the StrengthNet model was taken directly from its GitHub repository [1] and is implemented using Tensorflow. Since was trained on the same dataset as the original paper, the default hyper parameters for StrengthNet were kept. The exception to this is the batch size. For the larger models, the batch size was kept at 64, for the models with 2000 and fewer samples, the batch size was set at 32 and for the smallest model the batchsize was set to 16. This implementation also generated the StrengthNet mel spectra input from the audio files. There are a variety of hyper parameters associated with the mel spectrum creation such as frame length and the distance between frames. Since the model was originally designed to be trained on the ESD dataset, the hyper parameters were not changed. All random sampling and text processing was done in Python, scripts for this are available on the GitHub page of this thesis [2]. The optimization problem was approximated using a pairwise Rank SVM as described in (Herbrich, Graepel, & Obermayer, 2000). An existing Python implementation was taken from [3]. This model creates the two pairwise matrices $\boldsymbol{O}$ and $\boldsymbol{S}$ to indicate whether two feature vectors belong to emotional speech or neutral speech. It then fits a linear support vector classifier using the sklearn implementation, taking the paired samples as input. It must be noted that this method of solving the optimization problem is quite memory intensive. This is inherent to the problem, as the model is required to keep both the $\boldsymbol{O}$ and $\boldsymbol{S}$ matrices in memory at all times. The hyper parameters for this model were kept at the default values of the sklearn model.

4 RESULTS

We trained two models with varying amounts of training data: a Zhu model and a StrengthNet model. Both of these models were evaluated on an emotional classification task. Moreover, the Zhu model was also evaluated on the emotion vs. neutral speech task. Lastly there is also the

[1] https://github.com/ttslr/StrengthNet
[2] https://github.com/haza97/Thesis-Project-Emotion-Embeddings
[3] https://gist.github.com/agramfort/2071994

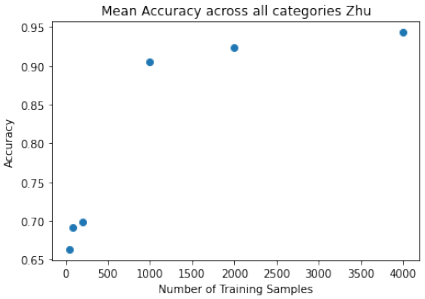


Figure 4: Mean accuracy of the Zhu model on the Emotion-Neutral Task

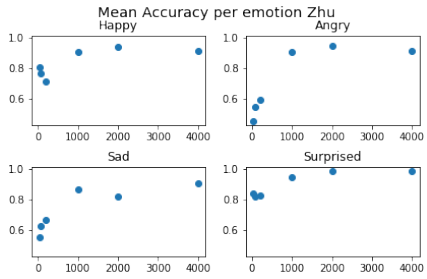


Figure 5: Per category accuracy of the Zhu model on the Emotion-Neutral Task

Number of training samples	40	120	200	1000	2000	4000
Happy Accuracy	0.807	0.769	0.714	0.903	0.942	0.969
Angry Accuracy	0.453	0.547	0.591	0.904	0.946	0.911
Sad Accuracy	0.553	0.627	0.668	0.867	0.821	0.903
Surprised Accuracy	0.841	0.822	0.823	0.947	0.983	0.988
Mean Accuracy	0.663	0.691	0.699	0.906	0.922	0.943

Table 2: Accuracy of the zhu model for the emotional vs. neutral task

lexical diversity model, which will also be evaluated on both an emotional vs. neutral task as well as an emotional classification task.

First the results for the Zhu model will be presented. Both table 3 and 2 clearly show that the model performs well on the emotion vs neutral task, with but only perform up to chance level at the emotion classification task. As the size of the training set increases, so does the performance on the emotional vs. neutral task. The same is not necessarily true for the emotional classification task; the performance shows no improvement with a larger training set; the mean performance is very constant. The per category accuracy shown in 6, where the red line stands for the performance of a random guess, is approximately the same, but shows higher variance.

Number of training samples	40	120	200	1000	2000	4000
Happy Accuracy	0.426	0.226	0.065	0.211	0.119	0.305
Angry Accuracy	0.045	0.304	0.136	0.213	0.042	0.197
Sad Accuracy	0.218	0.220	0.425	0.309	0.447	0.382
Surprised Accuracy	0.200	0.110	0.307	0.212	0.216	0.059
Mean Accuracy	0.222	0.215	0.233	0.236	0.206	0.237

Table 3: Accuracy of the zhu model for the emotional classification task

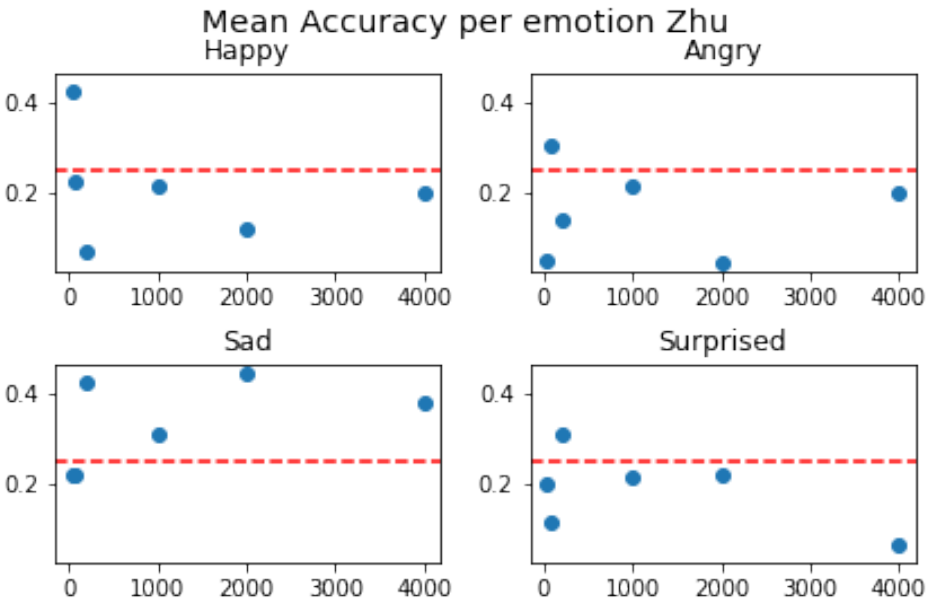


Figure 6: Per emotion accuracy on the emotion classification task

Number of training samples	40	80	200	1000	2000	4000
StrengthNet Accuracy	0.25	0.25	0.25	0.25	0.351	0.351

Table 4: StrengthNet Accuracy on emotion prediction task

The StrengthNet model was evaluated on an emotion prediction task, the results of which are visible in Figure 4 The results look strange, since all values for the smaller training sets are the same, and so are the values for the two biggest training sets. Upon closer inspection, it appears that for the smaller training sets, all the four weights for the emotion categories are very close to 0.25, more on this in the discussion section. The performance of the large training set models are slightly higher than chance, but still not as high as initially expected.

After taking the sentence samples and sorting by lexical diversity, the two most diverse samples had a TTR value of 0.658 and 0.654, while the least diverse samples had TTR values of 0.518 and 0.511 respectively.In the neutral-emotional task, the overall mean accuracy of the lexical diversity model shows no difference between high and low diversity models and both seem to perform the same across the different emotion categories. In the emotional classification task, both score around chance performance as is visible in table **??**

A simple error analysis was performed on the Zhu model emotional-neutral task. All confusion matrices show an approximately equal division

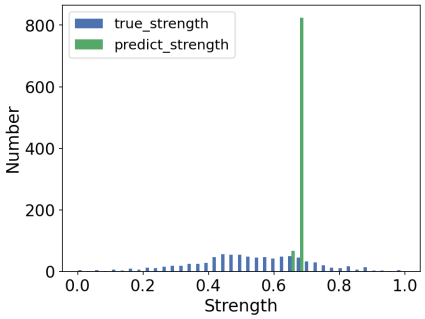


Figure 7: StrengthNet emotional strength prediction with a training set of 1000

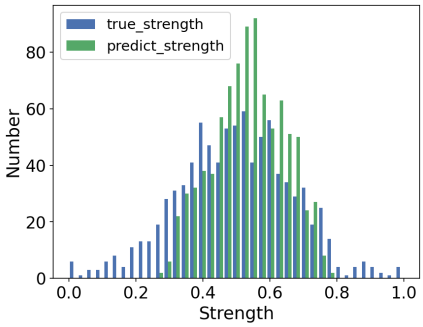


Figure 8: StrengthNet emotional strength prediction with a training set of 4000

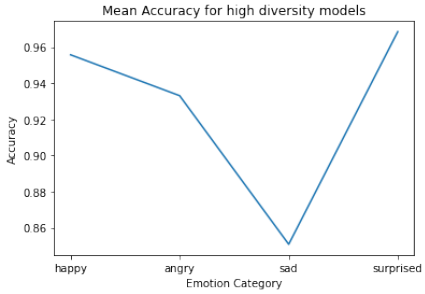


Figure 9: Mean accuracy of the high lexical diversity models on the Emotion-Neutral Task

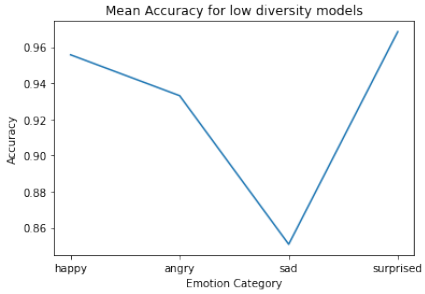


Figure 10: Mean accuracy of the low lexical diversity models on the Emotion-Neutral Task

Model Name	High div. 1	High div. 2	Low div. 1	Low div. 2
Happy Accuracy	0.284	0.067	0.378	0.135
Angry Accuracy	0.129	0.454	0.353	0.274
Sad Accuracy	0.351	0.242	0.250	0.238
Surprised Accuracy	0.103	0.217	0.077	0.170
Mean Accuracy	0.216	0.245	0.2644	0.204

Table 5: Accuracy of the lexical diversity models on the emotion classification task

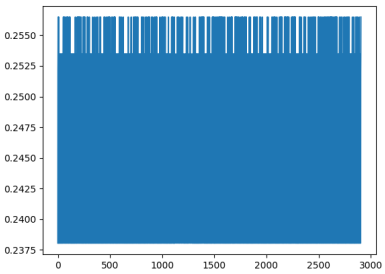


Figure 11: Strengthnet emotion scores for smaller training set models

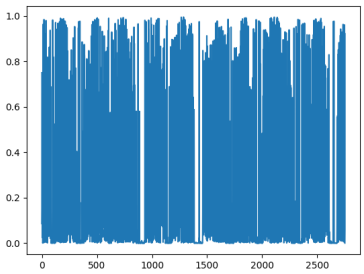


Figure 12: Strengthnet emotion scores for larger training set models

between True Positives and True negatives and between False Positives and False negatives. The confusion matrices can be found in appendix B .

5 DISCUSSION

The aim of this thesis was to determine the effect of training data limitations on the accuracy of two TTS embedding model architectures. The surprising finding in the results was the Zhu model and the linguistic diversity model performing poorly the emotion classification task while the performance on the emotional vs neutral speech was significantly better. This could indicates that learning the emotion vs. neutral speech task does not transfer well to the ranking problem. However there are a variety of other successful uses of the relative attributes methodology, so that does not seem likely (Singh & Lee, 2016; Souri, Noury, & Adeli, 2016). Regarding limiting the volume of training data, performance on the neutral vs. emotional speech task increased as the amount of input data increase. It appears that, for the low sample models (200 and fewer) the performance does not change as much as the number of samples increases, but once it reaches a certain threshold between 200 and 1000 samples, the model hits a high performance plateau, as is visible in Figure 4.

The results for StrenghtNet were also quite unexpected, that the result for the lower models were the exact same. Figure 11 shows all emotion scores for the StrengthNet model with 120 samples. It is clear that the variance in this signal is very small, especially when compared to 12. It may be that a training size this small is not close enough to any reasonable performance and that the model does not have enough data to learn to generalize anything. While training, all of the StrengthNet methods reached early stopping due to the validation error not shrinking past a certain point, indicating that run time was not a problem. One other problem here is that the performance of the Zhu model is directly related

to that of StrengthNet. StrengthNet uses the predictions from the Zhu model as its ground-truth, which will have negatively affected performance in this case. What is also of interest are figures 8 and 7. They compare the distribution of the predicted emotion intensity score for two different sizes of the training set. On the model with little training data, it overfits on one value, which is consistent with the poor accuracy results we saw for that model in the emotion classification task. However, the other model shows a larger variance in the predicted emotion strength, which is not reflected in the performance on the emotion classification task.

In the case of lexical diversity, there were no results that show that lexical diversity provides for higher quality training data. Reasons for this may be that the difference in lexical diversity between the samples was not high enough. This would be a hard problem to solve in this case, since the models trained here depend on parallel samples datasets, which are rare. Another possible explanation is that the model is unable to solve the problem for the reasons mentioned previously for the Zhu model, and that the quality of input data does not make a difference at such low levels of performance.

Coming back to the research questions, it is hard to answer the first question, due to the low overall performance of the models. It has at least been shown that there is a lower bound to the size of the training set in order for the model to achieve more than chance performance. At datasets of this size the lexical diversity does not affect the accuracy at all, although this also may be related to the poor overall performance.

5.1 *Limitations*

One could argue that StrengthNet and the Zhu et al. (2019) methodology could have been explored using more samples from the ESD. It is true that the ESD contains more samples that were not used, however this would lead to a large increase in required memory. As the dataset gets larger, the number of pairwise comparisons between the $\boldsymbol{O}$ and $\boldsymbol{S}$, increases quadratically. When applying Newton's method to the entire ESD, the total size in memory went up to 113 Gb of RAM, which was unavailable to me. It was attempted to move these computations to GPU memory using the Cupy package for Python, but time restrictions did not allow this to be completed. As noted in Parikh and Grauman (2011), the Relative Attributes optimization problem can be seen as a variation on ranking Support Vector Machines, but trained on pairwise data (Joachims, 2002). To solve the RAM problem, it could have been worthwhile to look into SVM literature to find a similar methodology with a more straightforward solution to batch learning.

5.2 *Future Research*

The current experiment leaves several directions unexplored. One is the effect of training time on the (Zhu et al., 2019) methodology. The optimization algorithm iterates over its samples in Newton's method, but while solving this also iterates over its samples to solve the non-linear least squares problem. As a result, the total algorithm is of O^2 complexity. This raises the question of how well this model could perform with tighter hyper parameters and more time. Empirically, the StrengthNet methodology does not appear to have a need for longer training time, as it never reached its maximum number of training epochs but always ended on early stopping. Previously, we described how using end-to-end models removes the requirement for explicitly labeled data. This is an advantage, but it does not take into account that the input data used in StrengthNet and the Zhu model are now restricted to datasets with parallel samples only, which are also rare. This seems like a restriction that future TTS models should take on.

6 CONCLUSION

This thesis investigated the performance of two emotional embedding models for TTS systems under limited training data. Training data was both limited in amount, as well as in quality, with lexical diversity being a proxy for quality. The two models trained were StrengthNet (Liu et al., 2021b) and a model based on Zhu et al. (2019). The Zhu model is trained using both a low amount of training data and low quality training data data, while the StrengthNet model is trained only on a low amount of training data. Both Zhu models perform well on their training task of classifying emotional and neutral speech, but perform at chance levels when classifying all four emotions. The StrengthNet model performs poorly, the low amount of training data is most likely too low to reach any good performance on.

REFERENCES

Akagi, M., Han, X., Elbarougy, R., Hamada, Y., & Li, J. (2014). Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages. In *Signal and information processing association annual summit and conference (apsipa), 2014 asia-pacific* (p. 1-10). doi: 10.1109/APSIPA.2014.7041623

Brunnstrom, K., Hands, D., Speranza, F., & Webster, A. (2009). Vqeg validation and itu standardization of objective perceptual video quality

metrics [standards in a nutshell]. *IEEE Signal processing magazine*, *26*(3), 96–101.

Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Isca tutorial and research workshop (itrw) on speech and emotion.*

Cai, X., Dai, D., Wu, Z., Li, X., Li, J., & Meng, H. (2020). *Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition.* arXiv. Retrieved from https://arxiv.org/abs/2010.13350 doi: 10.48550/ARXIV.2010.13350

Campbell, N., & Black, A. W. (1997). Prosody and the selection of source units for concatenative synthesis. *Progress in Speech Synthesis*, 279–292. doi: 10.1007/978-1-4612-1894-4_22

Chen, M., Chen, M., Liang, S., Ma, J., Chen, L., Wang, S., & Xiao, J. (2019). Cross-Lingual, Multi-Speaker Text-To-Speech Synthesis Using Neural Speaker Embedding. In *Proc. interspeech 2019* (pp. 2105–2109). doi: 10.21437/Interspeech.2019-1632

Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2018). *A tutorial on deep learning for music information retrieval.*

Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2018). Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357-366. doi: 10.1109/TASSP.1980.1163420

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv. Retrieved from https://arxiv.org/abs/1810.04805 doi: 10.48550/ARXIV.1810.04805

Ding, L., & O'Shaugnhnessy, D. (2003). 2.2 analysis based on linear predictive coding. In *O'shaugnhnessy* (p. 41–48). Marcel Dekker.

Edward, S. (2018, 07). Text-to-speech device for visually impaired people. *International Journal of Pure and Applied Mathematics*, *119*.

Effendi, J., Tjandra, A., Sakti, S., & Nakamura, S. (2020). *Augmenting images for asr and tts through single-loop and dual-loop multimodal chain framework.* arXiv. Retrieved from https://arxiv.org/abs/2011.02099 doi: 10.48550/ARXIV.2011.02099

Elyasi, M., & Bharaj, G. (2021). Flavored tacotron: Conditional learning for prosodic-linguistic features. *arXiv preprint arXiv:2104.04050*.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th acm international conference on multimedia* (pp. 1459–1462).

Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. (2021). A brief review of domain adaptation. *Advances in data science and information engineering*, 877–894.

Fujisaki, H. (2004, 01). Information, prosody, and modeling-with emphasis on tonal features of speech. *Scientific Programming - SP*.

Hayashi, T., Watanabe, S., Toda, T., Takeda, K., Toshniwal, S., & Livescu, K. (2019). Pre-trained text embeddings for enhanced text-to-speech synthesis. In *Interspeech* (pp. 4430–4434).

Herbrich, R., Graepel, T., & Obermayer, K. (2000, 01). Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, *88*.

Huybrechts, G., Merritt, T., Comini, G., Perz, B., Shah, R., & Lorenzo-Trueba, J. (2021). Low-resource expressive text-to-speech using data augmentation. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6593–6597).

Hwang, M.-J., Yamamoto, R., Song, E., & Kim, J.-M. (2021). Tts-by-tts: Tts-driven data augmentation for fast and high-quality speech synthesis. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6598–6602).

Ito, K., & Johnson, L. (2017). *The lj speech dataset.* https://keithito.com/LJ-Speech-Dataset/.

J, A. K., Trueman, T. E., & Cambria, E. (2021, Nov 01). A convolutional stacked bidirectional lstm with a multiplicative attention mechanism for aspect category and sentiment detection. *Cognitive Computation*, *13*(6), 1423-1432. Retrieved from https://doi.org/10.1007/s12559-021-09948-0 doi: 10.1007/s12559-021-09948-0

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (p. 133–142). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.tilburguniversity.idm.oclc.org/10.1145/775047.775067 doi: 10.1145/775047.775067

Kano, T., Sakti, S., Takamichi, S., Neubig, G., Toda, T., & Nakamura, S. (2012). A method for translation of paralinguistic information. In *Proceedings of the 9th international workshop on spoken language translation: Papers.*

Khatua, A., Khatua, A., & Cambria, E. (2020). Predicting political sentiments of voters from twitter in multi-party contexts. *Applied Soft Computing*, *97*, 106743. Retrieved from https://www.sciencedirect.com/science/article/pii/S1568494620306815 doi: https://doi.org/10.1016/j.asoc.2020.106743

Kwon, O., Jang, I., Ahn, C., & Kang, H.-G. (2019). An effective style

token weight control technique for end-to-end emotional speech synthesis. *IEEE Signal Processing Letters*, *26*(9), 1383-1387. doi: 10.1109/LSP.2019.2931673

Laptev, A., Korostik, R., Svischev, A., Andrusenko, A., Medennikov, I., & Rybin, S. (2020). You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)* (p. 439-444). doi: 10.1109/CISP-BMEI51763.2020.9263564

Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S., & Klimkov, V. (2019). Effect of data reduction on sequence-to-sequence neural tts. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 7075-7079). doi: 10.1109/ICASSP.2019.8682168

Liu, R., Sisman, B., & Li, H. (2021a). Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. *arXiv preprint arXiv:2104.01408*.

Liu, R., Sisman, B., & Li, H. (2021b). Strengthnet: Deep learning-based emotion strength assessment for emotional speech synthesis. *arXiv preprint arXiv:2110.03156*.

Livingstone, S. R., & Russo, F. A. (2018, April). *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.1188976 (Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak) doi: 10.5281/zenodo.1188976

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H.-M. (2019). Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). *Wavenet: A generative model for raw audio.* arXiv. Retrieved from https://arxiv.org/abs/1609.03499 doi: 10.48550/ARXIV.1609.03499

O'Shaughnessy, D., Barbeau, L., Bernardi, D., & Archambault, D. (1988). Diphone speech synthesis. *Speech Communication*, *7*(1), 55-65. Retrieved from https://www.sciencedirect.com/science/article/pii/0167639388900210 doi: https://doi.org/10.1016/0167-6393(88)90021-0

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345-1359. doi: 10.1109/TKDE.2009.191

Parikh, D., & Grauman, K. (2011, 11). Relative attributes. In (p. 503-510).

doi: 10.1109/ICCV.2011.6126281

Plutchik, R. (1984). Emotions and imagery. *Journal of Mental Imagery*.

Prajwal, K. R., & Jawahar, C. V. (2021). Data-efficient training strategies for neural tts systems. In *8th acm ikdd cods and 26th comad* (p. 223–227). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.tilburguniversity.idm.oclc.org/10.1145/3430984.3431034 doi: 10.1145/3430984.3431034

Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., & Wu, Z. (2019). Speech recognition with augmented synthesized speech. In *2019 ieee automatic speech recognition and understanding workshop (asru)* (pp. 996–1002).

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178. Retrieved from https://doi.org/10.1037/h0077714 doi: 10.1037/h0077714

Schuller, B., Steidl, S., & Batliner, A. (2009, 01). The interspeech 2009 emotion challenge. In (p. 312-315).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... others (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4779–4783).

Singh, K. K., & Lee, Y. J. (2016). End-to-end localization and ranking for relative attributes. In *European conference on computer vision* (pp. 753–769).

Sonobe, R., Takamichi, S., & Saruwatari, H. (2017). Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*.

Souri, Y., Noury, E., & Adeli, E. (2016). Deep relative attributes. In *Asian conference on computer vision* (pp. 118–133).

Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, *22*(2), 213–227.

Szekely, E., Steiner, I., Ahmed, Z., & Carson-Berndsen, J. (2013, 03). Facial expression-based affective speech translation. *Journal on Multimodal User Interfaces*, *8*. doi: 10.1007/s12193-013-0128-x

Székely, , Henter, G. E., & Gustafson, J. (2019). Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector. In *Icassp 2019 - 2019 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 6925-6929). doi: 10.1109/ICASSP.2019.8683846

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Tilly, S., Ebner, M., & Livan, G. (2021). Macroeconomic forecasting through

news, emotions and narrative. *Expert Systems with Applications*, *175*, 114760. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417421002013 doi: https://doi.org/10.1016/j.eswa.2021.114760

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, *101*(5), 1234-1252. doi: 10.1109/JPROC.2013.2251852

Turk, O., Schröder, M., Bozkurt, B., & Arslan, L. M. (2005). Voice quality interpolation for emotional text-to-speech synthesis. In *Ninth european conference on speech communication and technology.*

Tweedie, F. J., & Baayen, R. H. (1998, Sep 01). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323-352. Retrieved from https://doi.org/10.1023/A:1001749303137 doi: 10.1023/A:1001749303137

Ul-haq, S., & Jackson, P. (2010, 01). Multimodal emotion recognition. *Machine Audition: Principles, Algorithms and Systems*. doi: 10.4018/978-1-61520-919-4.ch017

Velsquez, J. (1997). Modeling emotions and other motivations in synthetic agents. *Aaai/iaai*, 10–15.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., ... Saurous, R. A. (2018, 10–15 Jul). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 5180–5189). PMLR. Retrieved from https://proceedings.mlr.press/v80/wang18h.html

Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and esd. *Speech Communication*, *137*, 1–18.

Zhu, X., Yang, S., Yang, G., & Xie, L. (2019). Controlling emotion strength with relative attribute for end-to-end speech synthesis. In *2019 ieee automatic speech recognition and understanding workshop (asru)* (p. 192-199). doi: 10.1109/ASRU46091.2019.9003829

APPENDIX A: OPENSMILE CONFIGURATION

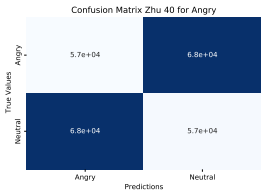
The OpenSMILE INTERSPEECH 2009 Emotion challenge feature set contains 16 low level descriptors, both in absolute values as well as their rate of change. Rather than expressing these for every frame, they are expressed using twelve functionals. Table 6 includes an overview of the feature set. This makes for a total of 2 * 24 * 12 = 384 features (Schuller et al., 2009).

Functionals (12)	Low Level Descriptors (16)
Mean	Zero-crossing-rate
Standard Deviation	Root mean square of frame energy
Kurtosis	Pitch frequency
Skewness	Harmonics to noise ratio
Minimum	Mel Frequency Coefficient (MFC) 1 through 16
Maximum	
Relative position	
Range	
Linear Regression Offset	
Linear Regression Slope	
Linear Regression Mean Squared Error	

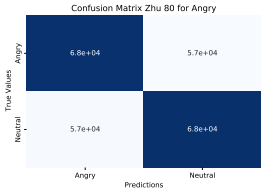
Table 6: OpenSMILE INTERSPEECH 2009 Emotion challenge features

APPENDIX B: CONFUSION MATRICES

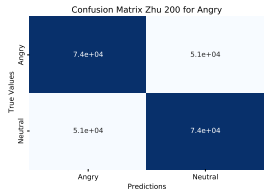
Confusion Matrix 1



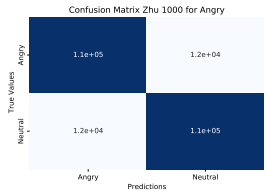
Confusion Matrix 2



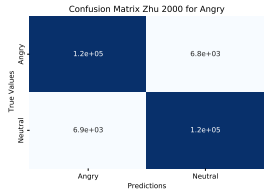
Confusion Matrix 3



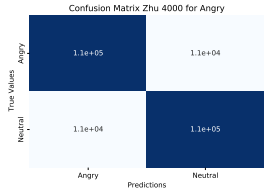
Confusion Matrix 4



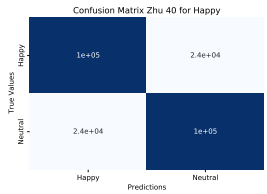
Confusion Matrix 5



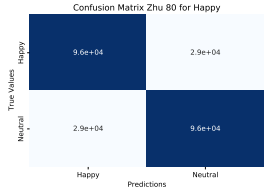
Confusion Matrix 6



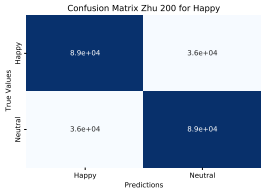
Confusion Matrix 7



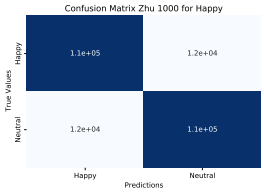
Confusion Matrix 8



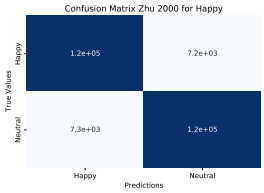
Confusion Matrix 9



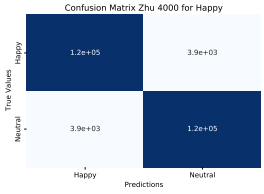
Confusion Matrix 10



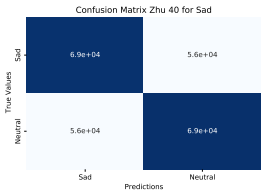
Confusion Matrix 11



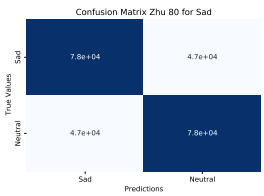
Confusion Matrix 12



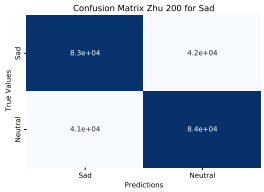
Confusion Matrix 13



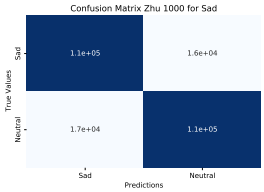
Confusion Matrix 14



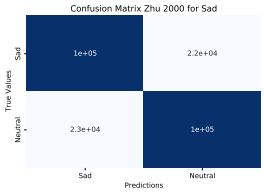
Confusion Matrix 15



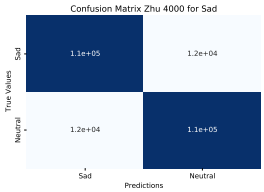
Confusion Matrix 16



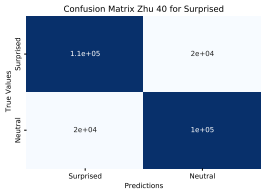
Confusion Matrix 17



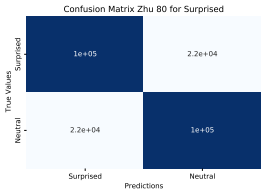
Confusion Matrix 18



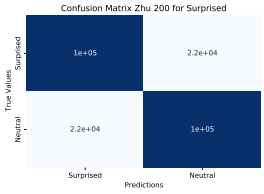
Confusion Matrix 19



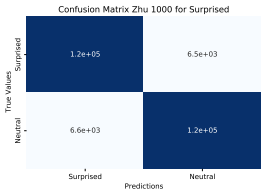
Confusion Matrix 20



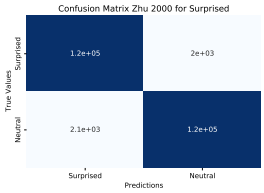
Confusion Matrix 21



Confusion Matrix 22



Confusion Matrix 23



Confusion Matrix 24