Text

Description automatically generated

|  |  |
| --- | --- |
| GROUP MEMBERS | |
| NAME | MATRIC NO. |
| ZUHARABIH SULAIMAN | S2032539 |
| A T M MANFAT ZAYEEM | 17221083/1 |
| E Z E FLASH | S2001402 |

**PROJECT TITLE: ANIMOJI EVOLUTION**

**COURSE NAME: ADVANCED ALGORITHMS**

**COURSE CODE: WOA 7001**

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

Logo, company name

Description automatically generated

Contents

[High level Project Plan 3](#_Toc73792997)

[Introduction 3](#_Toc73792998)

[1. Project Background 5](#_Toc73792999)

[1.1 Emotion Analysis 5](#_Toc73793000)

[1.2 Anger, Disgust Animoji analysis 6](#_Toc73793001)

[**Anger Animoji** 6](#_Toc73793002)

[**Levels of anger:** 7](#_Toc73793003)

[**Disgust Animoji** 7](#_Toc73793004)

[**Levels of disgust:** 8](#_Toc73793005)

[1.2 Animoji program design overview 8](#_Toc73793006)

[2. Animoji Creation 9](#_Toc73793007)

[2.1 Animoji created and saved in MP4 format 10](#_Toc73793008)

[2.1 Animoji converted into GIF format 10](#_Toc73793009)

[2.3 Animoji Classification/grouping 10](#_Toc73793010)

[3. Using the Dynamic Time Warp (DTW) algorithm Previous Studies 9](#_Toc73793011)

[3.1 Speech recognition & Processing flow 9](#_Toc73793012)

[3.2 Mel Frequency Cepstral Coefficients (MFCC Feature Extraction) 9](#_Toc73793013)

[DTW Background 10](#_Toc73793014)

[DTW algorithm 11](#_Toc73793015)

[DTW Weakness/Disadvantages 11](#_Toc73793016)

[FILA FORM 11](#_Toc73793017)

[References 12](#_Toc73793018)

# High level Project Plan

Below are the high-level project plan which is targeted to be completed within 4 weeks period.

|  |  |  |  |
| --- | --- | --- | --- |
| Week | Days | Period | Task Focus |
| 1 | 7 | 28/05/2021 -02/06/2021 | * Group contract alignment & signing * Project Planning * (DS1) Analyze, choose & define Animoji * (DS2) Analyze DS1 & Create/Reuse Animoji |
| 2 | 7 | 03/06/2021- 09/06/2021 | * (DS3) Develop Algorithms using Dynamic Time Warp (DTW) |
| 3 | 7 | 10/06/2021- 16/06/2021 | * (DS3) Develop Algorithms using Dynamic Time Warp (DTW) * Program |
| 4 | 7 | 17/06/2021 -23/06/2021 | * (DS3) Create an interface based on D3 * (DS4) Code Integration |

# Introduction

Speech Recognition has gained significant popularity over the years due to the rapid advancement of Information Technology. As human speech emits emotions, recognition of various emotions has also gained popularity in the domain of Automatic Speech Recognition or ASR. Speech or voice can be simply defined as one dimensional signal. Like any other signals, it has properties like amplitude, pitch, frequency etc. This widely available technology can now easily be seen in various applications which can extract texts from the spoken words if they are clearly spoken. For example, in python, Google, IBM etc. has built-in libraries already, which are useful for implementing speech to text or STT.

Extraction of various emotions from speech has profound impact on research not only in the domain of Computing Linguistics, but also covering areas such as speech recognition in car parking (Kexin, T., et al., 2019), psychological domain such as anxiety, glossophobia etc. (El-Yamri, M., et.al., 2019). Gamified VR based research conducted by El-Yamri, M., et.al. (2019) developed an algorithm to measure the anxiety level of the speaker in a VR environment based on the speech given by the speaker. The research mainly focused on the speakers’ voice tone. Thus, it can be claimed that the characteristic of speech has notable value in many domains.

In recent years, psychologists tried to segregate the complex emotion we have in our daily lives. A research conducted by Cowen, A. S., et. al. (2017) reports about 27 emotions that humans feel. Also, in recent times, an article written by Cherry, K. (2020) states six basic emotions namely, happiness, fear, anger, sadness, disgust, surprise.These basic emotions have also quantifiable characteristics in facial expressions. For example, happiness can be easily identified by a smiling face, fear can be identified by widening of the eyes, Anger can be identified by glaring, disgust can be identified by the curling of upper lips (Cherry, K., 2020).

Animoji or animated emoji was first introduced electronic devices by apple in 2017. Animojis are animated emojis. Emojis were developed in Japanese mobile phones back in 1997 and thus gained popularity in the coming years (Blagdon, J., 2013). The reason of emoji getting popularized because of gap of emotional cues in conversation (Evans, V., 2017). As animojis are animated, thus it can exert more emotions which fills up the emotional cues more. Speeches or voice are signals which has various quantifiable properties. The properties of voice can be calculated by MFCCs or Mel Frequency Cepstral Coefficients. Therefore, by the usage of this property, various emotional data can be extracted and matched (Likitha, M. S., 2017).

Dynamic Time warping or DTW is an algorithm to determine the similarity between two temporal signals. By combining the voice properties extracted by MFCCs and computing the temporal distance by DTW it is possible to develop a prototype of an animoji determining system. In this project, emphasize will be given to the basic emotions, specifically Anger and Disgust. In the first phase of the project, the animojis are created. Neutral and Unidentified Animojis are also considered because of the validation of the prototype.

# 1. Project Background

There have been many studies done previously to detect emotions based one the speech recognition algorithm. The ability to detect an individual's emotional state is one of the key components in human-machine interaction research while the aspects used in recognising emotion were influenced by facial mimicry and voice cues, emotion persists as a physiological reaction to experiences such as grief, anxiety, or enjoyment (Umamaheswar & Akila, 2019). In a research done by Umamaheswar and Akila (2019) a mix of KNN and PRNN emotion recognition systems was developed to detect the six primary emotions, including neutral, anger, pleasure, sorrow, surprise, and fear, are all separately explored and researched for their authenticity. Kexin et al., 2019 also did a research on speech recognition for an emergency parking instruction based on speech emotion recognition to assist novice drivers in an emergency situation. Speech emotion recognition been very useful for many applications and industries for their application in Human computer interfaces, Telecommunication, Assistive technologies, Audio mining, Security and so on.

## Emotion Analysis

Different Types of Emotions and How They Are Expressed:

**Anger:**violence, hostility, resentment, wrath, irritability, fury, and outrage.

**Shame:**  regret, guilt, contrition, chagrin, remorse, and embarrassment.

**Sadness:**depression, grief, melancholy, gloom, despair, sorrow, and loneliness.

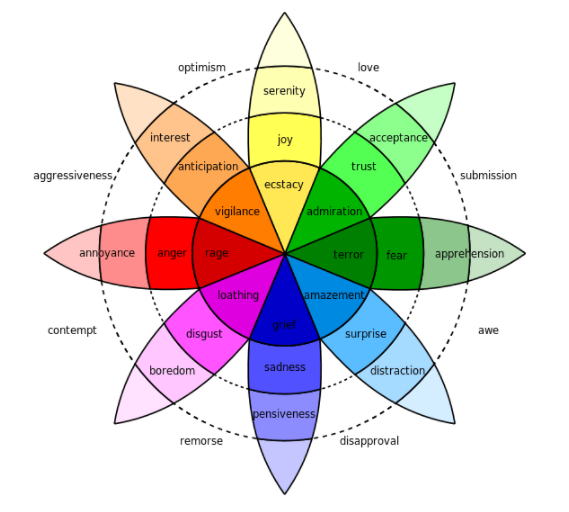
**Disgust:** scorn, contempt, distaste, disdain, revulsion, and aversion

**Fear:** anxiety, fright, nervousness, dread, apprehension, and panic.

**Surprise:** wonder, amazement, astonishment, astound, and shock.

**Joy:**enjoyment, thrill, delight, bliss, relief, pride, happiness, and ecstasy.

**Interest:** devotion, acceptance, affection, trust, kindness, love, and friendliness.



## 1.2 Anger, Disgust Animoji analysis

In order to design the system, each group has to implement an Animoji system that acts upon the type of speech sent in a recorded voice message. With the purpose to design a good algorithm, each group was required to choose a set of emotions and work on those particular emotions. Our group agreed and choose Set 1 out of 3 options which is anger and disgust

The study of emotions is greatly aided by digital signal processing hardware and software. However, machines cannot equal the performance of human equivalents in terms of accuracy and speed, especially with regards to speaker-independent emotion recognition. There are namely two phases of emotion recognition algorithms which are the testing phase and training phase.

### **Anger Animoji**

The image of a face with furrowed eyebrows and with its mouth curling downward is the emoji representing anger, upset or disapproval. It is typically used to emphasize that someone is upset or furious. Angry Face Emoji can mean “I am so upset right now!” or “This angers me so much! I hate it!”. The Angry Face Emoji appeared in 2010, and also known as the Mad Face. Sometimes it is mentioned as the Mad Emoji. Figure 1 Below is pictorial representation of each organization with their anger emoji.

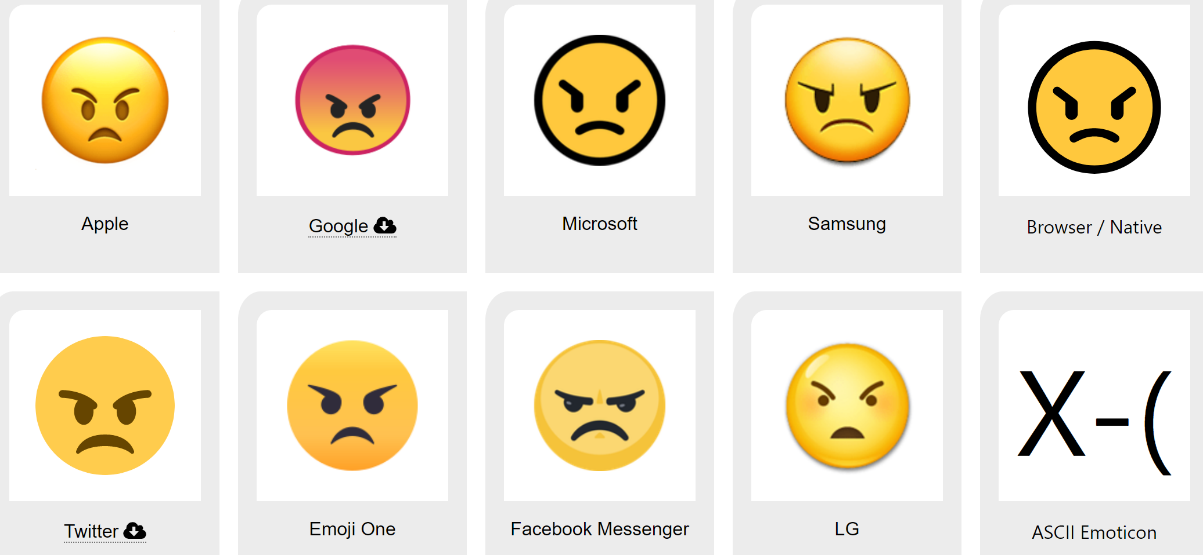


Figure 1 (Anger Emoji)

### **Levels of anger:**

1. **Annoyance** is the first level of anger when people get annoyed at various minute things, such as long lines, traffic jams, or hearing someone chew with their mouth open. Generally, annoyance is very mild and tends to subside quickly.
2. **Frustration** is the second level - come along once an annoyance has festered and lasted for too long. A frustrated individual may feel tense or otherwise have a hard time concentrating on certain matters, due to their current negative emotional state.
3. **Rage** is the final stage of anger. - enraged, they are often verbally confrontational, throwing objects, making threats, or even physically lashing out at others.

### **Disgust Animoji**

Ugh, nasty! That's the sentiment (and appearance) of the face with open mouth vomiting emoji, used to express literal and metaphorical disgust. Related words that are synonymous to disgust emoji are listed below

* 🤮 Spew
* 🤮 Throwing Up
* 🤮 Vomit

Figure 2 Below is pictorial representation of each organization with their disgust emoji

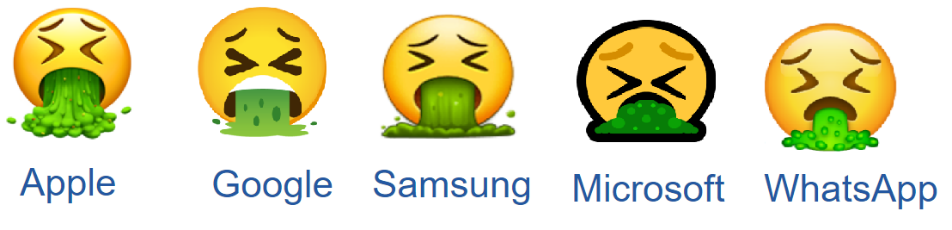
****

Figure 2 (Disgust Emoji)

### **Levels of disgust:**

1. **Boredom**: arises from survival instinct and, eventually, dread of death. ( link)
2. **Disgust** arises from a wish to avoid "biologically costly mates" and a thought of the results of their decisions.
3. **Loathing**: Moral distaste encourages avoidance of social ties with norm-violating persons because such ties jeopardise group cohesion.

For all the different levels of anger and disgust above voices are represented with higher or lower sound amplitude as compared to a neutral emotion’s voice. In terms of vocal emotional recognition, each level of emotions will be represented by different tone, pitch, amplitude, and frequencies. Hence normal sound volume of angry voices might not be enough to elicit differential emotional responses to angry tonality (Simon et al, 2016). Acoustic cues such as loudness help in deciphering spoken emotion. Sound intensity can have a quantifiable effect on the emotional impact of the sounds that it contributes to. Thus, sound intensity should not be treated as a simple control and should be included for vocal emotion investigations. (Chen et al, 2012).

## 1.2 Animoji program design overview

Stage 1

Stage 2

Repeat until achieve desired outcome

# 2. Animoji Creation

As per explained in the introduction above, we have created total of 8 Animoji which includes three levels of anger, three level of disgust emotions, one neutral and one unidentified Animoji. This is due the scope of project that only covers anger and disgust emotion leaving the rest of the emotions matched into unidentified emotion.

## 2.1 Animoji created and saved in MP4 format

In order to create an Animoji for this project, we have decided the approach to use Memoji in IPAD and create a video to show emotions expression and it is saved in MP4 format. This will the be converted into GIF format as shown in the next sections.

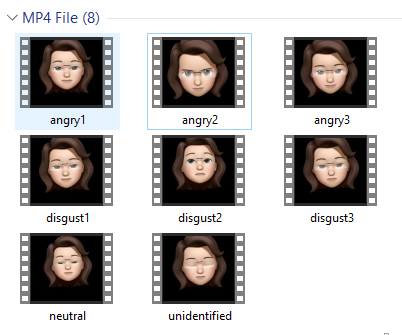
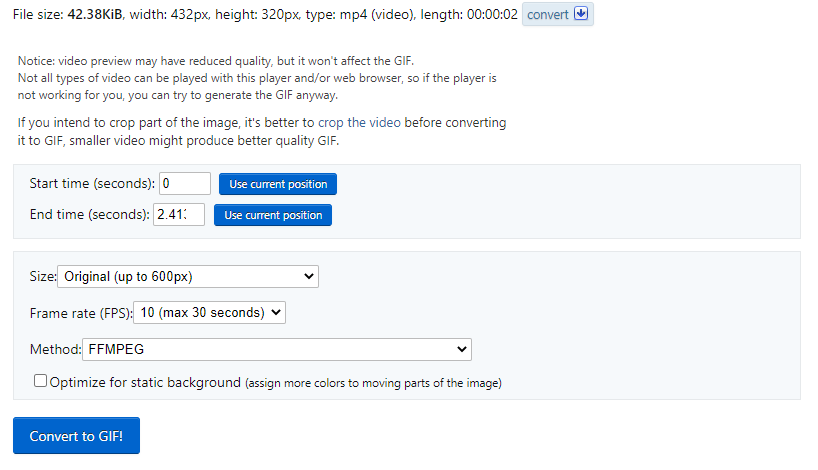


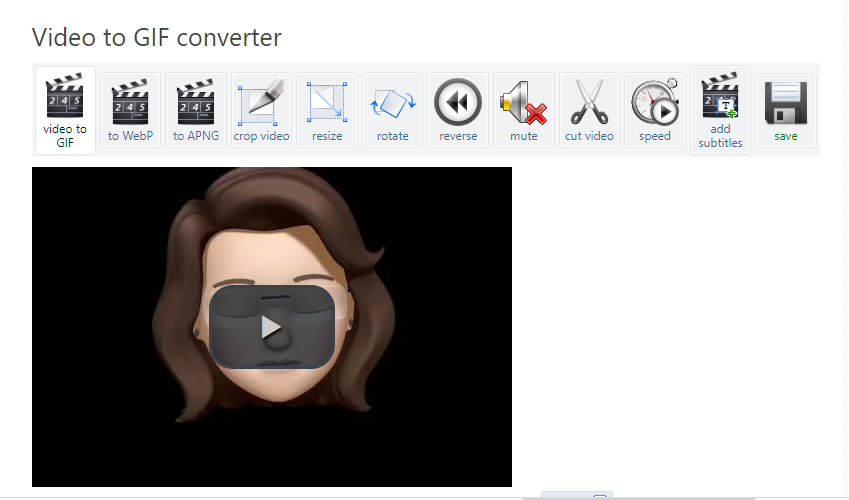
Figure 1: MP4 Memoji Videos

## 2.1 Animoji converted into GIF format

All the Memoji videos in the previous section is converted into GIF as shown in below figure accordingly. All the MP4 are converted into GIF using an online tool in <https://ezgif.com/> website.







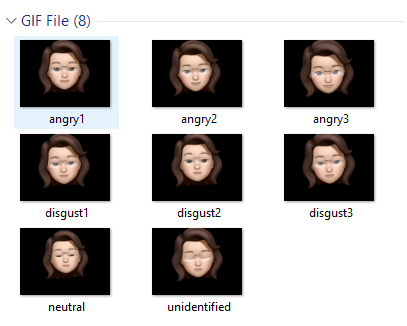


Figure 1: GIF Animoji

## 2.3 Animoji Classification/grouping

|  |  |  |  |
| --- | --- | --- | --- |
| Level  Emotion | Lightly affected | Medium affected | Highly affected |
| Anger |  |  |  |
| Disgust |  |  |  |
| Neutral |  | | |
| Unidentified |  | | |

# 3. Using the Dynamic Time Warp (DTW) algorithm Previous Studies

## 3.1 Speech recognition & Processing flow

In order to find out the emotion in the received voice signal, we need to have a large amount of information like energy, power spectral density (PSD) and so on in order to perform a statistical analysis. MFCC is enabling us to reach this goal. Also, in order to eliminate the disparity between the database and the input signal, we employ a method called Dynamic Time Warping.

An analysis of the user's voice is performed following the process of getting an input through a microphone. Manipulation of the input audio stream is an integral part of the architecture of the system. Additionally, at various levels, various processes are applied to the input signal such as Pre-emphasis, Framing, Windowing, and Mel Cepstrum analysis, all of which have the objective of recognising (matching) words spoken. The voice algorithms are made out of two distinct steps. In the first case, it is referred to as training, whereas in the second case, it is commonly known as a mission rehearsal or a mission exercise.

## Mel Frequency Cepstral Coefficients (MFCC Feature Extraction)

The Mel-frequency cepstrum (MFC) is a representation of a sound's short-term power spectrum in sound processing that is based on a linear cosine transform of a log power spectrum on a nonlinear mel frequency scale. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that make up an MFC collectively. They are derived from an audio clip's cepstral representation (a nonlinear "spectrum-of-aspectrum"). The distinction between the cepstrum and the Mel-frequency cepstrum is that the MFC uses equally spaced frequency bands on the Mel scale, which more closely approximates the human auditory system's response than the cepstrum's linearly spaced frequency bands. There are 13 cepstral coefficients for the Mel frequency. The Mel scale establishes a relationship between a pure tone's perceived frequency, or pitch, and its actual measured frequency. At low frequencies, humans are far more sensitive to subtle changes in pitch than they are at high frequencies. By using this scale, our features become more consistent with what humans hear.

## DTW Background

Dynamic Time Warping is needed to correctly identify a sound's compatibility (DTW). DTW is a pattern recognition method to compare different time zones. The closer the two sounds are, the more alike they are. Thus, both sound patterns are said to be the same. The preliminary speech recognition data is converted into frequencies. The distance of sound around the recording is affected by volume, pronunciation time, and noise. The smaller the effect, the smaller the distance.

An audio signal produces a time series. DTW projects each element in the series onto the temporal dimension. As such, DTW finds the optimum distance (i.e., the shortest route) while completing this mapping. Time will be warped. This table contains the distances between locations. Memoization is called that. When the shortest paths are calculated, a similarity metric between two time series is generated. As all of this happens, a warping path is made. Because the path has been altered, the two series have the same time levels. when the warp path shrinks, the similarity between the two-time series grows Warping paths have rules. Warping is defined by these rules. Warping is done to both series.

The DTW algorithm is a traditional algorithm that is easy to carry out, making it useful in speech recognition. The DTW algorithm uses the point matching method to determine the matching distance. At higher volumes, the reference template and test voice require more time to match, hence recognition times will increase. Feature extraction cannot be done directly because of the nonstationarity of the speech, as well as external noise. After pre-processing, the voice signal is used to extract distinctive parameters.

Signal characteristics metrics such as short time energy, short-term zero-crossing rate, and short-time autocorrelation coefficient are abundant. Since LPCC's speech parameters extraction is quite accurate, the computation speed is comparatively rapid and hardware-based. But LPCC in anti-noise performance, robustness, and the recognition rate and other aspects are below average, thus in practise, MFCC is utilised. The input voice is used to extract parameters, and the parameters are stored in a library of reference templates. Recognition speech's feature parameters are obtained in the same way and utilised to match with reference template library using the DTW algorithm, which yields the maximum similarity reference template in the library. DTW is a very successful recognition matching algorithm.

## DTW algorithm

## DTW Weakness/Disadvantages

The DTW has some weaknesses. First, O(n2v) complexity may not be sufficient for a larger vocabulary, which could lead to an increase in the recognition rate. Additionally, evaluating two parts from two different sequences is difficult because of the various channels with various properties.

# FILA FORM

Week 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FACTS** | **IDEAS** | **LEARNING ISSUES** | **ACTION** | **DATELINE** |
| What we know about the task | What do we need to find out? | | Who is going to do it? | 02/06/2021 |
| Planning | Define & Update group contract | | Zuha | 02/06/2021 |
| Define project planning | | Zuha/Zayeem/Flash | 02/06/2021 |
| Select emotion for the project | | Zuha/Zayeem/Flash | 02/06/2021 |
| Animoji program design process flow | | Zuha/ Zayeem/Flash | 02/06/2021 |
| Consolidate content in the report | | Zuha/ Zayeem | 02/06/2021 |
| Create FILA form | | Zuha | 02/06/2021 |
| DS1: Define& Analyze Animoji |  | |  |  |
| Project background/Literature Review | | Zuha/Zayeem | 02/06/2021 |
| Animoji Analysis | | Zuha/Zayeem/Flash | 02/06/2021 |
| References | | Zayeem | 02/06/2021 |
| DS2: Create Animoji | Create Anger, Disgust, Neutral, Unidentified Memoji | | Zuha | 02/06/2021 |
| Record Memoji Anger, Disgust, Neutral, and Unidentified | | Zuha | 02/06/2021 |
| Convert Memoji video into Animoji GIF | | Zuha | 02/06/2021 |
| DS3: DTW Algorithm | Research on DTW | | Zuha/Zayeem/Flash | 02/06/2021 |
| DS4: Program coding | Research on existing speech recognition program | | Zuha/Zayeem/Flash | 02/06/2021 |

# References

[1] Umamaheswari, J., & Akila, A. (2019, February). An enhanced human speech emotion recognition using hybrid of PRNN and KNN. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 177- 183). IEEE.

[2] Kexin, T., Yongming, H., Guobao, Z., & Lin, Z. (2019, November). Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition. In *2019 Chinese Automation Congress (CAC)* (pp. 2933-2937). IEEE.

[3] Cherry, K. (2020, January 13). *The 6 Types of Basic Emotions and Their Effect on Human Behavior.* verywellmind. <https://www.verywellmind.com/an-overview-of-the-types-> of-emotions-4163976

[4] El-Yamri, M., Romero-Hernandez, A., Gonzalez-Riojo, M., & Manero, B. (2019). Designing a VR game for public speaking based on speakers features: a case study. *Smart Learning Environments*, *6*(1), 1-15.

[5] Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900- E7909.

[6] Koops, T. “One More Thing…”–A critical approach to the Apple 2017 Keynote Presentation.

[7] Blagdon, J. (2013). How emoji conquered the world. *The Verge*, *4*.

[8] Evans,[V.(2017, August 17).](https://nypost.com/author/vyvyan-evans/)  *Emojis actually make our language better.* nypost. <https://nypost.com/2017/08/12/emojis-actually-make-our-language-way-better/>

[9] Likitha, M. S., Gupta, S. R. R., Hasitha, K., & Raju, A. U. (2017, March). Speech based human emotion recognition using MFCC. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)* (pp. 2257- 2260). IEEE.

[10] XinXing, J., Xu, S. (2012). Speech Recognition Based on Efficient DTW Algorithm and Its DSP Implementation. In International Workshop on Information and Electronics Engineering (IWIEE). Elsevier.

[11] Simon, D., Becker, M., Mothes-Lasch, M., Miltner, W.H.R., Starube, T. (2016). Loud and angry: sound intensity modulates amygdala activation to angry voices in social anxiety disorder. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5390751/>. NCBI.

[12] Chen, X., Yang, J.,Gan, S., Yang, Y. (2012). The Contribution of Sound Intensity in Vocal Emotion Perception: Behavioral and Electrophysiological Evidence. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3264585/>. NCBI.

<https://medium.com/walmartglobaltech/time-series-similarity-using-dynamic-time-warping-explained-9d09119e48ec>

<https://towardsdatascience.com/dynamic-time-warping-3933f25fcdd>