COSC 6480 Speech Emotion Recognition

Linchuan Tang, Neuman Alkhalil, and Tong Zhang

Georgetown University, Washington D.C., United States

May 10, 2024

Abstract

This project advances speech recognition by integrating emotional context analysis. It utilizes sophisticated infrastructure and pre-trained models like Whisper[1] and WavLM[2] to accurately recognize speech and its emotional nuances in real-time. This approach not only improves user experience by providing richer, more nuanced transcriptions but also opens substantial business opportunities in sectors like hearing aid, computer games, custom relationship management. The technology's potential for human-computer interaction makes it well for commercialization and widespread adoption across various markets.

1 Introduction

This project seeks to enhance traditional speech recognition by incorporating the analysis of emotional context in spoken language. By recognizing the emotional states of the speaker, we aim to enhance our understanding of spoken communication. This could significantly improve speech-to-text systems, particularly benefiting individuals with hearing impairments by offering richer, more nuanced transcriptions.

In addition, this project has the potential to enhance human-computer interaction by adding an emotional dimension. A voice assistant that can discern emotional nuances could interact in a more empathetic and context-sensitive manner, making conversations feel more natural and supportive. For instance, it could offer soothing responses to a user expressing frustration, based solely on vocal cues. Similarly, non-player characters (NPCs) in video games can interpret players' emotions and react dynamically, making gaming experiences more immersive and personalized.

To address this problem, we provide a specialized infrastructure to handle real-time audio streams. This setup includes a two-stages Voice Activity Detection (VAD) system to precisely capture speech, alongside a reactive programming framework to manage audio data asynchronously. Moreover, we employ advanced pretrained models: Whisper and WavLM, giving the capability to accurately recognize both speech and its emotional nuances. Finally, we discuss the market prospects and potential applications of this technology.

The code for this project are available at `https://github.com/NoomyWasTaken/Experimental_AI_project` .

2 Problem Statement

Speech emotion recognition (SER) is a task focused on identifying and categorizing the emotional states expressed in spoken language. This field combines techniques from linguistics, psychology, and computer science to analyze the acoustic and linguistic features of speech. The primary goal of SER is to enhance human-

machine interactions by enabling more intuitive, empathetic, and effective communication in applications. Ultimately, by understanding users' emotional states, these systems can tailor their responses, leading to more natural and user-friendly interactions.

Most recent, transformer models show potential in a various kind of tasks, including the speech recognition. Two pretrained models are widely used for SER tasks, namely, Wav2Vec 2.0[3] and HuBERT[4]. They share a same architecture. Wav2vec 2.0 uses a contrastive loss to learn the quantized speech representation from a masked segment of the audio, where HuBERT uses cross-entropy loss and an iterative clustering from MFCCs and labeling process to learn the speech representation. In addition, the models come in two sizes: base and large. More variant are derived by training on different dataset. For example, wav2vec2-large-960h-lv60-self was pretrained on LibriSpeech[5].

Wagner et al.[6] suggests an approach by finetuning on Wav2Vec 2.0 and HuBERT pretrained model. Rather than directly predicting emotional labels, it generates descriptors based on attributes such as Arousal, Dominance, and Valence, which are commonly utilized in psychological assessments of human emotions. The effectiveness of this method is demonstrated through validation across various datasets and the subsequent derivation for classifying emotions.

Antonuou et al.[7] provides a comprehensive review and suggests a framework addressing the methodological shortcomings in the field, such as the absence of a unified evaluation protocol and the lack of measures for speaker independence. It also highlights the constraints of existing datasets, including issues like class imbalance, restricted data availability, and the reliability of labels.

3 Datasets and Models

We list three popular datasets for SER and four representative pretrained models from Hugging Face. A comparative analysis of the various datasets is presented in Table 1. We employ RAVDESS[8] and IEMOCAP[9] as our test sets to assess the performance of these models, with the results given in Table 2.

Dataset	Type	Source	Emotions	Dimension	Hours	Commercial
RAVDESS	Audio, video	24 Actors	8	No	1	Yes
IEMOCAP	Audio, video, motion capture	10 Actors	8	Yes	7	No
MSP-Podcast	Audio	Filtered Podcasts	8	Yes	27	Yes

Table 1: Comparison of three commonly used datasets.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) has 1440 speech audio from 24 professional actors (12 female, 12 male) including calm, happy, sad, angry, fearful, surprise, and disgust expressions. Actors vocalizing two statements that are lexically matched, delivered in a neutral North American accent.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database has 7530 labeled speech audio from 10 actors including anger, happiness, excitement, sadness, frustration, fear, surprise, and neutral state. The audio files consist of segments from either improvised or scripted dialogues between two actors, specifically chosen to elicit emotional expressions.

The MSP-Podcast[10] corpus is the largest naturalistic speech emotional dataset, comprised of podcast segments annotated for emotions using crowdsourcing. Version 1.11 includes 151,654 speaking turns, totaling nearly 238 hours. Emotion labels include anger, sadness, happiness, surprise, fear, disgust, contempt, neutral.

As Antonuou et al.[7] pointed out, the IEMOCAP dataset may be hindered by the limited amount of speech and the potential imbalance in emotion labels. Moreover, the limited number of speech samples from actors could impede

[1] https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition

[2] https://huggingface.co/canlinzhang/wav2vec2_speech_emotion_recognition_trained_on_IEMOCAP

[3] https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Categorical

[4] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

Model	Based model	Finetune set	Reported accuracy	Accuracy on RAVDESS	Accuracy on IEMOCAP
ehcalabres/ wav2vec2-lg-xlsr-en -speech-emotion-recognition[1]	facebook/ wav2vec2-large -xlsr-53	RAVDESS	0.822	0.879*	0.266
canlinzhang/ wav2vec2_speech_emotion _recognition_trained _on_IEMOCAP[2]	facebook/ wav2vec2-base	IEMOCAP	0.67	0.342	0.894*
3loi/ SER-Odyssey -Baseline-WavLM -Categorical-Attributes[3]	microsoft/ wavlm-large	MSP-Podcast	NA	0.425	0.547
audeering/ wav2vec2-large-robust -12-ft-emotion-msp-dim (fine-tuned on IEMOCAP)[4]	facebook/ wav2vec2 -large-robust	IEMOCAP	NA	0.315	0.704*

Table 2: Selected models from Hugging Face with corresponding dummy test accuracy on RAVDESS and IEMOCAP. An asterisk (*) indicates models has trained on the same test set. Our testing approach is considered dummy because some models were trained on the test set. Additionally, we employ the entire dataset for testing due to the absence of a defined train-test split. We restrict our evaluation to four emotional labels (neutral, angry, happy, and sad) because various models generate different label sets.

the model's ability to generalize. Similarly, the RAVESS dataset includes a more diverse sample, but limited in amount of speech and variety in the sentences presented by actors.

It should be noted that although all three datasets feature eight emotion labels, the labels differ subtly. As we experiment with various pretrained model, we observe that it is highly common to make a biased prediction. This bias may be due to models struggling to recognize specific emotions, a lack of data for certain emotions in the training sets, or the speaker specific characteristics. Furthermore, the interpretation of emotions by a model varies depending on the dataset used for training.

We provide a straightforward cross-dataset evaluation in Table 2, which demonstrates models' performance on new data. While there are differences between the models, the key factor remains the training data, with the 3loi/SER-Odyssey-Baseline-WavLM-Categorical model outperforming the others. Although we use this model for our speech emotion recognition task, it is still far from perfect.

In our speech recognition experiments, Whisper significantly outperforms Wav2vec2 models in both accuracy and speed. Specifically, we utilize the faster-distil-large-v2 from faster-whisper[5]. This model is a reimplementation aimed at faster inference, and it is well-balanced for our task, optimizing for both speed and memory usage.

4 Implementations

The system is designed to be used with an audio input device, such as a microphone, and we use the WebRTC framework to handle the audio stream. The raw audio data is converted into a series of frames at a fixed sampling rate, which aligns with the specifications of the model's encoder.

After framing, the audio passes through a two-stage VAD process. The main role of VAD is to identify sections of the audio that likely contain speech. This segmentation is important as it organizes the raw audio stream into discrete buffers, making it more efficient and effective for following process.

The initial stage of the VAD process employs the built-in VAD function of WebRTC, which offers a real-time and efficient speech detection. However, it is not very accurate.

Therefore, we include Silero VAD[6] model, an advanced neural network model adept at more precise speech detection. Despite its complexity, the Silero model is also optimized for real-time applications.

The detected speech segments are transferred to a buffer managed by RxPY, a reactive programming framework allows us to focus on data streams and propagation, making it simpler to handle real-time audio data. The buffer acts as a queue for the speech segments, enabling asynchronous process for the following systems.

Whenever an audio buffer is available, it is automatically sent to the automatic speech recog-

[5]https://github.com/SYSTRAN/faster-whisper
[6]https://github.com/snakers4/silero-vad

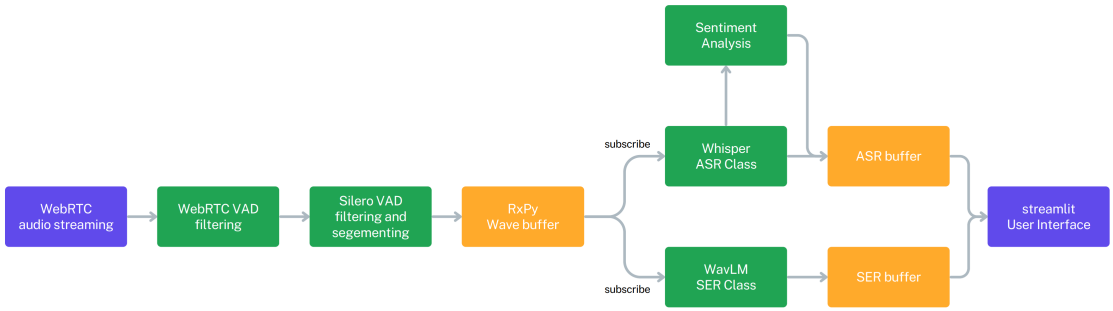

Figure 1: The high level pipeline of the speech emotion recognition system

nition (ASR) and SER systems. These systems utilize a common class diagram and receive the same inputs. With this arrangement, we can upgrade or replace our underlying model without modifying the overall architecture. As we need to evolve or as more advanced models become available, the system can integrate these improvements without major overhaul.

The outputs from both systems are collected and synchronized in the output buffer, generating a combined output of text transcription and emotion label. We use Streamlit as our user interface framework, which offers seamless integration between the Python backend and a modern web interface. It's notable that integrating it into our design requires only a few lines of code. An example of our interface is shown in Figure 2.

We also incorporate a simple sentiment analysis by VADER Sentiment[7] (positive, negative and neutral) as a supplement to our emotion analysis. This text based analysis helps us to further understand the context and emotion in speech.

5 Market Research

In this section, we will look at 3 potential markets for our product, which include the RPG (role-playing game) gaming market, hearing aid market, and customer relationship management (CRM) market.

[7]https://github.com/cjhutto/vaderSentiment

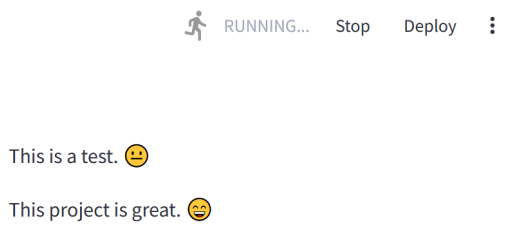

Figure 2: A sample user interface based on Streamlit.

5.1 RPG Gaming Market

The RPG gaming market was valued around $45 billion in 2022 with an annual growth rate of 8%, expected to reach a value of $70 billion by 2027. The most significant consumers in the market are in Asia and English-speaking countries, however our system currently only supports English.

5.2 Hearing Aid Market

Valued at approximately $8 billion in 2023 and projected to grow to over $20 billion by 2030, the

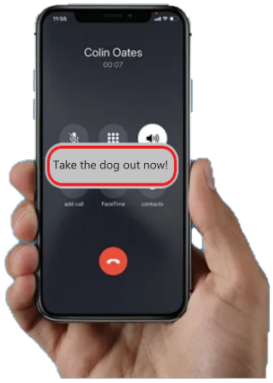

Figure 3: Mobile app during phone call

Figure 4: Mobile app during video call

hearing aid market is most prominent in Europe currently but is quickly expanding in the US, and our technology would benefit the community in need of hearing aids by allowing them to visually interpret emotional cues in real-time.

5.3 Customer Relationship Management (CRM) Market

The CRM market was worth approximately $65 billion in 2022, with projections of reaching $158 billion by 2030, growing at a staggering rate of 12% annually. Dominated by the US market, CRM's growth is driven by the need for improved customer interaction management, which our technology can revolutionize.

6 Business Model

In this section, we will talk about our various revenue streams, as well as different marketing strategies we will take in order to spread awareness of our technology.

6.1 Revenue Streams

The main revenue streams include licensing technology, API tokens, targeted advertising, and a subscription to the applications.

6.1.1 Licensing Technology

The first revenue stream we have is licensing the technology to companies mainly in communications, gaming, and customer service. Our technology helps enhance the experience for companies in communication such as Discord, Zoom, WhatsApp, and more, as well as giving their users more accessibility options. As for gaming companies, it's mostly targeted towards newer ones, which dynamically change NPC and environment interactions based on voice input and sentiment analysis on the tone. Customer service companies can also benefit greatly as our technology gives them the ability to automatically collect user experience data, which is often not provided or skipped by the users after their interaction with the company. This in turn can help improve customer experience management.

6.1.2 API Tokens

For this idea, we will implement a usage-based pricing model where developers are charged based on the hours of operation with their API token. This scales well as client applications gain popularity, increasing usage and, consequently, revenue.

Figure 5: Mobile app with voice messages

6.1.3 In-App Advertising

Since our primary focus is accessibility with the applications we develop which will be shown in the next section, we provide it for free, however we will run non-intrusive targeted ads in order to still generate income.

6.1.4 Premium App Features

A subscription will be available for users of the applications. The primary benefits of the subscription will be the absence of ads, being able to customize the font, and the representation of the emotions through different colors and emojis, and lastly subscribers will have early access to new features.

6.2 Marketing Strategies

6.2.1 Social Media Marketing

We will utilize platforms like Twitter, Instagram, LinkedIn, TikTok, Facebook, and YouTube to reach a diverse audience for a low cost, and in order to expand our reach we will be creating funny content which showcases the product and farms user engagement.

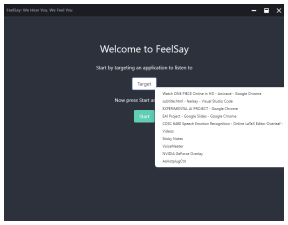

Figure 6: Desktop application: targeting an audio source

6.2.2 Strategic Partnerships

Collaboration with hardware and software companies, accessibility advocacy groups, and educational institutions will help reach a broader audience and validate the technology's utility and effectiveness.

6.2.3 Digital and Traditional Advertisement

The use of both online and offline advertisement will help raise awareness about our products. We will mainly use online advertisement as it is more targeted, however doing traditional advertisement can increase broad visibility.

6.2.4 Conferences

Presenting the technology at relevant tech and accessibility conferences helps showcase its capabilities to industry experts, which can help gather meaningful feedback. This also opens opportunities for networking with potential business partners.

7 Products

7.1 Mobile Application

Our mobile application has 2 main functionalities. First, it enhances phone calls by displaying a movable on-screen box that transcribes spoken words and indicates the speaker's emotions through different colors or emojis, which can be

seen in 3. This feature is valuable for individuals who are hard of hearing or in noisy environments, as it ensures no part of the conversation is missed. Additionally, this technology is highly beneficial for FaceTime/video calls with deaf individuals, facilitating communication by allowing them to read what is being said in real time so they don't have to resort to text communication with their friends and family which can be seen in 4.

The second functionality of our mobile app extends to voice messages. Users can transcribe and emotionally decode voice messages, making it easier to understand the context and sentiment of the message, especially when listening conditions are not ideal, such as in loud environments or for users with hearing impairments which can be seen in 5.

7.2 Desktop Application

Our desktop application is designed to work across multiple operating systems, offering users the ability to select an audio source from currently open applications, which is seen in 6. Once selected, a customizable subtitle box appears, which provides real-time transcriptions of the audio stream. Additionally, it enhances the transcription by color-coding or using emojis to represent the emotional content of the speech, as shown in 7. This is done with the help of Electron-Forge, a framework that helps package applications for cross-platform development.

This application is particularly useful for professionals who need to monitor or record meetings, webinars, or any digital communication where multiple applications may be used simultaneously. It ensures that users not only get a textual representation of the spoken content but also understand the emotional undertones, which can be crucial for effective communication and documentation.

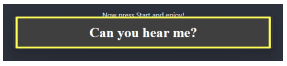

Figure 7: Desktop application: subtitle box

8 Conclusion

This project presents significant progress in integrating emotional context into speech recognition technologies. Utilizing specialized architecture and pre-trained models, the system effectively discerns both textual and emotional nuances in speech. Although challenges such as cross-dataset generalization and emotional diversity persist, the technology holds promising potential for applications in assistive technologies, gaming, and customer service.

Future efforts will concentrate on enhancing model accuracy through the incorporation of emerging datasets and refining language support to extend the technology's applicability. Collaborative initiatives with industry stakeholders are essential to tailor and refine models to meet specific market needs, striving to develop more empathetic and responsive systems that align with human communication.

References

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[2] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. 16(6):1505–1518.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations.

[4] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units.

[5] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. ISSN: 2379-190X.

[6] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. 45(9):10745–10759.

[7] Nikolaos Antoniou, Athanasios Katsamanis, Theodoros Giannakopoulos, and Shrikanth Narayanan. Designing and evaluating speech emotion recognition systems: A reality check case study with IEMOCAP. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

[8] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. 13(5):e0196391.

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. 42(4):335–359.

[10] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. 10(4):471–483.