

**Emotion Prediction Application Using Voice Recognition Built with Neural Networks**

**Progress Report (or Final Report) for**

**ENGR 498 Global Design Project II Synthesis**

**213911362 Eren Ali Aslangiray**

**215141728 Mehmet Enis İşgören**

**214050652 Meryem Şahin**

**215231692 Sümeyye Sena Eminmollaoğlu**

**Barış Arslan**

**College of Engineering**

**İstanbul Şehir University**

**29 March 2019**

# 

Abstract

This project is a graduation project in İstanbul Şehir University CS department. The main goal of this work is, with using unique data collecting techniques we come up with, and emotion recognition model that will take human speech as input and emotion status as an output. System will be also developed as a virtual assistant-like interaction with user so it will try to manipulate the user's current emotion stage while collecting the valuable human behavior data in its back stage. At the end of each contact with the user there will an update to the algorithm for letting it to suit more on user’s personality.

Table of Contents

**Abstract2**

**Table of Contents3**

**Introduction4**

**Literature Review5**

**Presentation of Model6**

**Pipeline of the Project6**

1st Model7

2nd Model7

Pre-Guided User Data8

User State Graph9

3rd Model9

**Assumptions10**

**Date Collected10**

**References11**

Introduction

As an individual, we all want someone or something that understands us and see what we are going through or how we feel. From the start of the civilization we humans come together to overcome the problems we have and socialize to make every individual’s life better. From that perspective, our aim arises. The mood of a person may get affected from various things such as weather, time of the day or even the day of the week, likewise life. It keeps on going. We are willing to give a better daily life for individuals to have a good impact on their mood. Our design aims to predict users’ emotions and their response to certain events with the help of previously recorded reactions of users.

Voice is the most common communication tool used to interact with one another. Today, voice recognition applications are available for limited areas, but these areas are increasing steadily. In these applications, the machine detects the sound and gives appropriate answer to it. When communicating with phone, people prefer the sound rather than using keyboard, because it is easy and fast. We designed an application which will improve the currently used assistants such as Siri or Alexa, and with the help of this we will have an assistant which will develop empathy with the user and this system will be able to make mutual conversations. The application is going to ask some general information about users (Pre-Guided User Data aka.PGUD) to learn what they like or dislike and their behaviors in certain situations at the beginning. Then it will get users’ voice and understand the meaning of the words and sentences in general. Not just the meaning, but also emotions will be recognized by the application from the voice. So that it will evaluate mood of the users at that time and return with some recommendations such as a movie, outside activity, food, product advertisement etc. Then Assistant will ask if the users are satisfied from that recommendation and according to the answer, it will update its decision tree (User-State Graph aka. USG) and adapt to upcoming user preferences.

In a nutshell, our goal is to improve people’s daily lives by learning their moods and thoughts at any time of the day. So that we can make suggestions to make them feel better and to make them evaluate their free times. By this motivation, we foresee that our application will make people feel less lonely. And with the outcome of usage of this application we collect very valuable human behavior data.

Literature Review

There are many studies in the area of speech recognition and there are basic approaches that used in this problem. As discussed Gevaert, Tsenov and Mladenov in their paper, general structure of a speech recognition consists five steps which are; speech, signal processing, feature extraction, speech classification and output. In addition, there are commonly used techniques to achieve this problem. Dynamic Time Warping (DTW) compares words with reference words, Hidden Markov Modelling (HMM) splits the speech into small entities and it compares with the best-suited model. Another technique is Neural Networks which we will also be using in our project are similar to HMM, but Neural Networks use connection strengths instead of probabilities for state transitions (Gevaert, Tsenov & Mladenov, 2010, p.2). The article is mainly focusing on Neural Networks so it is very useful for our project.

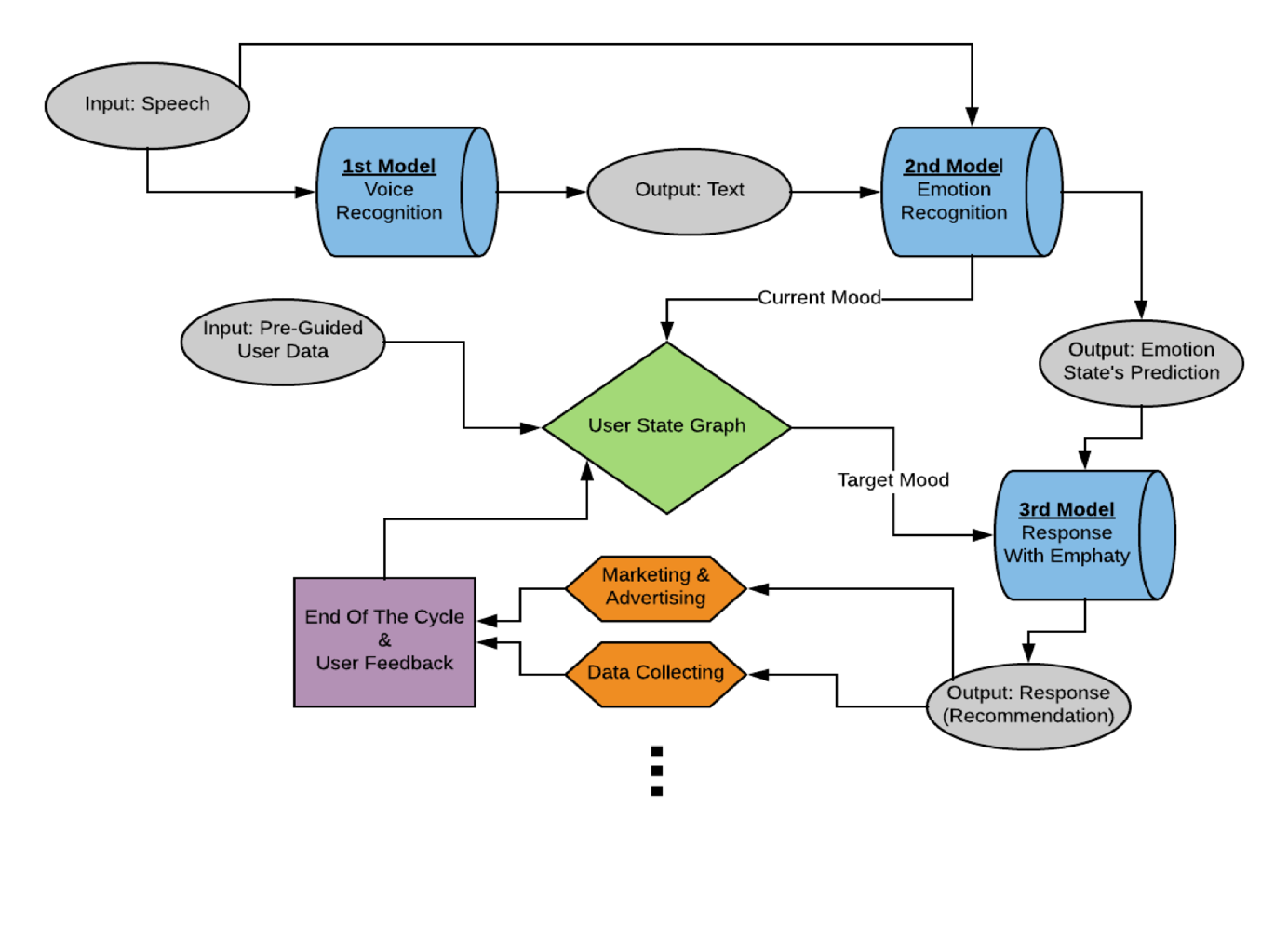
After the process of getting voice from the user and converting it to the text, with Natural Language Processing the meaning will be extracted from the text. There are 2 methods to do this which are Natural Language Processing for Speech Synthesis and Natural Language Processing for Speech Recognition. NLP for Speech Synthesis is based on text to speech conversion and it uses the sentence segmentation which deals with punctuation marks with a simple decision tree (Reshamwala, Mishra & Pawar, 2013, p.113). On the other hand, NLP for Speech Recognition is based on the grammar of a language (Trilla, 2009, p.3). It is needed to use Natural Language Processing in our project in terms of understanding user’s speech and return with some valuable suggestions.

Decision trees are a decision support tool for regression or classification models. According to Jordan, Ghahramani and Saul in their research, we have to know probabilistic decision trees and Hidden Markov models to understand Hidden Markov decision trees. In probabilistic decision tree, decisions are modeled probabilistically and recursion spreading upward and downward in the tree. In Hidden Markov model, the key calculation fit in it, recursion extending forward or backward in the chain. Hidden Markov decision trees is a probabilistic decision tree (upward and downward) with Hidden Markov model (forward or backward).

Presentation of the Model

This project contains Machine Learning, Deep Learning, NLP, Graph Theories, Decision Trees and Markov Models, Web Development and Signal Feature Extraction techniques. All the codings will be done in Python 3.6. Deep learning models will mainly use both TensorFlow and PyTorch libraries. For NLP and word vector representations we will use FastText. We will generate and use pseudo data from maps, weather and calendar. Moreover, we will combine all the algorithms in a single web application which will be implemented in Django.

Pipeline of the project:

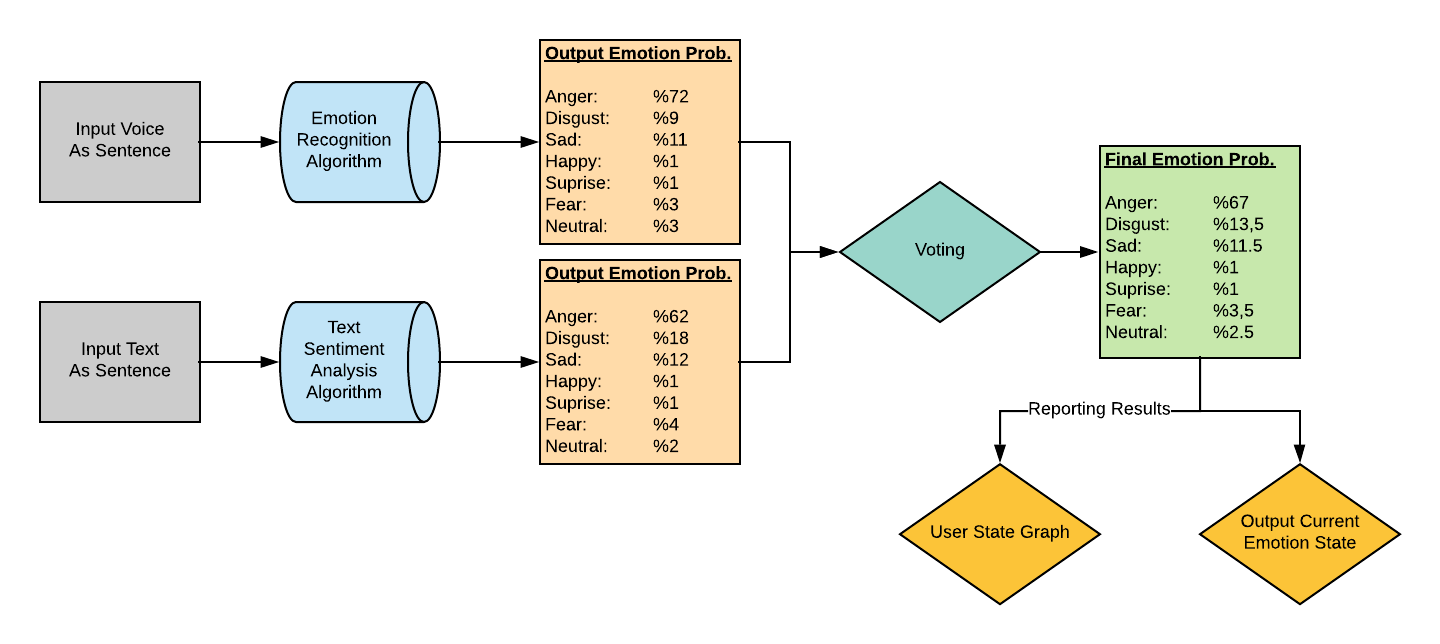


1st Model

First model is a basic speech recognition algorithm. Main goal is converting the human speech to text. In first period of our project we managed to implement speech recognition with over 90% accuracy with test dataset and 83% with our team members test data. It has total of 36 words in its library. Unfortunately, we are not able to scale it big enough to understand English language due to lack of computation power and dataset. Instead of it we will use open sourced Google’s speech recognition. But we still have a plan to use our own speech recognition algorithm at some parts of this application such has sound login system, or filling the PGUD with sound answer system.

2nd Model

Our second model is the heart of our project and its backbone has ensemble multi-model structure. We are using 7 main emotions as psychology science defined for basic emotions of a human. Pipeline of 2nd model is something like this:



We are planning to extract emotion from sound and also the meaning as coming in as text. We used Facebook’s FastText library to build text to emotion (T2E) model that works over 90% accuracy for given text input. Also, for the T2E algorithm it has 2 machine learning model inside of it. First of these models is doing binary classification such as “Positive Emotion” and “Negative Emotion”. Second of these is doing multi-label classification such as 7 emotion. They are all sensitive models that returns confidence levels of each output so that we combine these two models into one and return it for the voting part.

For sound to emotion (S2E) model again we will use two models that works and supports each other. First model will be trained with multiple information such as RMS, MFCC, DCT, etc. that is extracted from the online labeled dataset. The other model will use the unique method that our team come up with. The trick of this model is both simple and hard, we will use sound and music (Makam) for understanding the “color” of the voice and do predictions on it.

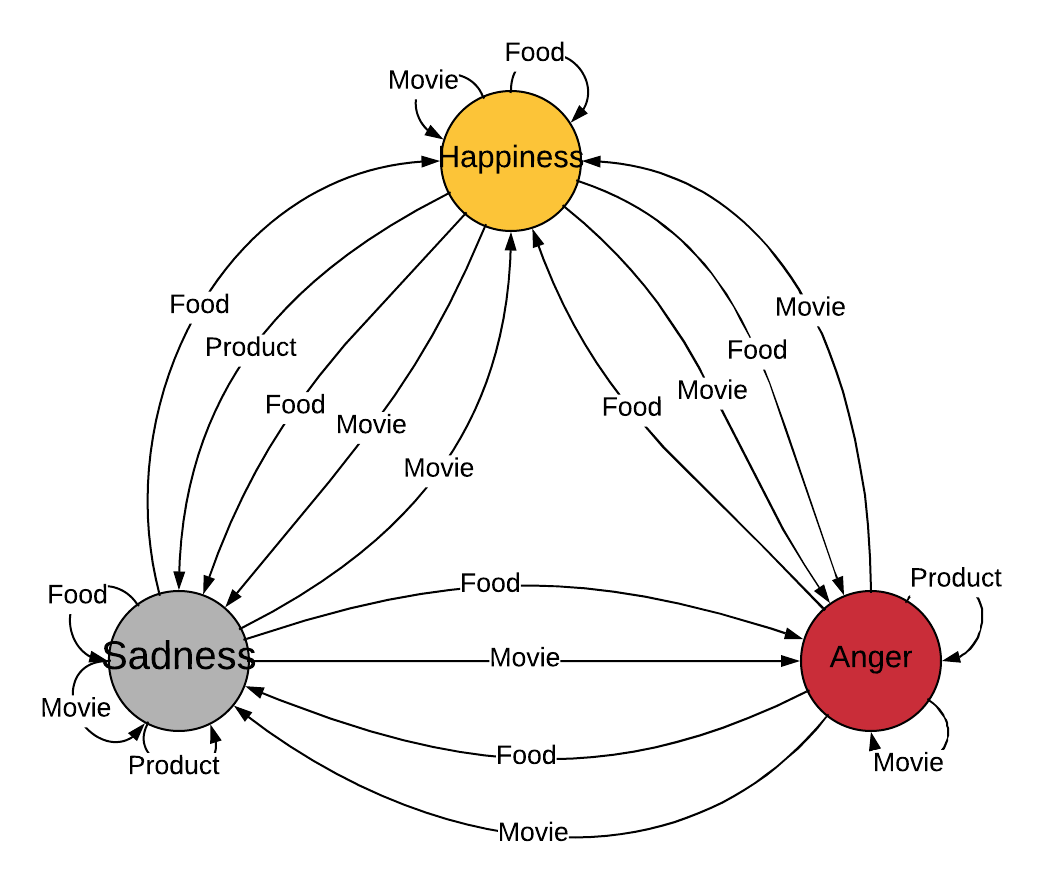
For the voting part we already started to implement our own soft-voting algorithm that will work sensitively to merge these 2 results with respect to their confidence level. After this step we will have the emotion predicted.

Pre-Guided User Data

For PGUD, we prepared set of questions with psychology professors in Sehir University to understand and build the users base case of initial USG. The questions will ask users gender, preferences, weather likes and dislikes, and so on. At each user’s first time usage of our system, we will ask these questions and with respect to the answers, we will build the USG. We will also do a survey at the target of 1000 people to get “Average” responses of the crowd. So that our USG will have a base case.

User State Graph

USG is an idea that we come up with while thinking of having some kind of decision tree and personality tree and tracker. Small example of it can be as following:



It is a directed and weighted graph which will track the user’s preferences and updates itself at each interaction with the user. The application will look this graph to decide what is should do. The graph will have its base case at initialization and this base case will be built with PGUD that user answered and also PGUD Survey.

3rd Model

3rd Model is a simple, non-complicated machine learning model that looks the PGUD and current emotion mood and also some third-party data such as weather, location, etc. to decide what to offer, or say back to user.

Assumptions

Our main assumption is that if music can alter people`s feelings one`s feelings can be detected from their voice. In Ottoman music there is a system called as Makam is being used to illustrate feelings we may be able to define patterns in that music system so that we may understand how a sound effect people or which feeling they are showing with their voice. Other assumption is having a good user-application connection and communication so that system will work healthy.

Data Collected

So far, we collected 7 emotion labeled text data, positive or negative labeled text data, 7 emotion labeled voice data. Also, we are going to start a survey with PGUD so we will have a base case for USG. We have a huge database around 150 GB

References

Gevaert, W., Tsenov, G., Mladenov, V. (2010, January). Neural networks used for speech recognition. Journal of Automatic Control. (20). 1-7.

Jordan M., Ghahramani Z., Saul L. Hidden Markov decision trees. University of Toronto, Canada.

Reshamwala, A., Mishra, D., Pawar, P. (2013, February). Review on natural language processing. Engineering Science and Technology: An International Journal. (3).113-116.

Trilla A. (2009). Natural Language Processing techniques in Text-To-Speech synthesis and Automatic Speech Recognition. Ramon Llull University, Spain.