TOWARDS BETTER UNDERSTANDING OF SPOKEN CONVERSATIONS: ASSESSMENT OF EMOTION AND SENTIMENT

by

Raghavendra Reddy Pappagari

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

February, 2022

Abstract

Emotions play a vital role in our daily life as they help us convey information impossible to express verbally to other parties. While humans can easily perceive emotions, these are notoriously difficult to define and recognize by machines. However, automatically detecting the emotion of a spoken conversation can be useful for a diverse range of applications such as human-machine interaction and conversation analysis. In this thesis, we present several approaches based on machine learning to recognize emotion from isolated utterances and long recordings.

Isolated utterances are usually shorter than 10s in duration and are assumed to contain only one major emotion. One of the main obstacles in achieving high emotion recognition accuracy is the lack of large annotated data. We propose to mitigate this problem by using transfer learning and data augmentation techniques. We show that x-vector representations extracted from speaker recognition models (x-vector models) contain emotion predictive information and adapting those models provide significant improvements in emotion recognition performance. To further improve the performance, we propose a novel perceptually motivated data augmentation method, Copy-Paste on isolated utterances. This method is based on the assumption that the

presence of emotions other than neutral dictates a speaker's overall perceived emotion in a recording.

As isolated utterances are assumed to contain only one emotion, the proposed models make predictions on the utterance level. However, these models can not be directly applied to conversations that can have multiple emotions unless we know the locations of emotion boundaries. In this work, we propose to recognize emotions in the conversations by doing frame-level classification where predictions are made at regular intervals. We compare models trained on isolated utterances and conversations. We propose a data augmentation method, DiverseCatAugment based on attention operation to improve the transformer models. To further improve the performance, we incorporate the turn-taking structure of the conversations into our models.

Annotating utterances with emotions is not a simple task and it depends on the number of emotions used for annotation. However, annotation schemes can be changed to reduce annotation efforts based on application. We consider one such application: predicting customer satisfaction (CSAT) in a call center conversation where the goal is to predict the overall sentiment of the customer. We conduct a comprehensive search for adequate acoustic and lexical representations at different granular levels of conversations. We show that the methods that use transfer learning (x-vectors and CSAT Tracker) perform best. Our error analysis shows that the calls where customers accomplished their goal but were still dissatisfied are the most difficult to predict correctly, and the customer's speech is more emotional compared to the agent's speech.

Thesis Committee

Najim Dehak (Primary Advisor)
Associate Professor
Department of Electrical and Computer Engineering
Johns Hopkins Whiting School of Engineering

Jesús Villalba (Co-Advisor)
Assistant Research Professor
Department of Electrical and Computer Engineering
Johns Hopkins Whiting School of Engineering

Hynek Hermansky
Professor
Department of Electrical and Computer Engineering
Johns Hopkins Whiting School of Engineering

Primary Readers

Najim Dehak (Primary Advisor)
Jesús Villalba (Co-Advisor)

Acknowledgments

There are many people without whom this journey would have been much more difficult. I am grateful to have them in my life. One person I received the greatest support from was my advisor Najim Dehak. He was always been encouraging and gave me the freedom to explore new ideas. I am greatly benefited by his attitude towards work and life in general. Another person who supported me in this journey was my co-advisor, Jesus Villalba. He was always my go-to man for any technical problems I faced. He patiently listens and makes an effort to help me and offer suggestions. When I mention my advisors, I should also mention two more persons Laureano Moro-Velazquez and Piotr Zelasko (who are my unofficial advisors) with whom I worked for most of my PhD. I learned a lot from them, especially the soft-skills aspect.

I extend my thanks to our collaborator Yishay Carmiel from Avaya. The project I worked on with him, customer satisfaction prediction, gave me a direction for my research and instilled a lot of confidence in me. He was always positive and provided a glimpse of research in the industry. I also thank Hynek Hermansky for having faith in me and recruiting me as a PhD student.

Apart from my advisors and collaborators, my life at Baltimore was greatly

influenced by my colleagues: Phani, JJ, Sonal, Nanxin, Bhati, Kataria, Jinyi, Ruizhi, and Arun. The technical discussions I had with them positively shaped my thinking and helped me when I am stuck at something. My friends Gayatri, Aditya, Valli, Lakshmi, and Juhi were a great help to me in my personal life. I always had their back in my tough times and learned a lot from them.

Finally, the most important people without whom I wouldn't have done my PhD is my family. Their constant support kept me going without a second thought on my career. They shielded me from any familial problems and always talked with a smile on their face. I cannot express my gratitude to them in words for their sacrifices.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Speech is one of the most important mediums of communication for humans while interacting with other humans. In general, human interaction using speech contains two channels: verbal and non-verbal (Cowie and Douglas-Cowie, 1995). The verbal channel transmits linguistic information – the message we utter explicitly to the partner. Whereas non-verbal channel encodes more implicit information such as emotion, intent, speaker identity, pauses, etc. There is significant evidence that non-verbal communication plays a crucial role in human interactions (Bambaeeroo and Shokrpour, 2017; Knapp, Hall, and Horgan, 2013; Mehrabian, 2017). Non-verbal communication helps to coordinate subjects and evoke appropriate responses (Cowie et al., 2001). One of the important factors in non-verbal communication is emotion. One's emotions have the capability to alter the conversational partner's responses whether positively or negatively (Schoenewolf, 1990). Research shows that people remember events with intense emotions more easily than events with neutral emotions suggesting that emotions play a role in our

memory and learning processes (Tyng et al., 2017). Take an example of motivational speeches. It is hard to imagine the audience connecting to the speaker and getting anything meaningful out of the speech if it does not have any emotions. Sometimes, our emotions (or other emotions in our life) can affect our decision-making and judgment too (Lerner et al., 2015).

As emotions are key to almost every part of our daily life, automatically recognizing them would help to improve the quality of human lives. In human-human interactions, automatic emotion recognition could help us to understand the mental state of the speakers. Building an emotional profile of the patients could help doctors to diagnose better in the case of mental health disorders. Authors in (Mäntylä et al., 2016) use emotion analysis to predict employee burnout and productivity in the software engineering field. In service-related applications, recognizing emotional segments (specifically negative regions) could help the companies to nudge/train the agents for better responses. Emotional profiles of customers/speakers could serve as a guideline for machines in human-machine interactions. They can be useful to provide personalized emotional responses from personal assistants like Alexa and Google Home. Additionally, there are a number of applications such as automatic analysis of emergency calls for quick response, providing appropriate recommendations to car drivers based on their mental state, synthesizing natural speech for a better experience, and so on. The main goal of all these applications is to improve the quality of human life.

Emotion has been studied extensively in multiple disciplines with the goal of understanding and recognizing it. The field of affective computing

deals with building automatic systems to recognize or synthesize emotions. This field considers the speaker's emotional display for recognition rather than an emotional experience. Emotional experience is mainly dealt with in psychology and neuroscience. The emotional display is what we observe through physiological changes or more generally through signals emanating from the subject.

Automatic speech emotion recognition (SER) concerns building automatic systems to recognize speakers' emotions from their speech. It can be broadly classified into two types: SER for isolated utterances and SER for long recordings. Isolated utterances are usually shorter than 10s and are assumed to contain single major emotion. Utterances longer than 10s can be considered as long recordings and contain more than one emotion. These long recordings can be monologues that contain only one speaker or dialogues between multiple parties. Examples of monologues include broadcast news, classes taught by teachers where the anchor/teacher speaks for a long time. In this thesis, we present several machine learning approaches for SER on both isolated utterances as well as long recordings. We considered three types of isolated utterances: 1) recorded in isolation using actors with targeted emotions, 2) cut from conversations that are meant to produce emotions in an induced manner, and 3) cut from spontaneous podcast conversations. For long recordings, we considered conversations between two speakers.

1.1 Current challenges and proposed approaches

The majority of the research on SER using machine learning is supervised i.e., it requires some data with emotion annotations. The important premise of this research is that there exist some emotional cues in the input signal which enable automatic recognition of emotion. For example, emotion correlates/attributes for acoustic signals include pitch, speaking rate, signal power among others, and for linguistic signals word meaning. Collecting data with emotion annotations requires noting down the listener's perception which could depend on a lot of factors. There is significant evidence that emotion perception can depend on the demographics of the speaker and listener, the relation between speaker and listener, and the context of the emotion expression (Cauldwell, 2000; Campbell et al., 2014; Lindquist, MacCormack, and Shablack, 2015). Due to this inherent lack of consistency in the annotation, building automatic systems is difficult which mainly depends on the consistent occurrence of emotional cues in the signal and the corresponding emotional label. To improve annotation consistency, annotation with multiple annotators is considered to build models. However, this process is very costly and time-consuming. Sometimes the annotators need special training too. In addition, current automatic systems which majorly use machine learning are data-hungry i.e., they perform better with more data (Hestness et al., 2017). Reliance on the annotations can be reduced if these automatic systems are efficient. One way to overcome this problem is exploiting advances in related tasks such as speech recognition (where annotated data is plenty) by transferring the learned knowledge to recognize emotions. The rationale behind

this approach is that the source task (from which we transfer knowledge) and the target task share some characteristics, and using that knowledge could simplify the learning process for the target task. This approach of transferring knowledge is referred to as transfer learning in the machine learning community. In this thesis, we propose to transfer knowledge from speaker recognition technology to emotion recognition (Chapter 3, 4) and show improvement on both isolated utterances as well as long recordings. Another machine learning technique that is commonly used in limited data scenarios to improve performance is data augmentation. (Tóth, Sztahó, and Vicsi, 2008) reports that the presence of emotion other than neutral in a speech utterance has more influence on the perception of the speaker's emotion. Based on this idea, we propose an augmentation method, CopyPaste to improve emotion recognition performance on isolated utterances (Chapter 3). The main principle behind CopyPaste is based on an observation that human emotion perception is mainly affected by the non-neutral parts of a speech signal. For conversational emotion recognition (CER), we propose the DiverseCatAugment (DCA) augmentation method motivated by the inner workings of the attention mechanism in transformer models.

It is assumed that only one major emotion exists in the isolated utterances and hence most systems operate in utterance-level classification framework i.e., the systems are built to predict one emotion for the input utterance. However, as multiple emotions can exist in long recordings, an utterance-level classification framework can not be applied unless we know locations of emotion segments. In case we know those boundaries, we can cut the recordings

into the segments and process each segment individually in utterance-level classification framework. But, obtaining those boundaries is not an easy task even with human annotators because often the emotion changes are gradual. In scenarios like friendly conversations, we can make assumptions like each speaker turn contains only one emotion as they are often shorter than 5s. But for other situations like broadcast news, there is no such assumption we can make when using only audio (visual change can be used as a heuristic if available). To overcome this limitation, we propose to achieve CER using frame-level classification. By formulating SER from isolated utterances as a frame-level classification task, we compare models trained on isolated utterances and conversations. Then, we present models that can exploit conversational structure (turn-taking patterns) when available.

As discussed above, emotion annotation is not a simple task – it is very expensive, time-consuming, and the emotion perception is not unique. However, annotation schemes can be changed to reduce annotation efforts based on application. For example, for some applications, the goal is to only know the polarity of emotion for the whole conversation i.e., positive or negative emotion instead of more detailed emotions like angry, happy, sad, and disgust. One such application is predicting customer satisfaction (CSAT) towards their interaction with an agent in customer care center calls. In this case, improvements can be made to the service by just knowing whether the customer is satisfied with the service. In this work, we address CSAT to answer questions such as 1) How useful is information existing at different granular levels of conversations to predict CSAT rating? 2) Agent's speech is enough for CSAT?

(useful when having privacy issues with storing customer's speech) 3) How well can we predict CSAT from just the last few seconds of the call? 4) Is it enough to resolve the customer's issue to keep the customer happy with the service? We present experiments aiming to answer these questions using real customer care center calls with self-reported satisfaction ratings.

1.2 Research contributions

- Exploring pre-trained models trained to discriminate speakers for emotion tasks on three datasets collected with different elicitation methods
- Adaptation of speaker recognition models for emotion recognition
- A novel perceptually motivated augmentation procedure, CopyPaste for emotion recognition
- A method for emotion recognition in conversations that do not require segmentation information
- Several methods to incorporate interlocutor information into emotion recognition models on segmented as well as unsegmented conversations
- A comprehensive analysis of feature representations at different granular levels for customer satisfaction prediction
- Customer satisfaction prediction from acoustic and linguistic modalities and their fusion

1.3 Thesis outline

In Chapter 2, we present a brief background on emotion and its automatic recognition from speech signals. First, we discuss different theories/perspectives of emotions (discrete Vs. dimensional) and their relevance to automatic emotion recognition. Then, we discuss several components of dataset preparation such as stimulus types, emotion elicitation methods (acted/induced/spontaneous), and evaluation of emotion. Then, we detail each part of the automatic emotion recognition systems pipeline and relevant literature.

In Chapter 3, we present techniques for emotion recognition on isolated utterances that contain single majority emotion. In particular, we explore transfer learning from speaker recognition models for emotion recognition. We show that speaker embeddings (x-vectors) (Snyder et al., 2018) do contain emotion-relevant information followed by an adaptation of the speaker recognition model for emotion recognition. To improve SER performance further, we propose a perceptually motivated data augmentation technique, referred to as CopyPaste. This technique operates on the idea that listeners are receptive to non-neutral emotions even if they occur for a short duration in an utterance. We present three CopyPaste schemes and show experiments using them. We compare with a widely used noise augmentation technique in both clean and noisy test conditions. one of the main limitations of the models trained on isolated utterances is that they may not be applicable for conversations or in general long recordings with multiple emotions.

Chapter 4 presents techniques for emotion recognition in conversations. Instead of an utterance-level classification framework that is used to recognize

emotion from isolated utterances, we perform frame-level classification to achieve conversational emotion recognition (CER). We propose to use transformers to achieve CER and compare with convolutional and LSTM based models. Based on insight from the inner workings of the self-attention mechanism, we propose an augmentation method, DiverseCatAugment (DCA), to train better transformer models. We evaluate the models trained with isolated utterances on conversations to quantify the importance of context and also to evaluate their robustness in the presence of multiple emotions from multiple speakers. As the speakers' emotions depend on partners' responses and their emotions, we hypothesize that infusing speaker information into the models improves CER performance. We present several techniques to infuse speaker information with and without ground truth segmentation information.

In Chapter 5, we address customer satisfaction prediction. We present a comprehensive analysis of feature representations at multiple granular levels that maximize sentiment prediction accuracy. Our analysis consists of two modalities – acoustic and linguistic. For acoustic modality, we evaluate features extracted from frame-, turn- and call-level for sentiment prediction. For linguistic modality, we evaluate features extracted from word-, turn-, segment- and document-level. Apart from the acoustic and linguistic modalities, we also present heuristic-based turn-taking features to predict sentiment. We show that through the fusion of the modalities and the turn-taking features, we can improve sentiment prediction accuracy. Then, we answer several important questions such as "Whose (agent or customer) data is most correlated with customer sentiment?", "Which part of the calls are more important

for sentiment prediction?", and "The knowledge of task completion status is useful to predict sentiment more accurately?".

Finally, in Chapter 6, we present conclusions of this thesis and future directions.

Chapter 2

Background

2.1 Emotion

Emotion is a complex phenomenon that happens in not only humans but also in other living organisms (Darwin, 2015). Emotional instincts help us to assess threats and react appropriately in order to survive and grow. Some argue that emotion is crucial for our evolution (Izard, 1993). Emotional experience and its display require coordination of several processes – cognitive, neural, physiological – for a given stimulus (Schachter and Singer, 1962). The stimulus can be a physical event, recalling past memories, or social interaction. Experiencing emotion can be voluntary or involuntary and it varies from person to person based on their own past experiences.

Different disciplines study emotion from different points of view. Emotion is mainly viewed as an individual experience in psychology where they study why emotion is experienced in a subject and what is its corresponding stimulus (James, 1948; Cannon, 1927; Schachter and Singer, 1962). There are several psychology theories – physiological, neurological, cognitive theories – each

arguing with different order of events that are responsible for emotions. In sociology, emotion is viewed as a social signal and studies its display mechanisms, their meanings, and effects on the observers in social life (Hatfield, Cacioppo, and Rapson, 1993; Bericat, 2016). In other words, psychology interprets emotion from a cause point of view and sociology from an effect point of view (Cowie and Cornelius, 2003). Note that, the display may not be what the person is actually experiencing. A common example is displaying happy emotion (smiling) when experiencing anger. Affective computing deals with developing mathematical models to recognize and synthesize emotions (Picard, 2000). This field majorly considers emotion as a social signal to develop models i.e., it attempts to deal with physiological/bodily reactions which are on display (facial changes, vocal changes) caused by stimulus (an activating event). It is impractical to consider the psychology (cause) point of view in affective computing because generally, we do not have access to the processes causing the emotions.

Below, we discuss the early conceptualization of emotion and its relevance in automatic emotion prediction. We first review three prominent theories of emotion – discrete model, dimensional model, and Plutchik's emotion wheel. Then, we discuss several types of emotions (acted/induced/spontaneous) and how they can be produced and collected in order to build datasets. Finally, we discuss automatic emotion recognition from the speech on the collected datasets.

2.1.1 Discrete model of emotions

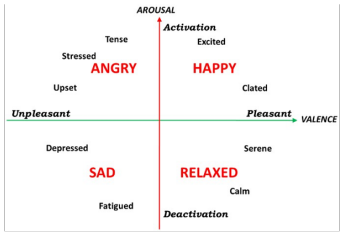
The discrete model of emotions assumes there are a set of basic emotions with which almost all emotions can be expressed. The most commonly considered emotions in this basic set are angry, happy, fear, disgust, sad, and surprise (shown in Figure 2.1a). The classical theory of emotions, proposed by Darwin (Darwin, 2015) in the 1890s is the main proponent of this model and supported and extended by many psychological theorists later (Ekman, 2006). Renowned researcher Paul Ekman in his 1969's work (Ekman, 1969) even claims that these basic sets of emotions are universal. Support for these basic sets of emotions is usually from two points of view: biological and psychological. The biological point of view assumes that these basic sets of emotions have biological fingerprints that cause these basic emotions with triggers from outside world objects or events. Whereas the psychological point of view assumes eliciting conditions are elementary for the basic set of emotions and also that other emotions can be derived solely from them.

Even though the discrete model of emotions is simple and highly useful for research, it has also drawn a lot of criticism mainly for its assumptions such as biological fingerprints existence, and the elementary nature of the basic emotions (Barrett, 2017). There has been a lot of disagreement about what basic emotions set should contain and why (Ortony and Turner, 1990). Several works use a wildly varying set of emotions from just 2 emotions to 7 emotions (Koolagudi and Rao, 2012). One argument often made against the basic set of emotions is by showing that emotion depends on culture (Scherer, Banse, and Wallbott, 2001). A culture can be defined as a set of concepts in

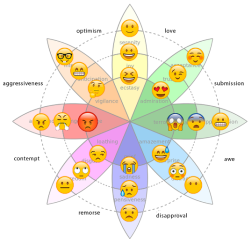
agreement with the community of people. Hence, a different group of people has different rules and new rules can be added with time leading to a variety of emotional signals. It suggests that we can not have a basic set of emotions that works across cultures. (Russell, 1991) argues that emotions are not categorized similarly across cultures and their definitions and boundaries vary depending on the culture. Some cultures might differentiate some emotions while others group them into one class.

Relevance to automatic speech emotion recognition (SER): The basic set of emotions are intuitive – uses everyday language – and hence easier for annotation in that annotators may not need special training. However, this type of simple annotation poses difficulties for automatic prediction. Each emotion class encapsulates a lot of similar emotions and there is no way to discriminate them when using basic set annotation. For example, angry can be hot anger or cold anger. Hot anger is usually loud or high arousal; cold anger sounds more like neutral. Similarly, sad can have several variants such as quiet sorrow and crying despair. These variants may not have similar vocal (or facial) characteristics making the automatic prediction challenging. Moreover, the lack of clear boundaries between these emotions poses problems for annotation. One study found that more classes for annotation lead to less agreement (Aman and Szpakowicz, 2007a) supporting a lack of boundaries between emotion classes. Even with 2-class and 3-class annotation (sentiment task), the inter-annotator agreement is not higher than 80%. Note that the agreement level can change depending on other factors such as spontaneity of the data, instructions to the annotators. The level of disagreement and the

(a) Ekman's basic emotion model



(b) Russel's circumplex model



(c) Plutchik's emotion wheel

Figure 2.1: Various ways of conceptualization of emotion. (Figures source: Wikicommons)

depending factors to some extent explain why different works use a different number of emotions for annotation and hence it is safe to say that annotation of the emotions keeping the application in mind is important. For example, consider a call center setting where the goal could be transferring to a human agent from an automated system as soon as the customer shows signs of dissatisfaction. In this application, we can group all the negative emotions angry, sad, disgust into one class and, happy and neutral into another class. Also, treating these basic emotions independently and classifying them into one of the classes may not be ideal (especially when considering just basic emotions because in many cases, these emotions occur together). From the analysis of a text dataset, XED, authors in (Öhman, 2020) report that anger and disgust occur together very often; anticipation, joy, and trust occur in combinations.

Irrespective of the disagreements, a discrete model of emotions with a basic set is widely used for automatic recognition mainly because of the annotation difficulties with more classes of emotion.

2.1.2 Dimensional model of emotions

The dimensional model proposes that there are abstract independent dimensions that can be used effectively to describe almost all emotions (refer to Figure 2.1b). Usage of the dimensional model can be traced back to the early 20th-century (Wundt and Judd, 1902). In this model, researchers formulate a set of questions to probe what a person feels when a stimulus is presented. These questions are aimed at revealing different aspects of the perception such that complete feeling can be described. The responses to the questions are usually the degree of experience/feeling. For example, a question could be like this "on a scale of 1-10, rate your urge to hit or break something after hearing/watching this stimulus". Then, using principal component analysis most important components can be extracted from the responses vector. Most often, researchers found that valence, arousal, and dominance correspond to the directions with maximum variance. Valence denotes the positivity/negativity level of perception for the given stimuli. For example, happy is a positive emotion and sad is a negative emotion. Some works use different adjectives such as pleasure/displeasure and happy/unhappy among others (Mehrabian and Russell, 1974). Arousal (calm/active, passive/active) describes the tendency to act. It signifies the level of activeness of the speaker. For example, anger has higher arousal and sad has lower arousal while both have negative valence. Dominance (weak/strong, control/power) describes the domination of the stimuli. It signifies the degree of control a person has over the corresponding situation.

One design choice in the dimension extraction process that could lead to

different results is formulating the questions and the corresponding stimuli. It requires thorough knowledge of emotions. If the stimuli or the questions mainly concern one dimension, say arousal, then the result would be just one dimension capturing nearly 100% data variation. (Russell, 1980) uses only 28 stimuli while (Morgan and Heise, 1988) uses 112 stimuli. Many studies report just 2 dimensions ignoring dominance (Russell, 1980; Kuppens, 2008). One limitation ignoring dominance is that anger and fear emotions overlap in arousal-valence space – anger has higher dominance whereas fear is more towards submissive (lower dominance). However, there are studies that argue even three dimensions are not enough (Fontaine et al., 2007; Cochrane, 2009). Most of the works have arousal, valence, and dominance dimensions in common in spite of the disagreements and we think that more research is needed to determine the optimal number of dimensions.

Advantages of the dimensional model are relative easiness to cover a large range of emotions compared to a discrete model with basic emotions and also its suitability for continuous annotation of emotion. In practice, our emotions vary continuously, and annotating with dimensions makes more sense compared to discrete emotions. The dimensions are not as intuitive as discrete emotions (uses everyday language for the class descriptions). Hence, it requires a bit of training for the annotators.

2.1.3 Discrete Vs. Dimensional and Plutchik's emotion wheel

The discrete model offers descriptions in everyday language and the dimensional model offers descriptions in abstract dimensions. Both models provide

ways to describe emotions but none of them may not be enough to completely represent all emotions in reality. There is some evidence that shows the best model of emotions for annotation could depend on individual annotators, adding to the already long list of annotation challenges (Barrett, 1998). The authors observed that individuals who focus mainly on valence when labeling their own emotional experiences are likely to group multiple discrete emotions together i.e., they report multiple emotions together more often. On the other hand, individuals who focus on both arousal and valence are likely to report discrete emotions with less co-occurrence. The authors suggest that it might be useful to view the dimensions (valence and arousal) as a function of discrete emotions or the other way around. In other words, one perspective can be expressed as a function of perspective, and the availability of both perspectives could enhance our understanding of the speaker's state. However, we think that care needs to be taken in transforming from one perspective to another. It is well known that the maximum level of loudness could vary among speakers. And, as loudness is one of the acoustic correlates of arousal, it creates disharmony among speakers when transformed from the dimensional model. This problem can be avoided in two ways: one is having access to the reference level of each speaker and another is using a different function for each speaker.

The drawback of categorical emotion labels is that the intensity of the emotion is not known from the label whereas dimensional labels describe intensity to some extent. On the other hand, the dimensional model is not intuitive and may need special training for annotators. Plutchik proposed an emotion

wheel considering both discrete and dimensional perspectives (Plutchik, 1980). In this wheel (shown in Figure 2.1c), emotions are arranged in a circumplex like in Russel's dimensional circumplex (Russell, 1980) based on the similarity of emotions. There are several layers of emotion wheels in a concentric manner to consider varying intensities (arousal dimension) and at the same time treat the emotions as discrete categories. Emotions in the upper wheels are represented as combinations of adjacent emotions in the inner wheels considering the basic emotion theory premise that new emotions can be formed by combining basic emotions. However, one major criticism of this theory is that emotions at the opposite ends of the wheel can not be combined. For example, joy and sad can not be combined whereas in reality, people do experience joy and sad at the same time (parents feeling when children leave them for college). Even though Plutchik's emotion wheel offers arguably better conceptualization compared to discrete- and dimensional-model of emotions, it is hard to adapt for automatic systems for practical reasons such as annotation difficulties. Next, we discuss several ways of building datasets using discrete and dimensional models' perspectives.

2.2 Building datasets

A number of parameters play an important role in building/collecting a dataset. From a data collection point of view, some of the parameters that need attention are emotion elicitation methods, recording conditions, the language of the dataset, and the demographics of the speakers. From a data

annotation point of view, those parameters can be a type of annotation (continuous/segmental, dimensional/categorical), set of emotion labels, annotator demographics, availability of meta-information to the annotators such as context and speaker profile. Choosing these parameters is mostly guided by the targeted purpose of the dataset.

Emotion datasets can be broadly classified into three types depending on the elicitation methods used to emotions in subjects: acted emotions, induced emotions, and spontaneous emotions. Below, we explain each of these types followed by evaluation/annotation of emotion and factors that influence the perception. For each type of emotion, we discuss general data collection procedures followed, suitable annotation mechanism (discrete-/dimensional-model), and some important characteristics.

2.2.1 Acted emotions

In this setup, the data collection group recruits few actors to just act out target emotions for a pre-defined set of phrases. Sometimes, actors are given prototypical examples of how an emotion sounds. Examples of acted emotion datasets are Crema-D dataset (Cao et al., 2014), EmoSpeech (Banga et al., 2019), MASC dataset (Wu et al., 2006). Acted datasets are more commonly annotated with discrete emotions because annotating with the dimensional model of emotions produces data points along the arousal-valence circle as the acted emotions are most often extreme. We can obtain higher agreement among annotators for two reasons: actors attempt to clearly express the emotions and most parameters such as context, stimuli are in our control.

As acting out emotions involve intentional emotional display, their characteristics may not match real-life spontaneous emotions. For example, raising the pitch for anger, high signal power (louder) for anger, the slow speaking rate for sad in acted emotions whereas their corresponding spontaneous emotions may not have these characteristics. In other words, actors, many times, overact the emotions. Hence, the conclusions drawn from the studies on acted emotions may be entirely different from reality (Batliner et al., 2003). But, there can be some situations acted emotions resemble real emotions. For example, life-threatening situations like fire and violence do produce intense emotions in humans.

The majority of the research on automatic emotion prediction focused on acted emotions due to several reasons. One reason could be that automatic prediction is a challenging task and acted emotions could simplify the task. It also eliminates the context parameter which arguably is the most important factor that influences emotion in real conditions. Another reason could be that natural data is very difficult to obtain.

2.2.2 Induced emotions

Unlike acted emotions, researchers attempted to induce emotions by putting the subjects through situations. This method produces emotions close to natural. Emotions in subjects can be evoked in several ways and they can be broadly classified into 5 methods: playing music, visual stimuli, autobiographical recall, imagery, situational procedures (Siedlecka and Denson, 2019). The

situational procedure often involves creating a social situation that often people face in real-life. For example, questioning the subject's self-worth, creating an uncomfortable situation like playing loud music, giving feedback on their performance, smelling odors. Visual stimuli include playing a video, showing an image. Recalling personal memories involving emotions is considered under autobiographic recall. Imagery includes imagining a scenario, reading emotionally provocative scripts, and usually, in an interactive manner.

Each of these methods evokes emotions using different means and hence they vary in terms of their effectiveness in evoking certain emotions. Authors in (Siedlecka and Denson, 2019) recommend autobiographical recall and imagery for anger elicitation; visual stimuli for disgust; situational procedures and visual stimuli for surprise; all methods except situational procedures for happy; situational procedure for fear; visual stimuli for sadness. However, authors in (Zhang, Yu, and Barrett, 2014) suggest using a combination of these methods would be more effective compared to any single procedure.

Datasets with induced emotions are annotated with both discrete emotions and dimensional attributes. In general, agreement among annotators in labeling these utterances is lower than acted emotional utterances. One of the example datasets for the induced emotions is IEMOCAP dataset (Busso et al., 2008). Even though the emotional utterances are less acted, the consciousness that subjects feel when they are being recorded could affect the way of expression (Labov, 1972). Also, the familiarity of the subjects involved could limit the naturalness of the responses. Unfamiliar subjects tend to be more formal and even if the subjects are familiar, the recording setup could influence them

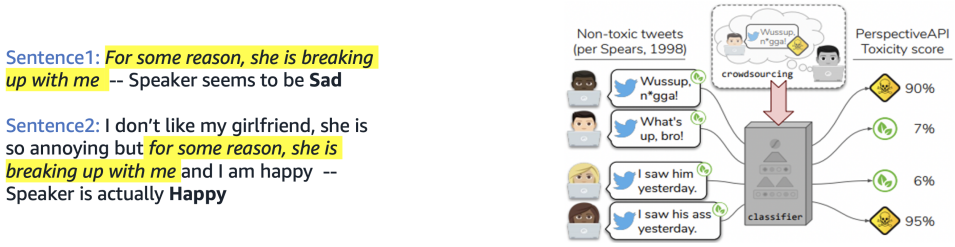
to be less friendly and more formal. Hence, this kind of data could be limited to only a few scenarios.

2.2.3 Spontaneous emotions

Spontaneous emotions are what we experience in our daily life. Recognizing these emotions is very hard compared to acted and induced emotions because of the subtle variations humans use to express them. Also, many times what we express and how we express could depend on a lot of factors that may not be available at the time of recognition. For example, the comfort level of two persons talking to each other could change the dynamics of the conversation. And, this comfort level is difficult to measure/consider for automatic systems or even for human evaluation.

Collecting natural data is highly difficult as the subjects are influenced if they know that they are being monitored (Observer's paradox) (Labov, 1972) and doing without their knowledge raises ethical problems. One problem that could arise in collecting datasets with spontaneous emotions is the skewed distribution of emotions as humans majority of the time are neutral. This problem is evident in the MSP-Podcast dataset (Lotfian and Busso, 2017) which is inundated with many neutral examples and very few other emotions.

From the emotion annotation perspective, spontaneous emotions are very hard compared to acted and induced emotions. One of the main reasons is the lack of full context for spontaneous utterances annotation whereas for acted and induced emotions context can be controlled to some extent. Annotator agreement is usually less compared to induced and acted emotions.



(a) Importance of context **(b)** Importance of speaker's race

Figure 2.2: Examples of importance of (a) context and (b) speaker's race to judge speaker's emotion. Example (b) is replicated from (Sap et al., 2019)

As many a time, the emotion in spontaneous recordings may not be full-blown, annotating with secondary emotion too would help (Lotfian and Busso, 2017; Sneddon et al., 2011). Even though secondary emotion may not be entirely contrastive from primary emotion, it is often useful to describe emotion adequately (Cowie and Cornelius, 2003). Spontaneous utterances are, in general, annotated with both discrete and dimensional model of emotions.

2.2.4 Factors that influence emotion perception

Emotion perception plays a crucial role in social interactions. Inability to perceive a partner's emotions correctly could lead to misunderstanding. Emotion perception depends on a lot of factors such as listener, the relation between listener-speaker, demographics of the listener and speaker (Campbell et al., 2014), context (Cauldwell, 2000), modality of emotion expression and language (Lindquist, MacCormack, and Shablack, 2015). Examples presented in Figure 2.2 demonstrate the importance of knowledge of context and speaker's race when judging the respective speakers' emotions. As can be seen in Figure 2.2a, it is likely that sentences with and without full context are perceived

differently. Similarly, Figure 2.2b demonstrates the importance of knowing speaker's race. (Paulmann, Pell, and Kotz, 2008) reports that emotion prosody comprehension abilities may decline with age and hence perception of emotion can be different with younger people.

Perception can also change based on the listener's past experiences. Studies show that individuals with post-traumatic stress disorder (PTSD) often process emotions differently compared to healthy individuals especially negative emotions (Buckley, Blanchard, and Neill, 2000) like anger, guilt, and shame. Literature shows that emotional cues used for emotion assessment might be different between different individuals (Barrett, 1998). Some individuals mainly focus on valence while others focus on both valence and arousal. All these factors of variability between individuals' perceptions pose a big challenge for building automatic systems.

2.2.5 Evaluation of emotion

As the original emotion of the speaker is usually difficult to obtain, most of the research in affective computing uses the perception of listeners as a proxy to the speaker's emotion. As discussed in the above section, the perception could change from person to person. It might affect the consistency of labels in the data i.e., the presence of the same cues in multiple utterances leads to different emotion labels. As a remedy, researchers usually use multiple annotators for each utterance and use majority or average label for building automatic systems. Even better, (Schuller and Batliner, 2013) uses a weighted average of annotations instead of majority or average. Self-reported annotation can be

used in the absence of annotators.

While annotating using a discrete model of emotions, the number of classes could affect the annotation. Forcing a choice from a limited set of labels may lead to inconsistency or unnecessary noise within each class. Also, it is useful to provide an option of mentioning primary and secondary emotions, especially when using discrete emotions because many times emotions co-occur(Öhman, 2020).

Traditional discrete and dimensional model of emotions offers absolute annotation in the sense that they do not offer any reference with which annotation needs to be carried out. If someone annotates an utterance as angry it is based on his/her own reference of what neutral emotion means. Similarly, if an utterance is said to have high arousal then it is based on his/her own reference of what low arousal means. Few studies (Wood and Ruder, 2016; Wood et al., 2018; Louviere, Flynn, and Marley, 2015; Yannakakis, Cowie, and Busso, 2018) explored using relative annotation with an intuition that we always judge/assess emotions w.r.t. an anchor. While the study in (Wood and Ruder, 2016) found annotation with relative values could be easier and provides consistent labels (Wood et al., 2018) found the opposite. Here, relative annotation consists of pair-wise comparison of sentences whereas absolute annotation consists of choosing a number on a 5-point scale.

Annotating certain modalities without inducing bias is quite difficult. In speech signals, acoustic and linguistic modalities are intertwined. Hence, during its annotation, the annotators do rely on both linguistic contents as well as acoustic content. In this case, attempting to detect emotion from only

speech signals may not be ideal. One way to avoid this problem could be choosing the annotators foreign to the language of the recordings (Kramer, 1964).

2.3 Automatic emotion recognition

Automatic emotion recognition includes building some kind of mathematical model that can process input speech recording to estimate the presence of emotion and detect its category. It mainly involves two steps as shown in Figure 2.3: signal representation and model building. Signal representation includes encoding acoustic (and possibly linguistic) information into a format suitable for the model. The model building includes prototyping and training an appropriate mathematical model that can extract relevant patterns from the signal representation. The ultimate goal would be to predict emotions on new data using the trained model. The trained model is evaluated on test data (that is not used for training) to get an understanding of the model efficacy. If the model is well trained and provides good performance on the test data then we can say that the model is optimally trained and generalizable. Usually, the generalizability of a model depends on many factors: choice of feature extraction algorithm, model, objective function, training procedure, and model hyper-parameters. Apart from these factors which can be controlled from an algorithm point of view, the choice of a dataset (size and its quality) can also affect the generalization ability of the models. Most of the research in affective computing focuses on extracting appropriate features and, building robust and efficient models. Below, we first discuss the extraction of emotional

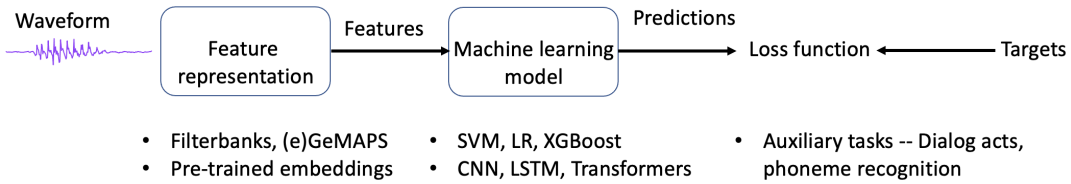


Figure 2.3: General framework for automatic emotion recognition. Some of the examples for feature representation and models are shown in bullet points

correlates in speech signals which simplifies the emotion recognition task. Then, we present commonly used feature sets and machine learning models to achieve automatic emotion recognition.

2.3.1 Correlates of emotions in speech signals

Several studies report that there are several acoustic and linguistic attributes that correlate with emotions in speech signals (Lieberman and Michaels, 1962; Burkhardt and Sendlmeier, 2000). Authors in (Lieberman and Michaels, 1962) experiment with isolating specific attributes (like fundamental frequency) and presenting them to the listeners for evaluation of perception. Whereas (Burkhardt and Sendlmeier, 2000) experiment with synthesizing speech with modified parameters and presents to the listeners. The findings in both studies are similar and provide a direction to perform automatic emotion recognition. Some of the acoustic correlates of emotion include pitch contour, pitch mean/range, speaking rate, phonation type (like breathy voice, tense voice) and, intensity. Fear emotion seems to often have a higher pitch with a wider range and also a faster-speaking rate. Wider pitch range or more specifically irregular pitch patterns could be explained by the tremor that happens when someone is feared. Anger does share similar characteristics as

fear in some aspects such as higher mean pitch and wider range. However, anger often has higher intensity and a slightly lesser speaking rate compared to fear. Within anger, hot anger usually has higher intensity compared to cold anger. Sad often has a narrow pitch range and a slower speaking rate. There can be different types of sad – crying and quiet sorrow. Crying associates with higher intensity compared to quiet sorrow type of sad. Also, articulation in sad emotion might be different compared to other emotions – speech is often slurry in sad. Utterances with happy seem to have both faster and slower speaking rates indicating that there can be sub-classes within happy. These sub-classes can be a loud laugh or a slight smile which is more close to neutral. Pitch changes in happy are usually smooth and upward inflections compared to sad.

In these studies, there is little to no emphasis on articulatory aspects of emotional speech (Kohler, 1995; Kienast, Paeschke, and Sendlmeier, 1999). (Kienast, Paeschke, and Sendlmeier, 1999) studies the effect of emotion on the duration of syllables and their accuracy of articulation. The authors report an articulatory reduction in sad and fear. Anger seems to have shorter consonants and long vowels and is also likely to have stressed syllables.

2.3.2 Speech signal representation

In general, the goal of representing a speech signal is to retain as much information as possible. Some of the characteristics of speech signals include spoken content, speaker characteristics, emotion, and noise. For a given task, we must encode task-relevant characteristics and should try to leave

out irrelevant information for the best results. What is relevant or irrelevant depends on the application. For example, speaker identity is not relevant for speech recognition (in most cases) and spoken content is not relevant for speaker recognition (except for text-dependent applications). For emotion recognition, spoken content is important and speaker characteristics are also useful. Broadly, we can categorize feature representation efforts into 3 classes: heuristic features, automatic feature learning, and pre-trained embeddings.

2.3.2.1 Heuristic features

Identifying relevant characteristics (vocal and linguistic) for emotion recognition is not an easy task in itself and designing algorithms to extract them is an added challenge. There are huge efforts in building emotion-specific features (Eyben et al., 2013; Eyben et al., 2015). Many versions of prosodic and spectral features are used in InterSpeech (IS) challenges (Schuller, Steidl, and Batliner, 2009; Schuller et al., 2010; Schuller et al., 2013; Schuller et al., 2020) which target either emotion recognition or some related tasks. Prosodic features include frequency- and energy/amplitude-related parameters. Some of the frequency-related parameters are pitch, jitter, formats center frequencies, and their bandwidths; and some of the energy parameters are shimmer, loudness, and harmonics-to-noise ratio. Spectral features include spectral slope in different frequency bands, harmonic difference, the ratio of the energy of spectral harmonic peak at formants to the energy of the spectral peak at F0, Mel-frequency cepstral coefficients (MFCC), spectral flux. A set of these features are usually referred to as low-level descriptors (LLDs) as they are extracted from the signal directly. Over these parameters, several functionals

such as average, max/min/median, and standard deviation are applied to create new features. Each year (2009-13) the feature set size kept increasing: IS2009 challenge feature set consists of 384 features (Schuller, Steidl, and Batliner, 2009), for IS2010 challenge 1582 features (Schuller et al., 2010) and for IS2013 6373 features (Schuller et al., 2013). Considering that the size of the datasets is small increasing the feature set size may not be optimal (it is an underdetermined system). To address this problem a group of scientists worked together and attempted to pick the most important features through experiments. This minimal set consisting of 58 features is referred to as Geneva Minimalistic Acoustic Parameter Set (GeMAPS) (Eyben et al., 2015). And, an extended version of it which contains 87 features is referred to as extended GeMAPS (eGeMAPS). Apart from these feature sets, many works show that using only MFCC features could also provide competitive performance for emotion recognition (Schuller, Rigoll, and Lang, 2003; Batliner and Huber, 2007).

To improve further, (Schmitt, Ringeval, and Schuller, 2016) propose to use bag-of-acoustic-words (BoAW) for emotion recognition inspired by its use in audio event detection. BoAW feature extraction technique quantizes the chosen feature set using clustering algorithms and replaces the features with the nearest cluster mean (codewords). The advantage of this technique is that it utilizes dataset global characteristics and also minimizes the variation in the features. However, the generalizability of the features is one main concern as it is extracted using the dataset global characteristics.

2.3.2.2 Automatic feature learning

Although the heuristic features (presented above) provide satisfactory performance, a lot of effort went into identifying them. They are designed to discriminate only a limited set of emotions and also use mainly acted datasets with prototypical emotional expression to identify acoustic cues and design feature sets. Hence, they may not be sufficient to detect subtle nuances in spontaneous speech. For example, authors in (Batliner et al., 2003) show that prosodic features are more effective on acted speech than on spontaneous speech. Authors suggest that it could be because actors emphasize and deliberately display emotions in their speech which is not the case in spontaneous speech. In addition, this method of identifying specific cues may not be scalable to detect more variety of emotions that are not easy to produce through acting. In this case, letting the model figure out the relevant cues could be the best choice from the scalability point of view. Several new studies propose techniques to automatically extract the features with the goal of maximizing the performance (Tzirakis et al., 2017; Sarma et al., 2018; Trigeorgis et al., 2016). They propose to use either raw-waveform or spectrogram as input to the models and show good performance. However, these models could be more sensitive to dataset-specific characteristics and impede generalizability as with the BoAW features. But, this problem could probably be mitigated with a lot of data, by building robust models, and/or by using augmentation techniques.

2.3.2.3 Pre-trained embeddings

Another set of approaches that aim to improve emotion recognition performance use pre-trained models to extract features (Cummins et al., 2017; Elshaer, Wisdom, and Mishra, 2019; Lakomkin et al., 2018a). (Cummins et al., 2017) proposes to use pre-trained image models to represent spectrograms and show that the representations can be used for emotion recognition. (Elshaer, Wisdom, and Mishra, 2019) uses audio event detection models and (Lakomkin et al., 2018a) uses speech recognition models. Generally, the extracted features consist of only one vector summarizing the whole utterance.

2.3.3 Model design and training

***General models*:** The early 2000s and before, only simple models such as logistic regression and SVMs were used to detect emotions (Koolagudi and Rao, 2012). They operate on just a vector representation of the input utterance ignoring sequence information. However, sequence information could be useful for better performance. For example, raising pitch is one of the important characteristics of anger emotion. Earlier, hidden Markov models (HMM) were used for emotion recognition to exploit sequence information (Schuller, Rigoll, and Lang, 2003). The resurgence of deep learning techniques enabled efficient use of the sequence information. (Cho et al., 2018; Zhao, Mao, and Chen, 2019; Huang et al., 2014; Lim, Jang, and Lee, 2016) explore CNN and LSTM based models from feature representations such as MFCC and (e)GeMAPS features. One common theme among these models is that the input representation is

processed with several convolutional/LSTM layers to obtain more contextual features and then some sort of pooling layers to summarize the entire sequence. Then, the application of fully-connected layers on the summary vector with appropriate activation function in the final layer produces the final output. Most of these methods use either max pooling or average pooling for summarization when using CNN-based models. Here, all the vectors in the sequence have equal priority. But, some frames could be more important containing more relevant information w.r.t. the corresponding class label. To exploit different levels of importance, attention operation (Vaswani et al., 2017) could be used which is similar to the weighted average. Studies (Zhang et al., 2018; Mirsamadi, Barsoum, and Zhang, 2017) use the attention operation for emotion recognition and show performance improvements. All these models directly optimize the target loss function which is usually cross-entropy. Some other paradigms such as adversarial learning have also been explored for emotion recognition to improve the robustness of the models. (Latif, Rana, and Qadir, 2018; Han et al., 2018; Parthasarathy et al., 2019; Sahu, Gupta, and Espy-Wilson, 2018) propose to use adversarial learning.

In general, deep learning models perform better with more data (i.e., data-hungry). However, emotion datasets are usually smaller, typically a few hours. Collecting more data with emotion annotation is expensive and its ambiguous nature makes it more difficult to collect. In such cases, three methods are generally helpful: semi-supervised training, transfer learning, and data augmentation. We review some of the past works that use these techniques for emotion recognition below.

***Semi-supervised learning*:** Semi-supervised learning paradigm aims to exploit unlabelled data along with labeled data to improve performance on the labeled data. There are several works that show exploiting unlabelled data is useful for speech emotion recognition (Liu et al., 2007; Deng et al., 2017; Zhang et al., 2021; Latif et al., 2020). (Liu et al., 2007) explores co-training procedure to exploit unlabelled data. In this procedure, two classifiers, trained with different feature sets, are used to select unlabelled data based on their predictions. Authors in (Deng et al., 2017) optimize unsupervised auxiliary objective function (reconstruction loss) along with emotion loss. For the unlabelled data, only the reconstruction loss is optimized and for the labeled data both losses are optimized. Unlabelled video data is used in (Zhang et al., 2021) to improve speech emotion recognition. Here, the authors enforce similarity constraints between predictions on audio and video. (Latif et al., 2020) uses several additional loss functions along with emotion loss such as speaker/gender classification loss, adversarial loss, and reconstruction loss.

***Transfer learning*:** Some of the past works that use transfer learning for emotion recognition include (Latif et al., 2018; Lakomkin et al., 2018a; Williams and King, 2019). It is shown in (Lakomkin et al., 2018a) that reusing an ASR model trained to predict phonemes is helpful for the SER task. Authors in (Williams and King, 2019) show that speaker-based utterance-level representations i-vectors and x-vectors encode speaking-style information and emotion. However, their experimental setup included overlapping speakers between training and testing data splits. We believe that speaker overlap

should be avoided in SER tasks, especially when using speaker-specific representations as input. Different from these two works, authors in (Latif et al., 2018) perform transfer learning between multiple emotion datasets.

***Data augmentation*:** Data augmentation techniques have been shown to improve emotion recognition performance (Lakomkin et al., 2018b; Etienne et al., 2018; Bao, Neumann, and Vu, 2019; Rizos et al., 2020). Authors in (Lakomkin et al., 2018b) show that adding noise to the clean recordings helps the model to better recognize emotions. Altering the speaking rate of speech (Lakomkin et al., 2018b) and vocal tract length perturbation (Etienne et al., 2018) is also shown to be useful for SER. Few recent studies (Bao, Neumann, and Vu, 2019; Rizos et al., 2020) ventured into generating emotional speech features using advanced techniques such as CycleGANs and StarGANs.

2.3.4 Training and evaluation metrics

If the goal is to discriminate emotions then classification is performed and if the goal is to predict emotion dimensions then regression is performed. For both classification and regression, formulating an appropriate objective function is important for model training. Usually, categorical cross-entropy is used for classification, and mean square error is used for regression. Optimization of the loss function can be done using standard gradient descent or advanced optimizers like Adam, Adadelta, RMSProp. We did not find studies related to the efficiency of optimizers specific to emotion recognition. Most of the studies use Adam optimizer to minimize the objective function.

Usually, emotion models are evaluated and compared using metrics such as precision, recall, and micro/macro-f1 score. The precision of an emotion class measures the fraction of relevant predictions out of all the predictions for that class. In other words, it is the ratio of true positives and the sum of true positives and false positives. Whereas recall of an emotion class measures the fraction of relevant predictions out of actual relevant (ground truth) instances of that class (ratio of true positives and sum of true positives and false negatives). F1-score is defined as the harmonic average of precision and recall. Micro-f1 score is calculated as the weighted average of class-wise f1-scores where the weight for a class is calculated as the ratio of the number of samples for that class and the total number of samples in the dataset. Macro-f1 score is an unweighted average of class-wise f1-scores (it does not depend on the size of the classes). However, we find no consistent metric that is reported in the literature. Reporting only precision or recall does not usually give a full picture of model ability as improving one often results in degradation of the other. We think reporting micro/macro-f1 score too would reflect the model efficacy more clearly.

2.3.5 Auxiliary tasks

Although optimizing the target task loss function is sufficient to realize the target task, it is often shown that the use of certain auxiliary tasks yields improvements (Bothe et al., 2020; Li et al., 2020a; Parthasarathy and Busso, 2018). The auxiliary task is optimized along with the target task often with less weight. It serves as a regularization for the model and avoids overfitting to

the target task. Some of the auxiliary tasks for emotion recognition are dialog acts (Bothe et al., 2020), phoneme recognition (Lakomkin et al., 2018a), GRL-based loss on speaker labels (Li et al., 2020a), autoencoding with reconstruction loss (Parthasarathy and Busso, 2018). Auxiliary dialog acts task could help in the disambiguation of some emotion classes. For example, forcing the model to predict *the appreciation* dialog act could help the model to easily disambiguate between happy and neutral. GRL-based losses force the model to discard some attributes specifically by maximizing the loss on the chosen auxiliary task (Ganin and Lempitsky, 2015). It is well known that the range of emotions between speakers could be different and the emotion model could form clusters of speakers. In this case, speaker identity could affect emotion models. By forcing the model to discard speaker identity, the model can be made more generalizable to new speakers. Whether to use GRL loss or not depends on whether the auxiliary task helps or degrades the target task. For example, usage of GRL loss on speaker labels makes sense when building speaker-independent emotion models whereas if speaker-dependent emotion models are desired (for example personal assistants) then it is better to not use GRL.

Chapter 3

Emotion recognition on isolated utterances

3.1 Introduction

In this chapter, we discuss speech emotion recognition (SER) from isolated utterances. Utterances containing only one emotion, generally shorter than 10s, are considered isolated utterances. In general, they are collected for emotion recognition in three methods: 1) recorded in isolation using actors with targeted emotions, 2) cut from conversations that are meant to produce emotions in an induced manner, and 3) cut from spontaneous podcast conversations. In the first method, the utterances are already short and targeted to contain only one emotion. Recordings collected in this manner, by design, contain only one speaker and are shorter than 10s in most cases. Data collected using the second and third methods i.e., using conversations between subjects do not result in isolated utterances by default. Hence, they are derived by segmenting conversations with respect to emotion.

In terms of the usefulness of the above mentioned data collection methods

for research, each has its own advantages and disadvantages. Recruiting actors to enact an emotion for a pre-defined set of phrases usually offer more control over the experiments. Because of the control on the experimental setup, this task could be a little simpler and more useful for analysis. For example, the emotion of the speaker is not contextual and hence only utterance characteristics impact a model's behavior. However, it rarely reflects a reality where context plays an important role in deciding the speaker's emotion. Also, this way of data collection is not scalable to large datasets as it is very expensive to design a data collection setup and recruit subjects. On the other hand, building isolated utterances datasets using conversations is a little simpler and can be automated using advanced technologies. For example, the MSP-Podcast dataset (Lotfian and Busso, 2017) is collected in this manner. The datasets collected in this style simulate reality in terms of the naturalness of emotion expression. However, both methods suffer from annotation costs. As the emotion of the speaker is highly subjective to the listener, annotation with multiple subjects is necessary to have a good estimate of the speaker's emotion. A single annotator for each utterance could result in a dataset with noisy labels i.e., the labels may not reflect the actual emotion of the speaker in the utterance. Fortunately, recent advancements in machine learning could enable us to build accurate models even with smaller datasets thereby minimizing annotation cost. In this chapter, we present two such machine learning methodologies to recognize emotion from isolated utterances: transfer learning (Bozinovski and Fulgosi, 1976) and data augmentation (Ramirez, Montalvo, and Calvo, 2019) techniques.

The transfer learning paradigm offers several benefits in terms of model robustness, the number of training samples, and label scarcity. This paradigm is usually associated with two domains, namely source and target. Source domains usually contain a large amount of annotated data. In most cases, source and target tasks share some common characteristics which help to achieve the best performance in each of the tasks. However, there are cases where both tasks/domains need not be related directly but are still useful for transfer learning. For example, transfer learning from image-related tasks to speech tasks. Our transfer learning approach for emotion recognition is motivated by several previous works (Lakomkin et al., 2018a; Raj et al., 2019; Williams and King, 2019). It is shown in (Lakomkin et al., 2018a) that reusing an ASR model trained to predict phonemes is helpful for the SER task. In (Raj et al., 2019), authors studied the applicability of speaker-based utterance representations such as i-vectors and x-vectors for several downstream tasks related to speech, speaker, and utterance meta information. However, they did not study for emotion-related tasks. Authors in (Williams and King, 2019) show that speaker-based utterance-level representations i-vectors and x-vectors encode speaking-style information and emotion. However, their experimental setup included overlapping speakers between training and testing data splits. We believe that speaker overlap should be avoided in SER tasks, especially when using speaker-specific representations as input. In this chapter, we present results using pre-trained as well as fine-tuned models which are not studied in (Williams and King, 2019).

Data augmentation technique to some extent can help us build efficient

models by artificially creating a lot of data from the available original data. The additional data is usually a perturbed version of the available data which often includes modifying selected acoustic characteristics. Data augmentation techniques have been shown to improve emotion recognition performance (Lakomkin et al., 2018b; Etienne et al., 2018; Bao, Neumann, and Vu, 2019; Rizos et al., 2020). Authors in (Lakomkin et al., 2018b) show that adding noise to the clean recordings helps the model to better recognize emotions. Altering the speaking rate of speech (Lakomkin et al., 2018b) and vocal tract length perturbation (Etienne et al., 2018) are also shown to help SER. Few recent studies (Bao, Neumann, and Vu, 2019; Rizos et al., 2020) ventured into generating emotional speech features using advanced techniques such as CycleGANs and StarGANs.

In this chapter, we present a transfer learning approach from speaker recognition models and a data augmentation procedure to improve SER performance on isolated utterances. First, we show that emotion-related information is encoded in x-vectors, and then we show that fine-tuning for emotion targets further improves the performance. We compare two pre-trained models for this study–one trained with augmentation and another without augmentation to understand the correlation between pre-trained models' performance in source task and their re-usability for SER (target task). Then, we propose an approach to adapt the pre-trained models to perform SER. To further improve our models, we propose the CopyPaste augmentation method for SER. This technique operates on the observation that the presence of emotions other than neutral affects the listener's perception. We propose three CopyPaste

schemes and compare them with widely used noise augmentation in both clean and noisy conditions.

The main contributions of this chapter are:

- Exploring pre-trained models trained to discriminate speakers for emotion tasks on 3 different types of datasets
- Fine-tuned models for SER task
- CopyPaste, a novel perceptually motivated data augmentation procedure for SER

The rest of the chapter is organized as follows: First, we present datasets used for this study in Section 3.2. Then, we discuss transfer learning from speaker recognition models in Section 3.3 followed by CopyPaste augmentation in Section 3.4. Finally, we discuss the conclusions of this chapter in Section 3.5.

3.2 Datasets

We validate our experiments on three different types of datasets: IEMOCAP (acted, no restriction on spoken content, induced emotions), MSP-Podcast (natural, no restriction on spoken content, spontaneous emotions), and Crema-D (acted, restricted to 12 sentences, prototypical emotions). The details of each dataset are as follows.

3.2.1 IEMOCAP

IEMOCAP dataset is a multimodal dyadic conversational dataset recorded with 5 female and 5 male actors Busso et al., 2008. It contains conversations from 5 sessions wherein each session one male and female actor converse about a pre-defined topic. Each session is segmented into utterances manually, and each utterance is annotated by at least 3 annotators to categorize into one of 8 emotion classes (angry, happy, neutral, sad, disgust, fear, excited). Conversations are scripted and improvisational in nature. In this work, we followed previous works in choosing data for our experiments. We combined happy and excited emotions into one class. We choose a subset of data consisting of 4 emotions: angry, sad, neutral, happy. As the number of speakers and utterances in this dataset is low, we opted for 5-fold cross-validation (CV) to obtain reliable results. As it was shown in Raj et al., 2019 that speaker verification models capture session variability along with speaker characteristics; we did leave-one-session-out training for 5-fold CV to avoid overlapping of speakers and sessions between training and testing. In each fold, we used the micro-f1 score (refer to Chapter 2 for definition) as our metric, and hence, we reported an average of micro-f1 scores of 5-fold CV for each experiment.

3.2.2 MSP-Podcast Dataset

MSP-Podcast corpus[1] Lotfian and Busso, 2017 is collected from podcast recordings. The recordings are processed with several tools before including them in the dataset. First, the speaker diarization tool is used to obtain segments for each speaker and remove all the segments shorter than 2.75 seconds and longer than 11 seconds. Then the segments with SNR less than 20dB, background music, telephone quality speech, and overlapping speech are removed. The remaining clean segments are annotated by crowd-sourcing workers after manual screening into one of 8 emotion classes (angry, happy, sad, surprise, fear, disgust, contempt, neutral) or other. In this work, we used 5 emotions: angry, happy, sad, neutral, disgust for classification as in Lotfian and Busso, 2019. We used the standard splits in Release 1.4 for training, development, and testing. This dataset has 610 speakers in the training split, 30 in the development, and 50 speakers in the test split.

3.2.3 Crema-D Dataset

Crema-D dataset[2] is a multimodal dataset (audio and visual) with 91 professional actors enacting a target emotion for a pre-defined list of 12 sentences. It includes 48 male and 48 female actors with a diverse ethnicity and age distribution. In this work, we use 4 emotion categories: angry, happy, sad, and neutral. We discarded disgust and fear to balance the dataset. We used 51

[1]Data provided by The University of Texas at Dallas through the Multimodal Signal Processing Lab. This material is based upon work supported by the National Science Foundation under Grants No. IIS-1453781 and CNS-1823166. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the University of Texas at Dallas.

[2]https://github.com/CheyneyComputerScience/CREMA-D

actors in training, 8 for development, and 32 for testing.

3.3 Transfer learning from speaker recognition models

In this section, we present details of the x-vector model reused for the SER task. Then, we explain the transfer learning approach followed to transfer knowledge to achieve the SER task. It is shown in the literature that i-vectors and x-vectors perform well on speaker-related tasks such as speaker verification (Villalba et al., 2019), speaker diarization (Shum et al., 2013; Sell and Garcia-Romero, 2014; Maciejewski et al., 2018; Sell et al., 2018). In this work, we only exploit the x-vector model because of its superiority over i-vectors (Snyder et al., 2018) and also because it is easy to adapt for downstream tasks.

3.3.1 x-Vector Model

In this work, we perform transfer learning from a state-of-the-art ResNet x-vector model reported in (Villalba et al., 2019). The network consisted of three parts: frame-level representation learning network, pooling network, and utterance-level classifier. Frame-level representation learning network uses ResNet-34 (He et al., 2016) structure, which consists of several 2D convolutional layers with short-cut connections between them. After that, we used a multi-head attention layer to summarize the whole utterance into a large embedding. This layer takes ResNet outputs $\mathbf{x}_t$ as input and computes

Component	Layer	Output Size
Frame-level Representation Learning	$7 \times 7, 16$	$T \times 23$
	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 3$	$T \times 23$
	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 4$, stride 2	$\frac{T}{2} \times 12$
	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 6$, stride 2	$\frac{T}{4} \times 6$
	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 3$, stride 2	$\frac{T}{8} \times 3$
	average pool 1×3	$\frac{T}{8}$
Pooling	32 heads attention	32×128
Utterance-level Classifier	FC	400
	FC	#spk:12,872

Table 3.1: ResNet architecture used in the x-vector model

its own attention scores $w_{h,t}$ for each head h:

$$w_{h,t} = \frac{\exp(-s_h \|\mathbf{x}_t - \boldsymbol{\mu}_h\|)}{\sum_{t=1}^{T} \exp(-s_h \|\mathbf{x}_t - \boldsymbol{\mu}_h\|)} . \tag{3.1}$$

Attention scores $w_{h,t}$ are normalized along time axis.

Output embedding for head h is the weighted average over its inputs:

$$\mathbf{e}_h = \sum_t w_{h,t} \mathbf{x}_t \tag{3.2}$$

Different heads are designed to capture different aspects of the input signal. Embedding from different heads is concatenated and projected by an affine transformation into the final embedding. From the pooling layer to output, there are two fully connected layers, and it predicts speaker identity in the training set. Angular softmax (Liu et al., 2017) loss was used to train the network. The whole network structure is illustrated in Table 3.1.

We trained the x-vector model using the following datasets:

- Switchboard phase1-3 and cellular1-2.
- NIST SRE04-10
- NIST SRE12 telephone data
- NIST SRE12 phone calls recorded through a far-field microphone
- MIXER6 telephone phone calls
- MIXER6 microphone phone calls
- VoxCeleb 1+2: We concatenated all examples from the same video into one file
- SITW-dev-core: single speaker segments from the Speakers in the Wild development set

SRE12 microphone, MIXER6 microphone, VoxCeleb, and SITW-dev-core were downsampled to 8 kHz. In total, there are 12, 872 speakers with 735, 018 utterances after removing utterances short than 8 seconds.

3.3.2 Speech Emotion Recognition (SER)

Generally, the performance of an x-vector model is a good indicator of its ability to discard speaker irrelevant information. That is, the embeddings extracted from a state-of-the-art x-vector model might have lesser emotion information compared to the embeddings of a slightly worse model. In this work, we perform transfer learning from two versions of pre-trained x-vector models: one trained with augmentation and another without augmentation.

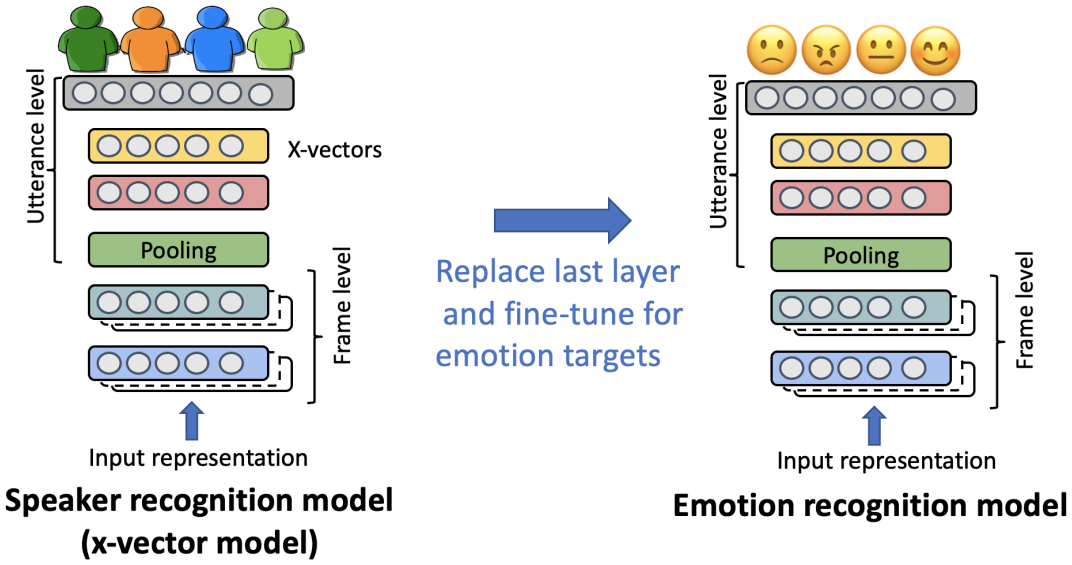


Figure 3.1: Transfer learning from x-vector model for SER

Augmentation is applied using MUSAN corpus (Snyder, Chen, and Povey, 2015). We refer to the model trained with augmentation as *ResNet-aug* and its speaker verification performance on the SITW dataset is 2.39. Similarly, the model trained without augmentation is denoted with *ResNet-clean* which stands at 3.89 EER on the SITW dataset. As expected, the speaker verification performance of *ResNet-aug* is better than *ResNet-clean* as the former model is trained with augmentation and hence more able to discard irrelevant information. Based on this observation, we hypothesize that embeddings extracted from *ResNet-aug* contain less emotion information compared to *ResNet-clean*.

From a pre-trained x-vector model, we can transfer knowledge to achieve SER in two ways:

- Extract x-vectors and train a standard linear model like logistic regression (LR) for emotion classification.

- Replace the speaker-discriminative output layer with the emotion-discriminative layer and fine-tune. In other words, use the weights learned in pre-training for all the layers except the last layer and then optimize all the weights for emotion classification (refer Figure 3.1).

We show experiments with both methods using the above mentioned x-vector models *ResNet-aug* and *ResNet-clean*. For emotion classification, we minimize cross-entropy loss function using Adam optimizer with default parameters in PyTorch. The epoch with the best micro-f1 score on the development set is chosen for evaluation on the test set. We report an average of micro-f1 scores from 3-runs on the test set for each emotion dataset considered.

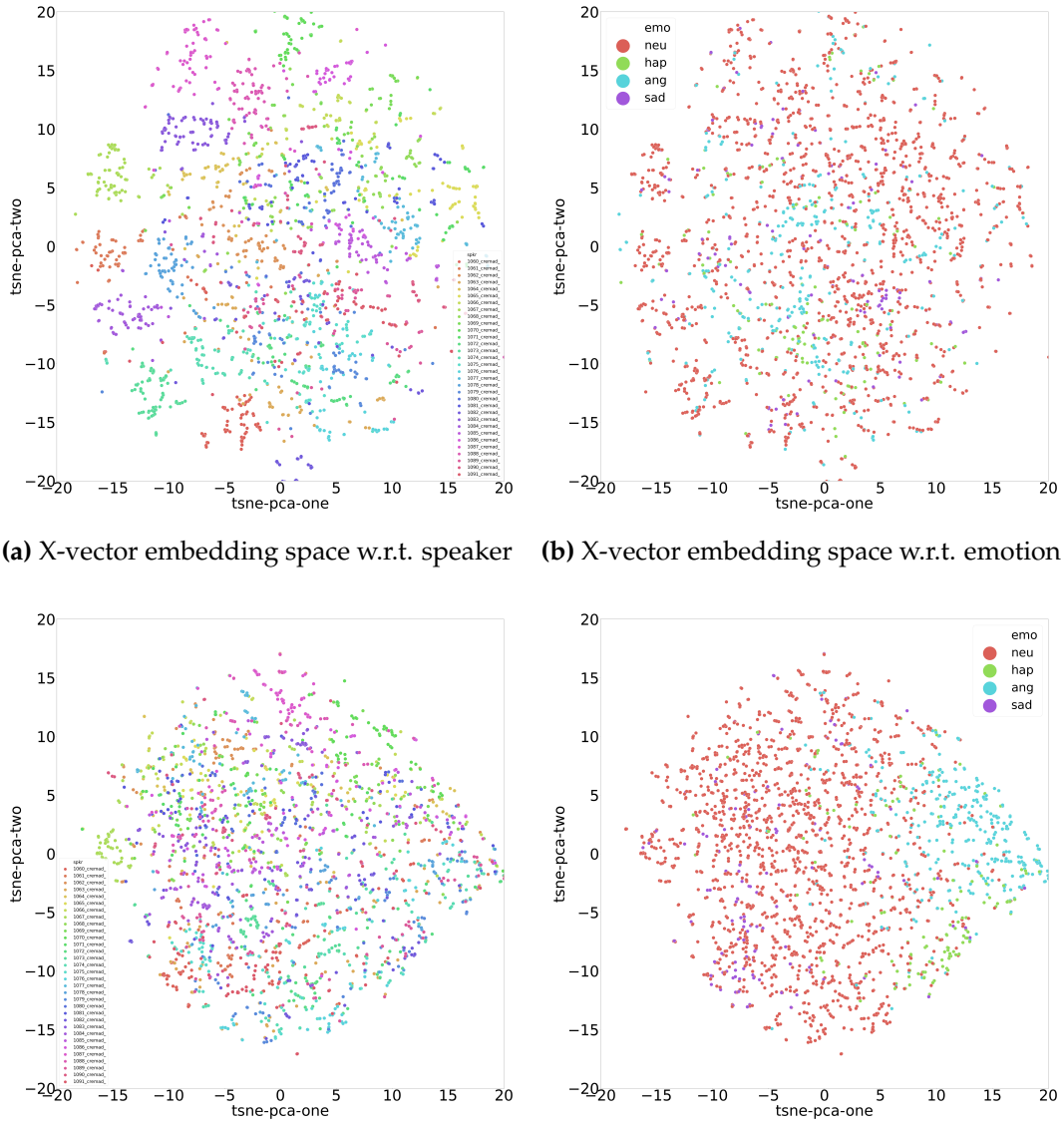
3.3.3 Results

Table 3.2 presents the results of the SER task with ResNet architecture on all three datasets. As noted in Section 3.3.1, *ResNet-clean* and *ResNet-aug* denotes unaugmented and augmented x-vector models.

Comparison of columns marked with Baselines suggests that eGeMAPS performed better than filter-bank features in most cases, but as our pre-trained models were trained with filter-bank features, we did not consider eGeMAPS for further experiments. Significant improvements were obtained on all the datasets by using pre-trained models compared to random initialization suggesting that pre-training is helpful. The performance of the LR model reflects the linear separability of the x-vector embedding space for emotion classification. X-vectors extracted from *ResNet-clean* provided better results compared to the x-vectors from *ResNet-Aug* (columns marked with Frozen models).

It suggests that x-vectors from *ResNet-clean* contain more emotion-related information which further implies the model is unable to discard emotion information as well as *ResNet-Aug*. This observation is in line with the corresponding EERs where *ResNet-clean* has a higher EER (3.89%) compared to *ResNet-Aug* (2.39%). A similar conclusion was reported in (Raj et al., 2019) for tasks such as prediction of the session, utterance length, gender, etc. Having observed the good performance with x-vector embeddings (frozen x-vector model), which are trained to discriminate speakers, we proceeded to fine-tune the pre-trained models for emotion recognition. By fine-tuning, we obtained improvements in all cases except when using *ResNet-clean* on MSP-Podcast and IEMOCAP.

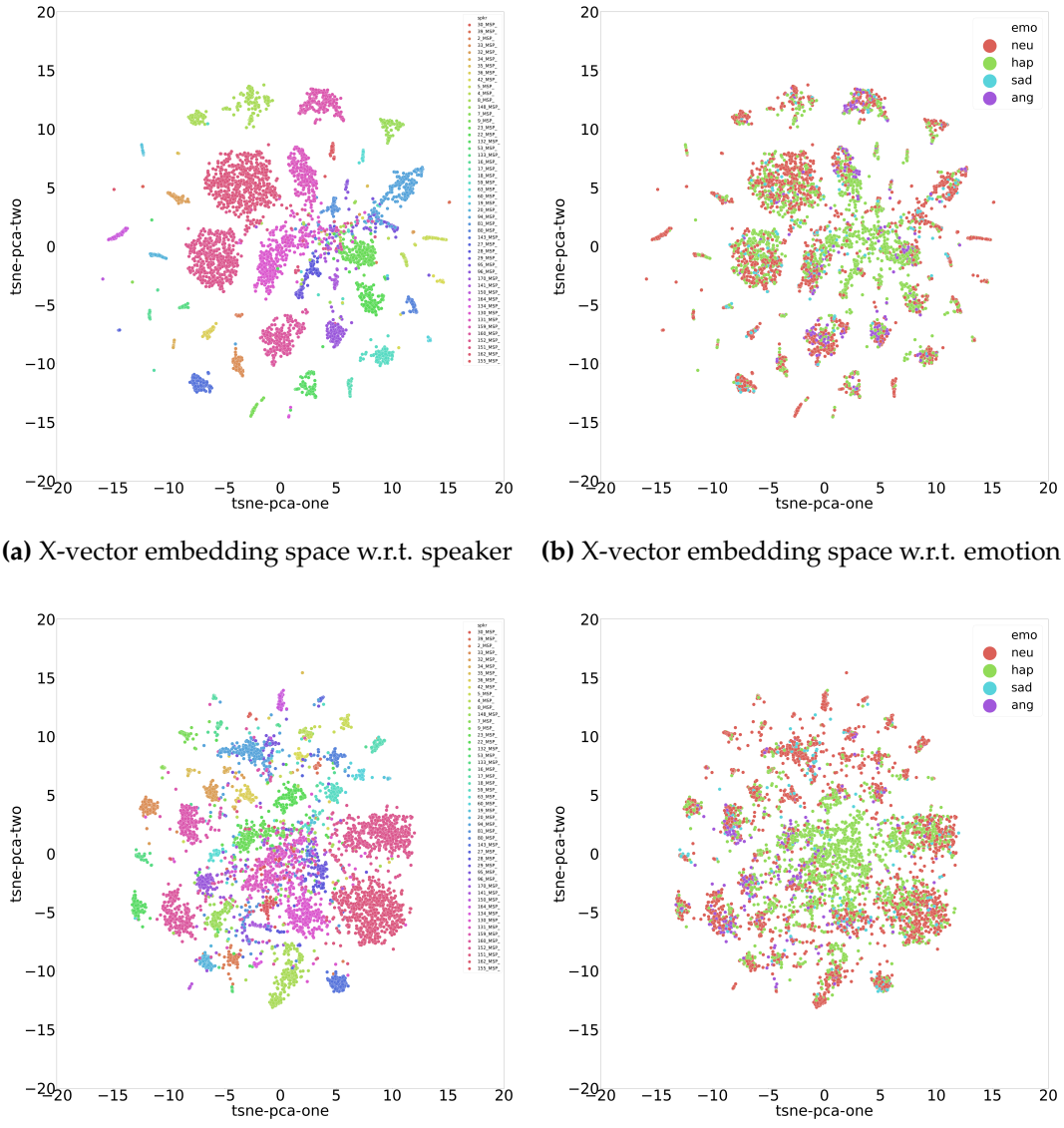
Overall, fine-tuned *ResNet-aug* model worked best with a micro-f1 score of 56.79%, 77.86%, and 61.18% on MSP-Podcast, Crema-D, and IEMOCAP respectively. Based on our experiments, we recommend using x-vector embeddings from *ResNet-clean* for SER if fine-tuning is not feasible otherwise fine-tuning *ResNet-Aug* is recommended. It is difficult to compare our results with previous works as there are no standard splits for IEMOCAP and Crema-D. In the case of MSP-Podcast, the dataset collection is an ongoing effort, and we did not find previous works on the current release yet.



(a) X-vector embedding space w.r.t. speaker **(b)** X-vector embedding space w.r.t. emotion

(c) Fine-tuned x-vector embedding space w.r.t. speaker **(d)** Fine-tuned x-vector embedding space w.r.t. emotion

Figure 3.2: Analysis of Crema-D embedding space before and after fine-tuning using t-SNE plots



(a) X-vector embedding space w.r.t. speaker

(b) X-vector embedding space w.r.t. emotion

(c) Fine-tuned x-vector embedding space w.r.t. speaker

(d) Fine-tuned x-vector embedding space w.r.t. emotion

Figure 3.3: Analysis of MSP-Podcast embedding space before and after fine-tuning using t-SNE plots

Table 3.2: SER results on three datasets. In the first column, *ResNet-clean* and *ResNet-aug* denotes unaugmented and augmented x-vector models. Text in the parenthesis denotes the feature set we used to train.

	Baselines		Frozen models		Fine-tuned models	
Dataset	Random Init. (eGeMAPS)	Random Init. (Filter Bank)	ResNet-Clean (Filter Bank)	ResNet-Aug (Filter Bank)	ResNet-Clean (Filter Bank)	ResNet-Aug (Filter Bank)
MSP-Podcast	49.85	47.36	56.75	52.58	55.71	56.79
Crema-D	73.52	71.46	76.00	74.35	76.54	77.86
IEMCOAP	46.20	43.03	57.40	55.40	54.57	61.18

3.3.4 Analysis

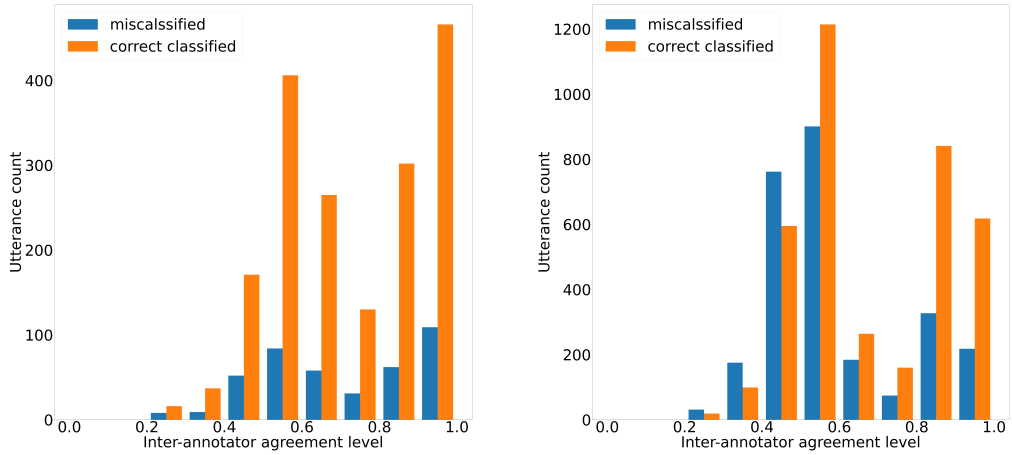
3.3.4.1 Embedding space analysis

Intuitively, the x-vector embeddings should have more speaker-discriminative and less emotion-discriminative information as the model is trained to discriminate speakers. Hence, we can expect to see clusters of speakers on x-vector embedding space as opposed to clusters of emotions. However, after fine-tuning/adapting for emotion classification, we expect to see clusters of emotions instead of speaker clusters. Figure. 3.2 and 3.3 show t-SNE plots of the embeddings for the Crema-D dataset and MSP-Podcast respectively. We choose these two datasets as they stand at extreme ends on the spectrum of acting to the naturalness of the spoken utterances. From Figure. 3.2a and 3.2b, we can observe clusters of speakers and somewhat arbitrary arrangement of embeddings w.r.t. emotion because the x-vectors are extracted from the speaker recognition model and contain speaker-specific information. After fine-tuning for emotion targets, the model successfully unlearned speaker information and learned emotion discriminatory information as is evident from Figure. 3.2c and 3.2d. We can observe clusters of emotions in Figure. 3.2d and no such clusters w.r.t. speaker as seen in Figure. 3.2c.

On the contrary, embedding space of a model fine-tuned on MSP-Podcast does not show clear clusters of emotions (Figure. 3.3d) and moreover, we can see visible clusters of speakers (Figure. 3.3c. Upon further investigation, we found that many speakers have single emotions for the majority of the time. This characteristic of the dataset could have affected model fine-tuning as the model is pre-trained to discriminate speakers. In other words, the model is unable to unlearn speaker information. Based on this experiment, we recommend maintaining diverse emotions per speaker during model training otherwise the model could potentially use speaker characteristics for SER leading to worse SER results on unseen speakers. It also implies that maintaining non-overlapped speakers in train and test sets is important for the realistic estimation of model abilities to perform SER on unknown speakers.

3.3.4.2 Model errors Vs. inter-annotator agreement

As discussed in Chapter 1, emotion perception of a speaker depends on a lot of factors. Hence, it is common to observe different emotion annotations among annotators. Higher agreement between annotators is usually achieved when the emotion of the speaker is very clear such as hot anger, crying. Similarly, the lower agreement can be observed when emotion is ambiguous. Figure 3.4 presents model correct classifications and misclassifications per each agreement level for Crema-D and MSP-Podcast datasets. On the spectrum of acted to naturalness, Crema-D stands very close to acted and MSP-Podcast to natural. Inter-annotator agreement for an utterance is calculated by taking the ratio of the maximum number of annotators agreed on emotion and the number of annotators for that utterance. For example, a value of 0.8 means



(a) Crema-D dataset **(b)** MSP-Podcast dataset

Figure 3.4: Model errors w.r.t. inter-annotator agreement

that 80% of the annotators agree on an emotion. MSP-Podcast recordings are annotated with a minimum of 5 crowd-sourced workers and more than 95% of the Crema-D utterances received a minimum of 7 crowd-sourced workers. For Crema-D, we can observe that number of misclassifications is less compared to correct classifications in every agreement level. Whereas for MSP-Podcast, the number of misclassifications dominates at lower agreement levels and gets better with a higher agreement. This experiment suggests a possible disparity between acted and natural datasets where the former might contain over-acted/over-emphasized emotions which most of the time, may not occur in real life.

3.4 CopyPaste data augmentation

In this section, we present a perceptually motivated data augmentation approach, CopyPaste, to improve SER performance. We compare the CopyPaste method with a widely used noise augmentation method on clean test sets (original recordings) and noisy test recordings as well.

3.4.1 CopyPaste approach

When trying to classify the emotion of an utterance, some segments might have more emotional information than others. To this respect, some authors observed that when an emotionally neutral speech segment and an emotional segment with emotion *E* are played in sequence, the human listeners commonly classify the whole sequence as emotion *E* (Tóth, Sztahó, and Vicsi, 2008). Therefore, the emotional segments (non-neutral) in an utterance might define the listener's perception. For example, consider a 10 s recording where the speaker is angry for the first 3 s and manifests a neutral emotion for the remaining 7 s. We surmise that a human annotator might label the utterance as angry even though the speaker expresses neutral emotion for the most part of the recording. In these cases, the recognition of the angry emotion by machine learning models might be difficult, as the neutral emotion dominates the overall statistics of the utterance referred to in the previous example. In this work, we address this problem by proposing the *CopyPaste* augmentation technique. This technique considers that a speaker is perceived to be expressing an emotion *E* (non-neutral) even if that emotion is exhibited for a short duration. Hence, we propose a data augmentation methodology consisting

of the concatenation of an emotional utterance with emotion *E* and a neutral utterance. The resulting concatenated utterance is then labeled with emotion *E* for model training. As we copy one utterance and paste (concatenate) it at the beginning or end of another utterance to produce a new one, we have called this process *CopyPaste* augmentation. Under this method, we present three data augmentation schemes for model training:

1. Concatenation of an emotional utterance (say emotion *E*) and a neutral utterance to produce another utterance with emotion *E*. We refer to this scheme as *Neutral CopyPaste (N-CP)*

2. Concatenation of two emotional utterances with same emotion *E* to produce another utterance with emotion *E*. We refer to this scheme as *Same Emotion CopyPaste (SE-CP)*

3. Using *N-CP* and *SE-CP* together during model training. We denote this scheme with *N+SE-CP*

Through *CopyPaste* schemes, we can produce a greater variety in the training data which can help the model generalize better. We expect that the *N-CP* scheme i.e., concatenating emotional utterances with neutral utterances forces the model to focus more on emotional parts of an utterance. For example, if the input is a concatenation of angry utterance and neutral utterance then *the N-CP* scheme forces the model to focus more on the angry part of the utterance compared to the neutral part of the utterance.

3.4.2 CopyPaste schemes implementation

During training, we randomly sample a batch of 128 utterances and perform *CopyPaste* based on the emotion class labels. For *the SE-CP* augmentation scheme, we pick the utterances with the same emotion labels and randomly pair them for concatenation. For *the N-CP* augmentation scheme, we pick utterances with neutral emotion and randomly pair them with all utterances in the batch including neutral utterances. In this scenario, there is a risk that the resulting models are biased against the neutral emotion, as 50% of each augmented utterance is of neutral emotion, and yet we force the model to predict the emotion of the other 50% of the augmented utterance. To avoid that danger, we perform *CopyPaste* augmentation only for 80% of the batches in each epoch. With the same premises, in the *N+SE-CP* scheme, we follow each of *N-CP* and *SE-CP* schemes for 40% of the batches in each epoch amounting to 80% of batches with *CopyPaste* augmentation. To avoid overfitting, we randomly pick 4 s from each recording for concatenation instead of the whole recording. We note that the average length of the training recordings in our datasets is less than 6 s. Hence, our hypothesis is affected only with negligible likelihood by picking only 4 s of each recording for concatenation.

3.4.3 Comparison with noise augmentation

In this work, we augment the training data by adding noise and music from MUSAN corpus (Snyder, Chen, and Povey, 2015). Our augmented data contains six copies of the training set with SNRs of 10 dB, 5 dB, and 0 dB after

Table 3.3: SER results (micro-f1 scores) with randomly initialized ResNet model. *Clean+Noise* and *Clean* denote SER model training is on clean and noise augmented data, and clean data respectively. In parenthesis, an absolute improvement compared to the model trained without *CopyPaste* (No *CP*) is shown.

Dataset	Emotion data	No *CP*	*SE-CP*	*N-CP*	*N+SE-CP*
MSP-Podcast	Clean	47.36	48.34	49.14	**49.69** (+2.33)
	Clean+Noise	48.15	50.61	49.25	**50.71** (+2.56)
Crema-D	Clean	71.46	71.80	**74.34** (+2.88)	73.79
	Clean+Noise	70.59	72.83	**75.87** (+5.28)	74.55
IEMOCAP	Clean	43.03	45.84	44.19	**45.88** (+2.85)
	Clean+Noise	43.65	49.49	**52.34** (+8.69)	51.41

Table 3.4: SER results (micro-f1 scores) with ResNet model pre-trained for speaker classification. *Clean+Noise* and *Clean* denote SER model training is on clean and noise augmented data, and clean data respectively. In parenthesis, an absolute improvement compared to a model trained without *CopyPaste* (No *CP*) is shown.

Dataset	Emotion data	No *CP*	*SE-CP*	*N-CP*	*N+SE-CP*
MSP-Podcast	Clean	56.79	**58.68** (+1.89)	57.71	58.22
	Clean+Noise	57.91	**58.62** (+0.71)	57.82	58.13
Crema-D	Clean	77.86	78.54	**80.18** (+2.32)	79.21
	Clean+Noise	79.60	79.98	**80.17** (+0.57)	79.88
IEMOCAP	Clean	61.18	**62.15** (+0.98)	61.21	61.90
	Clean+Noise	62.57	63.08	63.48	**63.78** (+1.21)

adding noise and music. We denote the models trained with clean and augmented data as *Clean+Noise*. As researchers showed that the effectiveness of adding noise to the training data is more evident on noisy test data compared to clean test data (Hsiao et al., 2015), we compare noise augmentation with *CopyPaste* in noisy test conditions. As emotion datasets are usually clean and have higher SNR, adding noise to the test data is considered. We create two sets of test data, one with an SNR level of 10 dB and another with 0 dB for comparison with *CopyPaste*.

Table 3.5: Class-wise f1-scores on Crema-D dataset with *CopyPaste (CP)* schemes. We used the ResNet model pre-trained for speaker classification and trained on clean data; No *CP* denotes model trained without *CopyPaste*

Emotion class	No *CP*	*SE-CP*	*N-CP*	*N+SE-CP*
Sad	20.48	20.59	21.11	**22.37**
Happy	37.17	46.19	**54.47**	46.7
Angry	70.53	71.4	**75.85**	73.61
Neutral	87.62	87.55	**88.11**	87.83

3.4.4 Results

We report the micro-f1 score as a metric (higher the better) to measure emotion model classification performance. We first show the effectiveness of *CopyPaste* schemes on clean data and noise augmented data. Then, we present results on artificially created noisy test data to compare *CopyPaste* and noise augmentation.

General considerations: Tables 3.3 and 3.4 show the results of *CopyPaste* schemes on randomly initialized ResNet model and speaker pre-trained ResNet model respectively. Comparing both tables, we can observe that pre-training improves the model performance significantly on all datasets. Models trained with noise augmented data perform better compared to models trained only on clean data corroborating with previous research (Lakomkin et al., 2018b). Comparison of models trained with and without *CopyPaste* schemes (4th-6th columns vs. 3rd column) reveals that our models perform better on all datasets with all schemes. Though the application of *CopyPaste* schemes provides performance improvement in most cases, we do not observe

a single best scheme across datasets and models except on Crema-D where *N-CP* scheme consistently performs best. We can observe that *CopyPaste* schemes are effective on both clean data as well as noise augmented data. We note that the improvements obtained with *CopyPaste* schemes on the randomly initialized ResNet model are relatively higher compared to the improvements on the pre-trained ResNet model.

Per-class analysis: As noted in Section 3.3.2, there is a risk that the model can get biased to not predict neutral when *N-CP* scheme is employed during model training. Hence, we examined class-wise f1-scores of our models to identify the main source of improvements and observed that in most cases performance improved for all emotion classes. As an example, we show in Table 3.5 class-wise f1-scores of emotion classes on the Crema-D dataset. These scores are obtained with the ResNet model pre-trained for speaker classification and trained on clean data. We can observe improvements for all emotion classes with *CopyPaste* schemes during training. Among *CopyPaste* schemes, *N-CP* is performing best for all classes except for *sad* emotion for which *N+SE-CP* performs best.

Noise augmentation: Comparing the augmentation techniques, *CopyPaste* and noise augmentation, we can observe from Tables 3.3 and 3.4 that *Copy-Paste* schemes perform better in most cases suggesting that concatenating utterances based on emotion helps the model generalize better compared to adding noise to the training data. As noted in Section 3.4.3, we compare noise augmentation and *CopyPaste* in noisy test conditions too. Tables 3.6 and 3.7 show the results on the noisy test data with SNR levels of 10 dB and 0 dB

Table 3.6: SER results (micro-f1 scores) on noisy test data with $SNR = 10dB$ with ResNet model pre-trained for speaker classification. *Clean+Noise* and *Clean* denote SER model training is on clean and augmented data, and clean data respectively; No *CP* denotes model trained without *CopyPaste*

Dataset	Emotion data	No *CP*	*SE-CP*	*N-CP*	*N+SE-CP*
MSP-Podcast	Clean	55.25	**57.39**	55.54	56.61
	Clean+Noise	57.09	**57.52**	56.63	**57.52**
Crema-D	Clean	72.76	73.47	**77.06**	74.06
	Clean+Noise	78.48	78.79	79.10	**79.30**
IEMOCAP	Clean	58.82	**59.30**	58.93	59.01
	Clean+Noise	61.47	61.80	62.03	**62.63**

respectively. We used the model pre-trained with speaker classification for this experiment as it is performing the best on all the datasets. As expected, SER performance degraded on the noisy test data suggesting that our models are sensitive to noisy test conditions. Models trained with noise augmentation are more robust compared to models trained with only clean data which illustrates the benefits of augmenting training data with noise. We can also observe that noise augmentation, in most cases, outperforms *CopyPaste* in noisy conditions. However, our best models on all the datasets are when used both augmentations together which showcases the effectiveness of proposed *CopyPaste* schemes even in noisy test conditions.

3.5 Conclusion

In this chapter, we presented two approaches based on transfer learning and data augmentation to improve emotion predictions from speech. From transfer learning experiments, we found that embeddings extracted (x-vectors) from pre-trained speaker recognition models do contain emotion predictive

Table 3.7: SER results (micro-f1 scores) on noisy test data with $SNR = 0dB$ with ResNet model pre-trained for speaker classification. *Clean+Noise* and *Clean* denote SER model training is on clean and augmented data, and clean data respectively; No *CP* denotes model trained without *CopyPaste*

Dataset	Emotion data	No *CP*	*SE-CP*	*N-CP*	*N+SE-CP*
MSP-Podcast	Clean	52.65	**55.15**	52.68	53.91
	Clean+Noise	55.88	56.40	55.28	**56.44**
Crema-D	Clean	64.95	66.21	**71.40**	65.53
	Clean+Noise	76.44	76.38	**76.83**	76.60
IEMOCAP	Clean	**52.05**	51.73	51.58	51.62
	Clean+Noise	58.55	58.52	59.32	**59.69**

information. Further, adapting the entire pre-trained model boosted SER performance on all three datasets considered. Our experiments suggested that the SER performance on x-vectors is inversely proportional to the speaker verification performance i.e., the better the x-vector model the less suitable the embeddings are for SER task. However, fine-tuning experiments revealed that fine-tuning the best x-vector model provides better results on SER task.

For data augmentation, we proposed three CopyPaste schemes to improve SER performance. We found that CopyPaste schemes improve SER performance and outperform noise augmentation in clean conditions. However, in noisy conditions, noise augmentation performed better than CopyPaste. We obtained best results when using both CopyPaste and noise augmentation on all three datasets.

Chapter 4

Beyond isolated utterances: Conversational emotion recognition

4.1 Introduction

In the previous chapter, we presented speech emotion recognition (SER) from isolated utterances. The proposed approaches can be applied to conversational speech, provided that an utterance-level emotion segmentation is available either from another system or a human annotator. If the segmentation is available we can pass each segment through the models built on isolated utterances. However, in this case we might not be exploiting the context in which the segment's emotion is produced. In this chapter, we present multiple techniques to recognize emotions in the conversations and also show that accuracy can be improved using segmented recordings.

Most of the past work on conversational emotion recognition (CER) can be broadly classified into two categories: the ones using segmented recordings (Hazarika et al., 2018; Majumder et al., 2019; Li et al., 2020b; Zhang et al.,

2019; Grimm et al., 2007; Metallinou, Katsamanis, and Narayanan, 2013; Eyben et al., 2010; Schmitt, Cummins, and Schuller, 2019) and the other without using them (Grimm et al., 2007; Metallinou, Katsamanis, and Narayanan, 2013; Eyben et al., 2010; Schmitt, Cummins, and Schuller, 2019). Authors in (Hazarika et al., 2018) explore a fixed context (4 recent utterances) and speaker-specific modeling using gated recurrent unit (GRU) architecture. Their model, referred to as conversational memory network, uses attention mechanism and memory hopping to combine information from multiple streams of representations and to attend to history. The main limitation of this approach is its fixed context and a lack of extensibility to multi-party conversations. Model proposed in (Majumder et al., 2019), referred to as DialogueRNN, overcomes limitations of the fixed context and also proposes to use separate GRUs to model speaker, emotion and global context. Authors in (Zhang et al., 2019) propose to use graph based neural net by defining utterances and speakers as nodes to exploit context and speaker dependencies. A transformer model with pairwise speaker verification as auxiliary task is proposed in (Li et al., 2020b) to encode context and speaker information into the model hidden representations. Even though the above approaches provide good CER performance, all of them are fundamentally limited by their reliance on the availability of a segmentation of the recording/transcript, and their strong assumptions about each speaker turn consisting of just a single emotion.

CER without requiring segmented recordings is explored in (Grimm et al., 2007; Metallinou, Katsamanis, and Narayanan, 2013; Eyben et al., 2010; Schmitt, Cummins, and Schuller, 2019) by predicting emotional attributes on

a frame-level. Authors in (Grimm et al., 2007) use a fuzzy logic estimator while (Metallinou, Katsamanis, and Narayanan, 2013) propose an optimal statistical mapping between audiovisual features and emotion attributes using Gaussian mixture model (GMM). Deep learning models such as CNN and LSTM are used in (Eyben et al., 2010; Schmitt, Cummins, and Schuller, 2019) for frame-level prediction.

In this work, we present transformer-based models for CER by treating it as a sequence labeling task, where short duration frames of the speech signal are assigned emotion labels by a model that looks at the broader context. Based on self-attention operation, we proposed *DiverseCatAugment* (*DCA*), an augmentation scheme to improve transformer model performance. We quantified the effect of context by comparing models trained on isolated utterances and conversations. We compared transformer architecture with several neural architectures: ResNet-34, which models context locally in each layer and globally with stacked layers and; BiLSTM, which captures context sequential manner. To leverage both the local and global context modeling strengths of ResNet-34 and transformer architectures, we explore their joint training. The resulting model is further enhanced by incorporating interlocutor information to exploit speaker dependencies in the conversations. We present models that can work even without segmentation information while most of the past works require it to exploit interlocutor information. The proposed models can deal with multi-party conversations and do not assume one emotion per turn.

The rest of the chapter is organized as follows. First, we present our

models in Section 4.2 and the proposed *DCA* augmentation scheme in Section 4.3. Then, the experimental setup and results are detailed in Section 4.4 and 4.5 respectively. Finally, conclusions and future directions are discussed in Section 4.6.

4.2 Conversational emotion recognition

We present transformer-based models that predict emotion on a frame-level. Our experiments include the use of a basic transformer architecture and also a combination of transformer and a CNN architecture. For all the approaches in this work, we employ filter-bank features as input, with 25 ms frame length and 10 ms shift.

Baseline models: The proposed transformer models are compared with two baseline models employed in previous studies: one using a BiLSTM architecture (Lee and Tashev, 2015) and the other employing CNN (same as in Chapter 3). We use BiLSTM and CNN models as the mechanism of exploiting context is different in them compared to the transformer. BiLSTM learns context information in a sequential manner; CNN exploits local context in each layer and global context with a stack of layers; in contrast, the transformer has access to the entire conversational context in every self-attention layer. Also, self-attention operation in the transformer model allows attending other frames with the same emotion in the sequence while convolutional operation treats all frames inside a receptive field in a similar manner disregarding their class label. Our BiLSTM architecture contains a sequence of 6 bi-directional LSTM layers followed by a dropout layer and 2 fully connected layers to

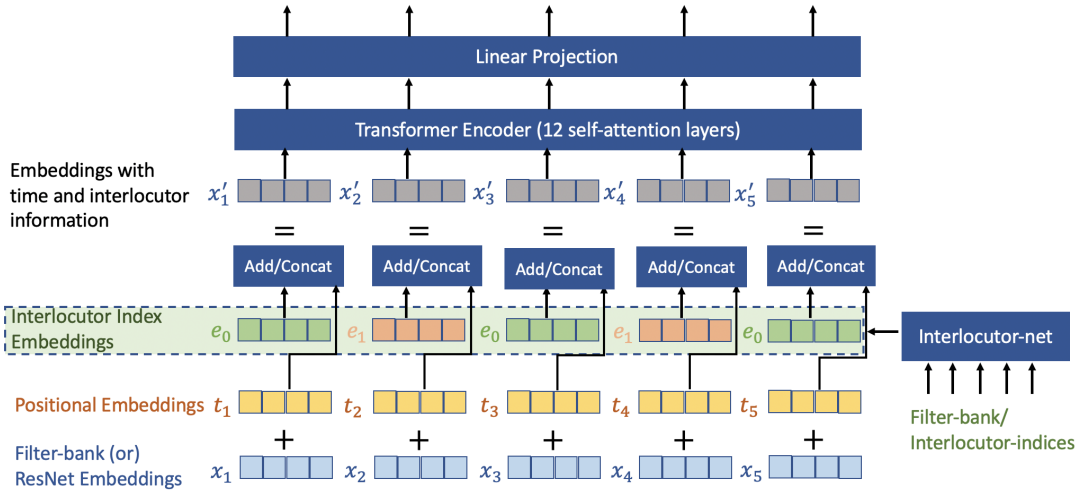


Figure 4.1: Transformer block diagram. Interlocutor index embeddings are used only with ResNet embeddings input in ResNet+Transformer model

obtain logits. For CNN, we use ResNet-34 model architecture reported in Chapter 3 without pooling layer.

4.2.1 Transformer model

As the employed frame length is very small to predict an emotion, the context plays a crucial role in deciding the emotion of a frame. Hence, an architecture that can use context efficiently is crucial. Recently, transformer architecture has shown to outperform other neural architectures in several speech and NLP tasks (Vaswani et al., 2017; Devlin et al., 2019; Wolf et al., 2020; Karita et al., 2019). It contains a sequence of self-attention operations which are designed to exploit long-range dependencies in the input sequence. Our architecture contains a sequence of 12 self-attention layers as in the standard BERT base model (Devlin et al., 2019). A schematic of the transformer model is shown in Figure 4.1. For this model, we use filter-bank features as input. As the entire

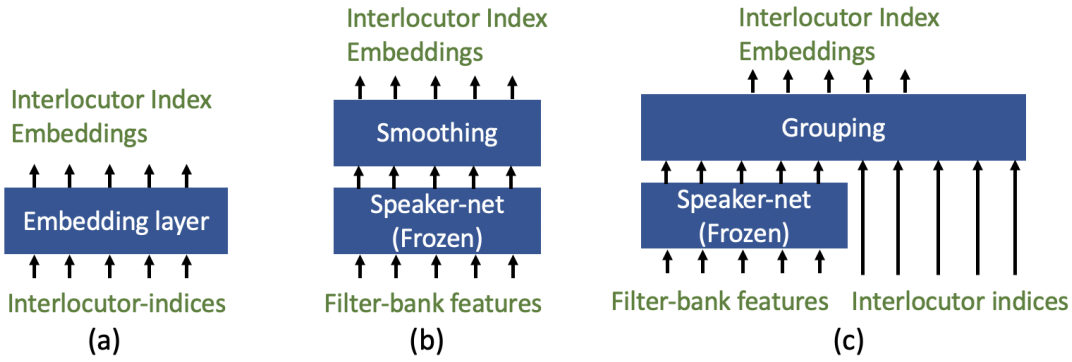


Figure 4.2: Proposed methods to interlocutor-net shown in Figure 4.1. Each of these are referred to as (a) Embedding layer (b) Speaker-net+Smoothing (c) Speaker-net+Grouping

input sequence is processed simultaneously in self-attention layers, the order of the input sequence does not matter. However, the sequence information could be useful for the CER task. We encode position information by learning a set of positional embeddings during training and adding them to the input sequence. For training, we use an input sequence length of 2048 hence we learn 2048 positional embeddings during training.

4.2.2 ResNet+Transformer model

Several studies suggest that down-sampling input representation using convolutional layers before processing with transformer layers provides better results for ASR (Lu et al., 2020; Mohamed, Okhonko, and Zettlemoyer, 2019). Intuitively, convolutional layers use local context to produce better contextual features. In this work, we used a pre-trained ResNet-34 to process input filter-bank features and fed its output to the transformer layers. ResNet-34 is pre-trained on the speaker classification task. We jointly trained ResNet-34 and transformer to exploit the benefits of both transfer learning and transformer

model capabilities.

4.2.3 Interlocutor-aware ResNet+Transformer model

A conversation is structured as a sequence of turns by all participating speakers. The emotion of a speaker in each turn could depend on that speaker's emotions in previous turns and also on the interlocutor's emotions (Hatfield, Cacioppo, and Rapson, 1993; Smirnov et al., 2019). Hence, we expect the model to perform better when the model knows who is speaking when in the conversation. Authors in (Majumder et al., 2019) show that distinguishing speaker and listener parties improves emotion prediction. However, they predict emotion on a segment/turn-level and also requires speaker diarized recording. In this work, we propose three methods to overcome these limitations by making frame-level predictions and using pre-trained speaker recognition model.

Our model schematic with interlocutor-net which produces interlocutor index embeddings is shown in Figure 4.1. We propose three methods to design interlocutor-net and they are shown in Figure 4.2. First method Figure 4.2(a) assumes the availability of speaker segmented conversations. In this method, we indexed the speakers in a conversation and represented them with one-hot encoding. Indices are assigned following the order in which the interlocutors appear in the conversation. We passed the one-hot encoding through an embedding layer to get interlocutor index embedding. The embedding layer learns a dictionary of embeddings which acts as a lookup table. The dictionary size is set to the maximum number of speakers that can appear in a training

sample. Then, we added interlocutor index embeddings to the ResNet output to incorporate interlocutor information into the transformer layers. This method can deal with multi-party conversations but during test time it is limited by the maximum number of speakers seen in a training conversation.

As the main goal of incorporating interlocutor information is to be able to distinguish speakers in the conversation, we propose to use a pre-trained speaker recognition model (Speaker-net) to extract speaker-specific representations. Interlocutor-net using speaker-net is shown in Figure 4.2(b). As the output of speaker-net is on frame-level, it could introduce more noise into the transformer model. This noise could be reduced by performing smoothing operation (moving average with a window of 0.8s duration). We experimented with adding and concatenating the output of interlocutor-net to the ResNet output in order to introduce interlocutor information.

Speaker-net represents different instances of a speaker in the conversation differently based on the spoken content in each instance. However, as the primary task of interlocutor-net in the transformer model is to introduce speaker-specific information, representing all instances of a given speaker using one vector would be more efficient. For this purpose, we use ground truth speaker diarization information and pool all the instances of each speaker in the conversation to represent with just one vector (refer Figure 4.2(c). Grouping in Figure 4.2(c) denotes picking all instances of each speaker and averaging. Even though this method requires speaker diarization information, it does not have any limitation on the number of speakers in unseen conversations like the method in Figure 4.2(a).

4.3 Diverse Category Augment Scheme

In this section, we present a data augmentation scheme, named as *Diverse-CatAugment* (*DCA*), motivated by the inner workings of the self-attention operation. Given an input sequence of vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, we perform self-attention operation and obtain a sequence of vectors $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$. Self-attention operation (dot-product variation) as shown in (4.1) includes finding dot-product between every vector in the sequence i.e., $\mathbf{X} \cdot \mathbf{X}^{\mathrm{T}}$. On the dot-product matrix, the softmax operation is employed to obtain normalized similarities for each vector with other vectors in the sequence. Then, the dot product matrix is multiplied with the input sequence $\mathbf{X}$ to obtain $\mathbf{Y}$. In essence, every vector in $\mathbf{Y}$ is a weighted sum of vectors in $\mathbf{X}$, as shown in (4.2) with weights being the normalized similarities with other vectors in the sequence as shown in (4.3).

$$\text{Let } \mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N],\ \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$$

$$\text{where } \mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d,\ \forall i \in [1, N]$$

$$\mathbf{Y} = \text{softmax}(\frac{\mathbf{X} \cdot \mathbf{X}^{\mathrm{T}}}{\sqrt{d}})\,\mathbf{X} \tag{4.1}$$

$$\mathbf{y}_i = \sum_{j=1}^{N} w_{ij}\mathbf{x}_j, \forall i \in [1, N] \tag{4.2}$$

$$[w_{i1}, w_{i2}, ..., w_{iN}] = \text{softmax}(\frac{\mathbf{x}_i \cdot \mathbf{x}_1}{\sqrt{d}}, \frac{\mathbf{x}_i \cdot \mathbf{x}_2}{\sqrt{d}}, ..., \frac{\mathbf{x}_i \cdot \mathbf{x}_N}{\sqrt{d}}) \tag{4.3}$$

Attention operation allows us to attend relevant vectors in the sequence by

assigning higher weights and discard irrelevant vectors using lower weights. For a given vector, say $\mathbf{x}_i$, if all the weights/similarities ($[w_{i1}, w_{i2}, ..., w_{iN}]$) are in a narrow range, it implies that all the vectors in $\mathbf{X}$ are equally relevant to $\mathbf{x}_i$. This could happen if all the vectors in the sequence belong to the same class. In this case, the attention operation acts as, effectively, an averaging operation instead of a weighted average. Consequently, we may not be exploiting transformer abilities to the maximum level. Based on this insight, we hypothesize that input sequences with less categorical variety hinder transformer model performance. Equivalently, training data with input sequences containing diverse emotion classes provide better performance compared to sequences with less emotional diversity.

We validate our hypothesis by proposing a data augmentation scheme, referred to as *DiverseCatAugment* (*DCA*), which improves the diversity of categories/emotions in the input sequences. We apply *DCA* on conversations as well as isolated utterances. When applying *DCA* to conversations, we choose two conversations and concatenate them for model training. For example, assume one conversation is filled with angry for most of the time and another with happy category. Concatenation of the two conversations results in a sequence with both angry and happy. It is easy to see that the concatenated conversations have a more diverse composition of emotions. According to the *DCA* hypothesis, proposed transformer models perform better if input sequences have diverse categories.

4.4 Experimental setup

4.4.1 Dataset

We performed CER on the widely used IEMOCAP dataset, which contains 150 dyadic conversations between 5 female and 5 male speakers. Each conversation is set up between one male and one female, and are approximately 5 min long. The scripts and topics for spontaneous conversations were selected to elicit emotions. Even though only 5 emotions – *Angry*, *Frustration*, *Happy*, *Neutral*, and *Sad* – are targeted for elicitation, more emotions albeit less frequently are found in the annotation process. In this work, we used only the most frequent emotions, – *Angry*, *Frustration*, *Happy*, *Neutral*, and *Sad* – for classification. We merged *Excitation* emotion with *Happy* as is commonly done for this dataset. The conversations are annotated in crowd-sourced manner with 3 annotators. Even though only 5 emotions were under consideration for the dataset, they found more emotions in the annotation process albeit less frequently. To facilitate comparison between models trained with isolated utterances and conversations, we discarded segments in the conversations which have labels other than the considered emotions. Training data setup is same as in Chapter 3, i.e., we use 3 sessions for training, 1 for development, and 1 for testing; we perform a 5-fold cross-validation (leave-one-out-session); report the micro-f1 score (refer to Chapter 2 for definition).

4.4.2 DCA implementation

We implement *DCA* augmentation during the formation of the batch for model training. We first choose a batch of 6 conversations and pick a sequence of length 1024 from each of the conversations. Then, we randomly pair each conversation with one of the other 5 conversations to form a sequence of length 2048 for model training. We note that to maximize *DCA* utility, conversations with distinct emotions should be selected for concatenation but as we train the model for 100 epochs, the model sees a fairly high number of sequences with diverse emotions. *DCA* on conversations produces sequences with conversational context preserved for most of the sequence and adds a bit of random context. When applying *DCA* on isolated utterances, we concatenate multiple isolated utterances until we obtain 2048 length sequences. We choose isolated utterances for concatenation randomly to result in a sequence with diverse emotions expressed by multiple speakers. *DCA* on isolated utterances results in sequences similar to conversations but without conversational context.

4.4.3 Impact of the context

To gain insights into the model capabilities and importance of context, we compare the transformer model with ResNet-34 and BiLSTM using 4 types of training data:

1. *Isolated utterances* (no context)
2. *Conversations* (original conversational context)
3. *DCA Isolated utterances* (random context)

4. *DCA Conversations* (original conversational context + random context)

We evaluate all the models on conversations. We compare models trained with *Isolated utterances* (no context) and *Conversations* to understand the impact of conversational context on the CER performance. To further improve the performance, we employ *DCA* on *Isolated utterances* and *Conversations*.

Based on the *DCA* method hypothesis, we expect *DCA Conversations* and *DCA Isolated utterances* to perform better than *Conversations* and *Isolated utterances* respectively. Also, as context could help to disambiguate emotions, we expect models trained on conversational data (2^{nd} and 4^{th} types) to perform better than models trained on isolated utterances data (1^{st} and 3^{rd} types). The performance of models trained with isolated utterances enables us to answer the question of "how well can we perform CER without access to the conversational data?". The answer to this question is important because most of the current datasets have only isolated utterances and a lot of past research efforts focused on them.

4.5 Results

4.5.1 Results with *DCA* augmentation and context

Table 4.1 shows the results with *DCA* augmentation and context. The first and second rows, denoted with *Isolated* and *Conversations*, show the results of models trained with isolated utterances and conversations. We can observe that models trained on isolated utterances perform worse than the models trained on conversations suggesting the importance of context. BiLSTM

seems to predict just a little better than chance when trained on isolated utterances. The impact of conversational context on the BiLSTM model is comparatively higher than ResNet and transformer. Among the architectures, the transformer model outperformed ResNet and BiLSTM in every case with the best performance of 42% when trained on conversations.

Models trained with *DCA* augmentation are denoted with *DCA Isolated utterances* and *DCA Conversations*. We can observe that along with the transformer model, ResNet and BiLSTM also perform better with *DCA* augmentation on isolated utterances suggesting that emotional variety in the training sequences helps to discriminate emotions well. On isolated utterances, ResNet and BiLSTM models perform 3.9% and 13.1% absolute better with *DCA* augmentation. However, they perform worse in comparison to *Conversations* suggesting that original conversational context is more important than categorical/emotional variety in the training sequences. Interestingly, the transformer model trained with *DCA Isolated utterances* performs better than *Conversations*. Upon further investigation into the conversations, we found that many conversations are dominated by a single emotion. Figure 4.3 shows proportions of emotions for a subset of 38 conversations (25% of the dataset) in the IEMOCAP dataset. Each bar represents the proportion of emotions in a single conversation. We can observe that these conversations have only one emotion dominating for more than 75% of the conversation time. In literature, this phenomenon is referred to as emotional inertia (Kuppens, Allen, and Sheeber, 2010) which states humans naturally tend to resist changing emotions.

Table 4.1: Effect of context on the CER performance (micro-f1). Conv. context means the original conversational context; *DCA Isolated utterances* – *DCA* augmentation on isolated utterances; *DCA Conversations* – *DCA* augmentation on conversations

Training data type	Context type	ResNet	BiLSTM	Transformer
Isolated utterances	No	34.3	27.1	39.1
Conversations	Conv.	39.2	41.6	42.0
DCA Isolated utterances	Random	38.1	40.2	42.7
DCA Conversations	Random+Conv.	37.5	41.6	45.3

Emotional inertia in the conversations explains the better performance with *DCA Isolated utterances* compared to *Conversations* even though the latter has conversational context. It also implies that emotional variety in the training sequences is important for the transformer model confirming the *DCA* augmentation hypothesis. Better (3.3% absolute) performance with *DCA Conversations* over *Conversations* further strengthens the *DCA* augmentation hypothesis.

Overall, we observed that training the models with random context is better than no context. Access to the conversational context further improved our models' performance. Transformer model trained with conversations and *DCA* augmentation performed best with a micro-f1 of 45.3%.

4.5.2 Results with ResNet+Transformer and its analysis per emotion

Table 4.2 compares the results of the ResNet+Tranformer model with only the transformer model. We can observe 4.5% absolute improvement in CER performance suggesting that processing with convolutional layers helps. For this model, we employed *DCA* augmentation on conversations as it yielded

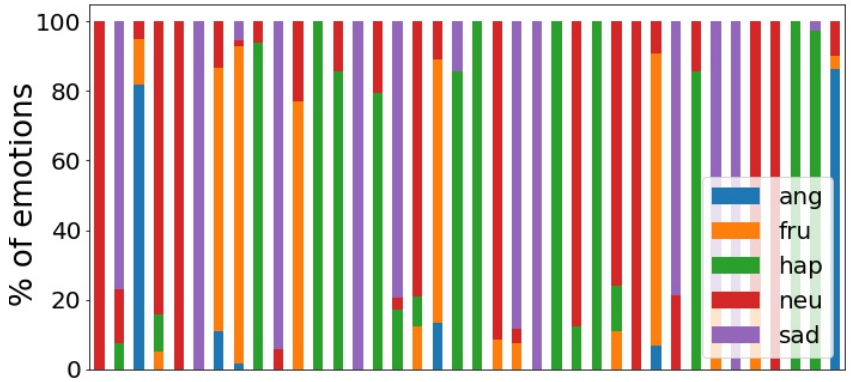


Figure 4.3: Proportion of emotions in a subset of 38 IEMOCAP dataset conversations (25% of the dataset). These conversations have one emotion occurring for more than 75% of the conversation. Each bar corresponds to one conversation

Table 4.2: Results of joint ResNet and transformer training. *DCA* on conversations is employed for model training

Model	micro-f1
Transformer	45.3
ResNet+Transformer	49.8

the best results. To understand our model errors, we show an analysis of our model's row-normalized confusion matrix in Figure 4.4. We can observe that our model is confusing *Angry* with *Frustration* 37.6% of the frames and *Neutral* with other emotions 67.9% of the frames. *Angry* and *Frustration* seem much more similar to each other than to any other emotion in the label set, hence we wondered whether there could be some confusion between them for annotators too. Looking at inter-annotator agreement, we found that when the annotation of each crowd-sourced worker is matched against their majority-voted annotation, *Angry* is found to be confused with *Frustration* 17% and *Frustration* with *Angry* 11% of total segments (Busso et al., 2008) which are

significantly high compared to any other emotion. These confusion rates in the ground-truth annotations would explain our model's confusion to some extent.

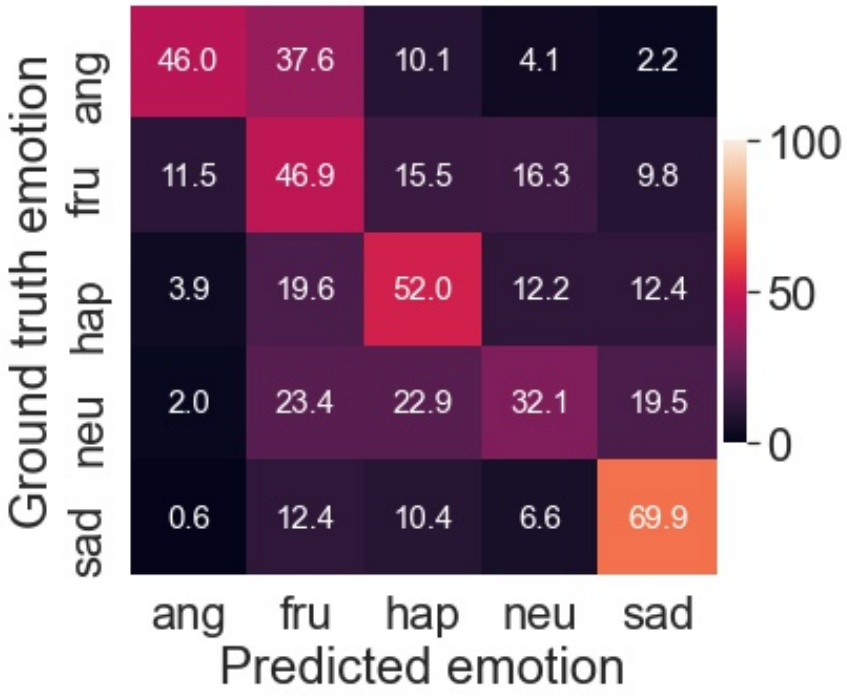


Figure 4.4: Confusion matrix of ResNet+Transformer model

To understand the confusion of *Neutral* emotion with others, we investigated the trigram probabilities of emotions in the conversations. Figure 4.5 shows dataset statistics for a subset of trigrams of the form (neighbor-emotion, central-emotion, neighbor-emotion). These statistics are computed from sequences of turn/segment emotions. Each row is normalized for analysis purposes. If central-emotion is equal to neighbor-emotion then we call the trigram as homogeneous, and heterogeneous otherwise. From Figure 4.5, we can observe that the majority of trigrams are homogeneous (diagonal

values) except when the central-emotion is *Neutral*. Approximately 54.3% (100%-45.7%) of the trigrams are heterogeneous for *Neutral* emotion compared to 38.5%, 37.5%, 7.5%, and 13.1% for *Angry*, *Frustration*, *Happy*, and *Sad* respectively. We speculate that the heterogeneous nature of *Neutral* in this dataset could be one reason why our model confuses with other emotions more often – it simply "prefers" to recognize longer contiguous segments with a single emotion, mislabeling *Neutral* in the process. This observation is consistent with the hypothesis presented in previous chapter (Chapter 3) that neutral utterances are perceived as emotional when presented in the context of another emotional utterance. However, whether this behavior is because of the dataset characteristics or the acoustic characteristics of *Neutral* emotion warrants further analysis which we plan to address in future work.

4.5.3 Results with interlocutor-aware ResNet+Transformer

Table 4.3 presents the results of models trained with various types of interlocutor-nets. We found that infusion of interlocutor information does improve CER performance. Learning interlocutor index embeddings using only speaker segmentation information (Figure. 4.2(a)) provided 2.98% improvement when infusion is done by addition. In contrast, 3.8% degradation is observed with infusion by concatenation. We suspect that the model could not learn to exploit concatenated information as it is easy to discard the extra dimensions by assigning small weights whereas with infusion by addition the model is forced to learn embeddings for interlocutors. Replacing the embedding layer with speaker-net as in Figure. 4.2(b) does not degrade the performance

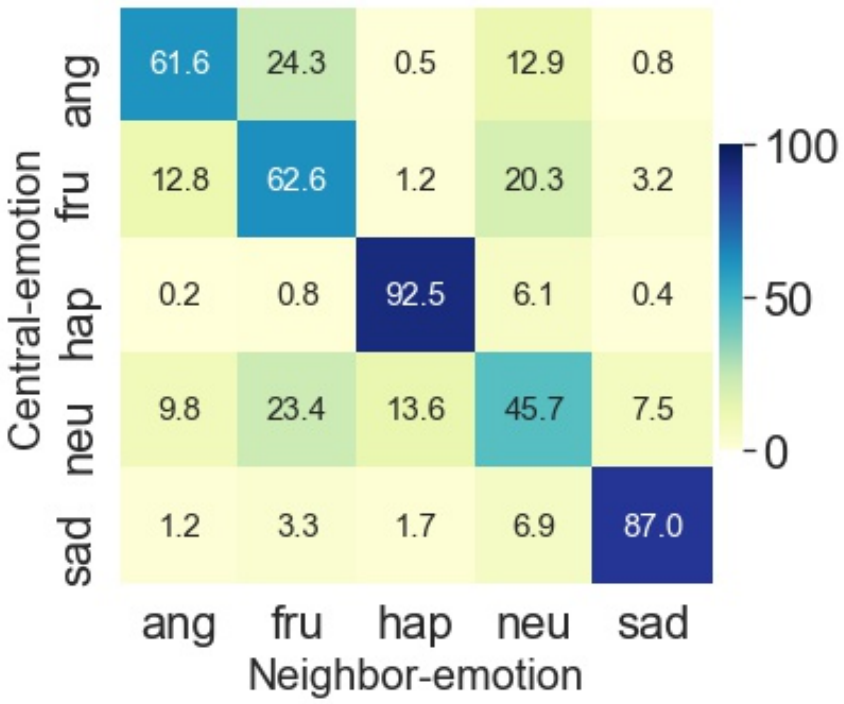


Figure 4.5: Probabilities of trigrams of the form (neighbor-emotion, central-emotion, neighbor-emotion). The labels ang, fru, hap, neu, and *sad* stand for *Angry*, *Frustration*, *Happy*, *Neutral*, and *Sad* respectively.

with both infusion by addition and concatenation. Even though speaker-net focuses on speaker-specific information, it also encodes other details such as emotion, channel and noise specific to the input conversation. For this reason, we think the model learned to exploit concatenated information as opposed to discarding unlike the case of Figure. 4.2(a). We obtained 1.4% improvement with infusion by concatenation where as no change in the performance with infusion by addition. Using interlocutor-net shown in Figure 4.2(c) which combines both Figure 4.2(a) and (b) provided improvements with both infusion by addition as well as concatenation. We obtained best performance of 53.3% with infusion by concatenation using interlocutor-net shown in Figure 4.2(c).

Table 4.3: Influence of interlocutor information on the performance of ResNet+Transformer model. Training without interlocutor-net is baseline for this experiment which provided 49.8% micro-f1 as shown in Table. 4.2

Interlocutor-net	Infuse by addition	Infuse by Concatenation
Embedding layer (Figure. 4.2(a))	53.0	46.0
Speaker-Net+Smoothing (Figure. 4.2(b))	49.8	51.2
Speaker-Net+Grouping (Figure. 4.2(c))	52.5	53.3

The interlocutor embeddings obtained with Figure 4.2(c) are more speaker specific as they are obtained by averaging all segments of the speaker in the conversation compared to the ones extracted using Figure 4.2(b) and hence the better performance with the former. Lastly, we think interlocutor-net in Figure 4.2(c) performed best because the interlocutor embeddings from it are speaker-specific and also specific to the conversation.

4.6 Conclusions and future work

In this work, we presented transformer-based models for conversational emotion recognition (CER). Our analysis on the impact of context showed that models trained with random conversational context perform better on conversations than those trained without context from other speakers. We found that less diversity of emotions/categories in the input sequences limits the transformer model performance. Our proposed data augmentation scheme which aims to improve diversity has helped to discriminate the emotions better. Conversational context and diversity of emotions provided the best results when using transformers. The proposed transformer-based approaches always outperformed the baseline architectures ResNet-34 and BiLSTM. We

presented a model combining ResNet-34 and transformer architecture to exploit local and global context, that provides better results than the model based on transformer only. We proposed three methods to incorporate interlocutor information to improve the CER performance. Two of which expect speaker diarized recordings at the test time and the other one method does not require them.

In this work, we evaluated the proposed methods on the IEMOCAP corpus. Some of the shortcomings of this corpus are its limited number of speakers and it is collected in controlled settings. We plan to evaluate our models on more spontaneous conversations data with more speakers such as MELD (Poria et al., 2019). Also, we look to investigate our model behaviour more closely especially on *Neutral* emotion as it is confused with other emotions more often compared to other emotions.

Chapter 5

Customer Satisfaction Prediction

5.1 Introduction

Today's markets often rely on customer feedback to improve their customer support. However, very few customers rate their experience with the company services. Thus, automatically predicting customer satisfaction (CSAT) plays a vital role in the company's businesses. This task consists of predicting the overall sentiment of the customer in a conversation with an agent. The specific goal could be to predict the likelihood of customer being satisfied or dissatisfied. Some of the applications of CSAT prediction include evaluating the quality of spoken dialog systems (SDS) (Engelbrech et al., 2009), market analysis, employee management, employee efficiency evaluation, customer retention prediction (Sabbeh, 2018; Ranaweera and Prabhu, 2003) and customer loyalty evaluation (Ansari and Riasi, 2016; Hallowell, 1996).

Evaluating the sentiment of conversations is challenging due to several reasons and it is different from evaluating written text like movie reviews or product reviews. Usually, sentiment in the call center conversations is

dynamic, i.e., it varies from time to time as the conversation goes on because it depends on how each person in the conversation responds. In contrast, movie reviews are written with a clear intent of expressing a particular sentiment. Vocal conversations involve multiple speakers talking to each other by taking turns for smooth flow of information. In customer care center (CCC) calls, one channel represents the agent who is trained and expected to speak in a particular manner, limiting the variety of their vocabulary and phrasing. The other channel represents the customer, whose speech is spontaneous and varies in terms of accent, emotions and expressions.

Obtaining data for CSAT prediction task is challenging because very few customers rate their experience. Alternative to customer's self-reported rating is annotating with expert annotators. However, sentiment/emotion of an audio/text utterance is subjective to each individual based on demographics, context, gender and culture. The work in (Aman and Szpakowicz, 2007b), sentiment analysis on blog posts, reported that the average inter-annotator agreement was only 0.76 in labeling a sentence as emotional or neutral, and only 0.6 to 0.79 for labeling emotion categories using crowd-sourcing methods which further illustrates the difficulty of data collection as well as building machine learning models for this problem.

In this work, we focus on designing and investigating several acoustic and lexical feature representations for CSAT prediction. Lexical features are extracted from automatic speech recognition (ASR) transcriptions of the speech signal. Acoustic cues are extracted from speech signal directly. We

propose to extract features at three-levels from speech: frame-, turn-, and call-level. Similarly, from transcript, we extract features at word-, turn-, segment- and document-level. Each feature-level representation encodes information available in the conversation at different levels of granularity. For each feature-level, we present stand-alone models as well as models using transfer learning. Along with the acoustic and lexical features, we experiment with turn-taking features which are based on conversation cues such as speaking rate or call duration. Then, we present systems for the fusion of all the information sources namely, acoustic, lexical and turn-taking features. We experimented with model-level fusion and score-level fusion to exploit complementary information present in these sources. Then, analysis of our classification results with respect to task completion metric (whether the task is completed), importance of agent's vs. customer's data to the CSAT rating and the effect of dataset size is presented.

The rest of the chapter is organized as follows. We present a review of literature related to CSAT prediction in Section 5.2. Then, we introduce the dataset used in this work in Section 5.3 followed by feature extraction methods and overview of our approaches on this dataset in Section 5.4 and 5.5 respectively. We present our experimental setup in Section 5.3 and our experiments on ASR transcripts, acoustic signal and turn-taking features in Section 5.7, 5.8 and 5.9 respectively. After presenting individual models and results on the transcripts, acoustic signal and turn-taking features, we present fusion methods in Section 5.10. Then, we discuss about the importance of each speaker's data for accurate CSAT prediction and the need for more data in Section 5.11. Finally,

ethical considerations using the models discussed in this work are presented in Section 5.12 followed by conclusion and future work in Section 5.13.

The main contributions of this work are:

- Extensive analysis using four-levels of document representation: word-, turn-, segment- and document level.
- Extensive analysis using three levels of speech representations: frame-, turn- and call level.
- Application of state-of-the-art pre-trained sentence encoder models ELMo, USE, and BERT
- Proposing the application of pre-trained x-vector model for CSAT prediction from speech signal
- Novel turn-taking features: Terse dialogue metrics
- Model-based and score-based fusion of acoustic, lexical and turn-taking feature representations
- Task completion metric based analysis of our results
- Analysis of the importance of agent and customer through their spoken content and acoustic cues

5.2 Related work

CSAT prediction was one of the essential goals of the DARPA Communicator program which targeted mainly travel planning human-machine conversations (Walker, Hirschman, and Aberdeen, 2000). Turn-taking features were extensively used in this program (Walker, Hirschman, and Aberdeen, 2000; Walker, Passonneau, and Boland, 2001). The authors in (Walker, Passonneau, and Boland, 2001) evaluated customer satisfaction on human-machine conversations using PARADISE (Walker et al., 1997) framework. Features extracted from dialog act tagging along with turn-taking features were used in this work. In (Yang, Levow, and Meng, 2012), authors used collaborative filtering models to evaluate customer satisfaction in PARADISE framework. Currently, machines involved in human-machine conversations usually can deal with a limited set of dialogue topics and are incapable of conversing with the human in a natural way, which makes these conversations very different from human-human conversations.

CSAT prediction on human-human conversations is explored in several works (Chowdhury, Stepanov, Riccardi, et al., 2016; Luque et al., 2017; Park and Gates, 2009; Meinzer et al., 2016). In (Chowdhury, Stepanov, Riccardi, et al., 2016), turn segmentation and labelling system is used to extract more accurate turn-taking features. The authors compare turn-taking features with prosodic features and lexical features, using a support vector machine (SVM) classifier. They use bag of words (BOW) representation and find that turn-taking features outperform the other features. Authors in (Luque et al., 2017) use principal component analysis (PCA) and CNN on lexical features and

XGBoost (Chen and Guestrin, 2016) on prosodic features and observe that lexical features have more relevant information compared to prosodic features. Word level lexical features are fused only with corresponding fundamental frequency and loudness in the audio signal at the feature level to process using CNN which needs word level boundaries.

Authors in (Park and Gates, 2009), explored several fundamental machine learning techniques such as decision trees, Naive Bayes, SVM, and logistic regression for the CSAT task on a feature set derived from various sources. The feature sets were categorized as structural, prosodic, lexical, and contextual features. To extract structural and some of the contextual features, external sources other than call transcripts were used and, prosodic and lexical features were extracted from call transcripts. Note that authors only considered talking speed, call dominance, long pause as a set of prosodic features and did not consider acoustic related features as opposed to standard literature. In (Meinzer et al., 2016), authors worked on the dataset collected from the automotive industry with similar goals, but the representation was obtained from discrete sources like warranty, vehicle type, problem type, etc. They used SVM and Random Forest to find dissatisfied customers.

Some of the highly correlated metrics to CSAT include net promoter score and task success rate. In (Auguste et al., 2019), a closely related label, net promoter score, is used to study agents and customers behaviour. The authors only use transcripts for their experiments. Task success was predicted in (Reitter and Moore, 2007) on a dataset which involves interactions between two subjects where one guides another to a destination using a pre-defined

route. Lexical and syntactic repetitions in the conversation were used on specific structural phrases. For classification, logistic regression and SVM was employed. In (Noseworthy, Cheung, and Pineau, 2017), task success is evaluated on human-human text conversations extracted from *stackoverflow*, an online forum, employing turn-based RNN models to predict success. In this work, we analyze the use of task success metric for CSAT prediction.

In most of these works, either the features are derived from external sources and/or employ only basic machine learning methods with simple input representations. In contrast, we investigate several feature representations for speech and text along with turn-taking features hand-curated from audio segmentation information. Also, transfer learning mechanism is not explored in the past works except for word embeddings. In this paper, we present methods using features obtained from several pre-trained word, sentence and speech utterance embedding models.

5.3 CSAT dataset

Our dataset comprises of US English telephone speech from call centers. Most of the calls in the dataset are about technical support, customer complaints and general inquiries. Dataset consists of 4331 audio calls with an average call duration of 8 minutes. Each call recording contains two channels, one for agent and another for customer. At the end of the call, customers are asked to rate their experience on a scale of 1-9, 9 being extremely satisfied. Figure 5.1 shows the histogram of customer ratings. We can observe that most of the calls are rated either as extremely dissatisfying (rating 1) or fully

satisfying (rating 9) as is the case with similar datasets (Schoenmüller, Netzer, and Stahl, 2019). In this work, we quantize the ratings above 4.5 and below 4.5 to positive and negative calls, respectively to obtain an almost balanced dataset for experiments and we present experiments to predict whether the call is positive or negative. Customers are also asked to tell if their issue or inquiry is resolved by pressing 1 for Yes or 2 for No. We denote these labels with task completion in this work. Figure 5.2 shows the histogram of the task completion responses for positive and negative rated calls. For ease of reading, we mapped the customer responses 1 and 2 to *Successful* and *Unsuccessful* respectively. We can observe that almost all positive calls have successful task completion and, an almost equal number of negative calls with successful and unsuccessful task completion.

For analysis of spoken content, we obtain ASR transcripts by employing an ASR system trained on Fisher and Switchboard datasets with lattice-free maximum mutual information criterion (Povey et al., 2016). The word error rates using four-gram language models are 9.2% and 17.3% respectively on Switchboard and CallHome portions of Eval2000 dataset[1]. The word error rate on a held-out dataset of 20 conversations is 21.4%, which means approximately one word in every five words is recognized incorrectly. The ASR transcript is supplemented with the information regarding who is speaking (i.e., agent or customer) at a given time instant.

As the length of the documents affect the choice of the machine learning model, we present few statistics of our dataset w.r.t. word and turn count.

[1]https://catalog.ldc.upenn.edu/LDC2002T43

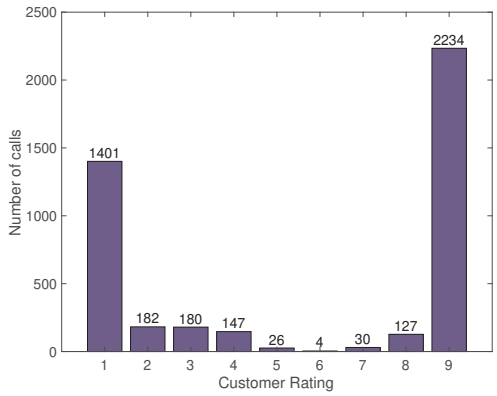


Figure 5.1: Histogram of customer ratings. Rating 9 corresponds to extremely satisfied and 1 to extremely dissatisfied

Table 5.1: Dataset statistics

	mean	median	min	max
words	821.9	582	15	10932
turns	90.1	62	6	1764

Table 5.1 shows the word and turn statistics and Figure 5.3 displays the corresponding cumulative distributions. We can observe that more than half of the dataset has calls longer than 582 words and 62 turns. Also, approximately 8-10% of the calls are very long (longer than 2000 words and 200 turns). The number of unique words found in these ASR transcripts, i.e., dataset vocabulary size is 23699.

In the next section, we present feature extraction methods we use in this work followed by an overview of our methodology for CSAT prediction on this dataset.

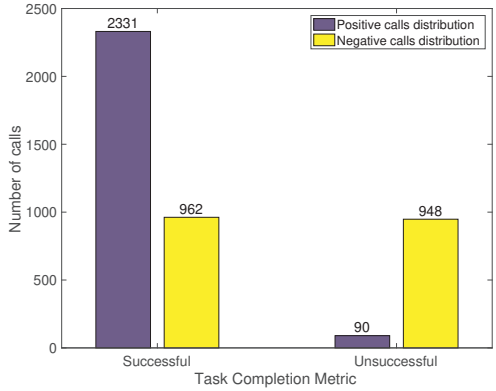


Figure 5.2: Histogram of positive and negative calls with respect to task completion metric

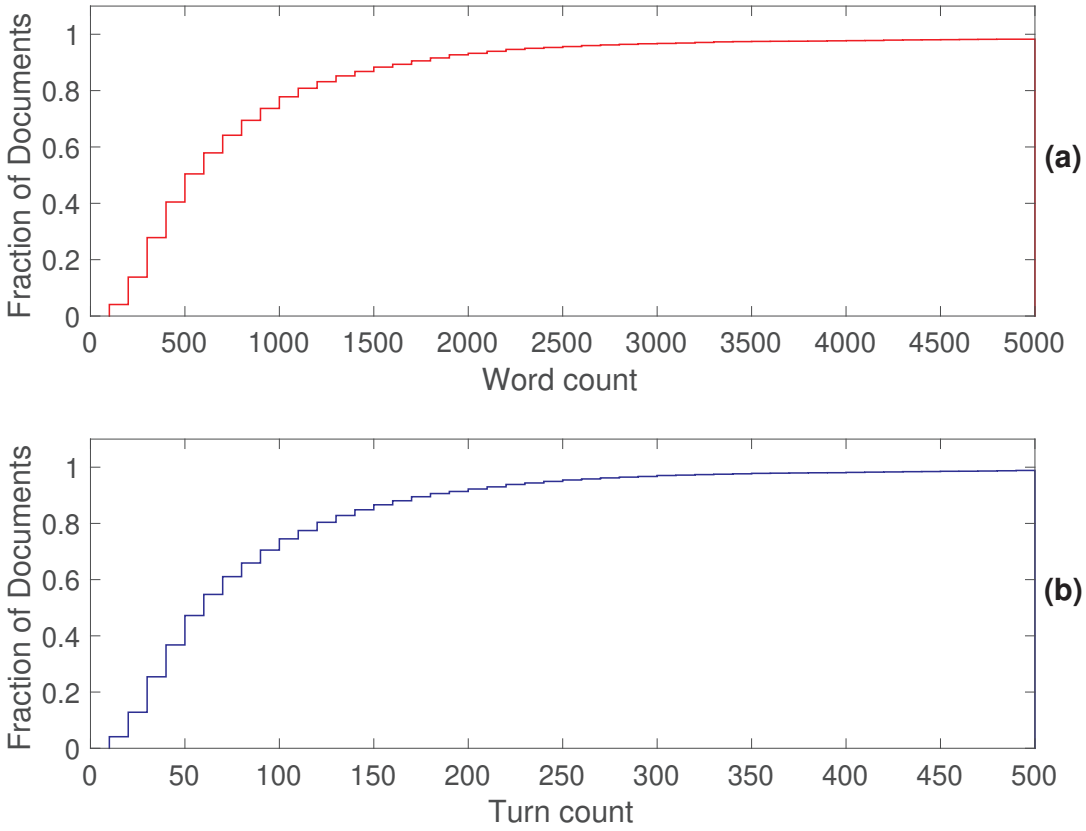


Figure 5.3: Cumulative distribution of document (a) word count (b) turn count

5.4 Feature extraction

In this section, we present the feature extraction methods used for our experiments. First, we give a brief introduction of pre-trained models used in this work to extract text features. Then, we explain acoustic feature extraction which includes the widely-used OpenSMILE features and the proposed x-vector embeddings.

5.4.1 Text feature extraction

It is shown in the literature that using pre-trained models for encoding words and sentences yields a significant improvement in the target task performance, especially when the target task has a small training dataset (Zhang and Wallace, 2015; Xu et al., 2016). In this work, we experimented with four types of pre-trained embedding methods:

- GloVe embeddings (Pennington, Socher, and Manning, 2014): GloVe (short form for Global Vector) is an algorithm based on the global word co-occurrence statistics which obtains semantically and syntactically meaningful representation for each word. We use GloVe vectors trained on the Common Crawl dataset [2] in our experiments.

- ELMo (Embeddings from Language Models) (Peters et al., 2018): This model obtains word embeddings based on context using two bidirectional language models. We use ELMo model trained on 1 Billion Word Benchmark [3].

[2] http://nlp.stanford.edu/data/wordvecs/glove.840B.300d.zip
[3] https://tfhub.dev/google/elmo/1

- Google's Universal Sentence Encoder (USE) (Cer et al., 2018): This model uses a deep averaging network to obtain embeddings for sentences, phrases, or even paragraphs. It is trained in multi-task style on a variety of datasets like Wikipedia, web news, and discussion forums[4].

- Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2018): This is a recently introduced state-of-the-art model for sentence representations shown to work well in several downstream tasks (Devlin et al., 2018). This model is based on a sequence of self-attention layers trained to predict next sentence and also minimize masked language model objective. It is trained on BooksCorpus and English Wikipedia datasets. In this work, we experiment with BERT model to improve the performance of CSAT prediction.

5.4.2 Acoustic Feature extraction

In this work, we explore two kinds of features: 1) OpenSMILE features from the acoustic signal and 2) Embeddings extracted from a pre-trained x-vector model which is trained to discriminate speaker identity (refer to Chapter 3 for details).

5.4.2.1 OpenSMILE features

OpenSMILE features are well studied for sentiment and emotion related tasks in the past (Chowdhury, Stepanov, Riccardi, et al., 2016; Luque et al., 2017; Cho et al., 2019; Eyben et al., 2015). In this paper, we follow the recommendations

[4]https://tfhub.dev/google/universal-sentence-encoder-lite/1

in (Eyben et al., 2015) to extract these features. First, we represent the entire speech signal as a sequence of short windows called frames. We used 1 second of speech as a frame length and we extract features for every 0.3 seconds. Then, each frame is represented with an 88-dimensional minimal set of features. These features quantify acoustic cues such as pitch, loudness and formant energies, among others.

5.4.2.2 x-Vector Embeddings

x-Vector model is a neural network trained to recognize speaker of a spoken utterance. Usually, this model contains a sequence of convolutional layers followed by a pooling layer and feed-forward layers to recognize speaker. Features extracted from the fully connected layers are called x-vectors. Since, the model is trained to discriminate speakers, we expect them to contain only speaker information. However, as shown in Chapter 3, x-vectors do encode other information such as speaking rate, channel information, spoken content and speaker emotion. In this work, we explore x-vectors for CSAT prediction.

In this paper, we use state-of-the-art ResNet x-vector model reported in (Villalba et al., 2019) for utterance level speaker embedding extraction. The network consisted of three parts: frame-level representation learning network, pooling network, and utterance-level classifier. Frame-level representation learning network uses ResNet-34 (He et al., 2016) structure, which consists of several 2D convolutional layers with short-cut connections between them. After that, we used a multi-head attention layer to summarize the whole utterance into a large embedding. This layer takes ResNet outputs $\mathbf{x}_t$ as input

and computes its own attention scores $w_{h,t}$ for each head h:

$$w_{h,t} = \frac{\exp(-s_h \|\mathbf{x}_t - \boldsymbol{\mu}_h\|)}{\sum_{t=1}^{T} \exp(-s_h \|\mathbf{x}_t - \boldsymbol{\mu}_h\|)} . \tag{5.1}$$

Attention scores $w_{h,t}$ are normalized along time axis.

Output embedding for head h is the weighted average over its inputs:

$$\mathbf{e}_h = \sum_t w_{h,t} \mathbf{x}_t \tag{5.2}$$

Different heads are designed to capture different aspects of input signal. Embedding from different heads are concatenated and projected by an affine transformation into the final embedding. From the pooling layer to output, there are two fully connected layers, and it predicts speaker identity in the training set. Angular softmax (Liu et al., 2017) loss was used to train the network. The whole network structure is illustrated in Table 5.2.

Datasets used to train this model include VoxCeleb, Switchboard, NIST SRE04-10, SRE12 and MIXER6. All the utterances were downsampled to 8 kHz. Our model was trained on mel-frequency cepstral coefficient (MFCC) features which are well-known in speech community. We extracted 23-dimensional MFCC features for every 10ms on a 25ms speech window. For more details, please refer to (Villalba et al., 2019). In this work, we extracted 400-dimensional features from the pre-final layer.

5.5 Methodology overview

Having introduced the feature extraction methods in the previous section, now we discuss how we use them to represent transcripts and acoustic signal

Component	Layer	Output Size
Frame-level Representation Learning	$7 \times 7, 16$	$T \times 23$
	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 3$	$T \times 23$
	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 4$, stride 2	$\frac{T}{2} \times 12$
	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 6$, stride 2	$\frac{T}{4} \times 6$
	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 3$, stride 2	$\frac{T}{8} \times 3$
	average pool 1×3	$\frac{T}{8}$
Pooling	32 heads attention	32×128
Utterance-level Classifier	FC	400
	FC	#spk:12,872

Table 5.2: ResNet architecture used in the x-vector model

for CSAT prediction. Overview of the feature representations investigated in this work is presented in Figure 5.4. We extract the acoustic/transcript features at multiple levels and turn-taking features from the segmentation information. Features from various levels have different kinds of information relevant to CSAT. For example, word-level features do not encode speaker turn information. While turn-level features encode speaker turns but loses the word-level granularity. In this section, we present the procedure to extract features at various levels followed by a brief overview of neural net based methods used in this work for CSAT prediction from the extracted features.

5.5.1 Transcript representations for CSAT models

ASR transcripts contain the information of who spoke when and what i.e., it contains the sequence of words spoken by speakers and their timing information. We considered representing a given document at four levels: a

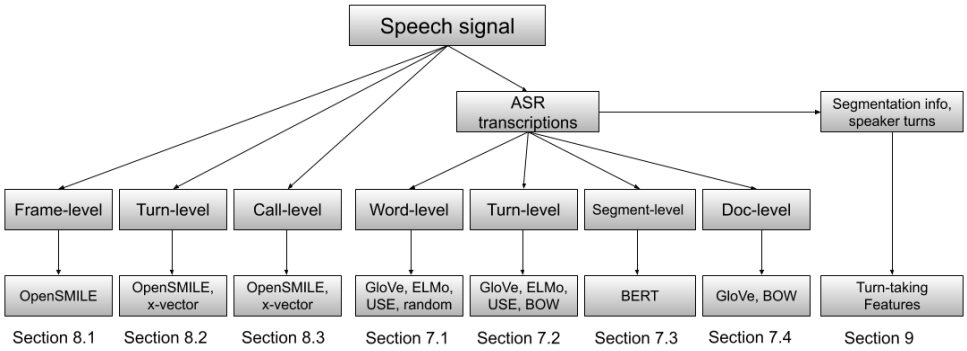


Figure 5.4: Overview of the feature representations

sequence of word vectors (word-level), a sequence of turn vectors (turn-level), a sequence of segment vectors (segment-level) and a single document vector (doc-level).

The word-level representation preserves the word order and ignores the speaker turns information. While the turn-level representation preserves the speaker turns order, it encodes only a summary of the words in that turn. Segment-level representation does not preserve word and turn order but we choose to explore it as it enables fine-tuning of the pre-trained models. The doc-level representation too does not preserve word and turn order, and is explored to understand if the sequence information is useful for the CSAT prediction.

- *At word level*: We represent every word with a single vector, which is usually referred as word embedding, thereby the transcript as a sequence of word embeddings. We can either learn the word embeddings for a specific task or use pre-trained word embeddings and then adapt for the task at hand. In this work, we use pre-trained word embeddings from models GloVe, ELMo and USE as they were shown to work well

for down-stream tasks. We compared the pre-trained word embeddings with the embeddings learnt from scratch. Our methods and experiments on the word level representations are presented in Section 5.7.1. Drawback of representing transcripts with word-level features is speaker turn information is not used which could be important for our task.

- *At turn level*: We represent each turn of a speaker with a single vector, thereby representing the transcript as a sequence of turns. In this representation, speaker turn information is preserved as the speakers take turns to converse each other. As with the word embeddings, we can use sentence/turn embedding from models such as ELMo, USE and BERT. As it was shown in (Adi et al., 2016), the turn embeddings can encode information about spoken content as well as other meta information such as word order and turn length which could be useful for our task. We compare the pre-trained turn embeddings with a simple word count based representation which does not involve pre-training. Our methods and experiments on turn-level representations are presented in Section 5.7.2.

- *At segment level*: We represent a small chunk of text, referred as a segment, with a single vector thereby representing the transcript as a sequence of segment vectors. In this work, we use pre-trained BERT model to represent segments. We also adapt/fine-tune the BERT model to our task by following a two-stage method to overcome limitations of BERT. Our methods and experiments on the segment-level representation are presented in Section 5.7.3.

- *At document level*: We represent the whole document with just a single vector as opposed to word-, turn-, segment-level representations. In this work, we used a simple BOW representation, which is a sum of one-hot encoding of words appeared in the document. It can be compared to the other representations investigated in this work as to whether the context is really important for a better CSAT prediction. Our methods and experiments on the document-level representation are presented in Section 5.7.4.

5.5.2 Acoustic representations for CSAT models

Acoustic signals contain the information of who spoke when and what along with how they spoke which is not available in ASR transcripts. We considered representing a given acoustic signal at three levels: a sequence of frame vectors (frame-level), a sequence of turn vectors (turn-level), and a single call vector (call-level). Similar to word-level representation of transcript, frame-level representation ignores speaker turns order. Turn-level representation preserves speaker turns order and it can encode only acoustic summary of the turn. Call-level representation does not preserve speaker turns order but we expect it to capture important events of the signal relevant to CSAT. Description of each representation level is as follows:

- *At frame level*: We represent a short window of speech, referred as a frame, with a single vector thereby representing the entire call as a sequence of frame vectors. We used 1 second of speech as a frame length and we extract features for every 0.3 seconds. Each frame is represented with an

88-dimensional minimal set of OpenSMILE features. Our methods and experiments on the acoustic frame-level representation are presented in Section 5.8.1. Similar to word-level representations, the main drawback of using this representation is speaker turn information is not used which could be important for our task.

- *At turn level*: Similar to the turn representations on text, we represent each turn of a speaker with a single vector to represent the whole call as a sequence of turn vectors. As with the frame-level representation, we used OpenSMILE features to represent turns. In addition, we propose to use x-vector embeddings to represent turns. We expect that emotion information encoded in x-vectors is helpful to find the overall sentiment of the call. Our methods and experiments on the acoustic turn-level representation are presented in Section 5.8.2.

- *At call level*: We represent the whole call with just a single vector as opposed to frame-, turn-level representations. Experiments with call-level representations reveal if we can ignore granular emotional changes along the call. In this work, we propose to use x-vector embeddings to represent the whole call. We expect the x-vector embeddings to retain the overall emotion in the call thereby predicting CSAT better. For comparison, we experiment with OpenSMILE features. Our methods and experiments on the call-level representation are presented in Section 5.8.3.

5.5.3 Turn-taking features for CSAT prediction

In addition to lexical/transcript and acoustic features, we explored using turn-taking features to predict CSAT rating. These features are extracted from the segmentation and speaker turns information in the ASR transcripts. They encode information related to duration of the call, speaking rate, overlapped speech duration etc.. Actual spoken content and acoustic cues are not used in these features and hence, these features could be complimentary to lexical and acoustic features. Our methods and results are presented in Section 5.9.

5.5.4 CSAT modeling from transcripts, acoustic signal, and turn-taking features

In the previous subsections, we presented feature extraction procedures we followed for CSAT. Now, we present a brief overview of machine learning models used on the extracted features. To predict CSAT rating from transcript, acoustic and turn-taking representations, we explored neural network based methods along with linear classifiers such as logistic regression and XGBoost. We used CNN based models on the sequential representations such as transcript word-, turn-level features and acoustic frame-, turn-level features. We found that LSTM based models did not work well as the input sequences are very long and hence we do not report them in this work. On document/call level representations, we explored using DNN, linear classifiers such as logistic regression and XGBoost.

5.5.4.1 Channel-aware CSAT models

In CCC calls, usually an agent and customer converse each other and, both of their speech/text may help us to predict CSAT. However, they may not have equal role as their speaking characteristics are different. In other words, agent is well-trained to speak to the customer while the latter is not trained at all. For example, usually the agent speaks formally while the customer need not be. Hence, we hypothesize that distinguishing agent's and customer's stream/channel of speech during model training will help to predict the CSAT better. In this work, we present the experiments showing the importance of distinguishing the agent's and customer's channel.

5.6 Experimental Setup

In this work, we divided our dataset of 4331 calls into 3 sub-sets: 2866, 362 and 1103 calls for training, validation and testing respectively, retaining the class balance in each set. We report the macro f1-score in all of our experiments. F1-score is computed as the harmonic average of precision and recall. After model training is finished, we chose an epoch with the best validation loss to report results on the test set. *Adam* optimizer (Kingma and Ba, 2014) was used to optimize the cross-entropy loss. Initial learning rate was set to 0.001 and reduced by a factor of 0.95 if validation loss did not decrease for 3-epochs. Randomness in neural network training causes some f1-score variance between runs of the same model. To alleviate this effect and make a better comparison between models, we report the f1-scores averaged over 5 runs.

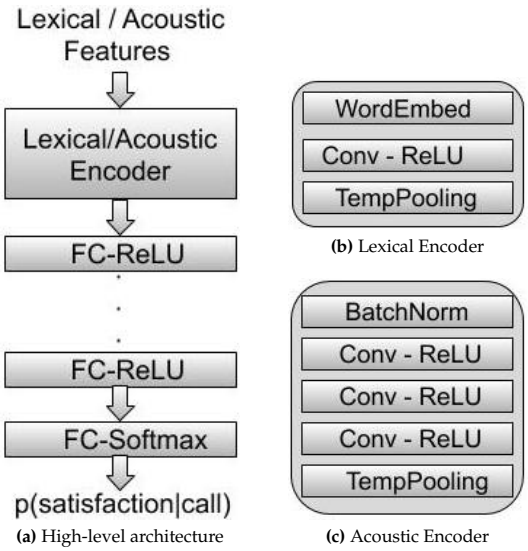
In the following sections, we present our methods and experiments on lexical, acoustic and turn-talking features. We also show experiments on fusing all these features to exploit complementary information. Then, we present analysis experiments on the importance of each speaker (agent vs. customer) and the need for more data.

5.7 CSAT on ASR transcriptions

Transcript representation plays a vital role in obtaining task-relevant information from transcripts. As explained in Section 5.5.1, we represent the transcript at four levels: word-, turn-, segment- and doc-level. In this section, we present details of models used on the transcript representations and corresponding results.

5.7.1 Modeling word-level transcript representations

Figure 5.5a shows the neural architecture of our model. We used one-hot encoding representation to represent each word. We obtained word embeddings by passing each word one hot encoding through a linear layer, referred to as the word embedding layer. We denoted this layer with *WordEmbed* in Figure 5.5b. In this work, we experimented with initializing *WordEmbed* layer with pre-trained embeddings and also compared with random initialization. The complete list of embeddings used in this work are presented in Table 5.3. The output of *WordEmbed* layer is passed through one convolutional layer with 100 filter maps and a kernel size of 7 followed by a global temporal pooling layer. The pooling layer's output is a single 100-dim vector representing



(a) High-level architecture **(b)** Lexical Encoder **(c)** Acoustic Encoder

Figure 5.5: Architecture of lexical and acoustic models. FC-ReLU and FC-Softmax: fully connected layer with ReLU and softmax activation, *WordEmbed*: word embedding, Conv-ReLU: convolution layer with ReLU activation, BatchNorm: batchnorm layer, TempPooling: average temporal pooling

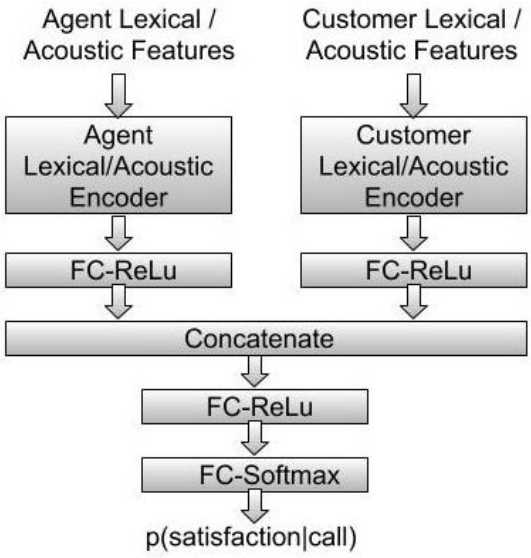


Figure 5.6: Channel-aware architecture for lexical and acoustic signals

Representation	Description
GloVe Word Embed	pre-trained GloVe word embeddings (300-dim)
ELMo Word Embed	pre-trained ELMo word embeddings (1024-dim)
USE Word Embed	pre-trained USE word embeddings (512-dim)
Random Word Embed	Random values sampled from uniform distribution (300-dim)

Table 5.3: Description of word embeddings

the whole document, which is referred as transcript embedding. We further refer to the sub-module used to obtain the transcript embedding as the lexical encoder (see Figure 5.5b). Then, the transcript embedding was processed with 30-dim and 2-dim fully connected layers with activation functions ReLU and softmax respectively to obtain the CSAT prediction. We applied dropout with 30% rate on the document embedding and fully connected layer to regularize the model.

Table 5.4 presents comparison of the results with different word embedding initializations. It can be observed that pre-trained word embeddings provided better performance compared to random initialization suggesting the importance of using pre-trained models. Among the pre-trained embeddings, GloVe embeddings performed best with 77.52% f1-score.

Table 5.4: Comparison of various word embedding initializations in CNN architecture

Word embeddings type	F1-score (%)
GloVe Word Embed	**77.52**
ELMo Word Embed	77.29
USE Word Embed	76.37
Random Word Embed	67.23

Channel-aware CSAT model: As discussed in Section 5.5.4.1, we experiment with distinguishing agent and customer during training. Specifically, we

Table 5.5: Comparison of channel-aware and channel-unaware models on ASR transcription

Type of input	F1-score
GloVe Word Embed (Figure 5.5a)	77.52
GloVe Word Embed Channel-Aware (Figure 5.6)	79.70

processed agent and customer transcript separately using lexical encoder to obtain agent transcript embedding and customer transcript embedding. Then we concatenate them to obtain one transcript embedding for the whole call, which will then be processed with two fully connected layers as noted above. Architecture for this experiment is shown in Figure 5.6. Table 5.5 compares the model performance for the case where the input is a single stream of customer and agent words (*GloVe Word Embed*); and the case where we separate the customer and agent word streams (*GloVe Word Embed Channel-Aware*). We noticed 2.18% absolute improvement in f1-score which could be attributed to style differences between speakers in expressing their sentiment.

5.7.2 Modeling turn-level transcript representations

For turn-level representations, we treated each turn as a sentence and used pre-trained sentence encoders to represent each turn. We experimented with three types of pre-trained sentence encoders: USE, ELMo and BERT. Apart from representations based on pre-trained sentence encoders, we used word counts of each turn (sum of one-hot encodings) as another representation. Table 5.6 presents the complete list of turn representations explored in this work.

For modeling the turn-level transcript representations we adopted the

Representation	Description
Turn-USE	Turn representation using USE
Turn-ELMo	Turn representation using ELMo
Turn-BERT	Turn Representation using BERT
Turn-BOW	Sum of one-hot encodings of words in the turn

Table 5.6: Description of Turn representations

same model architecture used in Section 5.7.1 (Figure 5.5a) but without the *WordEmbed* layer in the lexical encoder Figure 5.5b.

Channel-aware CSAT model: Boundaries of agent speech and customer speech are implicit in the turn representation. Hence, we did not perform separate experiments distinguishing agent and customer to predict CSAT rating.

Experiments with turn-level representations are shown in Table 5.7. Among the pre-trained turn representations, we observed that *Turn-USE* performed better than *Turn-ELMo* and *Turn-BERT*. Surprisingly, *Turn-BOW* outperformed all other turn representations with an f1-score of 79.86%. It could be because *Turn-BOW* representation encodes turn information in a simplistic manner (i.e., word counts) which our model is able to exploit well. From the Table 5.5 and 5.7, it can be observed that both the word-level and turn-level representations perform similarly.

5.7.3 Modeling sequence of segment representations

A long document can be represented as a sequence of segments. Segment representation can be seen as a compromise between turn representations and using a single vector for the whole document. Call center transcripts, while

Table 5.7: Comparison of sentence encoders for prediction based on turn level embeddings.

Input representation	F1-score
Turn-USE	78.41
Turn-ELMo	73.88
Turn-BERT	75.80
Turn-BOW	**79.86**

usually quite short and to the point, often involve agents trying to solve very complex issues that the customers experience, resulting in some calls taking even an hour or more. These transcripts sometimes exceed the length of 5000 words.

For this task, we use our previously introduced method that builds upon BERT's architecture (Pappagari et al., 2019). One important limitation of BERT model is it operates on only a limited context of symbols as their input (Dai et al., 2019). Hence, we split the input text sequence into shorter segments in order to obtain a representation for each of them using BERT. Then, we use a recurrent LSTM (Hochreiter and Schmidhuber, 1997) network to perform the actual classification. We call our technique as Recurrence over BERT (RoBERT).

Given a pre-trained BERT model, we can obtain features for a segment in several ways:

- Freeze the pre-trained BERT model and extract the features from the pre-final layer. We denote this with *BERT frozen features*.
- Fine-tune the pre-trained BERT model to CSAT data on segment-level

Table 5.8: Comparison of various BERT feature representations

Type of input \ Model	RoBERT
BERT frozen features	70.19
BERT fine-tuned features	83.33
BERT fine-tuned predictions	**83.38**

and extract the features for each segment from the pre-final layer. We denote this with *BERT fine-tuned features*.

- Fine-tune the pre-trained BERT model to CSAT data on segment-level and extract the predictions for each segment from the final layer. We denote this with *BERT fine-tuned predictions*

Results with RoBERT model on the extracted segment-level features are presented in Table 5.8. It can be observed that fine-tuned features/predictions outperform frozen features suggesting the importance of adaptation to our task. We obtained a best f1-score 83.38% using RoBERT model on fine-tuned predictions.

It is also possible to predict CSAT rating from *BERT fine-tuned predictions* by just taking their average or finding most-frequent class. In this work, we compare efficacy of these simple methods with RoBERT model. Table 5.9 presents the comparison of RoBERT with simple operations like averaging and most-frequent class. It can be observed that RoBERT performed better than most-frequent and averaging operations suggesting that temporal information is important for CSAT task. One example where temporal information useful is sometimes the customer is angry at the beginning of the call but if the agent

Table 5.9: Comparison of classification methods on the fine-tuned predictions

Average	Most-Frequent	RoBERT
82.18	80.25	83.38

addresses the customer's problem successfully then customer is happy at the end of the call.

5.7.4 Modeling document-level representation

BOW representation is obtained by taking summation over all one-hot encoding vectors in the temporal dimension, i.e., it is a vector with the counts of each word in that transcript. We denote BOW representation for the transcript/document with Doc-BOW. Since we take the summation over the temporal dimension, we lose relevant sequence information. Note that, the same vector can be used to represent two transcripts with the same words but in a different order. Instead of using the word counts to represent the document, we can also average word embeddings. For example, we can average GloVe embeddings of all words in the corresponding document – we denote this as *Avg-Doc-GloVe*.

As there is no temporal information in *Doc-BOW* and *Avg-Doc-GloVe* representations, it does not make sense to use CNN network models on them. Thus, we just used standard feed-forward DNN to predict the CSAT rating. We passed the *Doc-BOW*/*Avg-Doc-GloVe* vectors through a DNN with three hidden layers with 1000, 300, and 30 neurons with ReLu activations. The network configuration was selected empirically, though results were not very sensitive to the number or width of the layers. We compare this model with

shallow models such as logistic regression and XGBoost.

Channel-aware CSAT model: As discussed in Section 5.5.4.1, we hypothesize that distinguishing agent and customer during training helps to predict CSAT better. Specifically, we obtain BOW vector for agent and customer separately and then concatenate to represent the whole transcript. We denote this representation with *Doc-BOW channel-aware*.

Results with document-level representations are presented in Table 5.10. It can be observed that *Doc-BOW* performed better than *Doc-BOW channel-aware* with DNN model. This result suggests that distinguishing the speakers is not helpful which is opposite to our observation with word-level representation (Table 5.5). But, with linear classifiers LR and XGBoost, they perform similarly. *Avg-Doc-GloVe* performed significantly worse which could be because sentiment-related attributes can not be emphasized in the averaging operation of GloVe embeddings of words where the average is over all kinds of words. In other words, the overall sentiment is not just the summation of sentiment attribute of each word. From the comparison of Table 5.5 and 5.10, it can be observed that *Doc-BOW* representation is better than *GloVe Word Embed* but inferior compared to *GloVe Word Embed Channel-Aware*. It could be because CNN could not exploit sequential information when speech from multiple speakers is mixed. This result also suggests that semantics can be helpful for this task.

Table 5.10: Results with document-level representations. All numbers in this table are f1-scores (%)

Type of input \ Model	LR	XGBoost	DNN
Doc-BOW	73.80	78.35	78.41
Doc-BOW channel-aware	73.72	78.35	77.00
Avg-Doc-GloVe	64.20	66.70	66.62

5.8 CSAT on Acoustic signal

Models using ASR transcription utilize mainly spoken content, while the important acoustic cues such as change in tone, pitch and loudness of speech are not available. Hence, modeling acoustic signal can provide complimentary information to ASR transcriptions. In this section, we present methods based on acoustic signal to predict CSAT rating.

5.8.1 Modeling frame-level acoustic representations

As word segmentation is very difficult in speech signal, we represent speech signal using a sequence of short windows called as frames. In this work, we extract frames with 1 second windows sequentially shifted by 0.3 second. Each frame is represented with an 88-dimensional feature vector extracted using OpenSMILE tool.

We process frame-level acoustic call representation using a dilated 1D CNN model. Description of the model is as follows. Batch normalization layer is used as the first layer in model, as it has the implicit effect of normalizing the acoustic features in mean and variance. Then, the normalized input was passed through a sequence of 3 convolutional layers with dilation rate 1, 2 and

Table 5.11: Results with acoustic frame-based representation

Type of input	F1-score
frame-OpenSMILE (Figure 5.5a)	73.63
frame-OpenSMILE channel-aware (Figure 5.6)	76.10

3 respectively. We used 50 filter maps and kernel size 7 in each convolutional layer, leading to an effective context of 37 at the end of final convolutional layer. The output of the final convolutional layer was averaged temporally to obtain a single 50-dim vector representation for the whole call, referred to as acoustic embedding. We further refer to the submodule used to obtain an acoustic embedding from the acoustic features as the acoustic encoder (see Figure 5.5c). Then, the acoustic embedding was passed through 100-dim and 2-dim fully connected layers with ReLU and softmax activation functions respectively.

Channel-aware CSAT model: As discussed in Section 5.5.4.1, we hypothesize that distinguishing agent's and customer's speech during training helps. To validate our hypothesis, we process frame-level acoustic features for agent's and customer's speech separately to obtain their corresponding acoustic embeddings. Then, we concatenate them to process further to obtain CSAT prediction for the whole call. Architecture for this experiment is shown in Figure 5.6. We denote this model with *frame-OpenSMILE channel-aware*.

Results with frame-level acoustic features are presented in Table 5.11. We observed 2.47% absolute improvement in f1-score by distinguishing agent and customer speech, which can be attributed to differences in speaker characteristics related to expressing various emotions. Compared to word-level

transcript representation (Table 5.5), it performed worse by 3.6% indicating that spoken content has more information.

5.8.2 Modeling turn-level acoustic representations

Since, the goal of CSAT prediction task is predicting overall satisfaction of the customer, it is important to understand if the local variations affect final performance. In this section, we present experiments with acoustic turn representations. For each turn, we extracted one 88-dim vector of OpenSMILE features. We believe the overall acoustic turn statistics are important for this task compared to frame-level statistics as we usually can expect only one dominant emotion in each turn. In addition to OpenSMILE features, we propose to use x-vector embedding for each turn which is 400 dimensional. We believe the x-vector model attends to most informative parts of the turn.

For modeling the turn-level acoustic representations, we used same CNN model explained in Section 5.8.1 and corresponding results are presented in Table 5.12. It can be observed that OpenSMILE features and x-vectors perform similarly but worse than frame-level representation (Table 5.11). This result suggest that frame-level acoustic variations are important for this task.

Channel-aware CSAT model: Boundaries of agent speech and customer speech are implicit in the turn representation. Hence, we did not perform separate experiments distinguishing agent and customer to predict CSAT rating.

Table 5.12: Results with turn-based representations on acoustic signal

Type of input	F1-score
Turn-OpenSMILE	67.43
Turn-x-vectors	67.42

5.8.3 Modeling call-level acoustic representation

Similar to doc-level representation for transcripts, we represent entire call with a vector. In this work, we propose to use x-vector embedding to represent long audio calls for better CSAT prediction performance. We expect that the x-vector model does encode information relevant to sentiment of the speakers. We compare x-vector embedding with call-level OpenSMILE feature representation. Note that, it is not very common to use OpenSMILE tool to represent long calls with a single vector, as it mainly extracts statistics of acoustic features. Hence, we expect it to not work well compared to call-level x-vector embedding and also frame-level OpenSMILE representation. We used DNN for classification and compared with linear classifiers such as LR, XGBoost.

Table 5.13 presents results on call-level representations. It can be observed that x-vector representation perform significantly better than OpenSMILE features. Also, as we expected, call-level OpenSMILE features perform worse than frame-level features suggesting that important events are not emphasized in the call-level representation. Also, DNN outperformed linear classifiers, LR and XGBoost, in most cases.

Channel-aware CSAT model: As discussed in Section 5.5.4.1, we hypothesize that distinguishing agent and customer speech during training helps. To

Table 5.13: Results with call-level representations on acoustic signal

Type of input \ Model	LR	XGBoost	DNN
Call-OpenSMILE	67.95	68.99	70.36
Call-x-vectors	72.78	76.04	76.87
Call-OpenSMILE channel-aware	73.04	74.4	71.87
Call-x-vectors channel-aware	74.73	74.98	**78.21**

validate this hypothesis, we extract call-level acoustic features for agent and customer separately and concatenate them to represent the whole call. We denote these representations with *channel-aware*. It can be observed from Table 5.13 that in most cases *channel-aware* representations (rows 3 and 4) perform better than the representations without channel information (rows 1 and 2) suggesting the validity of our hypothesis. From the comparison of Table 5.11 and 5.13, it can be observed that call-level x-vectors with channel information performed best with 78.21% f1-score.

5.9 CSAT using turn-taking features

We have shown how to construct a CSAT model based on the verbal content and acoustic cues of the CCC call. However, human conversations are abundant with a variety of nonverbal cues, which are helpful to understand the dynamics of the dialogue. In this work, we attempt to capture some of these cues with turn-taking features to investigate their usefulness in CSAT prediction. We extracted 18 features and classified them into four types: dialogue efficiency metrics, dialogue quality metrics, task success/completion metrics (Chowdhury, Stepanov, Riccardi, et al., 2016; Park and Gates, 2009;

Walker, Passonneau, and Boland, 2001; Walker, Hirschman, and Aberdeen, 2000), terse dialogue metrics.

The dialog efficiency metrics were the duration of the call, the average number of speaker (agent and customer) words per turn, average speaker (agent and customer) turn duration, average speaker (agent and customer) talking rate and speaker (agent and customer) call dominance. Speaker call dominance is measured as a fraction of the speaking time of the speaker. We only considered customer call dominance as agent call dominance is derived as one minus customer call dominance.

For dialogue quality metrics, we considered the number of overlaps between customer and agent turns. For task success metrics, we used the task completion since satisfied customers often have their problem solved, as shown in Figure 5.2. For our dataset, this metric is available.

In addition to these metrics, which are adapted from literature, we propose eight other features in this work that are based on the number of words used in each speaker's turn. We call these features *terse dialogue metrics*, as they reflect conciseness and sometimes forceful expressiveness of the responses. For a smooth dialogue, cooperation between participants is essential, especially in the context of CCC calls. Depending on other factors, high density of short turns may suggest either conciseness and pragmatism of the dialog or lack of engagement of one or both of the interlocutors. In this work, we used the number of one-word, and two-word turns for each speaker and their average time duration as features. We define one-word turns as the turns in which the speaker utters only one word. Similarly, two-word turns are the turns with

exactly two words.

We used logistic regression (LR) to predict the satisfaction level of the customer using these features. We did not notice any significant improvement over LR when using DNN, SVM, or gradient boosted decision tree classifiers; hence, we preferred LR for its simplicity.

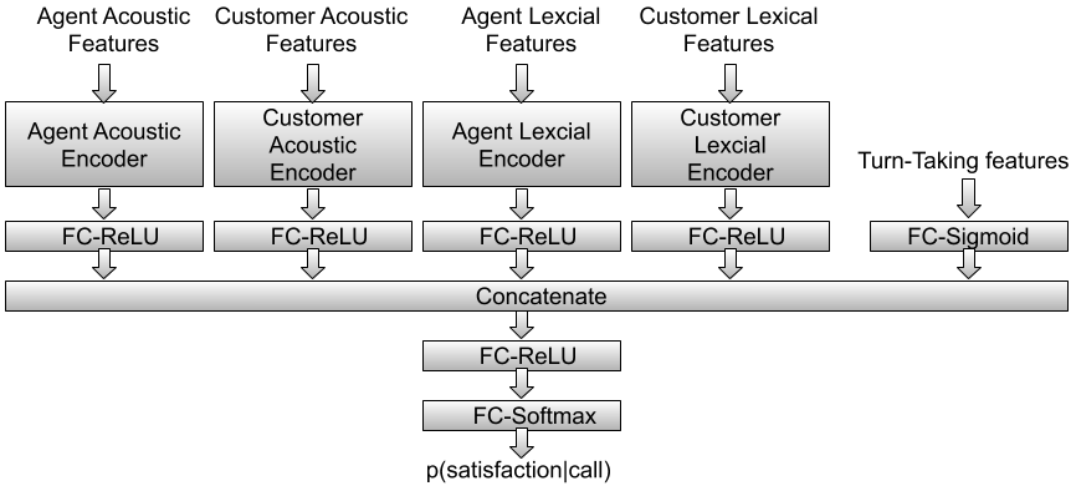
Results with with the *turn-taking features* are presented in Table 5.14. For simplicity, we denoted dialogue efficiency metrics with E, dialogue quality metrics with Q, *terse dialogue* metrics with T and task completion with TC. It can be observed that the turn-taking features yielded some information about the CSAT with an f1-score ranging between 66% and 67%. We also examined the improvement obtained from using the proposed *terse dialogue* metrics, based on the number of words used in each speaker's turn. Using the *terse dialogue* metrics, we obtained 1% improvement with and 0.5% without task completion metric as shown in Table 5.14. However, we observed that the improvements are not statistically significant.

It can be observed from the Table 5.14 that task completion metric alone contributes to 76.79% f1-score suggesting that it is highly correlated to CSAT rating. Figure 5.8(a) shows the distribution of positive and negative calls in the test dataset with respect to task completion metric. Almost all the satisfied customers/positive calls (599 calls) have successful task completion and there is an almost equal split between successfully (234 calls) and unsuccessfully (248 calls) completed calls when the customer was dissatisfied (negative calls). It can be implied that the task is successfully finished whenever the call is rated as positive. On the other hand, the converse is not valid, suggesting that

Table 5.14: F1-scores obtained with *turn-taking features*. EQ stands for features with only Efficiency(E) and Quality(Q) dialogue metrics and EQT is the same as EQ, but extend with *Terse dialogue*(T) features. TC stands for Task Completion

	EQ	EQT
Without TC	66.45	66.99
With TC	77.79	**78.78**
Only TC	76.79	

Figure 5.7: Architecture for fusion of lexical, acoustic and turn-taking features



the customer can be dissatisfied, even should the agent solve the customers' problem.

5.10 Fusion of lexical, acoustic and turn-taking features

As discussed in Section 5.5.1 and 5.7, text feature representations encode mainly the spoken content and the models based on them performed fairly well with a best f1-score of 83.38%. However, humans are capable of expressing multiple emotions such as angry, disgust, happiness etc.. for the

same content which can be seen only from the acoustic signal. Hence, we hypothesize that models based on transcripts and acoustic signal capture complementary information, and fusing them would enable us for more improvements. Also, as discussed in Section 5.9, turn-taking features are shown to work well for CSAT prediction albeit not as good as lexical and acoustic features. However, they capture meta information such as speaking rate, duration of the call etc.. which are not explicitly available in lexical and acoustic features. Hence, we hypothesize fusion of turn-taking features with both lexical and acoustic features helps us for better CSAT performance.

In this work, we present two methods of fusion: 1) model-based fusion and 2) score-based fusion. In model-based fusion, we train a single model which exploits multiple representations of a given call (lexical, acoustic and turn-taking features) and produces a single CSAT prediction. Figure 5.7 presents the architecture for model-based fusion. First, we obtain transcript and acoustic embeddings for both agent and customer, and pass them through a 100-dim fully connected layer separately. Alongside, turn-taking features are treated with 100-dim fully connected layer. Now, we have five embeddings at this point: 2 for agent (transcript and acoustic), 2 for customer (transcript and acoustic) and 1 from turn-taking features. Then, we fuse these five embeddings using two fully connected layers with 50-dim and 2-dim with ReLU and softmax activation functions respectively to obtain the CSAT prediction. Parameters for the multi-modal system were set based on the best transcript and acoustic models.

For score-based fusion, we follow a two-stage approach: we first extracted

the log-likelihoods from already trained transcript, acoustic and turn-taking features based models. Then, we concatenate the log-likelihoods to obtain a 6-dim vector and applied logistic regression to obtain CSAT prediction.

Table 5.15 presents results for model-based and score-based fusion. Fusion of transcript (word-level) and acoustic (frame-level) sources gave 2.17% improvement compared to transcript source alone (*GloVe Word Embed Channel-Aware*). Adding turn-taking features to our multi-modal system as an extra input degraded the result significantly. We observed that score fusion performed 2.45% better than model-based fusion . We expected model-based fusion to work better than score-based fusion as the former is trained to optimize in an end-to-end manner. As score-fusion performed better than model-fusion, we used score-fusion for further fusion experiments. Fusion with *Turn-BOW*, *Call-x-vectors channel-aware*, RoBERT and EQT provided more gains. We can observe that EQT features contribution is small in the score-fusion suggesting no complementary information in turn-taking features. In other words, the turn-taking features could have already been encoded in transcript and acoustic representations. For example, duration of the call can be roughly estimated from the number of words in the transcript and number of frames in the acoustic signal as they are directly proportional. We obtained the best f1-score of 88.35% with the score fusion of the data-driven features i.e, transcript features at word-, turn-, segment-level (RoBERT), and acoustic features at frame- and call-level. With the addition of hand-crafted features (EQT) we obtained an f1-score of 88.46%. Also, adding TC metric provided minimal gains suggesting we do not need it given data-driven models which

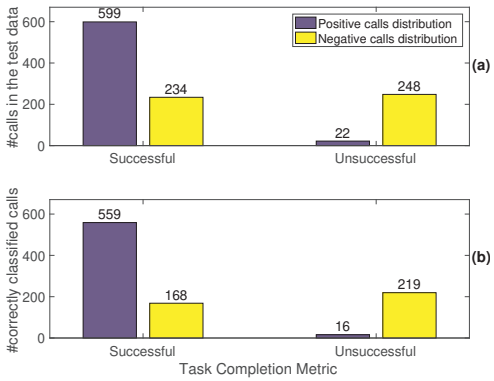


Figure 5.8: (a) Histogram of number of calls in the test data w.r.t TC metric, (b) Histogram of number of correctly classified calls in the score-fusion system (word-, turn-, frame-, call-level, RoBERT, EQT)

is a good sign as it is self-reported by customers.

In this work, we attempted to analyze fusion model decisions with respect to task completion metric. Figure 5.8(a) shows the histogram of positive and negative calls in the test dataset with respect to TC metric where as Figure 5.8(b) shows the histogram of only correctly classified calls in score-fusion system. It can be observed that, our fusion model is able to correctly classify most of the positive calls with successful task completion (6.7% error, 559 out of 599) and negative calls with unsuccessful task completion (11.7% error, 219 out of 248). However, our model is having difficulty in correctly classifying the negative calls with successful task completion (28.2% error, 168 out of 234). It is possible that these calls might have subtle cues compared to the calls corresponding to other cases and we might need more efficient models and more information to classify them correctly.

Table 5.15: Comparison of stand alone systems and fusion systems. For model fusion, word- and frame-level features are used and trained using Figure 5.7 architecture.

	F1-score
GloVe Word Embed Channel-Aware	79.70
Turn-BOW	79.86
RoBERT (Segment-level)	83.38
frame-OpenSMILE channel-aware	76.10
Call-x-vectors channel-aware	78.21
EQT	66.99
TC	76.79
EQT + TC	78.78
Model Fusion (Word-, frame-level)	81.87
Model Fusion (Word-, frame-level, EQT)	77.24
Score Fusion (word-, frame-level)	84.32
Score Fusion (word-, turn-, frame-, call-level)	86.69
Score Fusion (word-, turn-, frame-, call-level, RoBERT)	88.35
Score Fusion (word-, turn-, frame-, call-level, RoBERT, EQT)	**88.46**
Score Fusion (word-, turn-, frame-, call-level, RoBERT, EQT+TC)	**88.55**

5.11 Analysis

5.11.1 Learning Curves

Learning curves reveal the effect of dataset size used for training on the CSAT performance. Since our dataset is relatively small compared to usual machine learning tasks, these learning curves help us to assess the need for more data. We experimented with dataset sizes starting from using 10% of the data in increments of 10% up to full dataset. For this experiment, we used ASR transcriptions. We randomly sample the corresponding proportion of the training data from each class for these experiments while maintaining validation and test data splits fixed. Dataset proportion is plotted in logarithmic scale w.r.t test data f1-score in Figure 5.9. We used CNN on the sequence of word vectors for this analysis. A rapid increase in performance up to 50% of the data can be observed from Figure 5.9. With more than 50% of the data, the rate of improvement in performance is relatively less, but it improved with more data.

Also, we observed that for the dataset proportions 0-50%, the performance differences between *GloVe Word Embed*, *ELMo Word Embed* and *USE Word Embed* are not statistically significant at 90% significance level. We used student's one-sample t-test for statistical tests. The evidence from Figure 5.9 suggests that with more data for this kind of task, the technology can be more accurate.

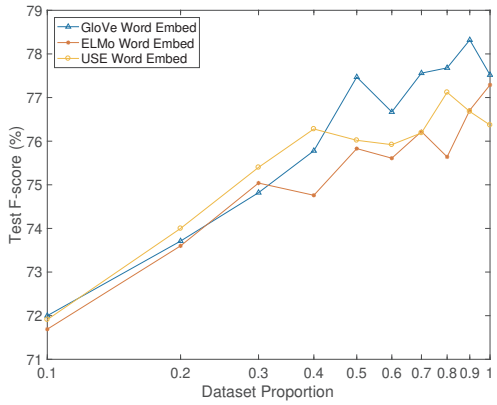


Figure 5.9: Effect of dataset size with learning curves.

5.11.2 Whose data, agent's or customer's, is more important for CSAT prediction?

As the customers were asked to rate their experience with the agents during the call, the rating depends on how the agent speaks, which in turn (at least partly) depends on the customer. Therefore, it is useful to quantify the importance of agent's and customer's speech to the satisfaction level of a customer. It is also important for practical applications like employee management and employee skill assessment. In order to assess each speaker importance, we separated agent's speech from the customer's. Then, we considered different training and testing scenarios with agent and customer acoustic features and ASR transcripts.

Experiments with transcript and acoustic models are shown in Tables 5.16 and 5.17 respectively. Good performance can be observed with the transcripts when training and testing scenarios match and diminished performance for mismatched scenarios as shown in Table 5.16 suggesting that agents and customers could be using a different set of expressions to express the same

Table 5.16: Comparison of f1-scores in different training and testing scenarios. We used ASR transcriptions for these experiments. *GloVe Word Embed* denotes original transcript of the call which does not differentiate between agent and customer.

Train \ Test	Agent	Customer	*GloVe Word Embed*
Agent	76.48	40.65	75.68
Customer	62.52	76	70.92
GloVe Word Embed	75.83	62.25	77.52

Table 5.17: Comparison of f1-scores in different training and testing scenarios. We used acoustic signal for these experiments. *frame-OpenSMILE* denotes original acoustic features of the call which do not differentiate between agent and customer.

Train \ Test	Agent	Customer	*frame-OpenSMILE*
Agent	66.26	51.54	55.38
Customer	48.39	75.76	61.36
frame-OpenSMILE	44.41	45.85	73.63

sentiment. Similar performance for agent vs. agent and customer vs. customer implies that both agent and customer's verbal content is equally important to the satisfaction level of the customer. Best performance was observed when we use both the agent and the customer transcripts (denoted with *GloVe Word Embed*).

Experiments with acoustic features in Table 5.17 lead to a different set of conclusions. We observed a good performance only when we are training and testing with customers acoustic features. Poor performance when training and testing with agent acoustic features suggest that the agents' acoustic cues do not have much information about CSAT score unlike their transcripts. Also,

Table 5.18: Classifying agent vs customer from ASR transcription and acoustic signal

Type of input	F1-score
ASR transcription	99.46
Acoustic signal	99

drop in the performance can be observed with *frame-OpenSMILE* suggesting that model could not differentiate agent and customer styles automatically. Conclusions drawn from Table 5.17 imply that differentiating between the agent and the customer is important and can also be observed same from Table 5.11 where we obtained 76.10% f1-score, denoted with *frame-OpenSMILE channel-aware*, by fusing channels.

From the comparison of both, Table 5.16 and 5.17, we can say that on average, a customer's acoustic cues and spoken content yield similar information regarding his own satisfaction. However, agents acoustic cues and spoken content do not seem to have similar CSAT relevant information, which could be because of their skill to not express too many emotions with the customers. Based on these observations, we speculate that customer speech is more emotional than agents speech.

Having observed the weak performance when the training and testing in mis-matched scenarios, we speculated that distributions of agent and customer data are different. Now, to strengthen this speculation, we provide evidence by classifying agent and customer based on their transcript and speech using XGBoost model. Table 5.18 presents results for classification of agent and customer from transcript and acoustic signal. It can be observed that the f1-score is more than 99% for both transcript and acoustic signals suggesting

that distributions of agent and customer data are indeed different.

5.12 Ethical Considerations

The models described in this paper have been developed to automate the customer satisfaction evaluation of a conversation with a call center agent. Therefore, an application of this model results in an automated rating of human performance. It is important to keep in mind that the proposed methods are based on statistical models and have a margin for error. Unsupervised reliance on the output of such models would likely result in an unfair assessment of call center agents performance. As these are black box models, the interpretation, or explanation, of a particular model decision is a challenging task, that is itself an active field of research. These issues would be further exacerbated when such models are applied in a domain with different characteristics than the training data set. One example is when the agents have a different ethnicity, unknown to the model - their accent and speaking manner might pose a confusing factor for the model. Another example is when the kinds of problems discussed during the call are different (as in: calls about refunds vs. calls about healthcare advice). In this scenario, the vocabulary between these domains is different, as is the typical flow of the conversation (or *conversation etiquette*).

To minimize these issues, we advise to use these models in a semi-automated manner. An example of such use could be ranking the calls given the scores from model predictions and pushing the calls predicted as unsatisfactory into a queue for human review. Such an application is of high practical importance,

as it enhances the process of manual review by identifying the calls with highest risk factor, thus saving reviewers' time. Should the model be used in a fully automated process, we advise to implement a mechanism for the agent to appeal to the models' decision before he/she suffers any consequences.

5.13 Conclusion and Future Work

In this paper, we have investigated several lexical and acoustic feature representations for CSAT prediction. We explored four-levels of representations for transcript and three-levels for acoustic signal at word/frame-level, turn-level, segment-level and call/document-level. In this work, we proposed using x-vectors for CSAT task and obtained best results on the acoustic signal. For transcript, word-, turn-level features performed competitively and for acoustic features, call-level x-vector representation worked best. We found that exploiting semantics is helpful for CSAT task. Features extracted from pre-trained encoders ELMo, USE and BERT for transcripts did not provide any improvements. However, fine-tuning BERT model provided significant gains. We observed that models perform better by distinguishing agent and customer channels. Overall, we observed best f1-score of 83.38% on ASR transcription using BERT and 78.21% on acoustic signal using x-vector which are both using pre-trained models. Turn-taking features performed worst with an f1-score of 66.99%. With score-fusion of lexical, acoustic and turn-taking features, we obtained 88.46% f1-score.

Analysis of model decisions with respect to task completion metric revealed that negative calls with successful task completion are the primary

source of errors. Our analysis on the importance of each speaker data to the CSAT rating revealed that customers' acoustic and lexical content have comparable significance, whereas agents' acoustic cues are less significant than their lexical content. Also, we observed that agents' and customers' acoustic/lexical content can be classified with more than 99% f1-score suggesting their distinct distributions.

Limitation of the models presented in this work is lack of the ability to exploit the temporal dependencies between speaker turns and also between modalities. We plan to address this limitation in the future work. Also, our models perform well in a single call center domain however, inter-domain generalization remains to be investigated in the future work.

Chapter 6

Conclusions and future work

6.1 Conclusions

In this thesis, we proposed several machine learning models to address some of the problems that the current automatic speech emotion recognition (SER) field is facing. To ensure the generalizability of our methods/hypotheses, we experimented on three datasets each collected with different emotion elicitation methods: Crema-D (acted), IEMOCAP (induced), and MSP-Podcast (spontaneous). All these datasets come with isolated utterances whereas only IEMOCAP contains conversations too. We show experiments on both isolated utterances and conversations.

The lack of large emotion datasets is one of the main problems that impede accurate automatic emotion recognition. To mitigate this problem, we proposed to transfer knowledge from the speaker recognition field where annotated data is plenty and also relatively simple to collect. We showed that representations/embeddings extracted from pre-trained speaker recognition models (x-vector models) do contain emotion predictive information and they

also outperform traditional representations such as eGeMAPS and MFCC. Further, adapting the entire pre-trained model boosted SER performance on all three datasets considering only isolated utterances. We found that the SER performance on x-vector embeddings is inversely proportional to the speaker verification performance i.e., the better the x-vector model the less suitable the embeddings are for the SER task. However, adapting the best x-vector model seems to provide the best results for the SER task. We also proposed a perceptually motivated data augmentation method, CopyPaste, on isolated utterances to further improve SER performance. The main idea of this technique is the observation that the presence of emotions other than neutral in a recording alters the listeners' perception. We found that the proposed three CopyPaste schemes improve SER performance and outperform the standard noise augmentation in clean conditions. Additionally, we obtained the best results using both CopyPaste and noise augmentation on all three datasets.

Models built on isolated utterances make predictions on the utterance level i.e., one prediction for the whole utterance. However, these models can not be applied to conversations unless we segment them based on emotions which is a very hard problem to solve even for humans. To avoid requiring segmentation, we proposed to build models that can make frame-level predictions. We showed that models trained with conversations outperform those trained with isolated utterances suggesting the importance of context. We compared several architectures based on CNN, LSTM, and transformer to understand their effectiveness to exploit context. We found that the transformer outperformed CNN and LSTM and was also more robust to the mismatch in training and

testing data. To further improve the performance, we proposed an augmentation scheme, DiverseCatAugment (DCA) based on the inner workings of the attention operation. In this technique, we diversify input sequences w.r.t. class labels to enable efficient use of attention operation. We found that diverse input sequences w.r.t. class labels are more important than conversational context for the best performance. With DCA, not only transformers but also CNN and LSTM models are more robust to the mismatched training and testing scenarios. However, these models do not have access to turn-taking structure in conversations. We proposed interlocutor-aware models that can exploit turn-taking structure even without speaker segmentation information. The superior performance of interlocutor-aware models supports the evidence from the literature that interlocutors do affect each other's emotions.

Finally, we considered one real-world application, predicting customer satisfaction (CSAT), with which we can still obtain insights on speakers' emotions but requires reduced annotation efforts. For this application, we considered US English telephone speech from call centers and the goal is to predict whether customers are satisfied after interacting with the agents. We presented an extensive analysis of the suitable feature representations extracted at multiple granular levels for CSAT prediction. We explored four levels of representations for transcript and three levels for the acoustic signal at word/frame-level, turn-level, segment-level, and call/document-level. We proposed using x-vectors for the CSAT task and obtained the best results on the acoustic signal. For transcript, word-, turn-level features performed competitively and for acoustic features, call-level x-vector representation worked

best. We found that exploiting semantics is helpful for the CSAT task. Features extracted from pre-trained encoders ELMo, USE, and BERT for transcripts did not provide any improvements. However, fine-tuning the BERT model provided significant gains. We observed that models perform better by distinguishing agent and customer channels. Overall, we observed the best F-score of 83.38% on ASR transcription using BERT and 78.21% on the acoustic signal using x-vector which are both using pre-trained models. Turn-taking features performed worst with an F-score of 66.99%. With score-fusion of lexical, acoustic, and turn-taking features, we obtained an 88.46% F-score. Analysis on the importance of agent Vs. customer to predict CSAT rating revealed that customer's transcripts, as well as agent's transcripts, have similar importance suggesting both of their transcripts have useful cues. However, the agent's acoustic signal seems to be less correlated with the CSAT rating compared to the customer's acoustic signal suggesting that the customer's speech is more emotional. This result is useful in cases where companies do not have the authority to record customers' speech. We also found that the last 10% of the call is more important than other parts of the call. Task completion metric seems to be well correlated with CSAT rating and found that resolving the issue for the customer may not make the customer happy. Analysis of model decisions with respect to task completion metric showed that negative calls with successful task completion are the primary source of errors.

6.2 Future directions

Our work can be extended in multiple directions. From the modeling point of view, semi-supervised and unsupervised/self-supervised techniques can be explored in addition to transfer learning and data augmentation techniques, to mitigate the problem of limited annotated data. Especially, recent publications with self-supervised approaches show a lot of improvements on several speech tasks and look promising for emotion recognition too (Khare, Parthasarathy, and Sundaram, 2021). For conversational emotion recognition, we have shown experiments using the IEMOCAP corpus which contains induced emotions with a limited number of speakers. Evaluating the proposed methods on more spontaneous conversations with more speakers such as MELD (Poria et al., 2019) can be an important step towards the analysis of spontaneous conversations. Also, it was shown that detecting valence from the speech is very difficult and relatively easier to detect from text modality (Sahu, 2019). Hence, multi-modal approaches could significantly boost the performance of conversations. Predicting customer satisfaction in customer care center conversations can be extended to pointing out the problematic regions in the conversations. We expect that advances in explainable machine learning could help in this case where the models can potentially point out the parts of the conversation that they are relying on to make the predictions.

The majority of the current research focuses on emotion recognition from isolated utterances where the context in which the utterances are produced does not exist. However, the interpretation of an utterance emotion depends on the context and can vary from person to person. We believe that shifting the

focus to conversations would enable us towards exploiting the context at least to some extent. Also, the emotions in spontaneous speech are much subtler than acted emotions and can be very hard to detect. Hence, moving away from acted speech and towards spontaneous speech would open up emotion recognition to many more applications in real-life. Categorizing spontaneous emotions, especially when using only basic emotions is very difficult because many a time they co-occur (Öhman, 2020). We believe uncertainty modeling would be more helpful here from a modeling perspective. And from a data collection perspective, it would help annotation of secondary emotions too along with primary emotions to consider for the co-occurrence of emotions.

References

Cowie, Roddy and Ellen Douglas-Cowie (1995). "Speakers and hearers are people: Reflections on speech deterioration as a consequence of acquired deafness". In: *Profound deafness and speech communication*. Whurr, pp. 510–527.

Bambaeeroo, Fatemeh and Nasrin Shokrpour (2017). "The impact of the teachers' non-verbal communication on success in teaching". In: *Journal of advances in medical education & professionalism* 5.2, p. 51.

Knapp, Mark L, Judith A Hall, and Terrence G Horgan (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Mehrabian, Albert (2017). *Nonverbal communication*. Routledge.

Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor (2001). "Emotion recognition in human-computer interaction". In: *IEEE Signal processing magazine* 18.1, pp. 32–80.

Schoenewolf, Gerald (1990). "Emotional contagion: Behavioral induction in individuals and groups". In: *Modern Psychoanalysis* 15.1, pp. 49–61.

Tyng, Chai M, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik (2017). "The influences of emotion on learning and memory". In: *Frontiers in psychology* 8, p. 1454.

Lerner, Jennifer S, Ye Li, Piercarlo Valdesolo, and Karim S Kassam (2015). "Emotion and decision making". In: *Annual review of psychology* 66, pp. 799–823.

Mäntylä, Mika, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu (2016). "Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?" In: *Proceedings of the 13th international conference on mining software repositories*, pp. 247–258.

Cauldwell, Richard T (2000). "Where did the anger go? The role of context in interpreting emotion in speech". In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Campbell, Anna, Ted Ruffman, Janice E Murray, and Paul Glue (2014). "Oxytocin improves emotion recognition for older males". In: *Neurobiology of Aging* 35.10, pp. 2246–2248.

Lindquist, Kristen A, Jennifer K MacCormack, and Holly Shablack (2015). "The role of language in emotion: Predictions from psychological constructionism". In: *Frontiers in psychology* 6, p. 444.

Hestness, Joel, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou (2017). "Deep learning scaling is predictable, empirically". In: *arXiv preprint arXiv:1712.00409*.

Tóth, Szabolcs Levente, David Sztahó, and Klára Vicsi (2008). "Speech emotion perception by human and machine". In: *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer, pp. 213–224.

Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur (2018). "X-vectors: Robust dnn embeddings for speaker recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5329–5333.

Darwin, Charles (2015). *The expression of the emotions in man and animals*. University of Chicago press.

Izard, Carroll E (1993). "Four systems for emotion activation: Cognitive and noncognitive processes." In: *Psychological review* 100.1, p. 68.

Schachter, Stanley and Jerome Singer (1962). "Cognitive, social, and physiological determinants of emotional state." In: *Psychological review* 69.5, p. 379.

James, William (1948). "What is emotion? 1884." In:

Cannon, Walter B (1927). "The James-Lange theory of emotions: A critical examination and an alternative theory". In: *The American journal of psychology* 39.1/4, pp. 106–124.

Hatfield, Elaine, John T Cacioppo, and Richard L Rapson (1993). "Emotional contagion". In: *Current directions in psychological science* 2.3, pp. 96–100.

Bericat, Eduardo (2016). "The sociology of emotions: Four decades of progress". In: *Current sociology* 64.3, pp. 491–513.

Cowie, Roddy and Randolph R Cornelius (2003). "Describing the emotional states that are expressed in speech". In: *Speech communication* 40.1-2, pp. 5–32.

Picard, Rosalind W (2000). *Affective computing*. MIT press.

Ekman, Paul (2006). *Darwin and facial expression: A century of research in review*. Ishk.

Ekman, Paul (1969). "The Repertoire of Nonverbal Behavior; Categories, Origins, and Coding". In: *Semiotica* 1.1, pp. 54–58.
Barrett, Lisa Feldman (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.
Ortony, Andrew and Terence J Turner (1990). "What's basic about basic emotions?" In: *Psychological review* 97.3, p. 315.
Koolagudi, Shashidhar G and K Sreenivasa Rao (2012). "Emotion recognition from speech: a review". In: *International journal of speech technology* 15.2, pp. 99–117.
Scherer, Klaus R, Rainer Banse, and Harald G Wallbott (2001). "Emotion inferences from vocal expression correlate across languages and cultures". In: *Journal of Cross-cultural psychology* 32.1, pp. 76–92.
Russell, James A (1991). "Culture and the categorization of emotions." In: *Psychological bulletin* 110.3, p. 426.
Aman, Saima and Stan Szpakowicz (2007a). "Identifying expressions of emotion in text". In: *International Conference on Text, Speech and Dialogue*. Springer, pp. 196–205.
Öhman, Emily (2020). "Emotion Annotation: Rethinking Emotion Categorization." In: *DHN Post-Proceedings*, pp. 134–144.
Wundt, Wilhelm Max and Charles Hubbard Judd (1902). *Outlines of psychology*. W. Engelmann.
Mehrabian, Albert and James A Russell (1974). *An approach to environmental psychology.* the MIT Press.
Russell, James A (1980). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161.
Morgan, Rick L and David Heise (1988). "Structure of emotions". In: *Social Psychology Quarterly*, pp. 19–31.
Kuppens, Peter (2008). "Individual differences in the relationship between pleasure and arousal". In: *Journal of Research in Personality* 42.4, pp. 1053–1059.
Fontaine, Johnny RJ, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth (2007). "The world of emotions is not two-dimensional". In: *Psychological science* 18.12, pp. 1050–1057.
Cochrane, Tom (2009). "Eight dimensions for the emotions". In: *Social Science Information* 48.3, pp. 379–420.
Barrett, Lisa Feldman (1998). "Discrete emotions or dimensions? The role of valence focus and arousal focus". In: *Cognition & Emotion* 12.4, pp. 579–599.

Plutchik, Robert (1980). "A general psychoevolutionary theory of emotion". In: *Theories of emotion*. Elsevier, pp. 3–33.

Cao, Houwei, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma (2014). "Crema-d: Crowd-sourced emotional multimodal actors dataset". In: *IEEE transactions on affective computing* 5.4, pp. 377–390.

Banga, Subham, Ujjwal Upadhyay, Piyush Agarwal, Aniket Sharma, and Prerana Mukherjee (2019). "Indian EmoSpeech Command Dataset: A dataset for emotion based speech recognition in the wild". In: *arXiv preprint arXiv:1910.13801*.

Wu, Tian, Yingchun Yang, Zhaohui Wu, and Dongdong Li (2006). "Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition". In: *2006 IEEE Odyssey-the speaker and language recognition workshop*. IEEE, pp. 1–5.

Batliner, Anton, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth (2003). "How to find trouble in communication". In: *Speech communication* 40.1-2, pp. 117–143.

Siedlecka, Ewa and Thomas F Denson (2019). "Experimental methods for inducing basic emotions: A qualitative review". In: *Emotion Review* 11.1, pp. 87–97.

Zhang, Xuan, Hui W Yu, and Lisa F Barrett (2014). "How does this make you feel? A comparison of four affect induction procedures". In: *Frontiers in psychology* 5, p. 689.

Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan (2008). "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42.4, p. 335.

Labov, William (1972). *Sociolinguistic patterns*. 4. University of Pennsylvania press.

Lotfian, Reza and Carlos Busso (2017). "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings". In: *IEEE Transactions on Affective Computing*.

Sneddon, Ian, Margaret McRorie, Gary McKeown, and Jennifer Hanratty (2011). "The belfast induced natural emotion database". In: *IEEE Transactions on Affective Computing* 3.1, pp. 32–41.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith (2019). "The risk of racial bias in hate speech detection". In: *Proceedings of*

the 57th annual meeting of the association for computational linguistics, pp. 1668–1678.

Paulmann, Silke, Marc D Pell, and Sonja A Kotz (2008). "How aging affects the recognition of emotional speech". In: *Brain and language* 104.3, pp. 262–269.

Buckley, Todd C, Edward B Blanchard, and W Trammell Neill (2000). "Information processing and PTSD: A review of the empirical literature". In: *Clinical psychology review* 20.8, pp. 1041–1065.

Schuller, Björn and Anton Batliner (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Wood, Ian and Sebastian Ruder (2016). "Emoji as emotion tags for tweets". In: *Proceedings of the Emotion and Sentiment Analysis Workshop LREC2016, Portorož, Slovenia*, pp. 76–79.

Wood, Ian D, John P McCrae, Vladimir Andryushechkin, and Paul Buitelaar (2018). "A comparison of emotion annotation approaches for text". In: *Information* 9.5, p. 117.

Louviere, Jordan J, Terry N Flynn, and Anthony Alfred John Marley (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Yannakakis, Georgios N, Roddy Cowie, and Carlos Busso (2018). "The ordinal nature of emotions: An emerging approach". In: *IEEE Transactions on Affective Computing* 12.1, pp. 16–35.

Kramer, Ernest (1964). "Elimination of verbal cues in judgments of emotion from voice." In: *The Journal of Abnormal and Social Psychology* 68.4, p. 390.

Lieberman, Philip and Sheldon B Michaels (1962). "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech". In: *The Journal of the Acoustical Society of America* 34.7, pp. 922–927.

Burkhardt, Felix and Walter F Sendlmeier (2000). "Verification of acoustical correlates of emotional speech using formant-synthesis". In: *ISCA Tutorial and Research Workshop (ITRW) on speech and emotion*.

Kohler, Klaus J (1995). "Articulatory reduction in different speaking styles". In: *Proceedings of the 13th International Congress of Phonetic Sciences*. Vol. 12, pp. 12–19.

Kienast, Miriam, Astrid Paeschke, and Walter Sendlmeier (1999). "Articulatory reduction in emotional speech". In: *Sixth European Conference on Speech Communication and Technology*.

Eyben, Florian, Felix Weninger, Florian Gross, and Björn Schuller (2013). "Recent developments in opensmile, the munich open-source multimedia

feature extractor". In: *Proceedings of the 21st ACM international conference on Multimedia*, pp. 835–838.

Eyben, Florian, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. (2015). "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE transactions on affective computing* 7.2, pp. 190–202.

Schuller, Björn, Stefan Steidl, and Anton Batliner (2009). "The interspeech 2009 emotion challenge". In:

Schuller, Björn, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan (2010). "The INTERSPEECH 2010 paralinguistic challenge". In: *Proc. INTERSPEECH 2010, Makuhari, Japan*, pp. 2794–2797.

Schuller, Björn, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism". In: *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Schuller, Björn W, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. (2020). "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA, pp. 2042–2046.

Schuller, Björn, Gerhard Rigoll, and Manfred Lang (2003). "Hidden Markov model-based speech emotion recognition". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* Vol. 2. Ieee, pp. II–1.

Batliner, Anton and Richard Huber (2007). "Speaker characteristics and emotion classification". In: *Speaker classification I*. Springer, pp. 138–151.

Schmitt, Maximilian, Fabien Ringeval, and Björn W Schuller (2016). "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech." In: *Interspeech*, pp. 495–499.

Tzirakis, Panagiotis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou (2017). "End-to-end multimodal emotion recognition using deep neural networks". In: *IEEE Journal of Selected Topics in Signal Processing* 11.8, pp. 1301–1309.

Sarma, Mousmita, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak (2018). "Emotion Identification from Raw Speech Signals Using DNNs." In: *Interspeech*, pp. 3097–3101.

Trigeorgis, George, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou (2016). "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network". In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5200–5204.

Cummins, Nicholas, Shahin Amiriparian, Gerhard Hagerer, Anton Batliner, Stefan Steidl, and Björn W Schuller (2017). "An image-based deep spectrum feature representation for the recognition of emotional speech". In: *Proceedings of the 25th ACM international conference on Multimedia*, pp. 478–484.

Elshaer, Mohamed Ezzeldin A, Scott Wisdom, and Taniya Mishra (2019). "Transfer learning from sound representations for anger detection in speech". In: *arXiv preprint arXiv:1902.02120*.

Lakomkin, Egor, Cornelius Weber, Sven Magg, and Stefan Wermter (2018a). "Reusing neural speech representations for auditory emotion recognition". In: *arXiv preprint arXiv:1803.11508*.

Cho, Jaejin, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak (2018). "Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts." In: *Interspeech*, pp. 247–251.

Zhao, Jianfeng, Xia Mao, and Lijiang Chen (2019). "Speech emotion recognition using deep 1D & 2D CNN LSTM networks". In: *Biomedical Signal Processing and Control* 47, pp. 312–323.

Huang, Zhengwei, Ming Dong, Qirong Mao, and Yongzhao Zhan (2014). "Speech emotion recognition using CNN". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pp. 801–804.

Lim, Wootaek, Daeyoung Jang, and Taejin Lee (2016). "Speech emotion recognition using convolutional and recurrent neural networks". In: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, pp. 1–4.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Zhang, Yuanyuan, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu (2018). "Attention based fully convolutional network for speech emotion recognition". In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 1771–1775.

Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang (2017). "Automatic speech emotion recognition using recurrent neural networks with local attention". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2227–2231.

Latif, Siddique, Rajib Rana, and Junaid Qadir (2018). "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness". In: *arXiv preprint arXiv:1811.11402*.

Han, Jing, Zixing Zhang, Zhao Ren, Fabien Ringeval, and Björn Schuller (2018). "Towards conditional adversarial training for predicting emotions from speech". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6822–6826.

Parthasarathy, Srinivas, Viktor Rozgic, Ming Sun, and Chao Wang (2019). "Improving Emotion Classification through Variational Inference of Latent Variables". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7410–7414.

Sahu, Saurabh, Rahul Gupta, and Carol Espy-Wilson (2018). "On enhancing speech emotion recognition using generative adversarial networks". In: *arXiv preprint arXiv:1806.06626*.

Liu, Jia, Chun Chen, Jiajun Bu, Mingyu You, and Jianhua Tao (2007). "Speech emotion recognition using an enhanced co-training algorithm". In: *2007 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 999–1002.

Deng, Jun, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller (2017). "Semisupervised autoencoders for speech emotion recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.1, pp. 31–43.

Zhang, Sheng, Min Chen, Jincai Chen, Yuan-Fang Li, Yiling Wu, Minglei Li, and Chuanbo Zhu (2021). "Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition". In: *Knowledge-Based Systems* 229, p. 107340.

Latif, Siddique, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Bjórn Wolfgang Schuller (2020). "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition". In: *IEEE Transactions on Affective Computing*.

Latif, Siddique, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps (2018). "Transfer learning for improving speech emotion classification accuracy". In: *arXiv preprint arXiv:1801.06353*.

Williams, Jennifer and Simon King (2019). "Disentangling Style Factors from Speaker Representations". In: *Proc. Interspeech 2019*, pp. 3945–3949.

Lakomkin, Egor, Mohammad Ali Zamani, Cornelius Weber, Sven Magg, and Stefan Wermter (2018b). "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 854–860.

Etienne, Caroline, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch (2018). "Cnn+ lstm architecture for speech emotion recognition with data augmentation". In: *arXiv preprint arXiv:1802.05630*.

Bao, Fang, Michael Neumann, and Ngoc Thang Vu (2019). "CycleGAN-Based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition." In: *INTERSPEECH*, pp. 2828–2832.

Rizos, Georgios, Alice Baird, Max Elliott, and Björn Schuller (2020). "Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3502–3506.

Bothe, Chandrakant, Cornelius Weber, Sven Magg, and Stefan Wermter (2020). "EDA: Enriching Emotional Dialogue Acts using an Ensemble of Neural Annotators". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 620–627.

Li, Haoqi, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou (2020a). "Speaker-invariant affective representation learning via adversarial training". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7144–7148.

Parthasarathy, Srinivas and Carlos Busso (2018). "Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes". In: *Proc. Interspeech 2018*, pp. 3698–3702.

Ganin, Yaroslav and Victor Lempitsky (2015). "Unsupervised domain adaptation by backpropagation". In: *International conference on machine learning*. PMLR, pp. 1180–1189.

Bozinovski, S and A Fulgosi (1976). "The influence of pattern similarity and transfer learning upon training of a base perceptron b2". In: *Proceedings of Symposium Informatica*, pp. 3–121.

Ramirez, Jose Manuel, Ana Montalvo, and Jose Ramon Calvo (2019). "A Survey of the Effects of Data Augmentation for Automatic Speech Recognition Systems". In: *Iberoamerican Congress on Pattern Recognition*. Springer, pp. 669–678.

Raj, Desh, David Snyder, Daniel Povey, and Sanjeev Khudanpur (2019). "Probing the Information Encoded in x-vectors". In: *arXiv preprint arXiv:1909.06351*.

Lotfian, Reza and Carlos Busso (2019). "Curriculum learning for speech emotion recognition from crowdsourced labels". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 27.4, pp. 815–826.

Villalba, Jesús, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Fred Richardson, Suwon Shon, François Grondin, et al. (2019). "State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18". In: *Proc. Interspeech 2019*, pp. 1488–1492.

Shum, Stephen H, Najim Dehak, Réda Dehak, and James R Glass (2013). "Unsupervised methods for speaker diarization: An integrated and iterative approach". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10, pp. 2015–2028.

Sell, Gregory and Daniel Garcia-Romero (2014). "Speaker diarization with PLDA i-vector scoring and unsupervised calibration". In: *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 413–417.

Maciejewski, Matthew, David Snyder, Vimal Manohar, Najim Dehak, and Sanjeev Khudanpur (2018). "Characterizing performance of speaker diarization systems on far-field speech using standard methods". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5244–5248.

Sell, Gregory, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, et al. (2018). "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge." In: *Interspeech*, pp. 2808–2812.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Liu, Weiyang, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song (2017). "Sphereface: Deep hypersphere embedding for face recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220.

Snyder, David, Guoguo Chen, and Daniel Povey (2015). "Musan: A music, speech, and noise corpus". In: *arXiv preprint arXiv:1510.08484*.

Hsiao, Roger, Jeff Ma, William Hartmann, Martin Karafiát, František Grézl, Lukáš Burget, Igor Szöke, Jan Honza Černockỳ, Shinji Watanabe, Zhuo Chen, et al. (2015). "Robust speech recognition in unknown reverberant and noisy conditions". In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 533–538.

Hazarika, Devamanyu, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann (2018). "Conversational memory network for emotion recognition in dyadic dialogue videos". In: *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2018. NIH Public Access, p. 2122.

Majumder, Navonil, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria (2019). "Dialoguernn: An attentive rnn for emotion detection in conversations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 6818–6825.

Li, Jingye, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu (2020b). "HiTrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations". In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4190–4200.

Zhang, Dong, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou (2019). "Modeling both Context-and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations." In: *IJCAI*, pp. 5415–5421.

Grimm, Michael, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan (2007). "Primitives-based evaluation and estimation of emotions in speech". In: *Speech communication* 49.10-11, pp. 787–800.

Metallinou, Angeliki, Athanasios Katsamanis, and Shrikanth Narayanan (2013). "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information". In: *Image and Vision Computing* 31.2, pp. 137–152.

Eyben, Florian, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie (2010). "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues". In: *Journal on Multimodal User Interfaces* 3.1, pp. 7–19.

Schmitt, Maximilian, Nicholas Cummins, and Björn W Schuller (2019). "Continuous Emotion Recognition in Speech—Do We Need Recurrence?" In: *Proc. Interspeech 2019*, pp. 2808–2812.

Lee, Jinkyu and Ivan Tashev (2015). "High-level feature representation using recurrent neural network for speech emotion recognition". In: *Sixteenth annual conference of the international speech communication association*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Wolf, Thomas, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. (2020). "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45.

Karita, Shigeki, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. (2019). "A comparative study on transformer vs rnn in speech applications". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 449–456.

Lu, Liang, Changliang Liu, Jinyu Li, and Yifan Gong (2020). "Exploring transformers for large-scale speech recognition". In: *arXiv preprint arXiv:2005.09684*.

Mohamed, Abdelrahman, Dmytro Okhonko, and Luke Zettlemoyer (2019). "Transformers with convolutional context for ASR". In: *arXiv preprint arXiv:1904.11660*.

Smirnov, Dmitry, Heini Saarimäki, Enrico Glerean, Riitta Hari, Mikko Sams, and Lauri Nummenmaa (2019). "Emotions amplify speaker–listener neural alignment". In: *Human brain mapping* 40.16, pp. 4777–4788.

Kuppens, Peter, Nicholas B Allen, and Lisa B Sheeber (2010). "Emotional inertia and psychological maladjustment". In: *Psychological science* 21.7, pp. 984–991.

Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea (2019). "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536.

Engelbrech, Klaus-Peter, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller (2009). "Modeling user satisfaction with hidden markov model". In: *Proceedings of the SIGDIAL 2009 conference: the 10th annual*

meeting of the special interest group on discourse and dialogue. Association for Computational Linguistics, pp. 170–177.

Sabbeh, Sahar F (2018). "Machine-Learning Techniques for Customer Retention: A Comparative Study". In: *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 9.2, pp. 273–281.

Ranaweera, Chatura and Jaideep Prabhu (2003). "The influence of satisfaction, trust and switching barriers on customer retention in a continuous purchasing setting". In: *International journal of service industry management* 14.4, pp. 374–395.

Ansari, Azarnoush and Arash Riasi (2016). "Modelling and evaluating customer loyalty using neural networks: Evidence from startup insurance companies". In: *Future Business Journal* 2.1, pp. 15–30.

Hallowell, Roger (1996). "The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study". In: *International journal of service industry management* 7.4, pp. 27–42.

Aman, Saima and Stan Szpakowicz (2007b). "Identifying expressions of emotion in text". In: *International Conference on Text, Speech and Dialogue*. Springer, pp. 196–205.

Walker, Marilyn A, Lynette Hirschman, and John S Aberdeen (2000). "Evaluation for Darpa Communicator Spoken Dialogue Systems." In: *LREC*.

Walker, Marilyn A, Rebecca Passonneau, and Julie E Boland (2001). "Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems". In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 515–522.

Walker, Marilyn A, Diane J Litman, Candace A Kamm, and Alicia Abella (1997). "PARADISE: A framework for evaluating spoken dialogue agents". In: *arXiv preprint cmp-lg/9704004*.

Yang, Zhaojun, Gina-Anne Levow, and Helen Meng (2012). "Predicting user satisfaction in spoken dialog system evaluation with collaborative filtering". In: *IEEE Journal of Selected Topics in Signal Processing* 6.8, pp. 971–981.

Chowdhury, Shammur Absar, Evgeny A Stepanov, Giuseppe Riccardi, et al. (2016). "Predicting User Satisfaction from Turn-Taking in Spoken Conversations." In: *INTERSPEECH*, pp. 2910–2914.

Luque, Jordi, Carlos Segura, Ariadna Sánchez, Martı Umbert, and Luis Angel Galindo (2017). "The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls". In: *Proc. Interspeech 2017*, pp. 2346–2350.

Park, Youngja and Stephen C Gates (2009). "Towards real-time measurement of customer satisfaction using automatically generated call transcripts". In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, pp. 1387–1396.

Meinzer, Stefan, Ulf Jensen, Alexander Thamm, Joachim Hornegger, and Björn M Eskofier (2016). "Can machine learning techniques predict customer dissatisfaction? A feasibility study for the automotive industry". In: *Artificial Intelligence Research* 6.1, p. 80.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 785–794.

Auguste, Jeremy, Delphine Charlet, Geraldine Damnati, Frédéric Béchet, and Benoit Favre (2019). "Can we predict self-reported customer satisfaction from interactions?" In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7385–7389.

Reitter, David and Johanna D Moore (2007). "Predicting success in dialogue". In:

Noseworthy, Michael, Jackie Chi Kit Cheung, and Joelle Pineau (2017). "Predicting Success in Goal-Driven Human-Human Dialogues". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 253–262.

Schoenmüller, Verena, Oded Netzer, and Florian Stahl (2019). "The extreme distribution of online reviews: Prevalence, drivers and implications". In: *Columbia Business School Research Paper* 18-10.

Povey, Daniel, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur (2016). "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI." In: *Interspeech*, pp. 2751–2755.

Zhang, Ye and Byron Wallace (2015). "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1510.03820*.

Xu, Haotian, Ming Dong, Dongxiao Zhu, Alexander Kotov, April Idalski Carcone, and Sylvie Naar-King (2016). "Text Classification with Topic-based Word Embedding and Convolutional Neural Networks." In: *BCB*, pp. 88–97.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365*.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. (2018). "Universal sentence encoder". In: *arXiv preprint arXiv:1803.11175*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Cho, Jaejin, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak (2019). "Deep neural networks for emotion recognition combining audio and transcripts". In: *arXiv preprint arXiv:1911.00432*.

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg (2016). "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks". In: *arXiv preprint arXiv:1608.04207*.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Pappagari, Raghavendra, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak (2019). "Hierarchical Transformers for Long Document Classification". In: *arXiv preprint arXiv:1910.10781*.

Dai, Zihang, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov (2019). "Transformer-xl: Attentive language models beyond a fixed-length context". In: *arXiv preprint arXiv:1901.02860*.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Khare, Aparna, Srinivas Parthasarathy, and Shiva Sundaram (2021). "Self-Supervised learning with cross-modal transformers for emotion recognition". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 381–388.

Sahu, Saurabh (2019). "Towards Building Generalizable Speech Emotion Recognition Models". PhD thesis. University of Maryland, College Park.