A Survey on Speech Emotion Recognition

N V S Abhishek
Department of Computer Science and Engineering
IIT Bombay
nvsabhishek.india@gmail.com

Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT Bombay
pushpakbh@gmail.com

Abstract

Speech Emotion Recognition (SER) is the task of identifying the emotion expressed in a spoken utterance. Emotion recognition is essential in building robust conversational agents in domains such as law, healthcare, education, and customer support. Most of the studies published on SER use datasets created by employing professional actors in a noise-free environment. In natural settings such as a customer care conversation, the audio is often noisy with speakers regularly switching between different languages as they see fit. We have developed, in collaboration with a leading unicorn in the Conversational AI sector, Natural Speech Emotion Dataset (NSED). NSED is a natural code-mixed speech emotion dataset where each utterance in a conversation is annotated with emotion, sentiment, valence, arousal, and dominance (VAD) values. In this paper, we give the background for speech production, processing and emotion recognition. We also discuss the various techniques and datasets used for the problem of speech emotion recognition.

1 Introduction

Emotion recognition is the task of identifying the implied emotion from a piece of text, speech or image. A variety of emotions exist in nature with varying degree of intensity and polarity. There are two models for expressing emotions. The dimensional model of emotions express emotions along the dimensions of arousal, valence and dominance. Arousal, valence and dominance signify the intensity, polarity and control exerted by an emotion, respectively, in a conversation. For example, *anger* has high arousal, negative valence and high dominance whereas *fear* has low arousal, negative valence and low dominance. The categorical model of emotions defines a set of discrete emotion classes such as *anger, happy* and *sad* for various downstream tasks.

Conversational agents which can participate in a dialogue effectively have massive applications across multiple domains. Mensio et al. (2018) discussed three steps of evolution for conversational agents: textual interaction, vocal interaction and embodied interaction. Recently, OpenAI released ChatGPT, a multi-lingual textual conversational model based on the large language model (LLM) GPT 3.5. ChatGPT can "answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests" effectively while retaining knowledge from the conversational context as well as the pre-training phase (Bang et al., 2023). ChatGPT has outperformed state-of-the-art LLMs for various tasks in the zero-shot setting. It was found that, through interactivity, one can improve the performance of ChatGPT by 8% ROUGE-1 on summarization tasks and 2% ChrF++ on the machine translation tasks (Bang et al., 2023). With the integration of interactability, ChatGPT has leaped over traditional LLMs with applications across several domains such as law, healthcare, finance and education.

In many situations, conversation through the speech modality is favorable and convenient as compared to the textual modality. ChatGPT, while a great conversational agent, can only work with the textual modality. A conversational agent which can take speech input and give speech responses that are polite and empathetic, in an end-to-end fashion, is the next phase of evolution for interactive chatbots.

Conversational agents such as ChatGPT need to recognize the emotion of the human interlocutor correctly in order to give responses which are polite and empathetic in nature. Emotion recognition, when done efficiently by chatbots, make the conversations more human-like. Speech

emotion recognition is an important sub-task while developing speech-to-speech chatbots.

Our specific **problem statement** is to solve Speech Emotion Recognition (SER) where the input is the raw audio of a spoken utterance in a dyadic conversation and the output is its corresponding emotion label, valence, arousal and dominance.

1.1 Motivation

SER has been an important yet challenging task for researchers. Whenever there is a human-machine interaction in environments where only speech can be propagated, SER becomes a key step for the machine to generate an appropriate response. The task of Emotion Recognition in Conversation (ERC) has many controlling variables such as the context, topic, argumentation logic and speaker/listener personalities, describe the emotional states of the interlocutors.
A recent study (Catania and Garzotto, 2022) explored the benefits of using an emotion-aware chatbot to help people with alexithymia, a condition which makes it difficult to understand and express emotions. Alexithymia is common in people with neurodevelopmental disorders (NDD). The chatbot provided different utterances to the users and asked them to imitate those utterances by inducing some kind of emotion such as joy or anger. It was found that the interaction with the chatbot became more straightforward as users acquired familiarity: 17 of the 19 participants could perform all emotional activities with progressively decreasing help from the facilitator.
Most of the SER datasets available today are created by employing professional actors in a clean noise-free environment. In a natural setting, conversations are impromptu, often involving frequent code-mixing and code-switching between multiple languages such as Hindi, English, Marathi, etc. In a customer care setting, it is essential for conversational agents to be polite and empathetic in response to the emotion expressed by the customer. This leads to better overall customer satisfaction and customer retention rates.

Our **industry-partner** is a unicorn company in the Conversational AI sector which empowers over 45000 businesses across the world through their conversational messaging platform. This platform helps businesses engage with customers effectively across commerce, marketing and support with over 9 Billion messages per month. Their mission is to "build the most advanced and innovative platform for conversational engagement with a focus on delivering customer delight".

We are collaborating with them to work on speech emotion recognition. Through our discussions with them, we explored various ways to approach this problem. They gave us a clear picture of the real-world challenges that are existent in the conversational AI sector. Some of the major challenges are: frequent code-mixing, low-quality recordings and a lack of annotated natural conversational datasets.

As we will discuss further, the dataset annotated for our experiments, NSED, contains customer care conversations from the escalation department of a customer care service. High accuracy for negative emotion recognition is essential, because customers expressing negative opinions/views need to be pacified with urgency, lest complaints and dissatisfaction snowball and get out of hand. Escalation of negative opinions speedily is crucial for business interests. This tells us that a speech emotion recognition model operating for the escalation department should be very good in detecting negative emotions in conversations.

An SER model which is capable of capturing contextual information well and is robust to the variations introduced by a natural code-mixed conversation dataset needs to be developed. This model then can be utilised in making speech-to-speech conversational agents more polite and empathetic in an escalation department setting.

2 Background

In this section we provide the background knowledge required to work on the problem of speech emotion recognition.

2.1 Sentiment Analysis

Sentiment analysis is the field which deals with extracting sentiments and opinions of people towards entities such as products, services, other people, places, organizations and their attributes. Senti-

ment analysis is often referred to as opinion mining as well. According to (Liu, 2012), opinion can be represented in one of the following ways.

2.1.1 The Quadruple Representation of Opinion

The quadruple representation of opinion, (g, s, h, t), is defined as follows:

- **g** is the sentiment target.
- **s** is the sentiment about the sentiment target g.
- **h** is the opinion holder.
- **t** is the time at which the sentiment was expressed.

2.1.2 The Quintuple Representation of Opinion

The quintuple representation of opinion, $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, is defined as follows:

- e_i is the name of the entity.
- a_{ij} is the an aspect of the i^{th} entity.
- h_k is the opinion holder.
- t_l is the time at which the sentiment was expressed.
- s_{ijkl} is the actual sentiment expressed.

Example: *"Ravi thinks that the picture quality of the iPhone 12 is very good but its battery life is very poor."* Here two opinion quintuples, **(iPhone 12, Picture Quality, Positive, Ravi, 10 PM 04/05/22)** and **(iPhone 12, Battery Life, Negative, Ravi, 10 PM 04/05/22)**, can be extracted.

2.2 Emotion Analysis

Emotion analysis is an extension to the sentiment analysis task. Instead of classifying an utterance into positive, negative or neutral, we classify it into emotion classes such as anger, happiness, sad and excited. This tells a lot more about the human mind when it reacts to various things in nature. There are a number of emotion models which help us differentiate between emotions. A few of those models are described below.

2.2.1 Two-Dimensional Emotion Model

Dimensional model of emotions is based on parameters such as valence, arousal and dominance. Valence reflects if an emotion is positive or negative. Arousal reflects the intensity of the emotion. Dominance reflects the amount of control exerted by an emotion. The two-dimensional emotion model is determined by two factors, valence and arousal. For example, as shown in the Figure 2, **anger** has high arousal and negative valence whereas **calm** has low arousal and positive valence.

Figure 1: Two-dimensional emotion model (Byun and Lee, 2021)

2.2.2 NRC-VAD Lexicon

The NRC-VAD Lexicon includes a list of more than 20,000 English words and their valence, arousal, and dominance scores. For a given word and a dimension (V/A/D), the scores range from 0 (lowest V/A/D) to 1 (highest V/A/D). There are versions of this lexicon in over 100 languages facilitated by translating the English words using Google Translate. This lexicon can be utilized when we are doing multi-modal emotion recognition using both speech and text modalities. This lexicon is used to find the word-level VAD values in the experiments shown in future sections of this study.

2.2.3 Ekman Classes of Emotion

In 1992, Paul Ekman and his colleagues described six basic human emotions. The six basic emotions are **anger**, **disgust**, **happy**, **sad**, **surprise** and **fear**. Each of these basic emotions can exist in varying degrees of intensities to give rise to other complex emotions. For example, anger with less intensity is annoyance while with more intensity it can become outrage.

2.2.4 Navarasa Theory of Emotions

In ancient Indian scriptures, ***Navarasa*** refers to nine emotions as expressed and perceived in various traditional art forms. ***Nava*** means nine and ***rasa*** means emotion. The nine emotions are listed below.

- ***Shringara*** meaning love and beauty.
- ***Hasya*** meaning humor.
- ***Karuna*** meaning compassion.
- ***Raudra*** meaning rage.
- ***Veera*** meaning valor.
- ***Beebhatsa*** meaning disgust.
- ***Bhayanaka*** meaning fear.
- ***Adbhuta*** meaning wonder.
- ***Shanta*** meaning peace.

2.2.5 Plutchik's Wheel of Emotions

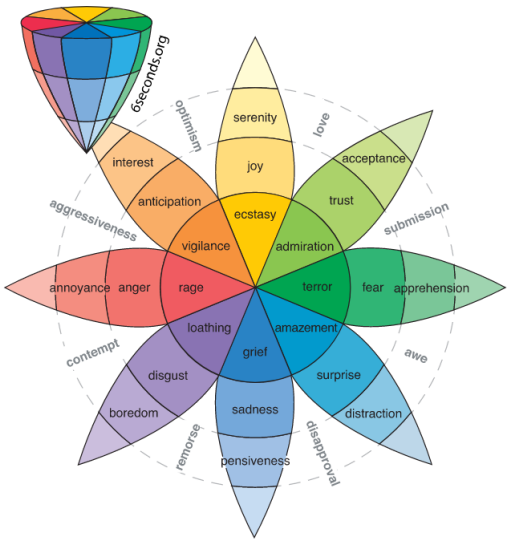


Figure 2: Plutchik's Wheel of Emotions (Plutchik, 2001)

Psychologist Robert Plutchik created a wheel of emotions to better understand various basic and complex emotions. Figure 1 depicts the Plutchik's wheel of emotions. The wheel of emotions can be interpreted as described below.

- **Basic Emotion Pairs:**
  - **Primary**: The eight sectors are designed to indicate that there are eight primary emotions: **anger**, **anticipation**, **joy**, **trust**, **fear**, **surprise**, and **disgust**. The emotions are arranged in the second concentric ring of the wheel.
  - **Opposites**: Each of the primary emotion lobes has a polar opposite lobe. Joy is the opposite of sadness. Fear is the opposite of anger.
- **Intensity**: The wheel's dimension represents intensity – more intense emotions are present on the inner circles and intensity decrease when we move towards the end of the lobes. The darker the shade, the more intense the emotion. For example, anger at its least level of intensity is annoyance. At its highest level of intensity, anger becomes rage.
- **Combination of Emotions:** More complex emotions can be formed with the combination of the basic ones. E.g. 'Serenity' and 'Acceptance' gives 'Love'.

2.3 Speech Emotion Recognition in Conversation

As mentioned in Poria et al. (2019), given the speech input and transcript of the conversation, the Emotion Recognition in Conversation (ERC) task can be formally defined as follows:

- $[(u_1, p_1), (u_2, p_2), ..., (u_N, p_N)]$ is a conversation with N utterances. Each utterance u_i is spoken by party p_i.
- $ui = [u_{i,1}, u_{i,2}, ..., u_{i,T}]$ consists of T words where $u_{i,j}$ is the j^{th} word in the i^{th} utterance.
- The task of ERC is to predict the emotion label e_i of each utterance u_i.

2.4 The Speech Signal

The speech signal has become an integral part of human communication after thousands of years of evolution. In this chapter we will discuss how speech is produced and the various features that are relevant to the task of sentiment analysis.

2.4.1 Speech Production

Figure 3 shows the anatomy of human speech production system. There are a number of mechanisms which should work synchronously for us to produce speech signals properly. Speech production can be

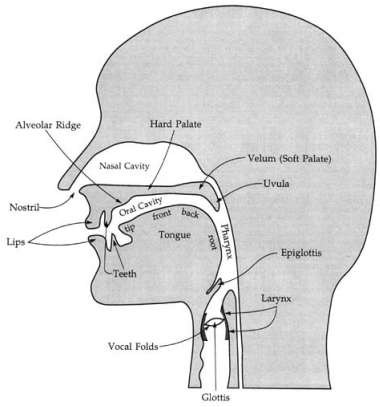


Figure 3: Speech Production Mechanism in Humans (L.Rabiner, 1993)

broken down into three steps, namely, **respiration**, **phonation** and **articulation** according to (Anderson, 2018).

- **Respiration**: When we breathe in, air gets rushed into our lungs. To produce sound, our lungs push air out to the larynx via the trachea.
- **Phonation**: The larynx has a pair of bands of smooth muscle tissues known as the vocal folds. The gap between the vocal folds is known as the glottis. The vocal folds are held wide open to produce voiceless sound. The vocal folds vibrate to produce voiced sound. This transformation of the sounds by the vocal folds is known as phonation.
- **Articulation**: Shaping the flow of air in the mouth by adjusting a number of "articulators" is known as articulation. Tongue, lips and velum are some of the articulators. The tongue can be moved into different positions relative to the Hard Palate to produce different types of sound. Velum (Soft Palate) acts as a valve between the oral cavity and the nasal cavity for air flow. The velum rests against the back of the throat to produce sound only via the oral cavity. Velum can also be opened up for air to flow from the nasal cavity. This produces nasal sounds. Lips also articulate sound based on how open they are.

2.4.2 The International Phonetic Alphabet

The English language is highly ambiguous in spelling speech sounds. Words like tray, weigh and they have the same vowel sound [a] but are spelled in three completely different ways. This is because the English language has a lot of vowel sounds. Words like cough, tough and though have the same spelling but produce completely different speech sound. The IPA is an repository of speech sound transcriptions. This standardization helps us to get unambiguous transcriptions of sounds which make research in linguistics and related areas a lot easier. Vowel Sounds Sounds can be differentiated based on their acoustic properties. Sonority is one such acoustic property. Sounds with high amount of air flow without any obstructions and high amount of vibrations in vocal folds are high in sonority. Vowel sounds are highly sonorous because of which they can be sustained for longer amount of time without any break in speech. Consonant Sounds Consonant sounds are usually low in sonority because the air flow is obstructed while producing consonant sounds. Consonant sounds can be both voiced and unvoiced. Consonant sounds are classified on the basis of two parameters, place of articulation and manner of articulation. Place of articulation tells us where the obstruction happens in the vocal tract. Manner of articulation tells us how the obstruction happened. Based on these there are a number of consonant sounds in the english language.

2.4.3 Speech Features

Speech features can be distinguished into two categories, spectral and temporal features. Temporal features are time domain based features. A spectrum is the representation of the audio signal in frequency domain. Spectral features are extracted from the spectrum such as MFCC, spectral centroid, spectral spread, sprectral roll-off and brightness. Some of the speech features are described below.

- **Mel Frequency Cepstral Coefficients:** MFCCs are widely used in speech recognition systems. These are the set of coefficients of Mel Frequency Cepstrum obtained by carrying out a variety of transformations on the raw audio signal as explained further in Section 2.4.4.
- **Spectral Features:** Spectral Centroid represents the center of gravity of a spectrum. Spectral Spread is the second central moment of a spectrum.
- **Pitch-related Features:** There are various pitch related features like pitch contour, pitch

period and harmonic-to-noise ratio.

- **Prosodic Features:** These features reflect the variations in speech related to stress, rhythm, and intonation. Examples include: Duration of phonetic units (e.g., phonemes, syllables, words), speech rate. pauses and silence durations.

2.4.4 Speech Feature Extraction Techniques

There are a lot of speech feature extraction techniques present today out of which Mel-frequency Cepstral Coefficients (MFCCs) feature extraction is the most popular. Its main idea is to come up with a set of features which represent human perception of speech and using which one can detect phones more efficiently. As shown in Figure 4, the following are the different stages of MFCC extraction.

- **A/D Conversion:** This stage converts the analog audio signal into discrete space by taking samples out of the analog signal with frame size of 20-50ms with a shift of 10ms for each frame.

- **Pre-Emphasis:** In this stage the energy in the higher frequencies. This is because of **spectral tilt** which refers to voiced sound like vowels having more energy in lower frequencies. This increases the accuracy by which phones are detected.

- **Windowing:** Windowing slices off parts of the pre-emphasised speech signal into frames. The basic windowing operation is shown in Equation 1. N is the feram size.

$$y[t] = W[t] * x[t] \tag{1}$$

Equation 2 shows rectangular windowing.

$$W[t] = \begin{cases} 1 & 0 \leq t \leq N-1 \\ 0 & othewise \end{cases} \tag{2}$$

Equation 3 shows sine windowing.

$$W[t] = \begin{cases} \sin \frac{\pi t}{N} & 0 \leq t \leq N-1 \\ 0 & othewise \end{cases} \tag{3}$$

Equation 4 shows hamming windowing.

$$W[t] = \begin{cases} 0.54 - 0.46 \cos \frac{2\pi t}{N} & 0 \leq t \leq N-1 \\ 0 & othewise \end{cases} \tag{4}$$

- **DFT:** Discrete Fourier Transform (DFT) is performed to convert the signal into frequency domain.

- **Mel-Filterbank:** Human hearing can not perceive higher frequencies of sound as well as it can perceive the lower frequencies. The spectrum generated by the DFT is converted into the mel-scale of frequencies. In the mel-scale the sounds which are perceived to be equidistant have the same number of mels between them. The mel-frequency of a frequency in the spectrum can be calculates as shown in the Equation 5.

$$mel(f) = 1127log(1 + \frac{f}{700}) \tag{5}$$

Triangular filters are applied to the spectrum in which the size of the triangular filter increases logarithmically for higher frequencies to include more number of frequencies in one bin.

- **IDFT:** Inverse Discrete Fourier Transform is performed on the log of the mel-spectogram to produce a cepstrum. Cepstrum helps us to distinguish between the information about the vocal filter and the glottal source as vocal filter is essential for phone production.

- **Time Domain Derivatives:** MFCCs have 39 values. First 12 values of the cepstrum are picked up as it is. The thirteenth value is the energy of the speech frame. Next 13 values are the first order derivatives of the cepstral values which show the rate of change of these values. The last 13 values are second order derivatives of the cepstral values.

2.5 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the task of converting a piece of speech audio signal into a sequence of tokens (words, syllables or characters). ASR techniques can be broadly classified into two categories. (a) Traditional ASR Systems (b) End-to-End ASR Systems.

2.5.1 Traditional ASR Systems

As shown in Figure 5, traditional ASR systems have four major components which are described below.

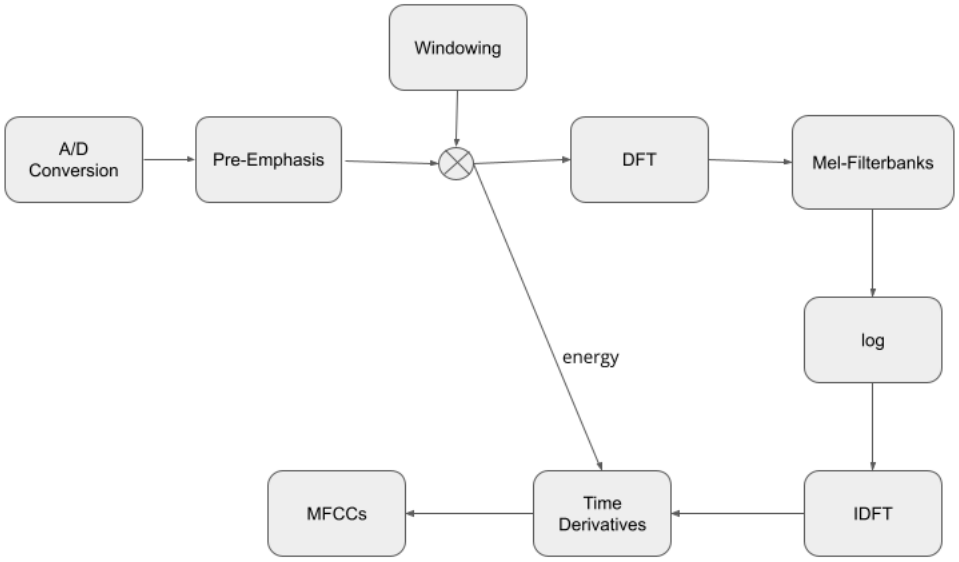


Figure 4: MFCC Extraction Pipeline

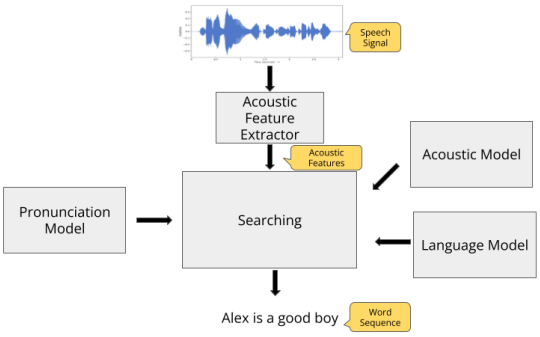


Figure 5: Traditional ASR System

- **Acoustic Model:** An acoustic model takes in the acoustic features generated by the acoustic feature extractor and converts them into a sequence of phonemes. Traditionally acoustic models were completely modelled using Hidden Markov Models (HMMs). Now various deep learning approaches have come up to aid the acoustic modelling process.

  1. Hybrid DNN-HMM system: Deep neural networks are used to calculate the HMM observation probabilities.
  2. Tandem system: Deep neural networks are used to generate the input features that is used by the acoustic model.

- **Pronunciation Model:** Pronunciation model/dictionary provides a set of pronunciations (phoneme sequence) for words. It is usually created by linguists and language experts. It is constructed by first deciding the vocabulary and then deciding the pronunciations to be included for each word in the vocabulary. Out-of-vocabulary (OOV) words are a problem when pronunciation models are used in ASR systems because the pronunciation dictionary doesn't have a entry for those words.

- **Language Model:** A language model gives the probability of the next word given the word sequence generated till now. Ngram language models are used extensively to reduce the amount of computation. In ngram language models, the next word's probability only depends on the last n words encountered in the word history.

- **Searching:** Searching/decoding refers to the task of finding the most probable word sequence given the observation sequence produced by the acoustic feature extractor. There

are two types of decoding techniques:

1. Greedy Decoding: In this type of decoding, the most probable next word is chosen greedily. This may result to a wrong prediction because there may exist a word sequence with higher overall probability.
2. Beam Search Decoding: In this type of decoding, a range/beam of next word probabilities are considered around the most probable next word. This gives us a good chance of ending up with the most probable word sequence. This requires us to maintain multiple possible word sequences/paths while computation because of which it is slower than the greedy decoding approach.

2.5.2 End-to-End ASR Systems

Figure 6 shows the schematic of an end-to-end ASR system. An end-to-end ASR system takes the acoustic features extracted and converts it to a corresponding token sequence without any additional processing.

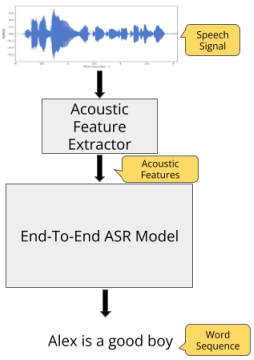


Figure 6: End-to-end ASR System

- **Encoder-Decoder with Attention:** As shown in Figure 7 Encoder-decoder is an end-to-end architecture where the encoder encodes the input features into an intermediate representation. The decoder takes this intermediate representation as input and generates the output. The encoder, implemented using recurrent neural networks (RNNs), produces a hidden state, h_i, in every timestep corresponding to every frame of the utterance. In a normal encoder-decoder architecture without attention the last timestep hidden state vector or the average of all the hidden state vectors is passed on to the decoder. But as number of frames increases this approach won't be able to remember long-term dependencies. Then comes the attention mechanism, as shown in Figure 6 where each hidden state is assigned an attention weight corresponding to their importance for computing the hidden state of the decoder, s. Attention weights are computed as a function of h_i and s normalized using the softmax function as shown in Equation 6

$$\alpha_i = \exp(h_i * s) / \sum_{j=1}^{t} \exp(h_j * s) \quad (6)$$

The weighted sum of the hidden states of the encoder gives the context vector c which is then fed to the decoder as shown in the Equation 7.

$$c = \sum_{i=1}^{t} \alpha_i * h_i \quad (7)$$

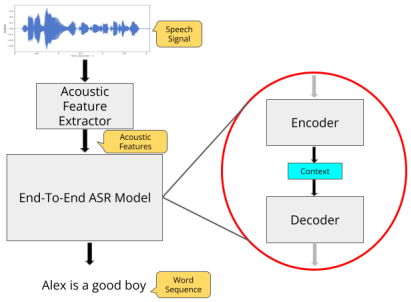


Figure 7: Encoder-Decoder end-to-end ASR system

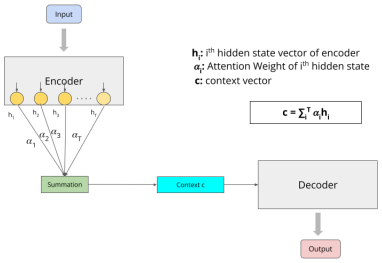


Figure 8: Encoder-Decoder with Attention

- **CTC Loss Function:** Connectionist Temporal Classification (CTC) is a loss function

which was introduced to overcome the need of alignments between speech utterances and word sequences while training RNNs. This is a loss function which considers all the possible alignments to produce the given word sequence. An augmented output sequence vocabulary, Y', is created by adding a blank symbol (_) to the output sequence vocabulary, Y. A 2-step operator $B : Y' - > Y$ is defined which converts an output sequence from the augmented output space to the real output space. In the first step, all the output symbols with consecutive occurrences are reduced to only a single symbol. In the second step, all the blank symbols are collapsed. For example, $B(a\_\_bb\_c) = (a\_b\_c) = (abc)$. This shows that a single output sequence can have a lot of possible alignments. The CTC objective function is defined in Equation 8. The probability of y being the output sequence given x as input is the summation of probabilities of all the possible alignments for y given x.

$$CTC(x, y) = Pr(y|x) = \sum_{a \in B^{-1}(y)} Pr(a|x) \tag{8}$$

The probability of each alignment is the product of the probabilities of all the output symbols in the alignment as shown in Equation 9. This is the assumption taken by CTC loss that all the output symbols are conditionally independent of each other. Using this objective function RNN-based end-to-end models can overcome the need of alignment of utterance and word sequences.

$$Pr(a|x) = \prod_{t=1}^{T} Pr(a_t|x) \tag{9}$$

2.5.3 Evaluating ASR Models

ASR models are usually evaluated using word error rate (WER). WER gives the minimum number of edits (insertions/deletions/substitutions) required to change the predicted word sequence, W^* to the target word sequence, W_{ref}. WER is formulated according to the Equation 10 for a test set with N instances. Here $Ins_j, Del_j and Sub_j$ are insertions, deletions and substitutions in j^{th} predicted word sequence and l_j is the length of the j^{th} word reference.

$$WER = \frac{\sum_{i=1}^{N} Ins_j + Del_j + Sub_j}{\sum_{1}^{N} l_j} \tag{10}$$

3 Literature Survey

In this section, we discuss the various types of techniques used for the task of speech emotion recognition.

3.1 Knowledge-based Techniques for SER

In Chakraborty et al. (2016) the proposed framework used contexts (derived from linguistic contents) and the knowledge regarding the time lapse of the spoken utterances in the context of an audio call. This shows the importance of context in predicting emotion of spontaneous speech. With the advent of machine learning and deep learning techniques knowledge-based systems took a backseat as it takes a lot of time to craft knowledge-based systems.

3.2 Statistical Machine Learning Techniques for SER

In machine learning techniques feature engineering is essential. Acoustic features such as MFCCs, energy, pitch and mel-spectrograms are extracted after a good amount of feature analysis. Then statistical machine learning models such as SVM are used to predict the appropriate emotions. Khalil et al. (2019) considered Speech emotion recognition as an exciting ingredient of Human Computer Interaction (HCI). The main approach for SER must be feature extraction and feature classification. Linear and nonlinear classifiers can be used for Feature classification. In linear classifiers, frequently used classifiers are Support Vector Machines (SVMs), Bayesian Networks (BN). Since, Speech signal is considered variating, thus, these types of classifiers work effectively for SER.

3.3 Deep Learning based Techniques for SER

With the boom in internet, high volumes of data have now become available with an increase in computation power. Artificial neural networks have enhanced the performance of various machine learning systems because of this boom. In Gulati et al. (2020) the Conformer architecture was introduced where Convolutional Neural Networks (CNNs) were integrated with parts of the Transformer architecture (Vaswani et al., 2017) for the task of speech recognition. This integrated model achieved a new state-of-the-art performance with a Word Error Rate (WER) of 1.9%/3.9% for the LibriSpeech test/testother datasets (Panayotov et al., 2015). (wen Yang et al., 2021) In Pepino

et al. (2021) learned speech representations from Wav2Vec 2.0 are utilized in a downstream model for speech emotion recognition. The proposed model outperformed the state-of-the-art for IEMOCAP (Busso et al., 2008a) and RAVDESS (Livingstone and Russo, 2018) datasets. The study also showed that combining low-level acoustic features with the Wav2Vec 2.0 speech representations resulted in performance gains. In Wang et al. (2020) the wav2vec2-conformer model was introduced where the attention-block is replaced by the conformer-block. This model achieved a better Word Error Rate (WER) as compared to the traditional Wav2Vec 2.0 model for speech recognition. WavLM (Chen et al., 2022) is a new SSL-based architecture, released this year by Microsoft, which has achieved the state-of-the-art in 14 speech processing tasks of the SUPERB (wen Yang et al., 2021) benchmark. To pre-train the WavLM model, 96 thousand hours of speech data were used for the tasks of masked audio prediction and denoising. In Speech Emotion Recognition (SER), the modalities that are available for any model to compute are the speech signal and its ASR transcritps. Li et al. (2022) showed that fusing ASR transcripts into speech emotion recognition achieved weighted accuracy of 63% which is close to the performance of the model using ground truth transcripts (64%) on the IEMOCAP dataset. This study also showed that the ASR text output showed better performance when compared to the ASR hidden layer outputs of the wav2vec2 architecture. This illustrates the strength of the textual modality even when it is taken from an ASR system for the task of emotion recognition.

3.3.1 Deep Learning based Speech Representation Learning Models

Self-supervised learning (SSL) based architectures learn powerful speech representations from unlabelled speech data. Some of the popular SSL-based architectures are described below.

- **Wav2Vec:** The wav2vec architecture explores unsupervised pre-training using large amounts of unlabelled audio clips to predict representations for these audio clips via a contrastive binary classification task. Figure 9 shows the architecture of wav2vec with two layers of convolutions.

- **Wav2Vec2:** Wav2vec2.0 is a self-supervised approach to learn powerful speech representations. Wav2vec2.0 has 3 major components: feature encoder, quantization module and the transformer block. The feature encoder takes the raw audio and converts them into latent representations. The quantization module takes these latent representations and generates their discretized representations. The architecture learns representations for masked audio segments via a contrastive loss over the contextualized representations from the transformer block and the quantized representations. Wav2vec2 is pre-trained on 56K hours of unlabelled Librispeech data. Figure 10 shows the architecture of wav2vec2.

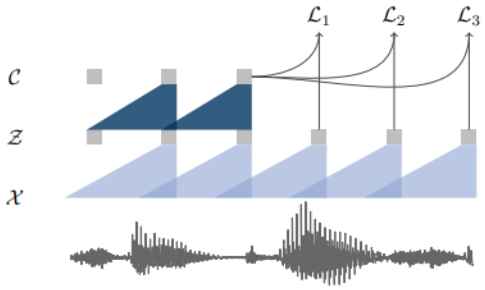


Figure 9: Wav2vec architecture (Schneider et al., 2019)

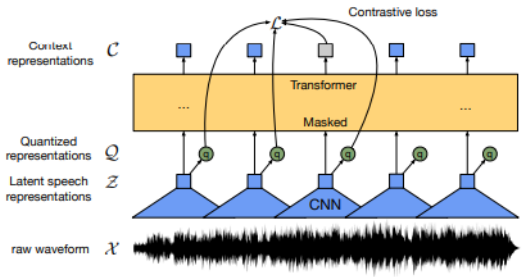


Figure 10: Wav2vec2 architecture (Baevski et al., 2020)

- **XLSR-Wav2Vec2:** XLSR-wav2vec2 is similar to the wav2vec2 architecture. One key difference between the two is that XLSR-wav2vec2 is pre-trained on 53 different languages to be able to generate representations which are language-agnostic and bridge gaps between any two languages by learning common representations for both. Figure 11 shows the architecture of xlsr-wav2vec2.

- **WavLM:** The SSL-based architectures discussed till now focus on generating speech representations for only the task of automatic

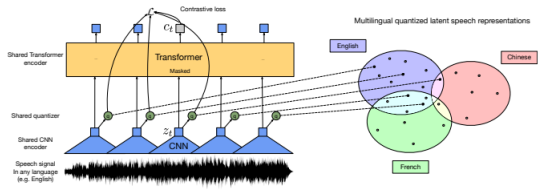


Figure 11: XLSR-Wav2vec2 architecture (Conneau et al., 2020)

speech recognition. WavLM is a new SSL-based framework which generated powerful speech representations which can be used for several speech processing tasks such as speaker identification, speech emotion recognition etc. WavLM has two learning tasks for pre-training: denoising and masked speech prediction. WavLM is trained on 94K hours of unlabelled speech data from multiple sources, not just from audio books. WavLM sets state-of-the-art performance in 14 subtasks of the SUPERB benchmark. Figure 12 shows the architecture of WavLM.

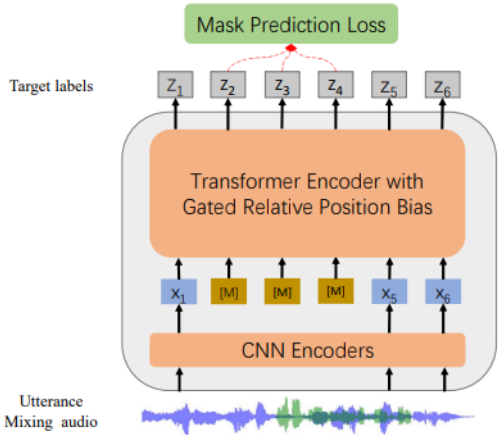


Figure 12: WavLM architecture (Chen et al., 2022)

4 Datasets

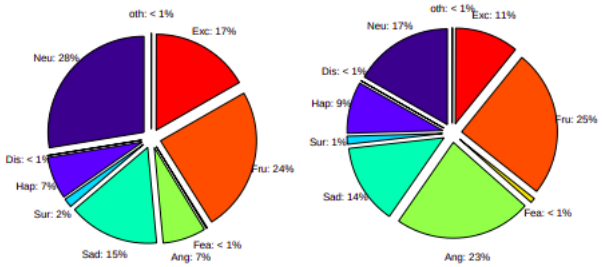
In this chapter, some of the public datasets, extensively used for the task of speech emotion recognition, are briefly described. Also, the annotation methodology used to annotate the Natural Speech Emotion Dataset (NSED) is described in detail.

4.1 Public Datasets

In this section, popular public speech emotion recognition datasets are descrined briefly.

4.1.1 IEMOCAP

Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) (Busso et al., 2008b) is a multi-modal emotion database captured through dyadic conversations of speakers. The distribution of all the emotions for scripted and unscripted conversations is shown in Figure 13 .



(a) Scripted session (b) Spontaneous sessions

Figure 13: IEMOCAP distribution of emotions for scripted and spontaneous conversations (Busso et al., 2008a)

4.1.2 RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo, 2018) is a multi-modal database of speech and song in the english language.

4.1.3 TESS

Toronto Emotional Speech Set (TESS) is a speech emotion dataset with data for seven basic emotions of anger, happy, sad, disgust, fear, neutral and surprise. Two actresses uttered 200 target words in these seven emotion categories giving rise to 2800 emotion-labelled utterances.

4.1.4 SUBESCO

SUST Bangla Emotional Speech Corpus (SUBESCO) is a speech emotion dataset in the bengali language. This dataset is one of the largest publicly available speech emotion dataset for Indian languages. Ten male and ten female actors utter sentences with seven different emotions of anger, happy, sad, disgust, surprise, neutral and fear. The dataset is almost 8 hours long.

4.1.5 Emo-DB

Emotion-Database (Burkhardt et al., 2005) is a database of emotional speech in the german language. 5 male and 5 female voice actors simulated emotional speech. The database consists of 800

Name	Language	# Speakers	Emotion Classes
IEMOCAP	English	10	anger, disgust, fear, frustration, sadness, excitement, happiness, surprise, neutral

Table 1: IEMOCAP dataset specification

Name	Language	# Speakers	Emotion Classes
RAVDESS	English	24	anger, disgust, fear, sadness, happiness, surprise, calm, neutral

Table 2: RAVDESS dataset specification

Name	Language	# Speakers	Emotion Classes
TESS	English	2	anger, disgust, fear, sadness, happiness, surprise, neutral

Table 3: TESS dataset specification

Name	Language	# Speakers	Emotion Classes
SUBESCO	Bengali	20	anger, disgust, fear, sadness, happiness, surprise, neutral

Table 4: SUBESCO dataset specification

sentences as each actor records 10 sentences for each emotion.

4.2 Natural Speech Emotion Dataset (NSED) Annotation

Natural Speech Emotion Dataset (NSED) is a code-mixed dyadic customer care conversation dataset created in collaboration with our industry partner. Below are the steps followed to create this dataset. Figure 14 shows the overall pipeline for annotating NSED.

4.2.1 Data Recording

Our industry partner provided us with over 18000 dyadic customer care audio recordings with duration ranging between a few seconds to about an hour and their corresponding machine-generated text transcripts. All the audio recordings were single-channel (mono) with a sampling rate of 8000Hz. The conversations are interactions between a customer and a customer care executive from the complaint escalation team of a car servicing company. Both the speakers, in most of the audio recordings, switch between Hindi and English freely with some occasional use of regional words in languages such as Marathi.

4.2.2 Data Processing

Thirty audio recordings were chosen, each of which was 8-10 minutes long making a total of 4.5 hours long audio recordings. The audacity tool was used to process audio files. Each of these audio recordings was clipped into smaller audio clips corresponding to each **speaking turn**. A speaking turn is defined as the utterance corresponding to a particular speaker before and after any other speaker speaks. Each of these audio clips were then aligned with their corresponding machine-generated transcripts and were tagged with either "customer" or "executive" depending on who was speaking. The machine-generated transcripts contained many crucial mistakes such as wrongly transcribing the word "escalation" as "cancellation". So, the transcripts were corrected, manually, in order to achieve a better quality of textual data. In some instances, the audio quality drops drastically, making it very difficult to understand the words that are being spoken. In this case, a tag, **<inaudible>** is used in

Name	Language	# Speakers	Emotion Classes
Emo-DB	German	10	anger, disgust, fear, sadness, happiness, boredom, neutral

Table 5: Emo-DB dataset specification

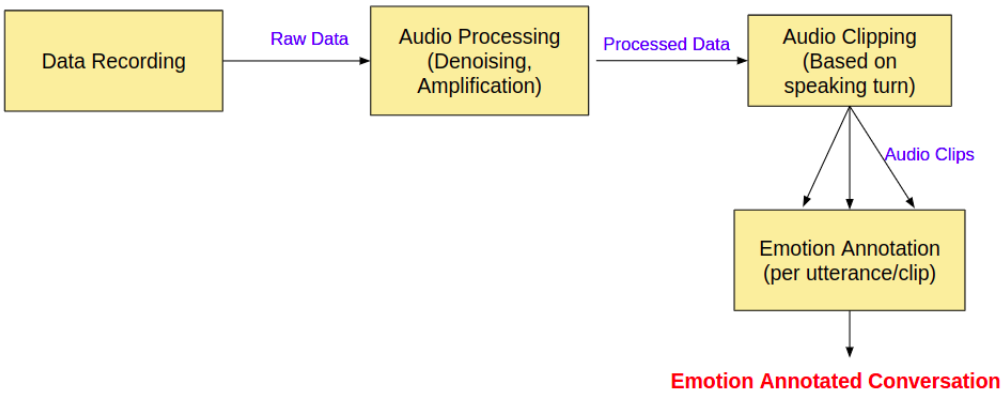


Figure 14: Annotation pipeline for NSED

place of its transcript and further annotations are not performed.

4.2.3 Emotion Annotation

The emotion annotations were performed by four annotators with a graduate degree, proficient in both English and Hindi. The annotators worked in pairs to listen and annotate these clips with emotion (neutral, happy, sad, excited, anger, fear, surprised, frustrated, disgust), sentiment (neutral, positive, negative), valence, arousal and dominance (VAD). VAD values were annotated in a scale from 1 to 10 where (5, 5, 5) corresponds to the VAD values of a completely neutral emotion. For VAD, 1 represents the minimum value and 10 represents the maximum value any of the dimensions can have e.g. for valence, 1 represents the most negative and 10 represents the most positive any emotion can get. As we can represent 1000 emotions using the VAD dimensional model and only 9 using the categorical emotion model, not all utterances tagged as "neutral" will have VAD values of (5, 5, 5). Each pair annotated the same data in order to calculate the agreement between them.

4.2.4 Dataset Distribution

Figure 15 shows the dataset distribution based on emotion classes. Emotion classes considered for the dataset are: neutral, anger, frustration, disgust, sad, happy, excited, surprise and fear. The neutral emotion class constitutes the majority of the dataset.

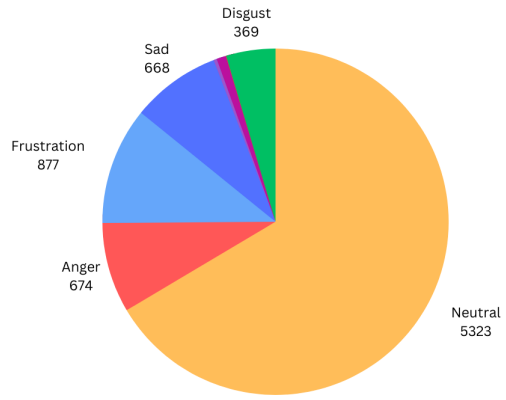


Figure 15: NSED distribution based on emotions

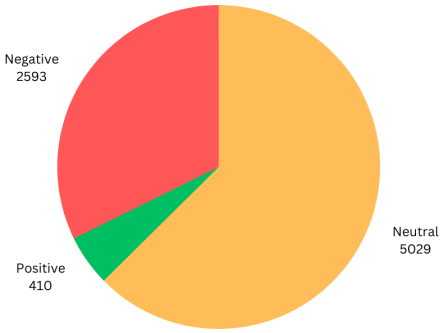


Figure 16: NSED distribution based on sentiments

Apart from that negative emotions like anger, frustration, sad and disgust constitute a fair amount of

Transcript	Speaker	Emotion	Sentiment	Valence	Arousal	Dominance
Do you want someone to get arrested? Haan?	customer	anger	negative	2	8	9
Mai samajhta hun aapko jo bhi problem hui hai. Aage se aapko ye nahi hoga nischint rahiye.	executive	neutral	positive	6	5	5

Table 6: NSED dataset examples

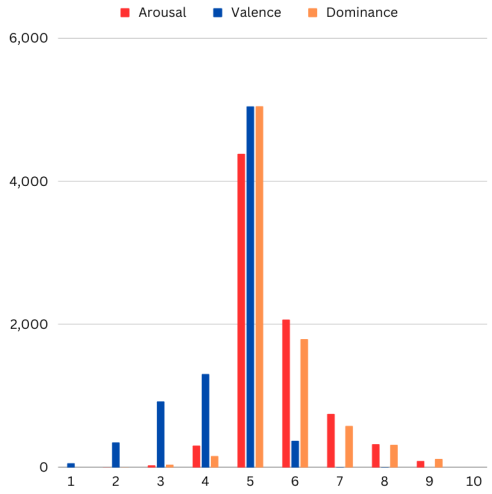


Figure 17: NSED distribution based on valence, arousal, dominance values

the rest of the dataset. Positive emotion classes like happy and excited constitute an insignificant amount of the dataset as these conversations are between a customer, who is dissatisfied with the services of a company and the customer care executive. Figure 16 shows the dataset distribution based on sentiments of positive, negative and neutral. The neutral class here is fewer in number as compared to the emotion label of neutral because there were instances where a particular utterance did not belong to any of the specified emotion classes. Figure 17 shows the distribution of the dataset based on valence, arousal and dominance in the range of 1 to 10. The valence, arousal and dominance values of 5, 5 and 5 respectively, constitute together a completely neutral emotion. Valence values are either 5 or lower as there are fewer positive utterances than negative. Arousal values are also either 5 or higher as negative emotions expressed are usually higer in intensity. Dominance values are also either 5 or higher as highly negative emotions usually exert a lot of dominance as well.

4.2.5 Dataset Examples

Table 6 shows two examples of conversation utterances from the NSED one from cutomer and executive each.

5 Conclusion and future work

Speech sentiment and emotion analysis is vital for coming up with good intelligent interaction systems. To classify a specific emotion several speech features have to be taken into account. Pre-trained transformer-based models have been proved to be useful for the task of speech emotion recognition. The annotation task of Natural Speech Emotion Dataset (NSED) was discussed in detail. NSED can be utilised by NLP researchers for tasks like speech recognition and emotion recognition. Most of the conversations today happen in a code-mixed setting where different languages are used to convey a message. NSED needs to be expanded further. Experiments to remove sensitive personal information possibly learnt by a model trained on NSED need to be explored. This study can be utilised to develop a speech-to-speech conversational chatbot for customer-care conversations. A transcript-correction step can be incorporated to handle the bottleneck of ASR. A Multi-Task Learning (MTL) setup can be incorporated to generate responses which are more emotion-aware (tasks: response generation+VAD prediction).

Ethics Statement

The Natural Speech Emotion Dataset (NSED) dataset used in our experiments was annotated by a team of 4 annotators. Each annotator had to listen to an audio conversation between a customer and a customer-care executive and annotate each speaking turn with emotion, sentiment, valence,

arousal, and dominance values. The conversational audio files were provided to us by our industry partner because of which NSED remains a proprietary dataset. We also acknowledge that the emotions annotated for each utterance might not be the exact emotion intended by the speaker. The emotion annotations are in accordance with the interpretations of the annotators. Consent was taken from both customers and customer-care executives before recording their conversations. The annotators were paid for the time and effort they spent on the annotation task.

References

Catherine Anderson. 2018. Essentials of linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. 2005. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008a. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008b. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Sung-Woo Byun and Seok-Pil Lee. 2021. A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences*, 11(4):1890.

Fabio Catania and Franca Garzotto. 2022. A conversational agent for emotion expression stimulation in persons with neurodevelopmental disorders. *Multimedia Tools and Applications*, pages 1–32.

Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kumar Kopparapu. 2016. Knowledge-based framework for intelligent emotion recognition in spontaneous speech. *Procedia Computer Science*, 96:587–596.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345.

Yuanchao Li, Peter Bell, and Catherine Lai. 2022. Fusing asr outputs in joint training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7362–7366. IEEE.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

B.H.Juang L.Rabiner. 1993.

Martino Mensio, Giuseppe Rizzo, and Maurizio Morisio. 2018. The rise of emotion-aware conversational agents: threats in digital emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1541–1544.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. 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.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.