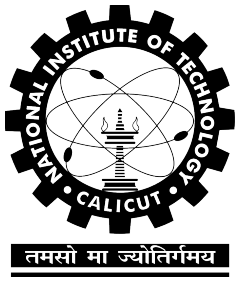
Natural and Emotional Linguistic Text to Speech Synthesis

CS4090 Project Final Report

Submitted by

TGDK Sumanathilaka	**B150413CS**
Uddhav Raj	**B150878CS**
Viggnah Selvaraj	**B150080CS**
Venkatesh Raju	**B151040CS**

Under the Guidance of
Dr. Jay Prakash

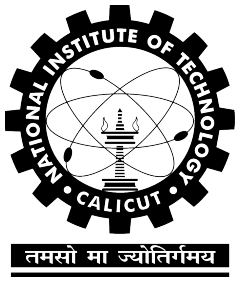


Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

April 26, 2019

NATIONAL INSTITUTE OF TECHNOLOGY CALICUT, KERALA, INDIA - 673 601

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



2019

CERTIFICATE

Certified that this is a bonafide record of the project work titled

NATURAL AND EMOTIONAL LINGUISTIC TEXT TO SPEECH SYNTHESIS

done by

TGDK Sumanathilaka
Uddhav Raj
Viggnah Selvaraj
Venkatesh Raju

of eighth semester B. Tech in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of the National Institute of Technology Calicut

Project Guide	**Head of Department**
Dr. Jay Prakash	Dr. Saleena N
Assisstant Professor	Associate Professor

DECLARATION

I hereby declare that the project titled, **Natural and Emotional Linguistic Text to Speech Synthesis**, is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or any other institute of higher learning, except where due acknowledgement and reference has been made in the text.

Place :
Date :

Signature :
Name :
Reg. No. :

Abstract

A real time platform to synthesize emotional and natural voice from unstructured text was built. The whole system has two parts. The first part deals with extracting emotion out of text using text mining and LSTMs which is the cutting-edge technology in the current scenario. The model thus built was flexible enough to obtain high accuracy over multiple datasets. The second part identifies the key features of speech and modifies those attributes in a neutral speech to give it an emotional and human-like sound. To identify these exclusive features, an SER(Speech Emotion Recognition) model was also built.

ACKNOWLEDGEMENT

We would like to thank our guide, Dr. Jay Prakash, for always being supportive throughtout the entire year. He listened patiently to all our concerns and queries, and never failed to guide us in the correct direction during confused moments.

We would also like to thank our panel members, Dr. Gopakumar and Dr. Saleena, for providing insightful comments and suggestions during the reviews and evaluations that have really helped us improve our project.

We gratefully acknowledge the time and effort invested by our guide, panel members and other faculty of our department, to help us finish this project successfully.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In affective computing, the field of involving emotions in computing, emotion detection from text has emerged as an important domain. Its applications are wide-ranging; measuring the emotional closeness of interpersonal ties using affective language in social networks, in marketing, to predict purchase intentions of customers and gauge brand reputation using emotional states [4], removing inappropriate posts from social media and, for online psychologists to assist their patients by analysing transcripts for affective content [1].

Emotion detection from text neatly fits as a crucial intermediate step in many applications, one particular application which we would like to highlight is emotional text-to-speech (TTS) synthesis since there is growing interest in the community about this task.

The idea of speech synthesis has been around for a long time. Text-to-speech synthesis (TTS) is to date a challenging task because voice produced by these systems sound robotic and are easily distinguishable from human voices. Even though there has been some considerable research to generate natural sounding voice, generating emotional speech is still a relatively new field. Emotional TTS has many applications like assisting the visually impaired, and emotion detection from text is an important module in this process.

The process of emotion detection starts from defining what emotions are exactly. There is no general consensus among psychologists as to how to define and categorise emotions [5]. There are various models like the categorical and dimensional models of emotions. We will attempt to generalise our model such that with only minor modifications it could be used on datasets which use both models.

Over the years many computational approaches for this problem have been proposed, such as lexicon-based, and machine learning based. Deep learning methods are becoming popular, since the rapid rise in computing power, and have shown promising results.

Chapter 2

Literature Survey

2.1 Emotion Detection from Text

The "emotion detection from text" pipeline primarily consists of three parts; 1) Choosing an emotion model to follow, 2) Identifying and aggregating relevant datasets for the emotion model chosen and, 3) Applying a computational approach to perform the task of accurately determining emotions given text.

2.1.1 Emotion Models

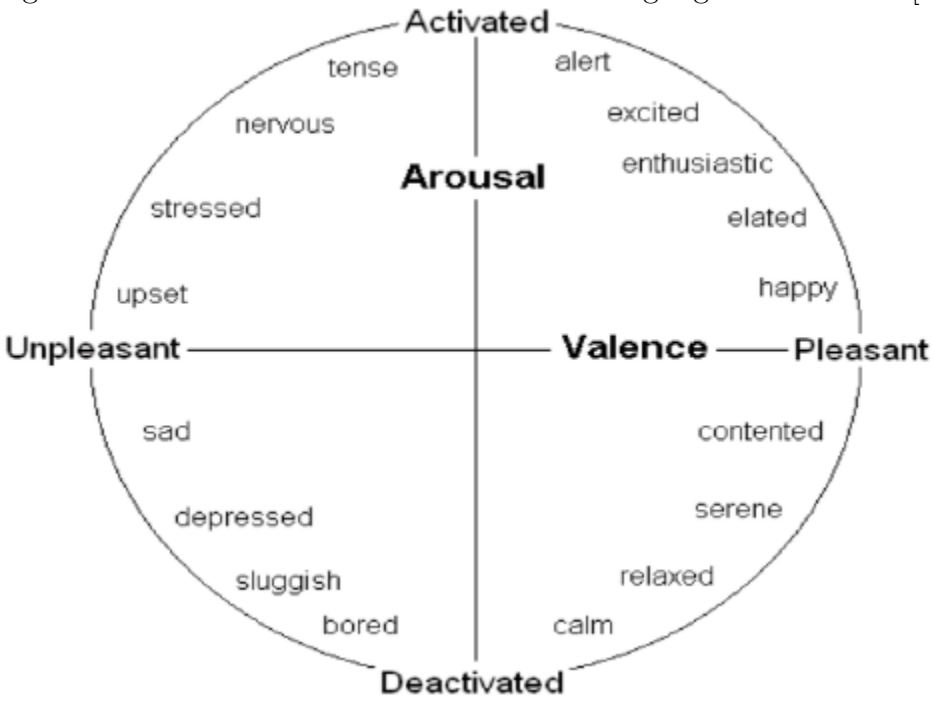
The emotional models in use today in the field of emotion detection can be broadly categorised into three types [5]:

- Categorical: Assumes humans capable of only a limited, finite set of emotions. Table 2.1 provides an overview of several categorical models.
- Dimensional: Emotions represented as points in dimensional space, subject to variables like valence and arousal. Eg: Russells Circumplex model (valence and arousal), Mehrabians model (pleasure, arousal and dominance). Figure 2.1 shows various emotions in the Russell model.
- Extended models: Takes into account personality, targets and desires of the communicating party, helpful for individual-based emotion recognition.

Table 2.1: Categories of emotions [1]

Author	Count	Emotions
Ekman	6	anger, disgust, fear, joy, sadness, surprise
Parrot	6	anger, fear, joy, love, sadness, surprise
Frijda	6	desire, happiness, interest, surprise, wonder, sorrow
Plutchik	8	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins	9	desire, happiness, interest, surprise, wonder, sorrow
Izard	12	interest, joy, surprise, sadness, anger, disgust, contempt, self-hostility, fear, shame, shyness, guilt
Extended Ekman	18	anger, disgust, fear, joy, sadness, surprise, amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure, and shame

Figure 2.1: Russell's dimensional model with highlighted emotions [1]



2.1.2 Datasets for Emotion Detection

Some datasets available for the task of emotion detection are given [1, 4]:

- SemEval2007-Task: News headlines tagged numerically (dimensional) with Ekmans six basic emotions.

- ISEAR: Self-reported experiences of people which includes seven emotions.
- Fairy Tales: Database of stories and fairy tales labelled at the sentence level.
- Election Tweets: Categorically labelled, based on 8 emotions.
- SemEval2018: Moderate and high emotion tweets annotated with a single dimension for four separate emotions.
- Facebook Posts: Annotated using the Circumplex model on two dimensions, valence and arousal

Preprocessing Data

In the literature, a number of authors also preprocess the data to so as to normalize it. The steps involved in the process are:

- Remove punctuation
- Remove repeated characters
- Negative expression replacer
- Stop word removal - removing words that occur very frequently like the, and etc.
- Stemming - Reducing inflectional forms of words
- Lemmatization

2.1.3 Computational Methods

In the literature, many computational methods have been proposed to detect emotions accurately given text. These methods can be broadly categorised into three approaches; 1) Lexicon-based, 2) Machine learning and, 3) Deep learning.

Lexicon-based Methods

The detection of the emotion of words is done primarily using words related to different emotions. In a keyword-based approach, an emotion has particular words associated with it that help in classifying the sentence to one of the emotions. Another method assigns a probabilistic affinity for a particular emotion to arbitrary words rather than detecting predefined emotional keywords from a text. This method called "lexical affinity" is an extension of the keyword-based method [6]. These methods are quite basic and face challenges such as: Ambiguity in the way keywords are defined, inability to recognize sentences without keywords and lack of linguistic information [7]. [8] is an example of a keyword-based implementation using POS (part of speech) tagging as well.

Machine Learning

Machine learning can be broadly defined as inference of decision rules from a database of labeled training samples for the task of recognizing emotions [4]. This overcomes challenges posed by lexicon-based methods. Random forests and support vector machines are commonly used as models in this approach. Usually, these models are used in conjunction with linguistic features to increase accuracy [9].

Supervised learning approaches rely on a labelled training data. The supervised learning algorithm analyses the training data and infers a function, which we use for mapping new examples. In unsupervised learning, the algorithms attempt to detect hidden structures in unlabeled data in order to build models for emotion classification [10]. It tries to compute an emotional vector for words based on semantic relatedness to other words [1].

Deep Learning

This a relatively new approach as the emergence of deep learning in the recent past has motivated researchers to try it out in a variety of domains, including natural language processing (NLP) [11].

Kratzwald et al. [4] attempt to use deep learning to improve the accuracy of emotion detection across various datasets. They use a recurrent neural network architecture (LSTM in uni and bidirectional), with dropout and a weighted loss function, trained on word embeddings (GloVe). Note that these word embeddings

Table 2.2: Emotion detection approaches (lexicon and learning-based) [3]

Papers	Categories	Emotion Model	Approaches
(Morinaga, et al.,2002)	Positive , Negative	Categorical	Rule-based
(Alm et al 2005)	Anger, Disgust, Fear, Happiness, Sadness, Positively Surprise, Negatively Surprise	Categorical	Supervised Learning based
(Neviarouskaya, etal., 2007)	Anger, Disgust, Fear, Guilt, Interest, Joy, Sadness, Shame, Surprise, Intensity	Hybrid	Rule-based
(S. Aman and S.Szpakowicz ,2007)	anger, disgust, fear, joy, sadness, surprise	Categorical	Hybrid
(Strapparava and Mihalcea 2008)	Anger, Disgust, Fear, Joy, Sadness, Surprise	Categorical	Lexical based
(Gill et al 2008)	Anger, Fear, Surprise, Joy, Anticipation, Acceptance, Sadness, Disgust	Categorical	Lexical based
(Strapparava and Mihalcea, 2008)	Anger, Disgust, Fear, Joy, Sadness, Surprise	Categorical	Unsupervised Learning based
(Strapparava and Mihalcea 2008)	Anger, Disgust, Fear, Joy, Sadness, Surprise	Categorical	Supervised Learning based
(Balahur et al 2011)	Anger, Disgust, Fear, Guilt, Joy, Sadness, Shame	Categorical	Lexical based
(Balabantaray et al 2012)	Anger, Disgust, Fear, Happiness, Sadness, Surprise	Categorical	Supervised Learning-based
(Roberts et al., 2012)	Anger, Disgust, Fear, Joy, Sadness, Surprise, Love	Categorical	Supervised Learning based
(Agrawal and An, 2012)	Anger, Disgust, Fear, Happiness, Sadness, Surprise	Categorical	Unsupervised Learning based
(Sykora et al., 2013)	Anger, Confusion, Disgust, Fear, Happiness, Sadness, Shame, Surprise	Categorical	Lexical based
(Wang and Zheng, 2013)	Anger, Disgust, Fear, Guilt, Joy, Sadness, Shame	Categorical	Lexical based
(Suttles and Ide, 2013)	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Trust, Anticipation	Categorical	Supervised Learning based
(Calvo and Kim, 2013)	Anger-Disgust, Fear, Joy, Sadness	Categorical	Unsupervised Learning based
(Calvo and Kim, 2013)	Anger-Disgust, Fear, Joy, Sadness	Dimensional	Unsupervised Learning based
(Hasan et al, 2014b)	Happy-Active, Happy-Inactive, Unhappy-Active, Unhappy-Inactive	Dimensional	Supervised Learning based

do not take sentiment into account unlike works such as [12–14] which attempt to incorporate sentiment features also into the embedding. They introduce a novel method in transfer learning where they transfer neural network parameters trained on sentiment analysis (only positive or negative emotion) to the task of emotion detection.

Majumder et al. [15] use word-level to sentence-level to document-level aggregation to classify documents into one of five personalities. They use a CNN to extract features, it is important to note that they trained five different networks for each of

Table 2.3: Emotion detection approaches (deep learning)

Papers	Categories	Emotion Models	Approach
(Potapova & Gordeev, 2016)	Aggression	Dimensional	Deep Learning - CNN
(Felbo et al., 2017)	Emoticons	Categorical	Deep Learning - LSTM
(U. Gupta et al., 2017)	Anger, Disgust, Fear, Joy, Sadness	Categorical	Deep Learning - custom LSTM
(Meisheri et al., 2017)	Anger, Fear, Joy, Sadness	Categorical - Intensity	Deep Learning - LSTM

the personality types.

2.2 Text-to-Speech Synthesis

The typical TTS pipeline consists of two parts: 1) Text analysis and 2) Speech synthesis [16]. Text analysis, also called the frontend deals with NLP (Natural Language Processing) to extract the semantic and syntactic meaning of the text to convert it into phonemes and other intermediate representations that can be consumed in the next stage to produce speech waveforms. The next stage, the backend, is done with a vocoder, which usually takes parametric input and generates speech.

2.2.1 Techniques used in TTS

Over the years many techniques have been used for TTS. Some of the major techniques are:

- Concatenative speech synthesis with unit selection, where small units of pre-recorded waveforms are tied together to form a single coherent waveform. Usually phonemes, diphones or phrase-based audio act as units. These are retrieved from a database according to the requirement of the input text [17].
- Statistical parametric speech synthesis, directly generates smooth trajectories of speech features to be synthesised by a vocoder. This attempts to solve boundary artifacts (at the point of concatenation) which occur with the use of concatenative speech synthesis [16].

Figure 2.2: Schematic diagram of neutral speech synthesis [2]

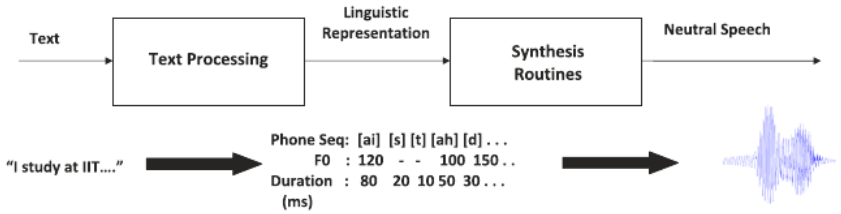
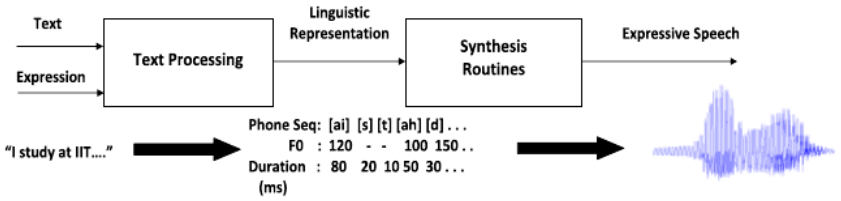


Figure 2.3: Schematic diagram of expressive speech synthesis [2]



- Since we now possess very high computing power, deep learning methods are becoming popular, and have shown promising results. Wavenet [16], Char2Wav [16] and Deep Voice 3 [18] are attempts at producing natural sounding speech from text using deep learning techniques.

2.2.2 Synthesizing Emotional Speech

After detecting the emotion, an approach for expressing emotion in the speech synthesizer is presented in [19]. Initially the model uses several linguistic resources which can be used to recognize emotions in a text and assigns appropriate parameters to the synthesizer to carry out a suitable speech synthesis. In order to incorporate the linguistic information an XML based markup language is used for audio script building [19].

In the paper [2] the authors review the current methods of emotional speech synthesis, with particular focus on "explicit control. In this technique, as shown in figures 2.2 and 2.3 a text processing stage is done and a linguistic representation obtained. Speech synthesis is done to obtain neutral speech which is then "explicitly controlled" to get emotional speech. The parameters upon which such speech can be modified are mainly prosodic and excitation parameters [2]. These parameters can be incorporated to the neutral speech using prosody modification algorithms like Over Lap Add (OLA), Synchronous Over Lap Add (SOLA) and Pitch Synchronous

Over Lap Add (PSOLA) [2].

Deep Learning Architecture

The paper [16] proposes a generative model for the second part of the TTS pipeline, speech synthesis. The model, "Wavenet", is conditioned on linguistic features such as phone, syllable, word, phrase, and utterance-level features (e.g. syllable stress, position of the current syllable in a phrase, the number of syllables in a word, and phone identities) which are derived from input texts and also on the logarithmic fundamental frequency ($\log F_0$.) Its architecture consisted of dilated causal convolutions in order to increase the receptive field and also to take care of long-range temporal dependencies [16].

There have been recent attempts to incorporate a neural netowork architecture for the whole piplie. Char2Wav is one of the first attempts at an end-to-end speech synthesis model that is capable of producing speech from text directly. It consists of two parts: 1) Reader and a 2) Neural vocoder. The main point of deviation from the classical pipeline of TTS, is that this model does not generate intermediate linguistic features, rather it directly generates vocoder acoustic features which are fed to the neural vocoder. The reader is a bidirectional recurrent neural network and the neural vocoder is a recurrent neural network (RNN) with attention [9].

Tacotron is another model that is an end-to-end speech synthesis model that is capable of producing speech from text without the need for intermediate generation of linguistic features. At a high level, the model takes as input characters and outputs spectrogram frames, which are then converted to waveforms. The architecture comprises of a convolution network which takes as input - character embeddings, and makes use of a seq2seq model with attention to output linear spectrograms which are reconstructed into speech waveforms using the Griffin-Lim algorithm. This model managed to outperform the parametric system of TTS, but was still short of the score achieved by the concatenative system [20].

Tacotron 2 is an improvement to the architecture of Tacotron, that showed immense improvement in producing natural sounding speech. The main difference from "Tacotron" was that the speech synthesis using the Griffin-Lim algorithm was replaced by a neural vocoder, namely a seq2seq network with attention that predicts mel spectrogram frames from an input character sequence, and 2) Wavenet that generates time-domain waveforms as it has been trained (conditioned) on mel

spectrogram frames. Despite its increased naturalness the authors noted that some synthesised sentences contained unnatural prosody, and that therefore there was room for improvement in prosody modelling [21].

Deep Voice 3 is a fully-convolutional attention-based neural TTS system unlike the previous attempts that made use of recurrent neural networks (RNN) [18]. The authors also note the shortcomings of attention-based speech synthesis networks, and compare several different waveform synthesis methods like Griffin-Lim vocoder, WORLD vocoder [22] and Wavenet [16].

The methods and papers described above tried to improve the synthesis of natural sounding speech. The architectures did not consider the incorporation of emotion to the speech. The authors of [23] tried to adapt Tacotron to emotional speech synthesis. They noted that the original Tacotron faces problems when producing long wave outputs as noise from time-steps are accumulated in subsequent steps. They try to overcome this problem by improving the attention alignment using monotonic attention from Raffel et al [24]. Their model contained a modified Tacotron, as an end-to-end emotional speech synthesizer that takes the required emotion and the character sequence as input and generates the corresponding wave signal. The authors also note that the model suffers from low speed due to the Griffin-Lim reconstruction; and that improvements can be made to produce a dynamic TTS capable of producing speech that display different personalities.

Chapter 3

Problem Definition

Synthesise Natural and human-like audio from unstructured text. Firstly, classify text into emotions across multiple datasets that use related annotation strategies by having a flexible model. The model should at least match or improve upon benchmarks currently set for these datasets. Next, classify audio into emotions, and analyse the features contributing to the presence of a particular emotion in the audio. Modify these features in a neutral audio to get a emotional and human-like sound.

Chapter 4

Methodology

4.1 Project Design

4.1.1 Emotion Detection from Text

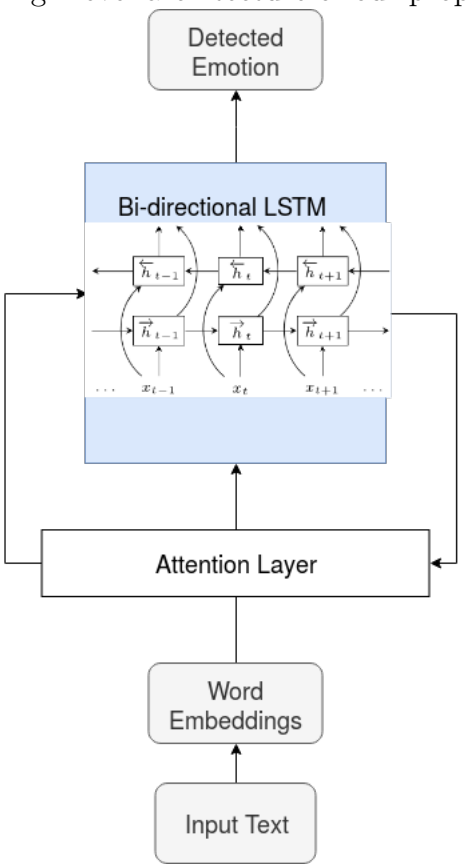
Our design was flexible enough take into account two of the predominant emotional models: categorical and dimensional. We used a deep learning architecture to accurately identify emotions in datasets.

In order to convert the words into a suitable intermediate representation that can be fed into the neural network, we used word embeddings. We also incorporated latent sentiment features into these embeddings as in [12–14] to improve performance. It is important to note that in text processing usually stop words (commonly occurring words) are removed and stemming is applied to standardise word representation. We did not follow this, instead, we used [12]s approach, because certain stop words may indicate sentiment and the model learns similar representations for words of the same stem when the data suggests it.

We identified that a recurrent neural network will be suited to take into account temporal dependencies, but due to potential instabilities during optimization, we followed common choices that advocate the use of long short-term memory networks. We extended the recurrent neural network architecture presented in [25] as in [4], to include dropout layers for regularization and use a weighted loss function to deal with class imbalance. We also incorporated the attention mechanism [26], to increase the accuracy.

Finally, a fully connected dense layer was used for the voting process, the output

Figure 4.1: High-level architecture of our proposed design.



of which was then passed through an activation function. Changes in this layer made our model flexible enough to adapt to both categorical (using a softmax activation function) and dimensional (using an affine transformation) emotion models.

4.1.2 Emotional Text-to-Speech Synthesis

We perform feature extraction and then principle component analysis (PCA) to identify the features that contribute the most to emotions in the audio. After identifying these important features, we then modify these very features in neutral audio to output emotional audio. This is achieved by mutating the audion waveform to obtain the appropriate wave for each separate emotion.

A speech emotion recognition model was designed in order to identify the various emotions present in our dataset. We tried different models to increase our overall accuracy, like naive bayes and random forest classifiers.

4.2 Work Done upto 7th Semester

- Did a comprehensive literature review.
- Familiarised ourselves with the relevant topics in the domain.
- Identified datasets that are being used for emotion recognition in the literature.
- Identified several innovative ideas in the field of emotion detection.
- Proposed a high-level design combining existing deep learning architectures.
- Did some research on applications of emotion detection in text (mainly emotional TTS), to look where we could expand our project after successful completion.
- Started implementation of a naive system to understand and familiarise ourselves with the building of machine learning models.

4.3 Work Done and Implementation in 8th Semester

We have implemented a system adhering to our proposed design. For the first part this implementation is an LSTM model which uses pretrained word embeddings (Glove – 840B 300D), and uses the ISEAR dataset to predict one of seven emotions present in the dataset. We have then extended this same model to work for the Election Tweets dataset which has 8 emotions with minimal modification. The only modification required was only for the final dense layer. We then proceeded to run one more datasets on the model in order to prove the flexibility of our model. All three of these datasets gave reasonable accuracy in the task of classifying text into emotions.

In order to understand the nature of audio data, we have analysed the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). We used tools such as jAudio and Praat to extract various parameters pertaining to the audio that help us classify them into various emotions. This has helped us to understand how to inject emotion to voice using the extracted parameters.

We then made use of various open source audio processing libraries in python for audio manipulation. By varying prosodic parameters in audio according to our

analysis we were able to inject emotion into the generated audio. The voices produced were still slightly robotic but definitely had discernable emotion in them.

We finally built an interface on top of our source code, that makes calls to our code, that is ready for general and public use.

Chapter 5

Results

Table 5.1 summarises the results of our first module, "emotion detection from text", which was run on three separate datasets.

Table 5.1: Emotion detection from text on various datasets.

Dataset	Accuracy without preprocessing (%)	Accuracy with preprocessing(%)
Isear Dataset	75.83	82.53
Twitter Tweet	71.06	79.81
Twitter Tweet (OLD)	59.22	62.47

The speech emotion recognition performed on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) helped us identify the key features in adding emotion to audio. Figures 5.1 and 5.2 demonstrate the variaton of these parameters for each of the eight emotions.

As the final step, we were able to successfully generate ten audio samples for each emotion and found them to be in line with the emotion expected from the test data.

Figure 5.1: Density plot after feature extraction for Spectral Rolloff Point

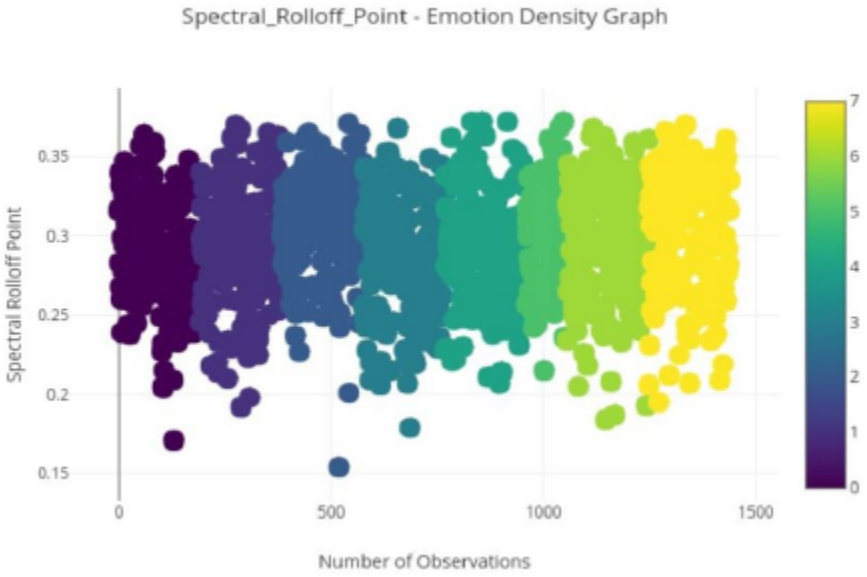
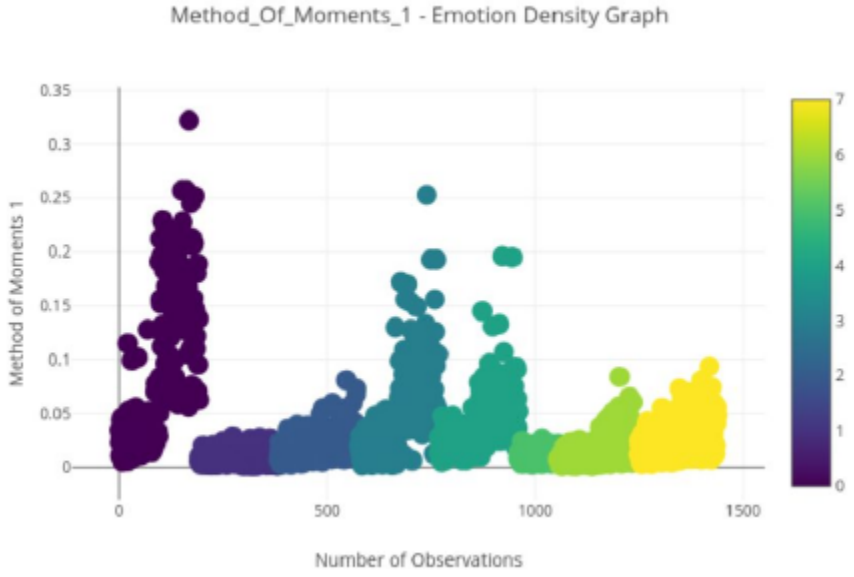


Figure 5.2: Density plot after feature extraction for Area Method of Moments



Chapter 6

Conclusion and Future work

6.1 Conclusion

We have designed an architecture and successfully implemented it to synthesize emotional and natural voice from unstructured text. The architecture consists of two separate parts. The first is a deep learning LSTM architecture to extract emotion from text. The model was flexible enough to obtain high accuracy over multiple datasets. The second part of the system consists of identifying the key features of speech that contribute to the emotional nature of speech using a speech emotion recognition (SER) model. Then these features were modified according to the analysis from the SER model to obtain emotional and human-like speech.

6.2 Future Work

An important parameter we found from speech emotion recognition was area method of moments, but we were not able to directly manipulate this feature. Manipulation of this feature may lead to more emotional sounding speech. Longer training times on GPUs and larger datasets will definitely improve the quality of the speech generated by our system; we hope that such datasets will be available in the future. Very recent, cutting edge research in text-to-speech suggests that an end to end system for generating speech from text, is able to generate higher quality audio, but at the expense of a much longer training and generation time [20]. But these architectures do not explicitly take an emotion label as input, they only take text and the corre-

sponding audio. We feel that by modifying this architecture to incorporate emotion label as an input also, we would be able to generate emotional audio in an end to end fashion.

References

[1] S. Elgayar, A. E. A. Abdelhamid, and Z. T. Fayed, "Emotion detection from text : Survey," 2017.

[2] D. Govind and S. R. M. Prasanna, "Expressive speech synthesis: a review," *International Journal of Speech Technology*, vol. 16, pp. 237–260, Jun 2013.

[3] L. Canales and P. Martínez-Barco, "Emotion detection from text : A survey," in *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Day (JISIC)*, pp. 37–43, 2014.

[4] B. Kratzwald, S. Ili, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24 – 35, 2018.

[5] O. Bruna, H. Avetisyan, and J. Holub, "Emotion models for textual emotion classification," *Journal of Physics: Conference Series*, vol. 772, no. 1, p. 012063, 2016.

[6] C. R. Chopade, "Text-based emotion recognition : A survey," 2015.

[7] K. Sailunaz, M. Dhaliwal, J. G. Rokne, and R. Alhajj, "Emotion detection from text and speech: a survey," *Social Network Analysis and Mining*, vol. 8, pp. 1–26, 2018.

[8] A. Hannan, "Emotion detection from text," *International Journal of Engineering Research and Development*, vol. 11, no. 07, pp. 23 – 34, 2015.

[9] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*,

HLT '05, (Stroudsburg, PA, USA), pp. 579–586, Association for Computational Linguistics, 2005.

[10] U. Jain and A. Sandhu, "A review on the emotion detection from text using machine learning techniques," in *International Journal of Current Engineering and Technology*, Inpressco, 2015.

[11] E. Saravia, H.-C. T. Liu, and Y.-S. Chen, "Deepemo: Learning and enriching pattern-based emotion representations," *CoRR*, vol. abs/1804.08847, 2018.

[12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, (Stroudsburg, PA, USA), pp. 142–150, Association for Computational Linguistics, 2011.

[13] D. Bespalov, B. Bai, Y. Qi, and A. Shokoufandeh, "Sentiment classification based on supervised latent n-gram analysis," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, (New York, NY, USA), pp. 375–382, ACM, 2011.

[14] I. Labutov and H. Lipson, "Re-embedding words," in *ACL*, 2013.

[15] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, pp. 74–79, Mar. 2017.

[16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[17] X. Gonzalvo, S. Tazari, C.-A. Chan, M. Becker, A. Gutkin, and H. Siln, "Recent advances in google real-time hmm-driven unit selection synthesizer," 06 2016.

[18] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *CoRR*, vol. abs/1710.07654, 2017.

[19] M. A. M. Shaikh, A. Ferreira Rebordao, K. Hirose, and M. Ishizuka, "Emotional speech synthesis by sensing affective information from text," in *Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*, p. 6, IEEE, 2009.

[20] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[21] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[22] M. Morise, F. YOKOMORI, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, pp. 1877–1884, 07 2016.

[23] Y. Lee, A. Rabiee, and S. Lee, "Emotional end-to-end neural speech synthesizer," *CoRR*, vol. abs/1711.05447, 2017.

[24] C. Raffel, T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," *CoRR*, vol. abs/1704.00784, 2017.

[25] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural sentiment classification with user and product attention," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1650–1659, Association for Computational Linguistics, 2016.

[26] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1 2018.