THE UNIVERSITY
of EDINBURGH

Recognizing Emotions in Spoken Dialogue with Acoustic and Lexical Cues

Leimin Tian

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2018

Abstract

Automatic emotion recognition has long been a focus of Affective Computing. It has become increasingly apparent that awareness of human emotions in Human-Computer Interaction (HCI) is crucial for advancing related technologies, such as dialogue systems. However, performance of current automatic emotion recognition is disappointing compared to human performance. Current research on emotion recognition in spoken dialogue focuses on identifying better feature representations and recognition models from a data-driven point of view. The goal of this thesis is to explore how incorporating prior knowledge of human emotion recognition in the automatic model can improve state-of-the-art performance of automatic emotion recognition in spoken dialogue. Specifically, we study this by proposing knowledge-inspired features representing occurrences of disfluency and non-verbal vocalisation in speech, and by building a multimodal recognition model that combines acoustic and lexical features in a knowledge-inspired hierarchical structure. In our study, emotions are represented with the Arousal, Expectancy, Power, and Valence emotion dimensions. We build unimodal and multimodal emotion recognition models to study the proposed features and modelling approach, and perform emotion recognition on both spontaneous and acted dialogue.

Psycholinguistic studies have suggested that DISfluency and Non-verbal Vocalisation (DIS-NV) in dialogue is related to emotions. However, these affective cues in spoken dialogue are overlooked by current automatic emotion recognition research. Thus, we propose features for recognizing emotions in spoken dialogue which describe five types of DIS-NV in utterances, namely filled pause, filler, stutter, laughter, and audible breath. Our experiments show that this small set of features is predictive of emotions. Our DIS-NV features achieve better performance than benchmark acoustic and lexical features for recognizing all emotion dimensions in spontaneous dialogue. Consistent with Psycholinguistic studies, the DIS-NV features are especially predictive of the Expectancy dimension of emotion, which relates to speaker uncertainty. Our study illustrates the relationship between DIS-NVs and emotions in dialogue, which contributes to Psycholinguistic understanding of them as well. Note that our DIS-NV features are based on manual annotations, yet our long-term goal is to apply our emotion recognition model to HCI systems. Thus, we conduct preliminary experiments on automatic detection of DIS-NVs, and on using automatically detected DIS-NV features for emotion recognition. Our results show that DIS-NVs can be automatically detected from speech with stable accuracy, and

auto-detected DIS-NV features remain predictive of emotions in spontaneous dialogue. This suggests that our emotion recognition model can be applied to a fully automatic system in the future, and holds the potential to improve the quality of emotional interaction in current HCI systems.

To study the robustness of the DIS-NV features, we conduct cross-corpora experiments on both spontaneous and acted dialogue. We identify how dialogue type influences the performance of DIS-NV features and emotion recognition models. DIS-NVs contain additional information beyond acoustic characteristics or lexical contents. Thus, we study the gain of modality fusion for emotion recognition with the DIS-NV features. Previous work combines different feature sets by fusing modalities at the same level using two types of fusion strategies: Feature-Level (FL) fusion, which concatenates feature sets before recognition; and Decision-Level (DL) fusion, which makes the final decision based on outputs of all unimodal models. However, features from different modalities may describe data at different time scales or levels of abstraction. Moreover, Cognitive Science research indicates that when perceiving emotions, humans make use of information from different modalities at different cognitive levels and time steps. Therefore, we propose a HierarchicaL (HL) fusion strategy for multimodal emotion recognition, which incorporates features that describe data at a longer time interval or which are more abstract at higher levels of its knowledge-inspired hierarchy. Compared to FL and DL fusion, HL fusion incorporates both inter- and intra-modality differences. Our experiments show that HL fusion consistently outperforms FL and DL fusion on multimodal emotion recognition in both spontaneous and acted dialogue. The HL model combining our DIS-NV features with benchmark acoustic and lexical features improves current performance of multimodal emotion recognition in spoken dialogue.

To study how other emotion-related tasks of spoken dialogue can benefit from the proposed approaches, we apply the DIS-NV features and the HL fusion strategy to recognize movie-induced emotions. Our experiments show that although designed for recognizing emotions in spoken dialogue, DIS-NV features and HL fusion remain effective for recognizing movie-induced emotions. This suggests that other emotion-related tasks can also benefit from the proposed features and model structure.

Acknowledgements

I would like to give my sincere thanks to my supervisors Prof. Johanna D. Moore and Dr. Catherine Lai for their dedicated guidance and insightful advice throughout my MSc and PhD study. Without Johanna and Catherine this thesis would not exist. They are models of exceptional researchers to whom I always look up to, and they continue to motivate me to proceed in academia.

I am in debt to Dr. Robin Lickley and Dr. Amos Storkey for their generous sharing of expertise knowledge on Psycholinguistics and Deep Learning. Discussion with them has inspired major ideas of this work. I am grateful for being able to collaborate with Michal Muszynski and the University of Geneva on movie emotion studies. It has been my pleasure co-supervising Xinran Shi, Chenyu Wang, and Stanley Wang on their MSc dissertations. I would also like to thank my examiners Prof. Catherine Pelachaud and Prof. Steve Renals for their perceptive comments.

I want to address my affection and appreciation for my parents Chuan Tian and Xiuqin Wang, as well as my boyfriend Frits de Nijs and his family, for their love and support. They have always been there for me through the ups and downs of life, and are the source of power and warmth in my heart. My special thanks to Frits and the Washoe County Sheriff office and Search and Rescue team for saving my life on the Mt. Rose trail of Lake Tahoe.

I would like to thank my best friend Xiao Wang for his support and all the fun together. I would also like to thank my friends in the Forum (in alphabetic order): Cheng Feng, Hadi Daneshvar Farzanegan, Lea Frermann, Weili Fu, Siva Reddy Gangireddy, Spandana Gella, Kathrin Haag, Xin He, Xuan Huang, Herman Kamper, Sam Ribeiro, Siva Reddy, Carina Silberer, Pawel Swietojanski, Zhengshuai Lin, Xingxing Zhang, and Yichuan Zhang; as well as my friends at the British Science Association Edinburgh branch (in alphabetic order): John Bradshaw, Madeleine Berg, Kaitlyn Hair, Nico Kroberg, and Frank Machin. It has been my privilege to spend time with my colleagues and friends in the Institute for Language, Cognition and Computation and the Centre for Speech Technology Research, and to attend all the inspiring discussion and presentations.

Last but not least, I appreciate the financial support provided by the Institute for Language, Cognition and Computation and the School of Informatics during my PhD study and conference trips. It has truly been an amazing and life-changing experience doing my PhD at School of Informatics, the University of Edinburgh.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Leimin Tian*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Speech, originally, was the device whereby Man learned, imperfectly, to transmit the thoughts and emotions of his mind.*
>
> — Isaac Asimov, *Second Foundation (1953)*

In this chapter, we provide an overview of our work on automatic emotion recognition in spoken dialogue with acoustic and lexical features. We discuss our motivation for why we are particularly interested in emotion recognition and what are the limitations of state-of-the-art studies in this filed. We then state our hypotheses addressing the identified issues in current emotion recognition research. We also lay out the roadmap of this thesis.

1.1 Motivation

Research in Cognitive Science has shown that emotions are vital in human cognition and communication processes (Picard, 2000). This has led to the establishment of the field of Affective Computing. The term *Affective* here refers to aspects of cognition relating to, resulting from, or influenced by human emotions, and *Affective Computing* is the study of developing emotion-aware technologies. There are three main topics in Affective Computing: how to recognize human emotions, how to model emotional interactions, and how to synthesize expressive reactions. It has become increasingly apparent that awareness of emotions is crucial for advancing technologies related to Human-Computer Interaction (HCI). For example, a virtual agent that is able to copy and adapt its laughter and expressive behaviour to the user's behaviour pattern has been shown to increase users' humour experience (Pecune et al., 2015). Similarly, in affective game design, Non-Player Characters that are aware of the emotional states

of the player and can generate emotional reactions have been shown to keep players engaged and improve the gaming experience (Popescu et al., 2014). In a teaching scenario, a robot lecturer expressing a positive mood while giving lectures to university classes was rated as having higher lecturing quality (Xu et al., 2014).

Automatic emotion recognition has long been a focus in Affective Computing because it is the first step towards emotion-aware technologies. In HCI systems, the user's emotions can either be described as categories such as joy or sorrow, or as values on emotional dimensions such as Arousal (excited/bored). The ability to accurately recognize emotions in spoken dialogue is crucial to building more natural and engaging dialogue systems. In earlier studies of automatic emotion recognition, experiments are often conducted on data produced by human participants portraying the emotions under artificial settings, such as reading the same short sentence with different emotions (Petrushin, 2000). This resulted in emotion recognition models that attain good performance on the specific database, sometimes even outperforming humans. However, these models are overfitting the data and are fragile to changes, such as different experiment participants, and are thus difficult to apply to HCI systems (Schuller et al., 2010a). Recent studies attempt to build more robust automatic emotion recognition models by training on natural data such as recordings of spontaneous dialogue. However, performance of these models is disappointing compared to human performance (Sauter et al., 2010b), and there is considerable room for improvement (Poria et al., 2017). For example, the correlation coefficient between the automatic predictions given by the best performing multimodal model and the human annotations on the AVEC2012 database is only 0.280 (Savran et al., 2012). Similarly, Zadeh et al. (2017) conducted experiments to predict binary Valence (positive/negative) values of videos, and their best accuracy achieved by the automatic model using multimodal information is 77.1%, while human performance is 85.7%. Therefore, the goal of this thesis is to develop an effective and robust emotion recognition model that can improve the performance of state-of-the-art automatic emotion recognition in spoken dialogue, with the long-term goal to apply the proposed model to current HCI systems and improve the quality of emotional interaction.

1.1.1 Limitations of Current Emotion Recognition Approaches

State-of-the-art approaches for emotion recognition have focused on identifying better feature representations and recognition models. However, how to identify effective

features and models remains an open question.

Features used in emotion recognition can be extracted from various modalities (e.g., audio, visual, lexical, and physiological). Typically better performance can be achieved when incorporating information from all available modalities, such as our previous work combining the audio, visual, and lexical modalities for emotion recognition (Moore et al., 2014). However, this thesis focuses on the acoustic and lexical modalities. This is because our task is to recognize emotions from spoken dialogue and the acoustic and lexical modalities are the major types of information most commonly available.

For the acoustic modality, previous studies extracted features describing the statistical characteristics of the speech signals for emotion recognition. By **statistical** we refer to features that are data-driven and do not require emotion-specific knowledge during the feature extraction. These features have been widely used in state-of-the-art emotion recognition models and have achieved robust performance (e.g., Song et al. (2016a)). However, recent studies indicate that acoustic features motivated by Psycholinguistic studies on relations between acoustic characteristics and speaker emotions (the **knowledge-inspired** features) may have comparable or better performance than a large set of statistical acoustic features. For example, three hand-selected features describing the pitch, energy, and quality of speech at the utterance level can achieve better performance than thousands of statistical acoustic features on predicting the Arousal (excited/bored) emotion dimension (Bone et al., 2014). For the lexical modality, state-of-the-art emotion recognition models have focused on features describing the lexical content, i.e., the words that are spoken.

Both the lexical and acoustic features used in state-of-the-art research on recognizing emotions in spoken dialogue have focused on speech in isolation, while specific characteristics of spoken dialogue compared to other forms of speech (e.g., monologue) are overlooked. Psycholinguistic studies have suggested that dialogue phenomena, especially DISfluencies and Non-verbal Vocalisations (DIS-NVs), may be related to human emotions (Shriberg, 2005; Vuilleumier, 2005; Lickley, 2015). However, such affective cues in spoken dialogue have been largely overlooked in feature extraction of state-of-the-art emotion recognition models. In addition, previous work has suggested that there are various data aspects that can greatly influence performance of emotion recognition models, such as dialogue type or database size (Zeng et al., 2009). In terms of DIS-NVs in spoken dialogue, Trouvain (2014) suggested that DIS-NVs are more common in spontaneous and unscripted dialogue

than in acted dialogue. Thus, a cross-corpora study using data from different types of dialogue is required to further study the robustness of using DIS-NVs for emotion recognition in spoken dialogue.

Beyond features, emotion recognition performance also depends on how those features are used to model emotions. To build emotion recognition models, widely used classification or regression models have been applied. There are two main types of machine learning models: shallow learning which has a flat model structure and uses the input features directly, and deep learning which has a multilayer network structure and learns an abstracted representation from the input features before performing recognition. Various shallow learning models have been applied to emotion recognition since the establishment of the field. For example, Support Vector Machines used by Lubis et al. (2016), Hidden Markov Models used by Ozkan et al. (2012), and Conditional Random Fields used by Baltrusaitis et al. (2013). There exist many different machine learning models and it is important to choose the appropriate one for a specific task. However, Forbes-Riley and Litman (2011) compared various shallow learning algorithms and their results show that performance differences are not significant when accounting for feature sets and other parameter settings.

In recent years, however, significant performance improvements have been obtained using deep learning models instead of shallow machine learning models in emotion recognition. Here the term **deep learning models** refers to neural networks with more than one hidden layer. The network structure of deep learning models allows flexible control when fusing multiple modalities and modelling temporal context. This enables the models to extract more effective features automatically. For example, deep and hierarchical neural networks have obtained top performance in detecting the Valence emotion dimension (positive/negative) and the level of conflict (Brueckner and Schuller, 2015), and the use of Autoencoders has improved unsupervised domain adaptation in affective speech analysis (Deng et al., 2014). Among different deep learning models, the Long Short-Term Memory Recurrent Neural Network (LSTM) model is especially powerful for emotion recognition because of its ability to model long-range temporal context. However, the small size of emotion databases compared to databases used for speech or image recognition tasks may limit optimization of the complex model structure of a deep learning model. The ability to generalize over different databases is also an issue for current deep learning models.

In addition to extracting more effective features with deep learning models, combining information from multiple modalities and building multimodal emotion

recognition models typically improves performance compared to unimodal emotion recognition approaches. However, the improvement is often limited (D'Mello and Kory, 2012). One reason may be that most multimodal models combine different modalities at the same level. Information from different modalities is either combined at the Feature Level by concatenating the feature sets (FL fusion), or at the Decision Level by fusing the predictions given by each unimodal model (DL fusion). However, different modalities may describe data at different time scales. For example, many statistical acoustic features describe data at the frame level, while knowledge-inspired prosodic features often describe data at the utterance level. Different features may also have different levels of abstraction. For example, the statistical acoustic features describe a wide range of acoustic characteristics, while knowledge-inspired prosodic features describe a small set of emotion-specific cues. Therefore, the statistical acoustic features may capture vocalisations unrelated to emotion and are thus less abstract than the knowledge-inspired prosodic features with respect to this task. Due to such differences, combining these features at different levels may improve the benefits gained by modality fusion. Cognitive Science studies also indicate that when perceiving emotions, humans make use of different types of information from different modalities at different cognitive levels and time steps (Grandjean et al., 2008). However, in state-of-the-art multimodal emotion recognition models, it is extremely rare to combine information at different stages in a knowledge-inspired hierarchy, and existing fusion strategies do not model both the inter- and intra-modality differences.

1.1.2 Hypotheses

This thesis attempts to address the issues we identified in current studies and improve on state-of-the-art of emotion recognition in spoken dialogue. Here we propose the two main hypotheses of this thesis.

H1: DIS-NV Features are Predictive of Emotions in Spoken Dialogue

In terms of feature representation, we propose to use dialogue phenomena beyond the traditional features representing acoustic characteristics and lexical content of speech. In particular, we propose novel features describing the occurrences of DISfluency and Non-verbal Vocalisation (DIS-NV) in utterances for recognizing emotions in dialogue. As discussed earlier, DIS-NVs are emotion related phenomena in speech which were overlooked by current emotion recognition models. Thus, we have good reason to

believe that the DIS-NV features will be predictive of emotions in dialogue, both when used on their own and when used in combination with other acoustic and lexical features. However, note that as a dialogue phenomenon, efficacy of the DIS-NV features may be influenced by the differences existing between spontaneous and acted dialogue.

H2: the HL Fusion Strategy will Outperform the FL and DL Fusion Strategies for Multimodal Emotion Recognition

In terms of recognition model, we propose a HierarchicaL (HL) fusion strategy which incorporates features that are more abstract or describe data at a longer time scale at higher layers of its knowledge-inspired hierarchical structure. Compared to existing fusion strategies, HL fusion can model both inter- and intra-modality differences. Thus, we expect the HL fusion to outperform the traditional FL and DL fusion strategies for multimodal emotion recognition.

1.2 Roadmap of This Thesis

Remainder of this thesis is arranged as follows: In Chapter 2, we review the theoretical background of human emotions and the Psycholinguistic studies that motivate our DIS-NV features and HL fusion strategy. In Chapter 3, we review state-of-the-art of automatic emotion recognition and provide the detailed methodology of our experiments, including the emotion databases of spoken dialogue we experiment on, the benchmark features and model approaches we compare our emotion recognition approaches with, and the evaluation metrics. In Chapter 4, we introduce our DIS-NV features and perform experiments investigating our first hypothesis on the effectiveness of the DIS-NV features for emotion recognition in spontaneous dialogue. To understand our findings in Chapter 4 on spontaneous dialogue in a wider context, in Chapter 5, we conduct cross-corpora studies on emotion recognition in spontaneous and acted dialogue. We first compare characteristics of spontaneous and acted dialogue by comparing distributions of emotions, DIS-NVs, and utterance-level prosodic measurements. We then continue to investigate the influence of dialogue type on emotion recognition approaches. In addition, we study the effectiveness of the deep LSTM model for emotion recognition compared to the shallow Support Vector Machines (SVMs). Note that Chapters 4 and 5 have mainly focused on unimodal

emotion recognition. In Chapter 6, we move on to multimodal emotion recognition with acoustic and lexical modalities and investigate our second hypothesis on the effectiveness of HL fusion compared to FL and DL fusion. Beyond recognizing speaker's emotions from spoken dialogue, we are curious to study whether or not other emotion-related tasks of spoken dialogue can also benefit from the DIS-NV features and HL fusion strategy. Thus, in Chapter 7, we apply our DIS-NV features and HL fusion strategy to predicting audience's emotions induced by movies. Finally, in Chapter 8, we summarize contributions and limitations of this thesis and discuss possible future directions. In particular, how to integrate our emotion recognition model into a HCI system and investigate its influence on the quality of emotional interaction.

Chapter 2

Human Emotion Theories

The history of science is full of revolutionary advances that required small insights that anyone might have had, but that, in fact, only one person did.

— Isaac Asimov, *The Three Numbers (1974)*

In this chapter, we review the theoretical background of human emotions. In particular, we discuss different theoretical frameworks for defining emotions and review Psycholinguistic findings on human emotion perception and induction in spoken dialogue. Note that the focus of this thesis is the computational aspect of emotion recognition in spoken dialogue. Therefore, here we do not attempt to answer fundamental questions in Psychology and Cognitive Science on human emotions, such as what are emotions or why do humans have emotions.

2.1 Emotion Theories

The question of how to define, study, and explain human emotions has been the subject of an on-going debate in current Psychology and Cognitive Science studies. Among different emotion definition theories, there are four major approaches that have influenced computational studies of emotions (Cornelius, 2000). The first theory is the Darwinian emotion perspective (e.g., Ekman et al. (1987)). The Darwinian perspective argues that emotions are products of evolution. Thus, it aims to identify a set of primitive and universal emotion categories. For example, the Darwinian perspective suggests that fear is a primitive and universal emotion because it was developed as an alarm system to increase chance of survival. The second theory is the Jamesian emotion perspective (e.g., Levenson (1992)). The Jamesian perspective argues that emotions are caused by physiological and bodily changes, and research

in this framework works towards identifying the physiological aspects of emotions, such as heart rate and neurological signals. The third theory is the cognitive emotion perspective (e.g., Ortony et al. (1990)). The cognitive perspective argues that changes in emotional states are induced by events and our perceptions on how these events influence us, such as fulfilment of personal goals. Thus, research in this framework focuses on identifying primitive emotional dimensions to represent complex and compound emotions, and on building emotional reaction models to describe the relations between events (appraisals) and emotions. The fourth theory is the social constructivist emotion perspective (e.g., Spelman (1989)). The social constructivist perspective studies the cultural, gender, and other social or individual differences of emotion perception and expression. These four emotion theories study different aspects of emotions and are equally important for understanding human emotions.

Considering the application of emotion theory in automatic emotion recognition, a large number of current studies have followed the Darwinian emotion theory, which defines emotions in terms of several primary and universal categories. The most widely used emotion categorisation is Ekman's Big-6 set of emotions, which identifies anger, disgust, fear, happiness, sadness, and surprise as the primitive and universal human emotion categories based on studies of facial expressions, as shown in Figure 2.1 (Ekman et al., 1987). However, this categorisation has been criticized as being biased towards negative emotions. Furthermore, the categories were based on a participant group lacking cultural diversity and, thus, their universality has been questioned (Ortony and Turner, 1990). What the basic emotion categories are remains an open question, and it is difficult to describe compound emotions or the intensity of different emotions with a limited set of emotion categories.

Instead of the Darwinian emotion theory, a number of automatic emotion recognition studies have been focused on the cognitive emotion theory. Such work associates emotions with specific appraisals (stimuli that evoke changes in emotional states) and describes emotions as vectors in a space defined by a set of primitive emotion dimensions (Ortony et al., 1990). For example, the Arousal (active/inactive) and Valence (positive/negative) dimensions (Fontaine et al., 2007). Compared to the emotion categorisations, emotion dimensions are able to describe emotions in a more flexible manner and capture subtle changes in complex emotions, as shown in Figure 2.2. The association between emotions and appraisals also leads to emotional interaction models which connects interaction events, recognized emotions, and expressed emotions (e.g., Marsella and Gratch (2009)).

Figure 2.1: The Big-6 Emotion Categorisation (Ekman et al., 1987)

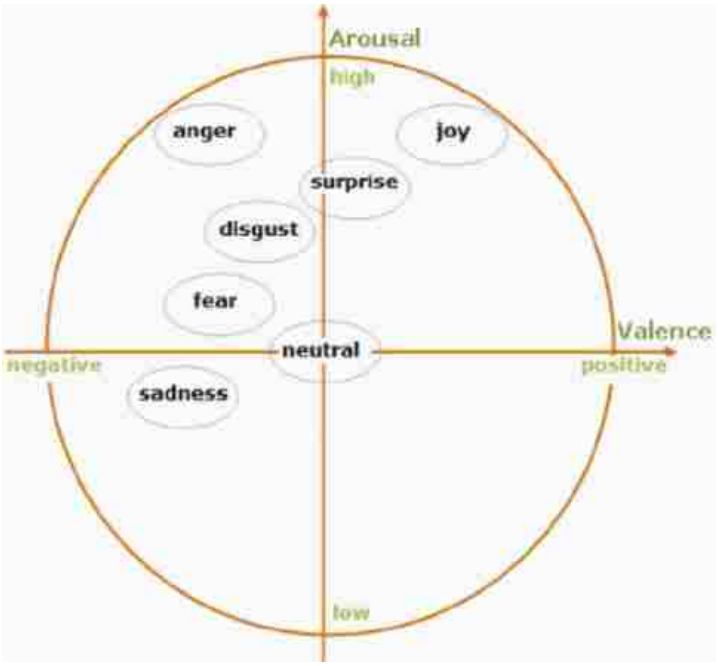


Figure 2.2: An Example of the Dimensional Emotion Definition

Note that the goal of this thesis is to build emotion recognition models which can potentially be applied to the emotional interaction module of Human-Computer Interaction (HCI) systems. The emotional interaction models of most current HCI

systems are developed with appraisal-based emotion models (e.g., Marsella et al. (2010)). Therefore, we adopt the cognitive emotion theory in this work and describe emotions as vectors in a multi-dimensional emotion space (see Section 3.2). The dimensional emotion definition also allows us to model more subtle and complex emotions in spoken dialogue.

2.2 Human Emotion Perception and Induction in Spoken Dialogue

Emotions play an important role in human cognition and communication. However, they are complex phenomena yet to be fully understood. Emotions have been traditionally studied from the biological and neuroscience perspectives, which view emotions as genetically determined and universally similar responses (Panksepp, 2004). In this thesis, we are more interested in emotions in social communications. From this perspective, emotions are socially constituted functions which vary among sociocultural systems, and are subjective experiences influenced by various biological, personal, and situational factors (Averill, 1980). Gordon (1990) identified five social dimensions of emotions: origin (primary or universal elements), time (short-term reactions or long-lasting emotions), structure (private or social experiences of emotions), change (micro-level self-regulations or macro-level historical changes), social situation (social relationship), and emotional culture (vocabulary, norms, or beliefs). It is important to study how emotions vary on each of these social dimensions. However, in this thesis we controlled the conditions of the emotional communication by selecting emotion databases of English dyadic dialogue collected under similar social dimensions, with the main difference being the type of spoken dialogue (spontaneous vs. acted) between the databases.

During a conversation, we perceive the emotions of our conversational partner (perceived emotions) which may evoke emotional experiences of our own (induced emotions). We can express such induced emotions during the interaction which our conversational partner can then perceive (Niedenthal and Brauer, 2012). In Chapter 7, we will discuss the perception and induction of emotions in more detail. However, in the majority of this thesis we will focus on human recognition of emotions expressed in spoken dialogue, i.e., the perceived emotions.

Humans convey and perceive emotions through all communicative modalities

during a social interaction, such as the audio, visual, or gestural modalities (Jaimes and Sebe, 2007). When recognizing emotions, humans are shown to have better performance when given information from multiple modalities (De Silva et al., 1997). However, spoken dialogue applications may only have access to the speech data. Because we are interested in emotions in spoken dialogue, in this thesis we focus on emotions communicated through the acoustic and lexical modalities. Among the different types of information conveyed through the acoustic and lexical modalities, Psycholinguistic studies have identified the prosodic and lexical content of the speech as being closely related to human emotions expressed in spoken dialogue (Zeng et al., 2009).

For the acoustic modality, identifying the optimal set of voice cues for human emotion recognition remains an open question. Previous studies indicate that human listeners can recognize emotions accurately with prosodic cues such as pitch, energy and speech rate (Juslin and Scherer, 2005). Spectral cues such as Mel-Frequency Cepstral Coefficients (MFCCs) also received great attention in previous research because of their ability to represent the short-term power spectrum of speech audio by modelling the human auditory system's response. Among different types of acoustic characteristics, empirical experiments studying correlations between emotions and different acoustic cues indicate that pitch and energy are the most relevant to emotions compared to other acoustic cues (Cowie et al., 2001; Kwon et al., 2003).

For the lexical modality, the affective state of the speaker is related to the words that have been spoken. Continuing efforts have been made by Psycholinguistic researchers to establish dictionaries of words with affective annotations, such as the WordNet Affect (Strapparava and Valitutti, 2004) and the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007). In addition to individual words, information about semantic context is also important for understanding emotions expressed in the whole utterance or the whole dialogue. The emotional state of a person during a conversation tends not to change rapidly and thus depends on the temporal context (Zeng et al., 2009). Although lexical content remains an important clue of emotions in dialogue, individual preferences and cultural differences can influence the correlation between lexical content and emotions greatly. Previous studies have suggested that in some cases, linguistic information can be unreliable for analysing human emotions, and the association between lexical content and emotions is hard to generalize over different types of dialogue, different languages, and different speaker personality types, etc (Ambady and Rosenthal, 1992; Furnas et al., 1987).

Besides the acoustic characteristics and lexical content, aspects of spoken dialogue such as DISfluencies and Non-verbal Vocalisations (DIS-NV) are also common and interesting phenomena in speech. However, the relationship between disfluencies and emotions has been largely overlooked in previous Psycholinguistic research. Emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention (Vuilleumier, 2005). This in turn influences speech processing and production, which may result in disfluencies (Lickley, 2015). Current studies on human-human dialogues suggest that disfluency conveys information such as speaker uncertainty (Lickley, 2015), level of conflict (Vidrascu and Devillers, 2005), or points of interest in meetings (Shriberg, 2005). Unlike disfluencies, non-verbal vocalisations, especially laughter, have been identified as universal and basic cues of human emotions (McGettigan et al., 2015). For example, Affective Bursts (short emotional non-speech expressions) have been shown to convey identifiable emotions when presented without context to participants (Schröder, 2003). Cognitive studies have suggested that listeners can perceive affective states such as distress, anxiety, and boredom from non-verbal vocalisations including laughter, cries, sighs and yawns (Russell et al., 2003; Petridis and Pantic, 2008). These indicate that DIS-NV may be predictive of emotions in spoken dialogue in addition to the prosodic cues and lexical content.

2.3 Discussion

In this chapter, we reviewed theoretical studies of human emotions. Our review indicates that speech prosody and lexical content are important cues of human emotions in spoken dialogue. However, factors such as individual differences and nature of the dialogue may influence the relationship between emotions and specific acoustic or lexical cues greatly. Beyond the acoustic characteristics and lexical contents, non-verbal vocalisations are strong indicators of speaker's emotions, and disfluencies are suggested to be related to emotions in dialogue by previous Psycholinguistic studies. In addition to different cues related to emotions in spoken dialogue, previous studies also suggest that emotions are dependent on temporal context. Humans make use of dialogue history and current information from all available modalities to better recognize the emotions of their conversational partner.

In Chapter 3, we will review computational studies on recognizing human emotions in spoken dialogue. We will then identify the misalignment between

the Psycholinguistic knowledge on human emotions in spoken dialogue and state-of-the-art automatic emotion recognition approaches, which motivate the main hypotheses of this thesis.

Chapter 3

Experimental Framework for Automatic Emotion Recognition

I believe in evidence. I believe in observation, measurement, and reasoning, confirmed by independent observers.

— Isaac Asimov, *The Roving Mind (1983)*

In the first part of this chapter, we review current approaches for automatically recognizing emotions in spoken dialogue using acoustic and lexical cues. We identify limitations in state-of-the-art approaches motivated by our review of Psycholinguistic studies of human emotions in Chapter 2. To address the limitation in feature extraction of missing affective cues in spoken dialogue, we propose novel DISfluency and Non-verbal Vocalisation (DIS-NV) features. To address the limitation of combining multimodal information at the same level, we propose the HierarchicaL (HL) fusion method for multimodal emotion recognition. The framework of our proposed emotion recognition model is shown in Figure 3.1.

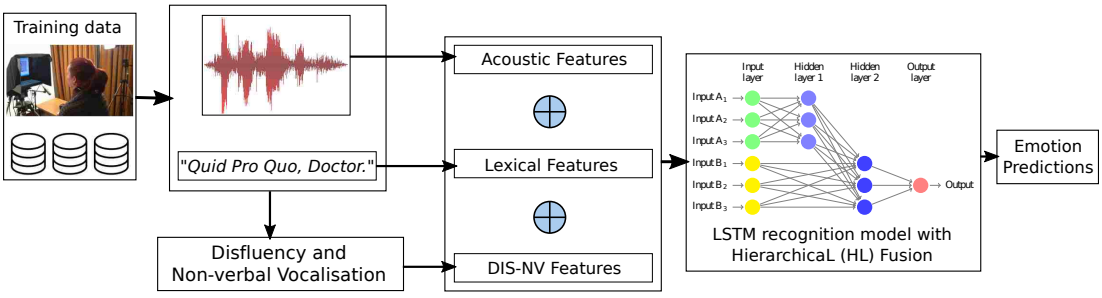


Figure 3.1: Framework of Our Emotion Recognition Model

To study the efficacy of the proposed features and modelling approach, we conduct experiments comparing our DIS-NV features and HL fusion with benchmark features

and emotion recognition models in this thesis. In the second part of this chapter we describe the methodology for our experiments. First, we provide definitions of the four dimensions we use to describe emotions. Second, we describe the two emotion databases of spoken dialogue we conduct our experiments on. Third, we describe five types of benchmark acoustic and lexical features which we use to compare our DIS-NV features with. Finally, we describe two widely used recognition models which we apply to build our emotion recognizers and the evaluation metrics we use.

3.1 Automatic Emotion Recognition in Spoken Dialogue

This section reviews different aspects of automatic emotion recognition in spoken dialogue. These include emotion database collection, acoustic and lexical features for representing the data, and Machine Learning models for recognizing emotions. We then discuss the limitations of state-of-the-art automatic emotion recognition approaches, and attempt to address such limitations motivated by our review of the Psycholinguistic studies of human emotions in Chapter 2.

3.1.1 Emotion Recognition as a Machine Learning Problem

The task of this thesis is to recognize dimensional emotions in spoken dialogue with acoustic and lexical features. We are especially interested in recognizing emotions in spontaneous dialogue because this leads to more robust automatic emotion recognition models that can be applied to more natural scenarios (Schuller et al., 2010a). Current performance of emotion recognition in spontaneous dialogue is disappointing compared to human performance, and there is room for improvement (Sauter et al., 2010b). For example, Zadeh et al. (2017) explored the efficacy of deep neural networks for predicting Valence (positive/negative) of videos. However, even the best performing automatic model using multimodal information is still far from human-like performance. For binary Valence classification, the best accuracy achieved by the automatic model is 77.1%, while human performance is 85.7%. To improve on current performance, the key research challenge is identifying more effective features and recognition models depending on the characteristics of the data set.

There are three main aspects for building an emotion recognition model, namely the data, the features, and the recognition model. Here we summarize major approaches

for each aspect in current work:

- Data:
  - Spontaneous: data collection by recording spontaneous dialogue.
  - Induced: data collection by using stimuli to evoke target emotions in the participants.
  - Acted: data collection by acting.
- Feature:
  - Knowledge-inspired: features describing affective cues identified by Psychological studies on human emotions.
  - Statistical: features describing statistical properties of the data.
- Model:
  - From a temporal point of view:
    - Non-contextual: models using information only from the current time step.
    - Contextual: models including temporal context either in the features or in the model structure.
  - From a structural point of view:
    - Shallow: models with a flat structure that uses the input feature representation directly.
    - Deep: models with a hierarchical structure that learns abstraction of the input features before performing recognition.
  - From a modality point of view:
    - Unimodal: models using information from a single modality.
    - Multimodal: models combining information from multiple modalities.

In the remainder of this section, we review state-of-the-art research on these different emotion recognition aspects. In addition, we discuss what are the suitable evaluation metrics for comparing different emotion recognition approaches.

3.1.2 Databases

A non-exhaustive list of recent emotion databases widely used for automatic emotion recognition is given in Appendix A. In general, in recent years there has been increasing interest in collecting multimodal emotion databases of spontaneous dialogue with both categorical and dimensional emotion annotations. This trend has been observed in various emotion database literature reviews, such as D'mello and Kory (2015). There have also been growing efforts in collecting benchmark emotion databases for hosting emotion recognition challenges where researchers can compare their models under the same experimental settings.

A large portion of existing emotion databases were collected in an acted or induced manner, especially unimodal databases. For example, the Acted Facial Expressions in the Wild (AFEW) database (Dhall et al., 2012) of movie clips, which has been used in the annual Emotion recognition in the Wild (EmotiW) challenge since 2013 (Dhall et al., 2017). Recently, more effort has been placed on collecting spontaneous dialogue in a more natural and realistic environment. For example, the Sentiment Analysis in the Wild (SEWA) database[1] contains dyadic conversations between acquainted people discussing a commercial they just viewed, which has been used in the AVEC2017 challenge (Ringeval et al., 2017). However, databases collected by recording spontaneous dialogue often contain unbalanced distributions of emotion labels: a large number of instances have neutral or mild emotion labels and a relatively small number of instances capture intense emotions. Emotion databases, especially those collected by recording spontaneous dialogue, are often small in size with a limited number of participants. For example, the SEMAINE database contains approximately 10 hours of audiovisual recording from 24 participants talking to virtual agents (McKeown et al., 2010).

The audio and visual modalities are the two most frequent modalities captured in emotion data collection. In more recent databases, other modalities, such as gestures and physiological signals, have also been recorded due to the development of wearable sensors. For example, the LIRIS-ACCEDE database records the physiological and behavioural measurements of movie audiences during movie screening (Baveye et al., 2015b).

For emotion annotation, the Big-6 emotion categorisation of Ekman et al. (1987) remains dominant in the field, while in recent databases dimensional emotions of

[1] https://sewaproject.eu/

Ortony et al. (1990) are often annotated in addition to the categorical emotions. Among different emotion dimensions, the Arousal (excited/bored) and Valence (positive/negative) emotion dimensions are the mostly widely annotated. For dimensional emotion annotation schemes, the emotion values are either annotated as continuous real values, as discrete classes (e.g., high/low), or Likert scales (e.g., 1 to 5 integer scores). In order to address the individual variation in emotion perception, recent work has employed ordinal rankings instead of absolute values to annotate emotions (Yannakakis et al., 2017).

Recall that our interest is recognizing dimensional emotions from spoken dialogue with acoustic and lexical features. Thus, we require emotion databases with audio recordings and transcriptions that are annotated with dimensional emotions. To study the influence of data aspects on different emotion recognition approaches, we need to experiment on multiple databases with different data collection strategies. To compare our study with state-of-the-art, we require databases widely used in current studies. Therefore, in this work, we choose the Audio/Visual Emotion Challenge 2012 (AVEC2012) database of spontaneous dialogue (Schuller et al., 2012) and the Interactive Emotional dyadic MOtion CAPture (IEMOCAP) database of acted dialogue (Busso et al., 2008). The AVEC2012 database contains recordings of participants having social conversations with virtual agents designed with different personality types. The IEMOCAP database contains recordings of pairs of participants acting scripted and improvised scenarios. Details of these databases are provided in Section 3.3.

3.1.3 Features

Features for emotion recognition can be extracted from multiple modalities, such as audio, visual, and physiological. For example, Chen et al. (2016) used peripheral signals and Electro-Encephalo-Gram outputs to recognize Arousal and Valence. However, in this work we focus on the acoustic and lexical modalities, because our task is to recognize emotions from spoken dialogue and these two modalities are the most commonly available in emotion databases of spoken dialogue.

3.1.3.1 Acoustic Features

For the acoustic modality, the majority of state-of-the-art studies have focused on statistical features describing spectral and prosodic characteristics of the speech signal

(Poria et al., 2017). For example, Wang et al. (2015) used frame-level statistical features for predicting Arousal and Valence in music. The most widely used statistical acoustic features are the Low-Level Descriptor (LLD) features (see Section 3.4.1.1), such as the mean of loudness over a 25ms segment of the speech. Various sets of LLD features have been used in previous work. For example, Eyben et al. (2015a) tested the GeMAPS LLD set (Eyben et al., 2015b), the InterSpeech 2009 LLD set (Schuller et al., 2009), and the Interspeech ComParE LLD set (Schuller et al., 2013), and achieved robust performance for predicting Arousal and Valence from speech in real time. Song et al. (2016a) extracted the InterSpeech 2010 LLD feature set (Schuller et al., 2010b) which achieved an average recognition rate of 52% for cross-domain categorical emotion recognition in their cross-corpora study. These results indicate that the LLD features are robust benchmark features for emotion recognition tasks and can be used as a baseline for studying performance of novel features.

Besides using the LLD features directly, previous studies also experimented on deriving more abstract feature representations from the LLD features. For example, log-Gabor filters were extracted from a self-defined subset of LLD features to improve the performance of categorical emotion recognition by Gu et al. (2015). Pokorny et al. (2015) transformed the InterSpeech 2009 LLD set (Schuller et al., 2009) to a bag-of-audio-words representation and their experiments show that these features can outperform the LLD features on predicting binary Valence.

Although the statistical LLD features have achieved good performance in state-of-the-art emotion recognition studies, the high dimensionality of the LLD feature set constrains the computational effectiveness of the emotion recognition model. These speech signal based features may also include variations unrelated to emotion. For example, the frame-level F0 envelope included in LLD features is largely influenced by different pronunciations of the speech content instead of the emotions expressed in speech. Thus, Eyben et al. (2015b) proposed a smaller set of Paralinguistic knowledge-inspired LLD features (the expanded Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)). Their cross-corpora experiments show that the eGeMAPS features have comparable or better performance than various statistical LLD feature sets containing thousands of features for predicting binary Arousal and Valence. Similarly, Bone et al. (2014) proposed three utterance-level speech prosody features (the Global Prosody features) which were shown to be more predictive than the statistical InterSpeech 2011 LLD feature set (Schuller et al., 2011) on predicting continuous Arousal values in multiple databases. Lefter et al. (2015) compared the

InterSpeech 2009 LLD set (Schuller et al., 2009) with 31 hand-crafted utterance-level acoustic features (e.g., duration, pitch, intensity, harmonics to noise ratio) on recognizing negative incidents in dialogue. Their cross-corpora experiments showed that the knowledge-inspired features achieved comparable or better performance. This indicates that knowledge-inspired acoustic features may achieve better performance than statistical acoustic features for emotion recognition.

To study the efficacy of our proposed emotion recognition approaches, we select the InterSpeech 2010 LLD feature set (Schuller et al., 2010b), the eGeMAPS features (Eyben et al., 2015b), and the Global Prosody features of Bone et al. (2014) as benchmark acoustic features to compare our approaches with. In Section 3.4.1, we describe the computation of the LLD features (Section 3.4.1.1), the eGeMAPS features (Section 3.4.1.2), and the Global Prosody features (Section 3.4.1.3) in detail.

3.1.3.2 Lexical Features

Sentiment analysis is a large subfield of Natural Language Processing. Various lexical features have been extracted to analyze opinions and affects in text. For example, word-level features representing the presence of key words such as negations or topic-specific words, or sentence-level syntax-based features to represent dependencies in the text (Yadollahi et al., 2017). However, lexical features are not as widely used as the acoustic features for emotion recognition in spoken dialogue. The majority of previous work has focused on frequency-based lexical features which describe the text as a vector of the frequency of dictionary items being repeated in the text, i.e., the sparse bag-of-words style features describing the lexical content of the speech. For example, Nazari et al. (2015) used features based on the Linguistic Inquiry and Word Count (LIWC) emotion lexicon (Pennebaker et al., 2007) to predict High vs. Low Machiavellian personalities of the speaker in a negotiation situation. Gievska et al. (2015) used lexical features based on the WordNetAffect dictionary (Strapparava and Valitutti, 2004) to detect categorical emotions in dialogue.

Although lexical features are less frequently used than acoustic features, previous work on emotion recognition in spontaneous dialogue has shown that bag-of-words style lexical features based on the correlations between words and binary emotion dimensions (the Point-wise Mutual Information (PMI) features) are more predictive of emotions than the LLD acoustic features. These lexical features achieved good emotion recognition performance both when used on their own and in combination with acoustic features (Savran et al., 2012). These indicate that lexical features are

predictive of emotions in spoken dialogue, and incorporating them with acoustic features may bring improvements to emotion recognition.

One issue with current lexical features used for emotion recognition is that they focus only on the lexical content. Both the lexical and acoustic features used in state-of-the-art research on recognizing emotions in spoken dialogue have focused on speech in isolation, while specific characteristics of spoken dialogue compared to other forms of speech (e.g., monologue) are overlooked. As discussed in Section 2.2, dialogue phenomena, especially disfluencies and non-verbal vocalisations, are cues of emotions in spoken dialogue. Therefore, features describing disfluencies and non-verbal vocalisations in speech may be predictive of emotions, and may bring additional benefits when incorporated with other acoustic and lexical features.

To study the efficacy of our proposed emotion recognition approaches, we examine PMI features (see Section 3.4.2.1) and Crowd-Sourced Annotation (CSA) features based on the emotion lexicon dictionary of Warriner et al. (2013) (see Section 3.4.2.2) as benchmark lexical features to compare our approaches with. In Section 3.4.2, we describe the computation of these benchmark lexical features.

Besides identifying predictive features, feature engineering is also important for developing accurate emotion recognition models. For example, multi-scale Gaussian kernels with a Fisher discriminant embedding graph were used to reduce dimensionality of the InterSpeech 2010 LLD feature set (Schuller et al., 2010b) in the work of Xu et al. (2015). Canonical Correlation Analysis (Kaya et al., 2014) and Correlation-based Feature-subset Selection (Hall, 1998) are also widely used feature dimension reduction methods in previous work. In these studies, significantly better performance has been achieved after feature engineering compared to using the original feature set.

3.1.4 Recognition Models

Beyond features, emotion recognition performance also depends on how those features are modelled. Here we review the main Machine Learning models that have been applied to emotion recognition.

3.1.4.1 Shallow Machine Learning Models

To build emotion recognition models, most widely used classification or regression models have been applied. For example, Pokorny et al. (2015) built a Naive Bayes

model for predicting binary Valence in dialogue; Huang and Epps (2016) built a Gaussian Mixture Model to track emotion changes in dialogue. Among previous studies, Support Vector Machines (SVMs) have been the most widely used shallow learning model for emotion recognition (e.g., Yang et al. (2017); Huang et al. (2015a); Lotfian and Busso (2015); Stolar et al. (2015)). Although many different algorithms exist and it is important to choose the appropriate one for a specific task, Forbes-Riley and Litman (2011) compared various shallow learning algorithms and suggested that performance differences were not significant when accounting for feature sets and other parameter settings. They reported an average recall of 57.8% using a linear logistic regression model for binary classification of student certainty/uncertainty in an Intelligent Tutoring System.

Most shallow emotion recognition models in previous work do not include temporal context information. However, Psycholinguistic studies suggest that the emotional state of a person during conversations tends not to change rapidly and thus depends on the temporal context. Similarly, models that included temporal context, in either the features extracted (Moore et al., 2014) or the recognition model used (e.g., Hidden Markov Model in Ozkan et al. (2012) or Particle Filtering in Savran et al. (2012)), have shown significantly better performance than non-contextual models in emotion recognition.

3.1.4.2 Deep Neural Network Models

In recent years, significant performance improvements have been obtained using neural network models with multiple hidden layers (i.e., deep learning models) in emotion recognition. The network structure of deep learning models allows flexible control when fusing multiple modalities and including temporal context. This enables the models to abstract more effective features automatically. For example, using features abstracted by Deep Belief Networks as inputs to a SVM model yields improved performance on predicting Arousal and Valence compared to using the features directly (Xia and Liu, 2015). Studies comparing Convolutional Neural Networks (CNNs) with SVMs showed that the CNN achieved significantly better performance on recognizing both dimensional and categorical emotions (Baveye et al., 2015a; Zheng et al., 2015). The CNN model also has the ability to automatically extract features from raw speech signals (Pini et al., 2017; Trigeorgis et al., 2016; Bertero et al., 2016), although the performance of such auto-learned features is still limited: Trigeorgis et al. (2016) achieved a Concordance Correlation Coefficient of 0.686 on Arousal

regression, 0.261 on Valence regression, for multimodal emotion recognition on the RECOLA database (Ringeval et al., 2013), which contains French dyadic spoken dialogue of web-based collaborative problem solving; Fung et al. (2016) achieved an average accuracy of 65.7% on recognizing categorical emotions of over 200 hours of TED talks. Banda et al. (2015) compared Recurrent Neural network (RNN) with Support Vector Regression models and showed that the RNN model performed better in predicting Arousal and Valence on the AVEC2012 database. Deep and hierarchical neural networks are also commonly used for emotion recognition in previous work (Cardinal et al., 2015) and Brueckner and Schuller (2015) have obtained the best reported results in detecting the Valence emotion dimension and level of conflict in a database of broadcasted Swiss political debates in French (average recall of 80.2%).

Among different deep learning models, the Long Short-Term Memory Recurrent Neural Network (LSTM) model is especially powerful for emotion recognition because of its ability to model long-range temporal context (Eyben et al., 2015a). Wei et al. (2014) applied the LSTM model to learn more effective features, and then applied Support Vector Regression over the outputs of the LSTM model for recognizing dimensional emotions in spontaneous dialogue. The LSTM model was also used directly for classification in previous work and obtained better performance than Hidden Markov Models for recognizing dimensional emotions in acted dialogue (Wöllmer et al., 2010, 2012). The Bidirectional-LSTM (BLSTM), a modification of LSTM that includes temporal context from both the past and the future, is also widely used in state-of-the-art emotion recognition studies (Han et al., 2017; He et al., 2015b). However, Chen and Jin (2015) argued that using BLSTM does not improve performance significantly compared to using standard LSTM in emotion recognition: With respect to emotion recognition on the RECOLA database of French dyadic spoken dialogue (Ringeval et al., 2013), for Arousal regression, BLSTM achieved correlation coefficient of 0.816 while LSTM achieved correlation coefficient of 0.810; For Valence regression, BLSTM achieved correlation coefficient of 0.573 while LSTM achieved correlation coefficient of 0.564. To improve performance, previous work has focused on identifying more predictive feature representations to input to the LSTM model (Chao et al., 2015) and stacking other Machine Learning models on top of the LSTM model (Lee and Tashev, 2015).

One issue with using deep learning models for emotion recognition is that the small size of emotion databases compared to databases used for speech or image recognition tasks may limit optimization of the complex model structure of a deep learning

model (Baveye et al., 2015a). The ability to generalize over different databases is also an issue for current deep learning models. To address the issue of insufficient training data in emotion recognition, previous work studied semi-supervised and unsupervised methods for emotion recognition. For example, Abdelwahab and Busso (2015) used domain adaptation method to adjust an emotion recognition model trained on non-scripted acted dialogue to fit spontaneous dialogue; Pini et al. (2017) used SoundNet CNN (Aytar et al., 2016) pre-trained with the eNTERFACE database (Martin et al., 2006) of acted emotional scenarios to extract audio features directly from the raw signal for predicting perceived emotions of movie clips; In the semi-supervised emotion recognition study of Zhang et al. (2016b), the model trained with manually labelled data at the supervised learning phase is used to re-evaluate the auto-labelled data provided by the unsupervised learning phase to correct possibly mislabelled data and enhance the overall confidence of the system's predictions. Compared to supervised emotion recognition, there are fewer studies on semi-supervised and unsupervised emotion recognition. In this thesis, we will focus on supervised emotion recognition approaches. This is because most previous studies on the two emotion databases we conduct our experiments on applied supervised methods for emotion recognition.

3.1.5 Modality Fusion

Consistent with human studies, multimodal models combining information from different modalities have consistently provided better performance than unimodal models in previous work of emotion recognition (e.g., Savran et al. (2012); Wei et al. (2014); Poria et al. (2017)). Modality fusion has received increasing attention in various recognition tasks. For example, speech recognition (Sun et al., 2016), video classification (Liu et al., 2016a), and image summarization (Camargo and González, 2016). There are mainly two types of fusion strategy for building multimodal models in state-of-the-art research, namely Feature-Level (FL) fusion (also known as "early fusion") which combines features from different modalities before performing recognition, and Decision-Level (DL) fusion (also known as "late fusion") which combines the predictions and their probabilities given by each unimodal model for the multimodal model to make the final decision (D'Mello and Kory, 2012).

FL and DL fusion are also the most widely used fusion strategies in multimodal emotion recognition. For example, Vielzeuf et al. (2017) combined the InterSpeech

2010 LLD acoustic feature set (Schuller et al., 2010b), VGG facial features (Parkhi et al., 2015), and temporal visual features at the decision level to predict the big-6 emotion categories of movie clips; Banda et al. (2015) combined acoustic and visual features at the feature level to detect Arousal and Valence in dialogue; Gao et al. (2016) combined prosodic and spectral features at the feature level to detect categorical emotion; Pei et al. (2015) combined acoustic and visual features at the decision level to predict Arousal, Power and Valence in monologue. Pre and post processing are sometimes applied to the FL or DL models to improve their performance (Kächele et al., 2015). For example, Gievska et al. (2015) combined acoustic and lexical features at the feature level after feature engineering to detect categorical emotions in dialogue. Previous studies comparing the FL and DL fusion have shown that DL fusion typically outperforms FL fusion, although when the features are highly effective on their own, DL fusion can result in worse performance than FL fusion. This is due to loss of feature level information at the final decision step of DL fusion and overlooking interactions between features from different modalities (Huang et al., 2015b; Jin et al., 2015; He et al., 2015a).

Both FL and DL fusion incorporate modalities at the same level. However, as discussed in Section 3.1.3, different features may describe data at different time scales or levels of abstraction. When perceiving emotions, humans make use of information from different modalities at different cognitive levels and time steps (Grandjean et al., 2008). This may be the reason that the improvements given by modality fusion are often limited in emotion recognition (D'Mello and Kory, 2012; Poria et al., 2017). Combining modalities at different levels is extremely rare in previous work. To the best of our knowledge, there are only three previous studies that performed multimodal emotion recognition in a hierarchical manner: Chen and Jin (2015) used features from the audio, visual and physiological modalities in different layers of a LSTM model for recognizing frame level continuous Arousal and Valence values from French dialogue. Performance of their hierarchical model is better than FL fusion, but worse than DL fusion; Wu et al. (2015) combined LLD features extracted at different time scales at the decision level to detect categorical emotions in dialogue; Kim et al. (2015) developed a logistic regression model incorporating features derived from prosody, spectral envelope, and glottal information in a tree structure, and this model outperformed a logistic regression model using all acoustic features at the input level. These indicate that a fusion strategy that has a hierarchical structure which can model both inter- and intra-modality differences may increase the gain of modality fusion.

3.1.6 Evaluation Metrics

For categorical emotions and dimensional emotions described with discrete scores, emotion recognition is essentially a classification task. For categorical emotions with intensity levels and dimensional emotions described with continuous values, emotion recognition is essentially a regression task. To evaluate performance of different emotion recognition approaches, metrics widely used in other machine learning tasks are applied to emotion recognition as well.

Common evaluation metrics for classification tasks are also applied to emotion classification, i.e., accuracy, recall, and F1-measure. The majority of previous work report accuracy for emotion classification (e.g., Pokorny et al. (2015); Xu et al. (2015); Xia and Liu (2015)), while some also use recall for the evaluation metric (e.g., Eyben et al. (2015b)). F1-measure is only reported in rare cases (e.g., Monkaresi et al. (2012); Koelstra et al. (2012)). However, in emotion databases, particularly for spontaneous dialogue, emotion annotation often suffers from the unbalanced class issue: intense emotions (e.g., fear or Arousal scoring 1 in a 1 to 5 integer score annotation scheme) are rare compared to the neutral state or mild emotions. When using only accuracy or recall for the evaluation metric, a model may be shown to have good performance when it only successfully predicts the majority classes, but fails on the smaller yet still important emotion classes. Therefore, F1-measure which combines accuracy and recall may be a better evaluation metric for emotion classification with unbalanced data.

For emotion regression, most previous work used correlation-based evaluation metrics. For example, the 2012 AudioVisual Emotion Challenge (Schuller et al., 2012) used Correlation Coefficient Score (CCS) to evaluate the results, which is calculated as the average of Pearson's Correlation Coefficient on each session of the test data partition. The 2016 AudioVisual Emotion Challenge (Valstar et al., 2016) used Concordance Correlation Coefficient (Lawrence and Lin, 1989), which is a common evaluation metric in Physiological and Neuroscience studies of emotions that combines correlation with value shifting errors (e.g., Ringeval et al. (2015)). Other regression evaluation metrics, such as the Mean Square Error, have also been reported in addition to the correlation-based metrics in previous studies (e.g., Chen and Jin (2015)).

To evaluate the performance difference between different approaches, the majority of previous studies compared results of different approaches directly without significance testing (e.g., Monkaresi et al. (2012)). For those who performed

significance tests, in emotion classification, the paired t-test (Snedecor and Cochran, 1989) is the most commonly reported (e.g., Soleymani et al. (2012b)), while some report the mean and standard deviation of results on each fold of cross-validation experiments (e.g., Metallinou et al. (2012)); In emotion regression, the z-test (Snedecor and Cochran, 1989) is the most commonly reported significance test (e.g., Wöllmer et al. (2013)).

3.1.7 Summary

Our review of state-of-the-art automatic emotion recognition shows that for the feature aspect, current studies have focused on using statistical acoustic features derived directly from the speech signal and sparse lexical features describing the speech content. However, knowledge-inspired features describing affective cues in dialogue suggested by Psycholinguistics studies are equally or more predictive than statistical features for recognizing categorical and dimensional emotions. For the model aspect, consistent with human studies, automatic emotion models that include temporal contexts and combine multiple modalities typically yield better performance than non-contextual and unimodal emotion recognition models. However, unlike humans, multimodal emotion recognition models incorporate information at the same level, which may have limited their performance. Because strong emotions are relatively rare in emotion databases, we find that the F1-measure is more suitable to use in evaluating classification experiments, as it takes both accuracy and recall into account. For regression experiments, both the correlation and the value shifts between predictions and annotations need to be considered. In the remainder of this chapter, we provide methodology of our experiments based on our review of current emotion recognition studies.

3.2 Emotion Dimensions

As discussed in Section 2.1, in this work, we define emotions as vectors in a multidimensional space following the cognitive emotion theory. We use four common emotional dimensions which have been identified as being able to describe most everyday human emotions (Fontaine et al., 2007), namely Arousal, Expectancy, Power, and Valence:

- The **Arousal** dimension describes whether the speaker feels excited or bored

- The **Expectancy** dimension describes whether the speaker feels certain or uncertain towards the discussion
- The **Power** dimension describes whether the speaker feels that (s)he dominates the conversation or (s)he is being dominated
- The **Valence** dimension describes whether the speaker has positive feelings or negative feelings towards the discussion

Note that values on each dimension can either be continuous real numbers or discrete scores or categories.

3.3 Emotion Databases

In this section, we describe the Audio/Visual Emotion Challenge 2012 (AVEC2012) database of spontaneous dialogue (Schuller et al., 2012) and the Interactive Emotional dyadic MOtion CAPture (IEMOCAP) database of acted dialogue (Busso et al., 2008), which we use for our experiments. As discussed in Section 3.1.2, we choose these databases because they are the most widely used emotion databases of English spoken dialogue with dimensional emotion annotations. Their different data collection strategies allow us to study how different aspects of the data, especially dialogue type, influence performance of emotion recognition models.

Note that the AVEC2012 and IEMOCAP databases have different data collection and emotion annotation schemes. The AVEC2012 database contains spontaneous dialogue, while the IEMOCAP database contains acted dialogue. The AVEC2012 database originally annotated Arousal, Expectancy, Power and Valence at the word-level as real values. The IEMOCAP database originally annotated Arousal, Power and Valence at the utterance-level as integer scores ranging from 1 to 5. Thus, performance of emotion recognition models on these two databases should be discussed separately.

3.3.1 The Audio/Visual Emotion Challenge 2012 (AVEC2012) Database

The AVEC2012 database (Schuller et al., 2012) contains the Solid-SAL (Sensitive Artificial Listener) part of the SEMAINE (Sustained Emotionally coloured

Machine-human Interaction using Nonverbal Expression) corpus (McKeown et al., 2010). It includes approximately 8 hours of audiovisual recordings of 24 participants conversing with 4 on-screen characters with specific personalities role-played by human operators. These virtual agents include Poppy who is always cheerful, Spike who is always aggressive, Obadiah who is always pessimistic, and Prudence who is always calm. Each dialogue session is approximately 5 minutes long and the participants are free to discuss any topic. Manual transcriptions are provided with word timings in the AVEC2012 database.

Emotions in the AVEC2012 database were annotated as real-value vectors in the Arousal-Expectancy-Power-Valence emotional space with value range of [−1,+1]. Annotations were provided at both the word-level and the frame-level. In this thesis, we use the word-level emotion annotations. There are 49,874 data instances in total in the word-level AVEC2012 database.

The inter-annotator agreement of emotion annotations has been an important issue in many spontaneous emotion databases including the AVEC2012 database. Perception of emotions can vary greatly due to individual differences. Current studies work on collecting reliable emotion annotations by having multiple annotators. One widely used measurement of inter-annotator agreement is the Chronbach's Alpha. It is based on the average inter-correlations of the item-pairs within the compared annotations. A higher Chronbach's Alpha represents a stronger inter-annotator agreement. A Chronbach's Alpha of $a < 0.5$ is often interpreted as unacceptable inter-annotator agreement, while $a > 0.7$ is often interpreted as acceptable inter-annotator agreement, with $a > 0.8$ being a desired good inter-annotator agreement. The inter-annotator agreement of the AVEC2012 database measured by the Chronbach's Alpha is shown in Table 3.1 (Nenkova, 2013). As we can see, emotion annotation is a challenging task, especially for dimensional emotions. The data annotated by 6 or 8 annotators is more reliable than the data annotated by only 2 annotators. This is consistent with the work of Vlasenko and Wendemuth (2015) which suggests that it is better to have multiple annotators when annotating an emotion database. To obtain emotion annotations from more annotators and increase the annotator agreement, crowd-sourced annotation has been employed in recent studies, such as the work of Burmania et al. (2016). In addition to averaging over annotations provided by multiple annotators, deriving ranking labels from absolute emotional labels can be used to increase reliability of emotion annotations (Lotfian and Busso, 2016). Using discrete scores as values on the emotion dimensions has

also shown to increase the inter-annotator agreement and is commonly used in current studies (e.g., Busso et al. (2008)).

Table 3.1: Inter-Annotator Agreement on the AVEC2012 Database (Nenkova, 2013)

Chronbach's Alpha	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)
2 Annotators				
a<0.5	28.6	64.3	28.6	35.7
a>0.7	28.6	7.1	35.7	35.7
a>0.8	0.0	0.0	14.3	28.6
6 Annotators				
a<0.5	25.8	29.0	6.5	9.7
a>0.7	51.6	32.3	87.1	77.4
a>0.8	22.5	9.7	38.7	45.2
8 Annotators				
a<0.5	8.3	8.3	0.0	0.0
a>0.7	75.0	41.7	66.7	83.3
a>0.8	41.7	33.3	58.3	66.7

3.3.2 The Interactive Emotional dyadic MOtion CAPture (IEMOCAP) Database

The IEMOCAP database (Busso et al., 2008) contains approximately 12 hours of audiovisual recordings from 5 mixed gender pairs of actors and actresses. Each conversation is approximately 5 minutes long. There are two types of dialogue collected in the IEMOCAP database, namely non-scripted dialogue and scripted dialogue. When collecting the non-scripted dialogue, the participants were instructed to improvise hypothetical scenarios without any scripts, such as someone telling his/her best friend that (s)he is getting married. These scenarios are designed to induce different target emotion categories, including neutral, anger, frustration, happiness, and sadness (See Table 1 of Busso et al. (2008) for a detailed description of each scenario). In contrast, when collecting the scripted dialogue, the participants follow pre-written lines. The target emotions of the scripts are the same with those for the non-scripted scenarios. Manual transcriptions are provided with utterance and word timings in the IEMOCAP database.

Both categorical and dimensional emotions were annotated in the IEMOCAP database. In this thesis, we only use the dimensional emotion annotations, which were annotated at the utterance-level with a 1 to 5 integer score on the Arousal, Power, and Valence dimensions. Note that the Expectancy dimension was not annotated in the IEMOCAP database. Emotion annotation was conducted by at least 6 annotators to achieve reliable annotator agreement. The inter-annotator agreement on the IEMOCAP database is shown in Table 3.2 (Busso et al., 2008). For Arousal, annotators agreed or were one point apart in their ratings of 85% of the utterances labelled. For Valence, annotators agreed or were one point apart in their ratings of 94% of the utterances labelled (Wöllmer et al., 2010). The average score over all the annotators was used as the gold-standard emotion annotation. There are 10,037 data instances in total in the IEMOCAP database.

Table 3.2: Inter-Annotator Agreement of IEMOCAP Database (Busso et al., 2008)

Chronbach's Alpha	Arousal	Power	Valence
Entire database	0.607	0.608	0.809
Scripted acting	0.602	0.663	0.783
Non-scripted acting	0.612	0.526	0.820

3.4 Features

Based on our review of features used in state-of-the-art emotion recognition in Section 3.1.3, we select three types of acoustic features and two types of lexical features as benchmark feature sets to compare with the proposed DIS-NV features. In this section, we describe how these benchmark features are computed on the AVEC2012 and the IEMOCAP databases.

3.4.1 Acoustic Features

The benchmark acoustic feature sets we use in this work include the statistical, frame-level Low-Level Descriptor (LLD) features, the knowledge-inspired, frame-level Expanded Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features, and the knowledge-inspired, utterance-level Global Prosodic (GP) features.

The LLD feature set contains over a thousand features describing detailed acoustic characteristics of the speech audio at the frame level. As discussed in Section 3.1.3,

the LLD feature set is widely used for emotion recognition with robust performance. However, part of the detailed information provided by the LLD features may be unrelated to emotions. For example, the frame-level F0 envelope is largely influenced by different pronunciations of the speech content instead of the emotions expressed in speech. Thus, the eGeMAPS feature set was proposed as a hand-tailored LLD set selected by correlation experiments and findings in studies of Paralinguistic phenomena, such as continuous voiced regions per second. The eGeMAPS feature set contains less than a hundred features. Both the LLD and the eGeMAPS features are based on frame level measurements. However, as discussed in Section 2.2, emotions are stable over a time interval longer than a single frame. Thus, Bone et al. (2014) proposed three utterance level Global Prosodic (GP) features (see Section 3.4.1.3), as the features which are the most related to emotions according to Paralinguistic studies.

In terms of feature abstraction level, the GP and the eGeMAPS features are knowledge-inspired and contain information specific to emotions. Thus, they have higher level of abstraction. The LLD features are statistical and data-driven, and include information unrelated to emotions. Thus, they have lower level of abstraction. In terms of the time scale at which the features are extracted, the GP features describe data at the utterance level and, thus, are calculated over a longer time scale. The LLD and the eGeMAPS features describe data mainly at frame level and have smaller time scale.

In the remainder of Section 3.4.1 we describe the computation of the LLD, eGeMAPS, and GP acoustic features.

3.4.1.1 Low-Level Descriptor (LLD) Features

LLD features are statistical features extracted using a frame-level (e.g., length of 25ms) sliding window over the speech audio. Feature values are calculated as functionals (e.g., mean) applied to LLDs (e.g., energy and spectral descriptors) and their corresponding delta coefficients. We use the OpenSMILE toolbox (Eyben et al., 2010b) to extract LLD features from audio recordings automatically.

For a reference set on emotion regression experiments with the AVEC2012 database in Chapter 4, we use the AVEC2012 challenge baseline LLD set containing 1842 features (Schuller et al., 2012). The AVEC2012 baseline LLD set is computed over word segments with the 25ms sliding window. These features include 42 functionals applied to 25 energy/spectral LLDs (e.g., loudness), 19 functionals applied to 25 delta coefficients of the energy/spectral LLDs, 32 functionals applied to 6

voicing related LLDs (e.g., jitter), 19 functionals applied to 6 delta coefficients of the voicing related LLDs, and 10 voiced/unvoiced durational features. Lists of LLDs and functionals can be found in Table B.1 and Table B.2 of Appendix B.

For a reference set on the IEMOCAP database, we use the InterSpeech2010 Paralinguistic Challenge (IS10) feature set (Schuller et al., 2010b), because this is a widely used benchmark set for emotion recognition in spoken dialogue. The IS10 LLD set is computed over segments of utterances with the 25ms sliding window. This feature set contains 1582 LLD features: 21 functionals applied to 34 LLDs (e.g., logarithmic power of Mel-frequency bands) with their corresponding delta coefficients, 19 functionals applied to 4 pitch-based LLDs (e.g., envelope of the smoothed F0 contour) and their corresponding delta coefficients, the number of pitch onsets (pseudo syllables) and the total duration of the input. Lists of LLDs and functionals can be found in Table B.3 and Table B.4 of Appendix B.

The reason we use the AVEC2012 challenge baseline LLD set containing 1842 features (Schuller et al., 2012) for emotion regression experiments in Chapter 4 is to compare our results with the AVEC2012 challenge results. For emotion classification experiments in Chapters 5 and 6, we use the InterSpeech 2010 LLD feature set (Schuller et al., 2010b) (IS10) for both the AVEC2012 and the IEMOCAP databases in order to make fair comparisons. Our Experiments 5 and 6 in Chapter 5 show that for classification experiments on the AVEC2012 database, performance of the IS10 LLD set is either better or has no significant difference compared to the AVEC2012 baseline LLD set.

3.4.1.2 Expanded Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) Features

The eGeMAPS features are frame-level, knowledge-inspired features. The eGeMAPS feature set contains LLD features that have been suggested as the most related to emotions by Paralinguistic studies (Eyben et al., 2015b). For example, only the 1-4 Mel-frequency cepstral coefficients (MFCCs) are used in the eGeMAPS feature set, while the AVEC2012 baseline LLD feature set contains MFCCs 1-10 and the IS10 LLD feature set contains MFCCs 0-14. Cross-corpora studies indicated that the eGeMAPS features have comparable or better performance than the large LLD feature sets while reducing the feature dimensionality greatly (Eyben et al., 2015b). The 88 eGeMAPS features include the arithmetic mean and coefficient of variation of 25 LLD, the arithmetic mean and coefficient of variation of the spectral flux and MFCC 1-4 in

voiced segments only, 8 functionals applied to pitch and loudness, 5 functionals applied to unvoiced segments, the equivalent sound level, and 6 temporal features (numbers of loudness peaks and continuous voiced regions per second, mean length and standard deviation of continuous voiced and unvoiced region). Lists of LLDs and functionals can be found in Table B.5 and Table B.6 of Appendix B. We use the OpenSMILE toolbox (Eyben et al., 2010b) to extract eGeMAPS features automatically from audio recordings.

3.4.1.3 Global Prosodic (GP) Features

The Global Prosodic (GP) features are utterance-level, knowledge-inspired prosodic features based on the work of Bone et al. (2014). These include three features: median of log pitch, intensity, and voice quality (HF500) over the utterance. The acoustic measurements are taken using a 25ms sliding window with 15ms overlapping between neighbouring windows. HF500 is a spectral-slope measurement. It is computed as the ratio between the total energy E above 500Hz and the lower-frequency energy (between 80Hz and 500Hz) in an utterance:

$$HF500 = \sum_{E_t>500}^{t} E_t \Big/ \sum_{80\leq E_t\leq 500}^{t} E_t \tag{3.1}$$

These features were highly predictive of Arousal in previous work on the IEMOCAP database (Bone et al., 2014). Compared to the frame-level LLD and eGeMAPS features, the GP features represent a small set of knowledge-inspired acoustic affective cues at utterance level.

3.4.2 Lexical Features

The reference lexical feature sets we use in this work include the Point-wise Mutual Information (PMI) based features (Savran et al., 2012) and the Crowd-Sourced Annotation (CSA) features (Warriner et al., 2013). Both are utterance-level, knowledge-inspired features. In the remainder of Section 3.4.2 we describe the computation of the PMI and CSA lexical features.

3.4.2.1 Point-wise Mutual Information (PMI) Features

PMI is a widely used measurement for the relationship of words and emotions in semantic analysis. It is based on the frequency of a word w being labelled as an emotion

class c:

$$PMI(c,w) = \log_2\left(\frac{P(c|w)}{P(c)}\right) \tag{3.2}$$

In previous work, 1000 bag-of-words style lexical features based on PMI values were the most predictive features in the AVEC2012 Challenge (Savran et al., 2012). Savran et al. (2012) computed the PMI features by first binarizing the emotion dimensions, followed by computing the PMI value of each word for each binarized emotion dimension. The 500 words with the highest PMI values for the positive and negative classes of each emotion dimension respectively are concatenated into a list of 1000 words for this emotion dimension. Each utterance is then represented as bag-of-words vector using this list of 1000 words.

The PMI features of Savran et al. (2012) is a highly sparse feature representation. To reduce the sparseness and high feature dimensionality, in this work, we propose to use non-sparse PMI features, which are calculated as the total PMI values of all the words in an utterance for each binarized emotion dimension. There are eight non-sparse PMI features for the AVEC2012 database, but only six non-sparse PMI features for the IEMOCAP database because only three emotion dimensions were annotated in the IEMOCAP database. Both the sparse PMI features proposed by Savran et al. (2012) and our non-sparse PMI features are utterance-level, knowledge-inspired features.

The PMI features were the most effective type of features in the AVEC2012 Challenge (Savran et al., 2012). However, the PMI features were calculated from the emotion annotations of the database being experimented on, which results in overfitting of the features and difficulty in predicting unseen data. Thus, we propose to use the CSA features, which are calculated from crowd-sourced emotion annotations of a large vocabulary. Compared to the PMI features, the CSA features are more robust because they are extracted from a non-domain-specific emotion lexicon dictionary instead of the specific database.

3.4.2.2 Crowd-Sourced Annotation (CSA) Features

Note that both sparse and non-sparse PMI features are calculated for specific databases that the emotion recognition task is performed on. Thus, they may not generalize well to unseen data. Therefore, we also compare the PMI features with 63 utterance-level, knowledge-inspired features based on statistics of crowd-sourced annotations of

Arousal, Power, Valence ratings of 13,915 English lemmas (Warriner et al., 2013). To compute the CSA features, we first remove the stop words in each utterance, lemmatize the remaining words, and then search for these lemmas in the entries of the dictionary of Warriner et al. (2013). For each entry, there are 63 statistics calculated over the crowd-sourced emotion ratings (21 for each emotion dimension). Sums of each of the 63 statistics for all the lemmas in each utterance are returned as the CSA feature values. For unseen words, the mean values of each of the statistics over the whole dictionary are used. The reason that we use mean values instead of zero vector for unseen words is to preserve the information of utterance length. Williams and Stevens (1972) conducted an empirical study on the relationship between emotions and aspects of speech, and found that the total duration of an utterance is an important indicator of emotions. We also compared the emotion recognition performance of using zero vector and using mean values for the unseen words when extracting the CSA features. We find that using the mean values instead of the zero vector gives better results.[2]

The reason we chose the dictionary of Warriner et al. (2013) is because compared to the WordNetAffect (Strapparava and Valitutti, 2004) or the LIWC (Pennebaker et al., 2007) emotion lexicon dictionaries used in previous work, the emotion lexicon dictionary of Warriner et al. (2013) has an expanded collection of lexical entries. It is also based on crowd-sourced emotion annotations rather than annotations from a few experts, which we expect will lead to more robust and non-domain specific annotations.

3.4.3 Pre-Processing of Features

Before performing the emotion recognition experiments, we conduct pre-processing of the extracted features, namely speaker normalization and feature synchronization. These are standard processes in state-of-the-art emotion recognition studies.

3.4.3.1 Speaker Normalization

Previous work has found affective similarity in acoustic and lexical expression of emotions across cultures (Chong et al., 2015; Palogiannidi et al., 2015; Sauter et al., 2010b). However, individual variance is common in emotion expression (Sagha et al., 2015). To account for such individual variance and improve emotion recognition performance, it is standard to apply speaker normalization to the features extracted

[2]We compute overall F1-measures of using the LSTM model with the CSA features to predict emotions on the IEMOCAP database. CSA features using mean values for unseen words result in F1 = 47.545%, CSA features using zero vector for unseen words result in F1 = 45.912% ($p << 0.0001$).

before performing emotion recognition. The most widely used normalization method is the z-score speaker normalization, which we also apply to all the features extracted in this thesis. Z-score speaker normalization is calculated as:

$$V_a^{'} = \frac{V_a - \bar{V}_a}{Std_a} \tag{3.3}$$

In Equation 3.3, $V_a^{'}$ is the normalized value of a feature attribute a; V_a is the original value of attribute a; $\bar{V}_a$ is the mean value of attribute a over all the samples extracted from a speaker; Std_a is the standard deviation of attribute a over all the samples extracted from a speaker. The original feature vectors are first grouped by speakers, then the z-score speaker normalization is applied to each group to calculate speaker-normalized feature vectors.

3.4.3.2 Feature Synchronization

Note that the features are extracted at different time scales. We synchronize all features to the time scale of the emotion annotations before performing the emotion recognition experiments, i.e., to word-level on the AVEC2012 database and to utterance-level on the IEMOCAP database. For the frame-level LLD and eGeMAPS features, the LLDs are measured at the frame-level, while the functionals are applied to word segments on the AVEC2012 database, and to utterance segments on the IEMOCAP database. For the utterance-level GP, PMI, and CSA features, if the utterance-level feature vector is $\vec{F}_U$ for utterance U, all the words $w_1, ..., w_n$ within the utterance U will each be given $\vec{F}_U$ to have word-level feature values for experiments on the AVEC2012 database.

3.5 Recognition Models

In this section, we describe the SVM model and the LSTM model as representatives of state-of-the-art emotion recognition models, and provide the parameter settings we use in our experiments.

As discussed in Section 3.1.4.1, the SVM model is a widely used benchmark recognition model which remains the most predictive in many current emotion recognition studies (e.g., Baveye et al. (2015a)). However, as a flat and non-contextual model, it is hard to incorporate history information or abstract more effective features with the SVM model. However, recent studies have shown improved performance of

SVM emotion recognition model by computing contextual features and performing feature engineering (e.g., Xu et al. (2015); Moore et al. (2014)).

As discussed in Section 3.1.4.2, compared to the flat and non-contextual SVM model, recently neural network models, especially the LSTM model, have achieved leading performance in emotion recognition. The multilayer structure of LSTM models enables automatic learning of more effective features, and the memory cell component of the LSTM model allows modelling of long-range temporal context. However, compared to the SVM model, the LSTM model has a more complex structure with a large number of parameters that require a large number of training instances to optimize. Considering the small size of current emotion databases (tens of thousands), the advantages of the LSTM model may be limited compared to a SVM model that incorporates temporal context.

3.5.1 Support Vector Machines (SVM)

The SVM model is a flat and non-sequential recognition model which is used in many classification tasks. The algorithm aims to find a hyperplane that can divide instances from different classes with the largest distance to the nearest instance of any class (maximizing of the margin cost function) (Boser et al., 1992).

Take a data set with binary class labels as an example:

$$\{x_i, y_i\}\ i = 1, \ldots, l, x_i \in \mathbb{R}^n, y_i \in \{-1, +1\} \tag{3.4}$$

An infinite number of potential hyperplanes separating data instances from the two classes exist. A hyperplane can be defined as $w \cdot x + b = 0$, where w is the normal to the hyperplane, and $\|w\|$ is the Euclidean norm of w. The distance between the hyperplane and the origin of the data space is then $\frac{|b|}{\|w\|}$.

The data instances closest to the hyperplane are called the support vectors and the SVM algorithm works by finding the hyperplane with maximum distance to the support vectors. If d_+, d_- are the distances from the support vectors of both classes to the hyperplane, the margin m will be:

$$m = d_+ + d_- = \frac{|1-b|}{\|w\|} + \frac{|-1-b|}{\|w\|} = \frac{2}{\|w\|} \tag{3.5}$$

Thus, the learning process of the SVM algorithm is maximizing $\frac{2}{\|w\|}$ subject to:

$$\begin{aligned} x_i \cdot w + b &\geq +1, \text{ if } y_i = +1 \\ x_i \cdot w + b &\leq -1, \text{ if } y_i = -1 \end{aligned} \tag{3.6}$$

The dot products in Equation 3.6 can be replaced by non-linear kernel functions for non-linear SVM classification, such as a Radial Basis Function (RBF) kernel.

In this work, we build SVM models using the LibSVM (Chang and Lin, 2011) toolbox with the WEKA (Hall et al., 2009) platform. Following previous studies (Savran et al., 2012), for regression experiments we implement the epsilon-SVR model with a linear kernel, for classification experiments we implement the C-SVC model with a RBF kernel. We normalized all features to [−1,+1] before regression or classification and replaced missing values with the mean. Grid search is used to identify the optimal parameter values, such as the cost parameter of the SVM model. We also use the probability estimation function and the shrinking option of LibSVM to improve the accuracy and efficiency of the SVM models.

3.5.2 Long Short-Term Memory Recurrent Neural Networks (LSTM)

The LSTM model (Hochreiter and Schmidhuber, 1997) is a neural network with multiple hidden layers and a special structure called "the memory cell" that can model long-range temporal context. Compared to conventional Recurrent Neural Networks, the LSTM model is able to learn from a longer history. A hidden layer in a LSTM model is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells. Each memory cell has three multiplicative "gate" units: the input, output, and forget gates. These gates perform the operations of reading, writing, and resetting, respectively. They allow the network to store and retrieve information over long periods of time. The structure of a LSTM memory cell is shown in Figure 3.2 (Schaul et al., 2010). "CEC" in the figure represents the "Constant Error Carousel", which is the central neuron that recycles status information from one time step to the next. The small blue circles with a cross inside indicate multiplicative connections. The peephole connection gives direct access to the central neuron.

The LSTM model stores information at each time step using a cell state vector c and a hidden vector h. They control the updating of states and the outputs. At time step t, in a LSTM memory cell, if input vector is x_t, c_t and h_t are computed as bellow

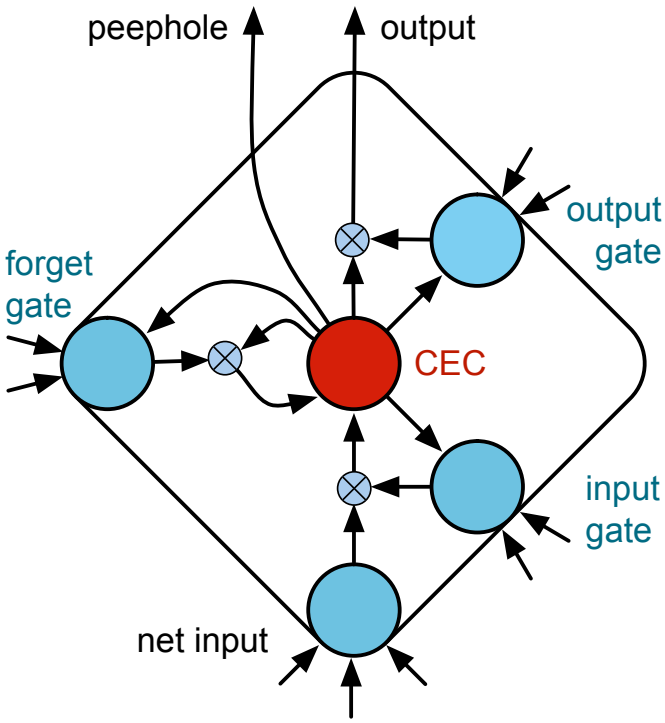


Figure 3.2: Structure of a LSTM Memory Cell (Schaul et al., 2010)

(Hochreiter and Schmidhuber, 1997):

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\ \mathbf{c}_t &= f_t \odot c_{t-1} + i_t \odot \sigma_c(W_c x_t + b_c) \\ \mathbf{h}_t &= o_t \odot \sigma_h(c_t) \end{aligned} \tag{3.7}$$

Where $\odot$ is entry-wise product. W, U, and b are parameter matrices. f_t is the forget gate vector, i_t is the input gate vector, o_t is the output gate vector.

We used the PyBrain (Schaul et al., 2010) toolbox to build the LSTM models for the emotion recognition experiments on the AVEC2012 and the IEMOCAP database.[3] There is one memory cell in each memory block of the LSTM model. We optimize the model structure with cross-validation experiments. The number of neurons in the input layer equals the total number of features used in the LSTM model. Following previous studies (Wöllmer et al., 2012), we used the R-Propagation-Minus (RMSprop) trainer with a learning rate of 10^{-5} during training. In the recurrent layers, the kernel initializer is Glorot uniform, the recurrent initializer is orthogonal, and the bias is

[3] In our experiments in Chapter 7 on the LIRIS-ACCEDE database we use the Keras (Chollet, 2015) toolbox instead.

initialized to zero (Glorot and Bengio, 2010). All data instances are assigned the same weight.

One important limitation of applying deep neural networks, such as LSTM, to emotion recognition is the small amount of training data available, which can cause over-fitting of the deep neural network. There have been several studies addressing this issue. For example, regularisation of the deep neural network. One regularisation technique that has been particularly successful for the LSTM model is the dropout technique. This technique probabilistically excludes input and recurrent connections to LSTM units from activation and weight updates during training (Zaremba et al., 2014). However, the PyBrain toolbox does not provide regularisation of the LSTM model. Thus, we used the early stopping strategy to prevent over-fitting instead.[4]

3.6 Evaluation Metrics

To evaluate performance of different emotion recognition models, for regression experiments on the AVEC2012 database, we follow the AVEC2012 challenge setting and use Correlation Coefficient Score (CCS) (see Section 3.1.6) as the evaluation metric. This enables us to compare our results with other studies working on the same database under the same experimental settings. For classification experiments on the AVEC2012 and IEMOCAP databases, as discussed in Section 3.1.6, using precision or recall alone may not be sufficient to evaluate performance of emotion recognition models due to the class imbalance issue of emotion databases. Thus, we report the F1-measures which combines both precision and recall:

$$\begin{aligned} \text{Precision} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \\ \text{Recall} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \\ \mathbf{F}_1 &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{3.8}$$

We use the weighted average of F1-measures for each emotion class (weighted F-measure) as the evaluation metric for our classification experiments.

In our cross-corpora experiments using both the AVEC2012 and the IEMOCAP database, in order to unite different annotation schemes of these two databases and

[4] In our experiments in Chapter 7 using the Keras toolbox, we applied dropout with the probability of 0.5 to the bottom hidden layer of the LSTM models. This dropout setting has been commonly used in current studies.

to address the class imbalance issue, we transform the original annotations of both databases to three discrete classes $low, medium, high$ on each emotion dimension. For the AVEC2012 database, the continuous annotations are first normalized to [−1,+1], and the value range of each discrete class is: [−1,−0.333) as low, [−0.333,+0.333] as $medium$, (+0.333,+1] as $high$. For the IEMOCAP database that has original annotations with value range [1,5], the value range of each discrete class is: [1,2.333) as low, [2.333,3.667] as $medium$, (3.667,5] as $high$.

Because of different settings in previous work, such as data pre-processing and focusing on different emotion annotations, it is difficult to compare our emotion recognition results to previous results directly. Thus, for experiments on each database, we build emotion recognition models using our proposed features and modelling approaches, and replicate state-of-the-art approaches under the same experimental setting. We compare these models to study the efficacy of the proposed approaches relative to the state-of-the-art approaches. The fact that emotions are annotated at different time scales on the AVEC2012 and the IEMOCAP databases also leads to the issue that results on these two databases have limited comparability. In order to improves the fairness of comparisons between different approaches, we perform 10-fold cross-validation experiments to have more robust results for the classification experiments on these databases. Averaging over cross-validation experiments also helps to reduce effects other than feature or model effectiveness, such as initialization of the LSTM models or outliers in the data.

To evaluate the level of significance in performance difference, in the regression experiments, following Savran et al. (2012), we use two-tailed z-test after Fisher's r-to-z transformation (Raghunathan et al., 1996) to compare the CCS given by each emotion recognition model. In the classification experiments, we use the paired Permutation test (Menke and Martinez, 2004) with 100,000 randomisations to compare predictions made by each emotion recognition model. Compared to the paired t-test used in previous work, the Permutation test is non-parametric. It obtains the p-value from a sample-specific permutation distribution (multiple times of randomization of the original distribution) instead of from a parametric assumption of the distribution. In all our significance tests, we use $p < 0.05$ to reject the null hypothesis, which is that the two emotion recognition models have no difference in their performance.

3.7 Discussion

In this chapter, we reviewed previous studies of automatic emotion recognition. Our survey reveals two issues in current automatic emotion recognition studies: For the feature aspect, non-lexical elements of dialogue such as disfluencies and non-verbal vocalisations are often overlooked when extracting features, yet Psycholinguistic studies suggest they are related to emotions. For the model aspect, when building multimodal models, modalities are often combined at the same level. Existing modality fusion strategies do not take into account the fact that different features may describe data at different time scales or different levels of abstraction. Unlike how humans perceive emotions hierarchically, current multimodal emotion recognition models fail to incorporate both inter- and intra-modality differences. We are motivated to address these issues in state-of-the-art of emotion recognition by using features describing occurrences of disfluency and non-verbal vocalisation in utterances, and by incorporating different information hierarchically in a knowledge-inspired structure. Our hypothesis is that the proposed knowledge-inspired features and model will improve state-of-the-art performance of automatic emotion recognition in spoken dialogue.

To investigate the gain of the proposed approaches, we compare our models with state-of-the-art models under the same experimental settings, as well as recreate state-of-the-art emotion recognition approaches and compare them with the proposed approaches under extended scenarios. We described the experimental setting of this thesis in this chapter. This includes two emotion databases of spoken dialogue (the AVEC2012 and the IEMOCAP databases), three types of benchmark acoustic features (LLD, eGeMAPS, and GP), two types of benchmark lexical features (PMI and CSA), and two state-of-the-art recognition models (SVM and LSTM). In the following chapters, we will present our experiments and discuss our major findings.

Chapter 4

Emotion Recognition with Disfluencies and Non-Verbal Vocalisations

"Oh... I wouldn't remember exactly. Besides I couldn't repeat it. You know how you get when you're excited." His embarrassed laugh was almost a giggle, "I sort of have a tendency to strong language."

— Isaac Asimov, *I, Robot (1950)*

Recall that most previous work on emotion recognition in spoken dialogue used features that describe the acoustic characteristics of the speech signal and its lexical content. However, these studies do not take into account that spoken dialogue contains richer phenomena other than simply uttering a sentence, especially in spontaneous dialogue. Psycholinguistic studies have suggested that speech disfluencies and non-verbal vocalisations are cues of the speaker's emotions in dialogue. Therefore, in this chapter, we present our studies on using features describing occurrences of DISfluency and Non-verbal Vocalisation (DIS-NV) in spoken utterances for emotion recognition in spontaneous dialogue.

In this work, we use the term "turn" to refer to the continuous speech spoken by a speaker without interruption from the other speaker. Note that each speaker turn may contain one or more utterances, and consecutive utterances of a speaker may or may not belong to the same speaker turn. Here our definition of speaker turn focuses on the sentiment and speech production intergrity, which differs from the "turn" under the context of a turn-taking system, which focues on the transitioning between different speakers.

4.1 DISfluencies and Non-verbal Vocalisations (DIS-NVs)

Here we define DIS-NVs and review the Psycholinguistic studies on DIS-NVs and emotions. We also explain our motivation of proposing the DIS-NV features for emotion recognition in spoken dialogue.

4.1.1 Definition of DIS-NVs

Disfluencies are phenomena in speech that "interrupt the flow of speech and do not add propositional content to an utterance" (Fox Tree, 1995). Previous Psycholinguistic studies on disfluencies have focused on disfluencies caused by speech disorders. However, recently disfluencies in normal speech have received increasing attention because they are common and important phenomena in dialogue, and have functions such as turn-holding (Lickley, 2015). Shriberg (2005) also argues that disfluencies are common in spontaneous speech and reflect cognitive aspects of both language production and interaction management. Psycholinguistic studies of spontaneous dialogue have shown that on average, for every 100 words the speaker produces, there are approximately 6 disfluencies (Finlayson, 2014). Speech production has three main phases: conceptualisation, planning, and articulation. Disfluencies can be generated at any of these three phases. For example, when the speaker is organizing responses to a complex question (the conceptualisation step), when the speaker is searching for a suitable word (the planning step), or when the speaker is having a problem pronouncing a syllable (the articulation step).

Non-verbal vocalisations are sounds the speaker produces in utterances other than the verbal content. Disfluencies are sometimes included as types of non-verbal vocalisation (Trouvain and Truong, 2012). However, it is more common to differentiate between disfluencies and non-verbal vocalisations (Finlayson, 2014). There are two main types of non-verbal vocalisations: voice qualifiers (e.g., audible breath or cough) and voice qualifications (e.g., laughter or cry) (Crystal, 1976). A cross corpora study comparing six different corpora showed that among different types of non-verbal vocalisations, laughter and audible breath are the most frequent in spontaneous dialogue (Trouvain and Truong, 2012).

4.1.2 DIS-NVs and Emotions

The relationship between disfluencies and emotions has been largely overlooked in previous Psycholinguistic research. However, emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention (Vuilleumier, 2005). This in turn influences speech processing and production, especially the conceptualisation and planning steps of speech production, which may result in disfluencies. Current studies on human-human dialogues also suggest that disfluencies convey information such as level of conflict (Vidrascu and Devillers, 2005), uncertainty of the speaker (Lickley, 2015), or points of interest in meetings (Shriberg, 2005). Thus, we expect more disfluencies in speech when the speaker is uncertain or when there is a hot spot in the dialogue.

Considering non-verbal vocalisations and emotions, previous work has identified laughter as a universal and basic cue in human emotion recognition (Sauter et al., 2010a). There are various types of laughter produced by humans, which relate to different emotional and social events (Szameitat et al., 2009b). For example, during a conversation, there can both be joyful laughter signalling amusement, and bonding laughter expressing affiliation and agreement (Scott, 2013). Note that laughter is a complex phenomenon and it can occur with negative emotions as well. For example, taunting laughter which often arises when the speaker is trying to humiliate his/her conversational partner (Szameitat et al., 2009a). In this work, we use the laughter annotations provided by the AVEC2012 and IEMOCAP databases directly, which did not differentiate between different types of laughter and only provide binary annotations of presence/absence of laughter. It would be interesting to study how different types of laughter relate to emotions in dialogue differently in the future with more detailed laughter annotations available.

There are relatively fewer studies on the relationship between emotions and voice qualifiers (e.g., audible breath) than on the relationship between emotions and voice qualifications (e.g., laughter). However, it has been suggested that audible breath is used consciously by the speaker to convey emotional arousal, or used unconsciously as involuntary reaction to such arousal (Roach et al., 1998). Besides laughter and audible breath, there are other non-verbal vocalisations that have been suggested as emotional indicators, e.g., sighs (Teigen, 2008). However, these non-verbal vocalisations are extremely rare in the databases we used (e.g., there are only two utterances containing sigh in the IEMOCAP database). Therefore, we only included laughter and audible

breath in our DIS-NV feature set.

Although Psycholinguistic studies suggest a potential relationship between DIS-NVs and emotions, previous work on automatic emotion recognition has rarely used DIS-NVs as input features. To the best of our knowledge, the only previous work using DIS-NVs for emotion detection is the work of Vidrascu and Devillers (2005), which included the number of filled pauses per utterance ("euh" in French) in their feature set for recognizing 20 emotion categories from recordings of a French Medical emergency call center. They compared the individual predictiveness of features and found that filled pause is the second most predictive feature (F0 range of the utterance is the most predictive feature). However, they did not report to what extent the emotion recognition model benefited from including the filled pause feature.

To address overlooking DIS-NVs in feature extraction of current emotion recognition studies, we propose features representing occurrences of DIS-NVs in spoken utterances for emotion recognition in spoken dialogue. In the following part of this chapter, we describe how these features are computed and conduct experiments studying their effectiveness for emotion recognition in spontaneous dialogue.

4.2 DIS-NV Features

In this section, we define the types of DIS-NV we annotated for extracting the DIS-NV features, and how the feature values are calculated.

4.2.1 Types of DIS-NV

To extract our DIS-NV features, we manually annotated three types of disfluencies, namely filled pauses, fillers, and stutters. We also used two types of non-verbal vocalisations provided in the manual transcriptions of the databases, namely laughter and audible breath. As discussed in Section 4.1, we focus on these specific types of DIS-NVs because they are emotion-related and are the most frequently occurring in spontaneous dialogue. We are aware that these five types of DIS-NV are only a subset of all DIS-NVs in speech. In addition, we also study the influence of including other common DIS-NVs, namely speech repairs, turn-taking times, and prolongations, in Section 5.3.1. However, our experiments show that including additional DIS-NVs does not improve the emotion recognition performance. Thus, in our emotion recognition experiments, the DIS-NV feature set contains the five selected DIS-NVs if not specified

otherwise.

- **Filled pauses**: non-lexical insertions in speech used by the speaker when (s)he pauses to think while trying to hold the turn. For example, "Hmm" in the utterance "Hmm... Maybe we should try another road". The three most common filled pauses we found in the AVEC2012 database of spontaneous dialogue are "em", "eh", and "oh".

- **Fillers**: lexical filled pauses. For example, "you know" in the utterance "I just want to, you know, get a drink and forget all about it". Some Psycholinguistic studies do not differentiate between filled pauses and fillers (e.g., Finlayson (2014)). In our work, we consider filled pauses and filler separately to have a more detailed understanding of their relationship with emotions. The three most common fillers we found in the AVEC2012 database of spontaneous dialogue are "well", "you know", and "I mean".

- **Stutters**: words or part of a word the speaker involuntarily repeats during speaking. For example, "Sa" in the utterance "Sa... Saturday will be fine", or the first "I didn't" in the utterance "I didn't, I didn't mean it".

- **Laughter**: a physical reaction consisting typically of rhythmical, often audible contractions of the diaphragm and other parts of the respiratory system. Laughter annotations were included in the manual transcriptions provided with both databases. Note that these are binary annotations of the presence/absence of laughter without differentiating different types of laughter.

- **Audible breath**: the sounds generated by the movement of air through the respiratory system. Audible breath annotations were included in the manual transcriptions provided with both databases.

4.2.2 Feature Extraction

We use a moving window with a length of 15 words to compute the disfluency features for word-level emotion recognition on the AVEC2012 database. We chose a window length of 15 words because this is the average length of an utterance in the AVEC2012 database. In our later experiments on utterance-level emotion recognition for the IEMOCAP database, we used the utterance duration instead of the moving window for computing the DIS-NV features. We have also tested using the utterance duration

instead of the moving window for word-level emotion recognition on the AVEC2012 database (Tian, 2013). However, the performance is worse than using the moving window and is not included here.

As shown in Figure 4.1, the window includes the current word and the 14 history words that precede it, and slides from the beginning of a dialogue session until its end. The feature value of word w for DIS-NV type D (D_w) is calculated as the ratio between the sum duration of DIS-NV type D appearing in the window of word w (T_D) and the total duration of the window of word w including silences between words (T_w). This results in five DIS-NV features for each word:

$$D_w = \frac{T_D}{T_w} \tag{4.1}$$

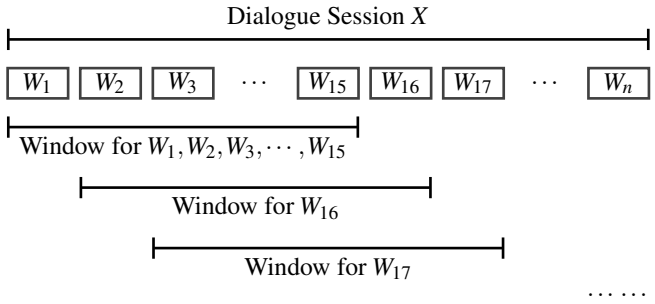


Figure 4.1: Window for Extracting DIS-NV Features from the AVEC2012 Database

4.2.3 Individual Effectiveness of the DIS-NV Features

We compare the individual effectiveness of the DIS-NV features on the AVEC2012 database using the Correlation-based Feature Selection (CFS) method (Hall, 1998).[1] The CFS method ranks individual effectiveness of features based on the recognition performance of each feature being used as a weak recognizer, as well as the degree of redundancy between features. Results are shown in Table 4.1, with smaller numbers representing ranking higher by the CFS method and thus higher individual effectiveness. As we can see, filled pause and laughter are the most effective types of DIS-NV for emotions in spontaneous dialogue. The fact that filler is not highly ranked supports considering filled pause and filler separately for studying the relationship between DIS-NVs and emotions.

[1] We only study the individual effectiveness of the DIS-NV features on the AVEC2012 database of spontaneous dialogue because there are much fewer DIS-NVs in the IEMOCAP database.

Table 4.1: Individual Effectiveness Rankings of DIS-NV Features

DIS-NV	Arousal	Expectancy	Power	Valence
Filled Pause	**1**	2	**1**	2
Filler	5	4	4	5
Stutter	4	5	5	3
Laughter	2	**1**	2	**1**
Audible Breath	3	3	3	4

4.3 Recognizing Emotions in Spontaneous Dialogue with DIS-NV Features

This section contains our experiments on effectiveness of the proposed DIS-NV features for recognizing emotions in spoken dialogue. These include four emotion regression experiments on the AVEC2012 database of spontaneous dialogue. In Experiment 1, we compare performance of the DIS-NV features with benchmark acoustic and lexical features. In Experiment 2, we study the gain from incorporating the DIS-NV features with acoustic and lexical features. In Experiment 3, we study the influence of temporal context for emotion recognition. In Experiment 4, we investigate automatic recognition of DIS-NVs and the effectiveness of the auto-detected DIS-NV features for emotion recognition. Here we follow the AVEC2012 challenge protocol and conduct our experiments on the AVEC2012 database of spontaneous dialogues.

4.3.1 Experiment 1: Emotion Regression in Spontaneous Dialogue with DIS-NV Features

Our goal for Experiment 1 is to study the effectiveness of using the DIS-NV features on their own for emotion recognition in spoken dialogue. To do so, we compare the performance of the proposed DIS-NV features with benchmark acoustic and lexical features and state-of-the-art emotion recognition studies.

4.3.1.1 Methodology

As described in Section 3.3.1, emotions were annotated at the word-level as real-value vectors on the Arousal, Expectancy, Power, and Valence emotion dimensions in the AVEC2012 database. Following the setting of the AVEC2012 challenge, we used

the Cross Correlation Score (CCS) as the evaluation metric. In the challenge, the AVEC2012 database is divided into three partitions, each containing 32 dialogue sessions: the training partition, the development partition, and the test partition. CCS is calculated as the average correlation-coefficients *CC* between the emotion predictions and the emotion annotations over the 32 dialogue sessions of the test partition of the AVEC2012 database:

$$CCS = \frac{1}{32}\sum_{i=1}^{32} CC_i \tag{4.2}$$

We evaluate significance of performance differences using the two-tailed z-test after Fisher's r-to-z transformation. Performance of the DIS-NV features are compared with the AVEC2012 baseline LLD acoustic features and the PMI lexical features (see Section 3.4). The LLD acoustic features were provided in the AVEC2012 challenge as a benchmark feature set, while the PMI lexical features were shown to be the most effective unimodal feature set in previous work on this task (Savran et al., 2012). We also compare the performance of our DIS-NV features with multimodal recognition results reported by other AVEC2012 challenge participants.

4.3.1.2 Results and Discussion

Results of Experiment 1 are reported in Table 4.2 (Moore et al., 2014). "Mean" represents the unweighted average of the results on the four emotion dimensions. Savran et al. (2012) represents the multimodal state-of-the-art results achieved by a model using thousands of audiovisual and lexical features. "DIS-NV" represents the model using the proposed 5 DIS-NV features. "S-PMI" represents the model using 1000 sparse PMI lexical features, which was the most effective feature set in previous work on the AVEC2012 database (Savran et al., 2012). "PMI" represents the model using the 8 non-sparse PMI lexical features we proposed (see Section 3.4.2.1). "LLD" represents the model using 1842 AVEC2012 baseline LLD features, as described in Section 3.4.1.1. We used the Support Vector Regression model described in Section 3.5.1 for building all the emotion recognition models. We also include a baseline model which predicts random numbers between [-1,1]. Because the evaluation metric is correlation based, we cannot use a baseline model which always predicts the mean.

As shown in Table 4.2, our five knowledge-inspired DIS-NV features achieved substantially higher scores than benchmark acoustic and lexical features for predicting every emotion dimension. The overall performance (average performance of all

emotion dimensions) of the model using only our DIS-NV features is tied with the best multimodal result reported on the AVEC2012 database (Savran et al., 2012).[2] This indicates the effectiveness of the DIS-NV features for recognizing emotions in spontaneous dialogue. The DIS-NV features also achieved the best reported performance on the Expectancy emotion dimension. This is consistent with the Psycholinguistic finding that disfluency is an indicator of speaker uncertainty (Lickley, 2015). Savran et al. (2012) outperformed the DIS-NV features the most on the Valence dimension, which may be due to the fact that the Savran's (2012) model incorporated visual features describing facial expressions that are specifically effective for disambiguating the Valence emotion dimension. The non-sparse PMI features we propose have results close to the sparse PMI features of Savran et al. (2012)[3] while reducing the feature dimensionality from 1000 to 8. The LLD features have extremely low performance here compared to the DIS-NV and PMI features. This may be due to the high dimensionality and the frame-level nature of the LLD feature compared to the utterance-level knowledge-inspired DIS-NV and PMI features. To study the influence of feature dimensionality and temporal context in more detail, later in Experiment 3 we performed feature engineering to reduce the dimensionality of the LLD feature set and included temporal context in the LLD features.

Table 4.2: Emotion Regression with DIS-NV Features on Spontaneous Dialogue

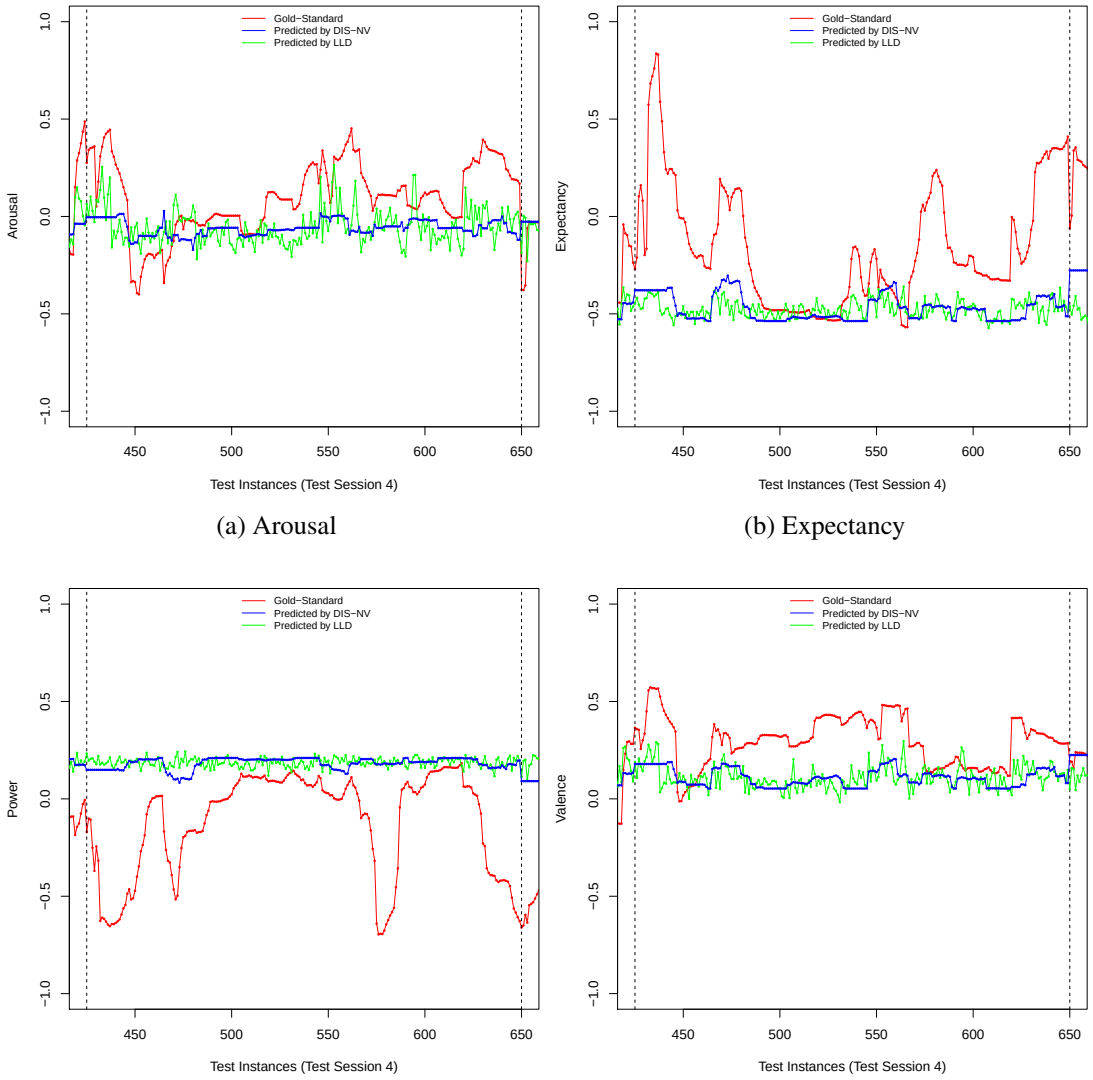
Models	Arousal	Expectancy	Power	Valence	Mean
Savran et al. (2012)	**0.302**	0.194	**0.293**	**0.331**	**0.280**
DIS-NV	0.250	**0.313**	0.288	0.235	*0.271*
S-PMI	0.131	0.285	0.254	0.188	0.214
PMI	0.152	0.216	0.220	0.186	0.193
LLD	0.014	0.038	0.016	0.040	0.027
Baseline	0.001	0.007	0.004	0.008	0.005

To study the performance of the proposed DIS-NV features in more detail, in Figure 4.2 we plot the predictions given by DIS-NV and LLD features compared to the gold-standard emotion annotations on test dialogue session 4 of the AVEC2012 database. As we can see, the predictions given by LLD features are more noisy and

[2] $p = 0.4237$, thus there is no significant difference between the overall performance of the DIS-NV model and the best multimodal model by Savran et al. (2012)

[3] $p = 0.0735$, thus there is no significant difference between the overall performance of the sparse PMI features and the non-sparse PMI features

have a flatter overall shape compared to those given by DIS-NV features. For the predictions given by DIS-NV features, there are segments which are straight lines due to absence of DIS-NV in the utterances. However, when DIS-NVs occur in the dialogue, the overall shape of DIS-NV predictions better captures the shape of the gold-standard emotion annotations with the predictions having smaller absolute values than the gold-standard annotations. Distributions of different emotion dimensions vary greatly, which indicates that performance of emotion recognition models on different emotion dimensions should be evaluated separately and the mean CCS over all emotion dimensions should be considered as an additional reference. Note that the gold-standard emotion annotations have a more smooth shape than the automatic predictions in Figure 4.2. Therefore, it may be beneficial to use a sliding window to smooth the predictions of the automatic emotion recognition model in the future.



(a) Arousal (b) Expectancy

(c) Power (d) Valence

Figure 4.2: Predictions vs. Annotations on the AVEC2012 Database

4.3.1.3 Summary

In Experiment 1, we compared the performance of our DIS-NV features with benchmark LLD and PMI features for emotion recognition on spontaneous dialogue. Our five knowledge-inspired DIS-NV features achieved substantially higher scores than benchmark acoustic and lexical features for predicting every emotion dimension. The DIS-NV features are especially predictive of the Expectancy dimension of emotion. This is consistent with the Psycholinguistic research by Lickley (2015) which indicated that disfluencies are related to the uncertainty of the speaker. The overall performance of our unimodal model using only the DIS-NV features is tied

with the best reported multimodal model for this task using thousands of audiovisual and lexical features. Results of Experiment 1 suggest that the proposed DIS-NV features are effective for emotion recognition in spontaneous dialogue, and they may contribute to multimodal emotion recognition by disambiguating the Expectancy emotion dimension better.

4.3.2 Experiment 2: Multimodal Emotion Regression on Spontaneous Dialogue with DIS-NV Features

Compared to the acoustic characteristics and lexical content, DIS-NVs contain additional information which may be related to emotions. Thus, in Experiment 2, we study whether or not the emotion recognition model can benefit from incorporating our DIS-NV features with benchmark acoustic and lexical features.

4.3.2.1 Methodology

To study the gain of including our DIS-NV features, we build multimodal emotion recognition models by concatenating the feature sets (i.e., Feature-Level (FL) fusion). As in Experiment 1, we build Support Vector Regression models for our unimodal and multimodal models and report CCS. Results of our experiment are shown in Table 4.3 (Moore et al., 2014). "LLD+PMI" represents the model using the concatenated feature set of the LLD and our non-sparse PMI features. In Table 4.3, Savran et al. (2012), Ozkan et al. (2012), and van der Maaten (2012) are the three best performing multimodal models on the AVEC2012 challenge, respectively, while Schuller et al. (2012) is the AVEC2012 baseline multimodal model.

4.3.2.2 Results and Discussion

As shown in Table 4.3, the LLD+PMI+DIS-NV model achieved significantly improved results compared to the LLD+PMI model on all emotion dimensions. Overall performance of the LLD+PMI+DIS-NV model is between the best (Savran et al. (2012)) and the second best (Ozkan et al. (2012)) performing multimodal model on the AVEC2012 challenge. The LLD+PMI+DIS-NV model also achieved the best result on predicting the Expectancy emotion dimension compared to reported challenge results. This verifies our conjecture that DIS-NVs contain information additional to the acoustic characteristics or lexical content of the speech that is predictive of emotions.

Thus, including DIS-NV features in existing models yields improved performance. However, the LLD+PMI+DIS-NV model has worse performance than the unimodal DIS-NV model on the Expectancy and Power emotion dimensions. The reason may be the imbalanced size of the feature sets. In simple feature concatenation (FL fusion), equal weights are assigned to each feature set. Thus, the highly predictive DIS-NV feature set with only 5 features may be overwhelmed by the noisy LLD feature set with over a thousand features. This indicates that a better fusion strategy is required to further improve the performance of the multimodal emotion recognition model. In Chapter 6 we will explore this in more detail.

Table 4.3: Multimodal Emotion Regression with DIS-NV Features on Spontaneous Dialogue

Models	Arousal	Expectancy	Power	Valence	Mean
DIS-NV	0.250	**0.313**	**0.288**	0.235	**0.271**
LLD+PMI	0.252	0.216	0.146	0.213	0.207
LLD+PMI+DIS-NV	**0.263**	0.269	0.162	**0.292**	0.247
Savran et al. (2012)	**0.302**	0.194	**0.293**	**0.331**	**0.280**
Ozkan et al. (2012)	0.210	0.240	0.289	0.208	0.237
van der Maaten (2012)	0.267	**0.241**	0.223	0.138	0.192
Schuller et al. (2012)	0.021	0.028	0.009	0.004	0.015

4.3.2.3 Summary

Experiment 2 shows that DIS-NVs contain information predictive of emotions beyond the acoustic characteristics and lexical content of spontaneous dialogue. However, simple feature concatenation may limit the gain of incorporating DIS-NV features in multimodal emotion recognition models.

4.3.3 Experiment 3: Influence of Contextual Information on Emotion Regression

In Experiments 1 and 2, both the DIS-NV features and the PMI features are utterance-level features that include temporal context, while the LLD features are non-contextual frame-level features. Thus, in Experiment 3 we study whether models based on the LLD features will benefit from including temporal context or not.

4.3.3.1 Methodology

To extract contextual LLD features, we first selected a subset of the LLD features using Correlation-based Feature-subset Selection (Hall, 1998), which results in 116 features (the CFS-LLD feature set). We then use the sliding window shown in Figure 4.1 to compute the min, max, mean, and standard deviations of the CFS-LLD feature values within the window, resulting in 464 (116×4) contextual LLD features.

4.3.3.2 Results and Discussion

CCS of unimodal models using the original non-contextual LLD feature set, the non-contextual CFS-LLD feature set, and the contextual LLD feature set are reported in Table 4.4. As we can see, the contextual LLD features have significantly better performance than the non-contextual LLD and CFS-LLD features on predicting all emotion dimensions. This verifies that emotion recognition can benefit from including temporal context, which is consistent with Psychological findings (Ortony et al., 1990). The improvement achieved by the CFS feature engineering compared to using the LLD features directly also indicates that learning a more abstract feature representation and reducing the feature dimensionality is helpful for emotion recognition.

Table 4.4: Influence of Temporal Context for Emotion Regression on Spontaneous Dialogue

Models	Arousal	Expectancy	Power	Valence	Mean
LLD	0.014	0.038	0.016	0.040	0.027
CFS-LLD	0.118	0.091	0.075	0.094	0.094
Contextual LLD	**0.252**	**0.216**	**0.146**	**0.213**	**0.207**

4.3.3.3 Summary

In Experiment 3 we verified that temporal context is predictive of emotions. Thus, building contextual models can improve performance of emotion recognition in spoken dialogue. Our results also show that feature engineering can be beneficial for emotion recognition models by reducing feature dimensionality.

4.3.4 Experiment 4: Automatic Detection of DIS-NVs

In this thesis, we focus on DIS-NV features based on manual annotations of DIS-NVs (gold-standard DIS-NV features) because we are interested in the effectiveness of DIS-NVs for emotion recognition in spoken dialogue. However, beyond improving the state-of-the-art of emotion recognition in spoken dialogue, our long-term goal is to improve the quality of emotional interaction in HCI systems. In a fully automatic emotion recognition model, the DIS-NV features will need to be extracted automatically, which may introduce noise to the DIS-NV feature set. Therefore, in Experiment 4, we conduct a preliminary study on the influence of using auto-detected DIS-NV features for emotion recognition. Note that automatic detection of disfluencies and non-verbal vocalisations in speech is an active research area itself. Thus, with improved DIS-NV recognition models, we will be able to further reduce the difference between the auto-detected and gold-standard DIS-NV features in the future.

4.3.4.1 Review on Automatic Detection of DIS-NVs

Automatic detection of DIS-NVs has attracted interest from both the speech recognition and Psycholinguistic communities. DIS-NV detection models can improve the performance of automatic speech recognition, as well as help researchers understand the speech generation process (e.g., Barczewska and Igras (2013)).

For automatic detection of disfluencies, various acoustic features and machine learning algorithms have been applied. Among different acoustic features, pitch and duration have been identified as highly predictive of disfluencies. For example, O'Shaughnessy and Gabrea (2000) classified filled pauses as vowels with durations longer than 120ms and F0 lower than the average F0 of the speaker. Formant stability is also used by Audhkhasi et al. (2009) and Barczewska and Igras (2013) for detecting disfluencies. Besides pitch, cepstral features, such as Mel-frequency cepstral coefficients (MFCCs), are also widely used in previous work (e.g., Stouten and Martens (2003)). Previous studies on disfluency detection have shown that contextual models are typically powerful for disfluency detection. For example, Yu et al. (2012) combined a Hidden Markov Model with a deep neural network and achieved a word error rate of 16.1% for disfluency detection. Similarly, Zayats et al. (2016) built a Bidirectional LSTM model and achieved state-of-the-art disfluency detection performance with a F1 measure of 85.9%.

For automatic detection of non-verbal vocalisations, the majority of previous

research has focused on binary laughter detection. Previous work on automatic detection of laughter has studied various types of acoustic features, such as prosodic features (Truong and Van Leeuwen, 2007) and MFCCs (Krikke and Truong, 2013). Similar to disfluency detection, among different types of acoustic features, pitch has been shown to be highly predictive of laughter (Salamin et al., 2013). Paralinguistic studies have found that F0 in laughter is higher than F0 in speech segments (Rothgänger et al., 1998; Bachorowski et al., 2001). For audible breath detection, previous work has focused on prosodic (e.g., Braunschweiler and Chen (2013)) and cepstral features (e.g., Ruinskiy and Lavner (2007)). Various machine learning algorithms have been applied to non-verbal vocalisation detection, such as Gaussian Mixture Models used by Krikke and Truong (2013), Multi-Layer Perceptrons used by Knox and Mirghafori (2007), and Support Vector Machines used by Kennedy and Ellis (2004). Dupont et al. (2016) achieved state-of-the-art performance for laughter detection with an accuracy of 79% by combining audio and visual information.

4.3.4.2 Effectiveness of Auto-Detected DIS-NV Features for Emotion Recognition

As discussed in Section 4.3.4, there is on-going research on automatic detection of disfluencies (e.g., Liu et al. (2006)) and non-verbal vocalisations (e.g., Niewiadomski et al. (2013)). In this thesis, we focus on the performance of gold-standard DIS-NV features for emotion recognition. Here we perform a preliminary experiment on the influence of auto detection of DIS-NVs. We use non-sparse PMI and the AVEC2012 baseline LLD features with a SVM model to predict the DIS-NV feature values. CCS of emotion recognition models using the auto-detected DIS-NV features are reported in Table 4.5.

Table 4.5: Using Auto-Detected DIS-NV Features for Emotion Regression on Spontaneous Dialogue

Models	Arousal	Expectancy	Power	Valence	Mean
Gold-standard DIS-NV	**0.250**	**0.313**	**0.288**	**0.235**	**0.271**
DIS-NV Predicted by PMI	0.133	0.191	0.192	0.161	0.169
DIS-NV Predicted by LLD	0.087	0.094	0.054	0.070	0.076
PMI	0.152	0.216	0.220	0.186	0.193
LLD	0.014	0.038	0.016	0.040	0.027

As shown in Table 4.5, performance of the auto-detected DIS-NV features has significant decreases on all emotion dimensions compared to the gold-standard DIS-NV features. However, both of the auto-detected DIS-NV features still have significantly better performance than the AVEC2012 baseline LLD features for emotion recognition in spontaneous dialogue. Our results suggest that besides acoustic features, lexical features are powerful predictors of DIS-NVs as well.

Note that we used a naive DIS-NV recognizer in this experiment. With an improved DIS-NV detection model, the performance difference between the auto-detected DIS-NV features and the gold-standard DIS-NV features can be further reduced. For example, Shi (2016) studied the auto-detection of disfluencies in the AVEC2012 database, Wang (2016) studied the auto-detection of non-verbal vocalisations in the AVEC2012 database.[4] Shi (2016) used an Auto-Encoder over the eGeMAPS features (see Section 3.4.1.2) and built a LSTM model with the encoded feature representation for disfluency detection. The highest F1-measures achieved are: filled pause = 77.0%, filler = 78.0%, stutter = 80.0%. Wang (2016) built a Deep Belief Network with Binary-Bernoulli Restricted Boltzmann Machine layers and combined the eGeMAPS feature set with 78 MFCC features for non-verbal vocalisation detection. The highest F1-measures achieved are: laughter = 69.8%, audible breath = 78.6%. These results indicate that automatic detection of DIS-NVs can be done with stable performance. Thus, the auto-detected DIS-NV features will remain predictive of emotions in spontaneous dialogue.

4.3.4.3 Summary

In Experiment 4, we showed that DIS-NVs in spontaneous dialogue can be automatically detected with stable accuracy, and the auto-detected DIS-NV features remain predictive of emotions in spontaneous dialogue. This indicates that our emotion recognition model using DIS-NV features has the potential to be applied to a fully automatic HCI system in the future.

4.4 Discussion

In this chapter, we proposed DIS-NV features for emotion recognition. We motivated the use of DIS-NVs for emotion recognition, and described calculation of the DIS-NV

[4] MSc dissertations co-supervised by the author

features. We performed experiments on the AVEC2012 database of spontaneous dialogue to study the effectiveness of the proposed DIS-NV features compared to benchmark acoustic and lexical features widely used in previous work on the same database.

Our results show that our DIS-NV features yield better performance than LLD or PMI features on predicting all emotion dimensions. The DIS-NV features are particularly predictive of the Expectancy emotion dimension which relates to speaker uncertainty, and achieved the best reported result. The emotion recognition model using only the 5 DIS-NV features achieved an overall performance that is tied with the best reported result achieved by a multimodal emotion recognition model using thousands of audiovisual and lexical features. These findings verified that the proposed DIS-NV features are predictive of emotions in spontaneous dialogue.

Our experiment on incorporating the DIS-NV features with other acoustic and lexical features indicates that DIS-NVs contain additional information related to emotions compared to the acoustic characteristics and the lexical content. However, a better fusion strategy than simple feature concatenation is required to increase the gain of modality fusion. We also verified that including temporal context is beneficial for emotion recognition. In addition, we perform preliminary experiments which showed that DIS-NVs can be automatically detected with robust accuracy, thus the DIS-NV features may remain effective in a fully automatic emotion recognition model.

One thing to notice is that the correlation-coefficient based evaluation metric reported by all our models and all previous work using the AVEC2012 continuous emotion annotations is extremely low.[5] This indicates that emotion recognition is a challenging task. As discussed in Section 3.3.1, using continuous emotion annotations also caused the issue of low inter-annotator agreement. Thus, in the following experiments, we map the original continuous emotion annotation of the AVEC2012 database into three discrete categories for each emotion dimension: *low* (original values within the range [−1,−0.333)), *medium* (original values within the range [−0.333,+0.333]), and *high* (original values within the range (+0.333,+1]).

[5] $CCS < 0.4$, meaning weak to mild correlation

Chapter 5

Spontaneous vs. Acted Dialogue

Inspect every piece of pseudoscience and you will find a security blanket, a thumb to suck, a skirt to hold. What does the scientist have to offer in exchange? Uncertainty! Insecurity!

— Isaac Asimov, *Asimov's Guide to Science (1972)*

In Chapter 4, we proposed DIS-NV features for emotion recognition in spoken dialogue, which describe occurrences of 5 types of disfluency and non-verbal vocalisation in utterances. Our experiments showed that these DIS-NV features are predictive of emotions in spontaneous dialogue. However, as described in Section 3.1.2, besides emotion databases of spontaneous dialogue, a large portion of existing emotion databases consist of acted dialogue. Cross-corpora studies suggest that fundamental differences exist between different types of dialogue, which may influence the effectiveness of features and models for recognizing emotions in different types of dialogue. In this chapter, to study how aspects of a given database influence the performance of different emotion recognition approaches, we conduct cross-corpora experiments on the AVEC2012 database of spontaneous dialogue and the IEMOCAP database of acted dialogue.

First, we study the differences in statistical distributions of emotions and the DIS-NV and GP features in the AVEC2012 and the IEMOCAP databases to illustrate the differences between spontaneous and acted dialogue. Note that the different data collection and annotation schemes used by the AVEC2012 and the IEMOCAP database may also result in differences in distributions of emotions and affective cues. Therefore, in order to have a better understanding of the differences caused by dialogue type, we compare the scripted and non-scripted acting subsets of the IEMOCAP database in addition. Second, we conduct emotion recognition experiments on both

databases using different features and models to study how differences in the data influence performance of emotion recognition approaches.

5.1 Cross-Corpora Studies in Emotion Recognition

Most previous work on emotion recognition focuses on experiments using a single database (e.g., Chen et al. (2016); Wang et al. (2015); Wöllmer et al. (2010)). Only a few studies have performed cross-corpora experiments to test the robustness of the features or models proposed (e.g., Bone et al. (2014); Eyben et al. (2015b)).

Previous cross-corpora studies on emotion recognition suggest that it is often hard to generalize the efficacy of features and models across different databases, especially when the dialogue type is different. For example, Eyben et al. (2015b) conducted experiments on the 6 most widely used emotion databases, and their results showed that the performance ranking of 7 standard acoustic feature sets varies greatly across databases. Similarly, Schuller et al. (2010a) built a SVM model with LLD features for detecting binary Arousal and Valence, and tested the emotion recognizer on multiple databases including both spontaneous and acted dialogue. The results of Schuller et al. (2010a) illustrate the large influence of dialogue type and the difficulty of predicting emotions in spontaneous dialogue. The survey of Zeng et al. (2009) also highlighted the fact that emotions in acted dialogue are more acoustically exaggerated than emotions in spontaneous dialogue, leading to the issue that performance of emotion recognizers trained on acted dialogue may decrease greatly when applied to a more natural scenario. Thus, they suggested collecting more emotion databases of spontaneous dialogue, which would result in emotion recognizers that can generalize better to natural interaction scenarios.

5.2 Distribution of Emotion Annotations

To study the difference between spontaneous and acted dialogue, we compare the distribution of emotion annotations on the spontaneous AVEC2012 database and the acted IEMOCAP database. As described in Section 3.1.2, the AVEC2012 and IEMOCAP databases annotated emotions with different schemes. As discussed in Section 4.4, we mapped the original continuous emotion annotation of the AVEC2012 database into three discrete categories for each emotion dimension: *low* (original values within the range [−1,−0.333)), *medium* (original values within the range

[−0.333,+0.333]), and *high* (original values within the range (+0.333,+1]). To unite the emotion annotations of both databases, we also group the original 1 to 5 scores based emotion annotation of the IEMOCAP database to three discrete categories on each emotion dimension. The value ranges of each discrete class here are: [1,2.333) for *low*, [2.333,3.667] for *medium*, (3.667,5] for *high*. Merging the original emotion annotations to discrete categories is also used in previous cross-corpora studies on emotion recognition (e.g., Eyben et al. (2015b)). To preserve information on mild vs. intensive emotions, we use three classes instead of binary classification.

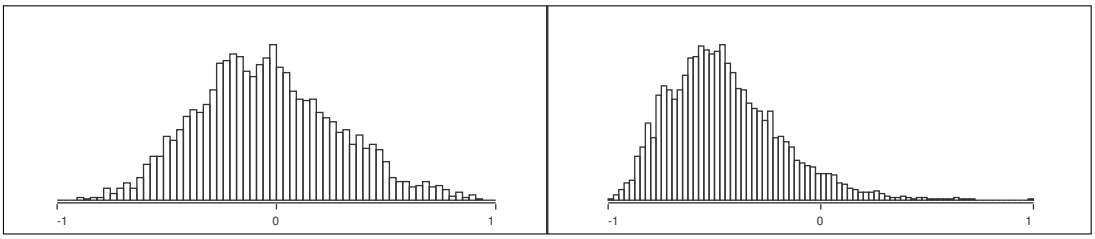
Note that the AVEC2012 database annotated emotions at the word level, while the IEMOCAP database annotated emotions at the utterance level. We keep the word level emotion annotation of the AVEC2012 database when studying the distributions of emotions and when conducting the emotion recognition experiments. However, when studying the distributions of DIS-NVs and acoustic characteristics, we down-sampled the word level AVEC2012 data to utterance level and plotted the descriptive statistics at the utterance level for both databases.

5.2.1 Distribution of Emotions in Spontaneous Dialogues

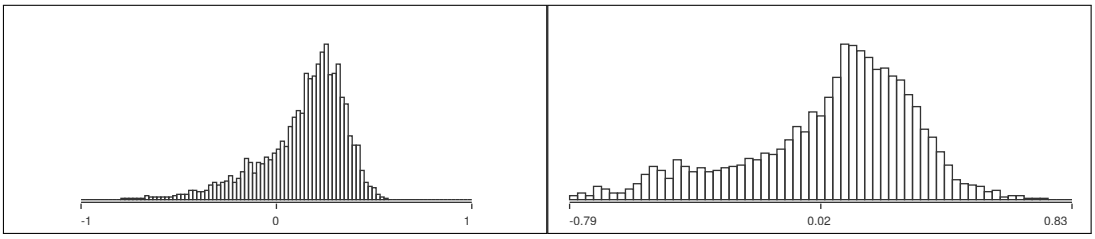
Figure 5.1 illustrates the distributions of the original word-level continuous emotion annotations on the AVEC2012 database. Figure 5.2 illustrates the distributions of the transformed discrete word-level emotion annotations on the AVEC2012 database. For each emotion dimension in Figure 5.2, the three bars from left to right represents the *low* (dark blue), *medium* (red), and *high* (light blue) categories, respectively.

As we can see, for the Arousal, Power, and Valence emotion dimensions, the *medium* category is significantly larger than the other two emotion categories. This indicates that emotions in spontaneous dialogue are mild or neutral most of the time, which is consistent with the previous finding that it is difficult to induce strong emotions in spontaneous dialogue (Zeng et al., 2009). Another interesting observation is that the majority of the data was annotated as having *low* or *medium* Expectancy, which indicates that speakers often show signs of uncertainty during spontaneous dialogue. This is consistent with the previous finding that DIS-NVs are common in spontaneous and non-scripted conversations (Lickley, 2015). When collecting the AVEC2012 database, four virtual agents with different personality designs were used (see Section 3.3.1). The use of Prudence the calm and neutral virtual agent may increase the number of neutral or non-emotional data instances in the AVEC2012

database. The peak of the Power and Valence distribution located in the positive value axis in Figure 5.1 also indicates that it is difficult to design a believable virtual agent that induces negative emotions in the participants.

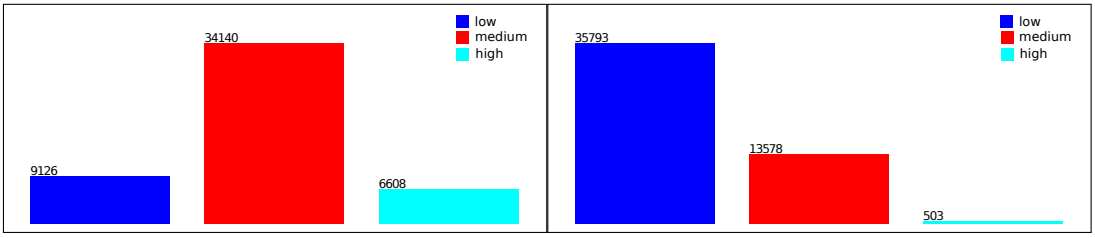


(a) Arousal (b) Expectancy

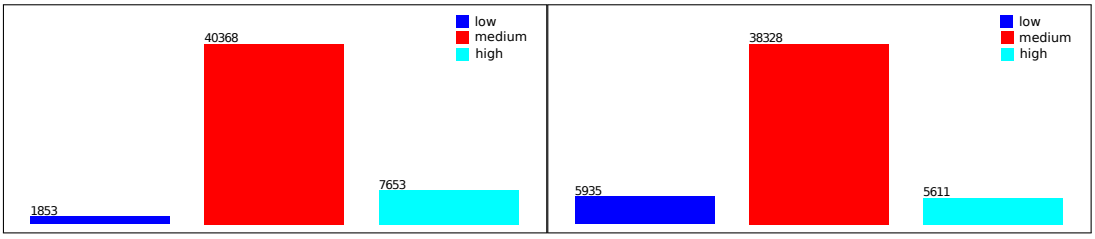


(c) Power (d) Valence

Figure 5.1: Word-Level Continuous Emotion Distribution on the Spontaneous AVEC2012 Database



(a) Arousal (b) Expectancy



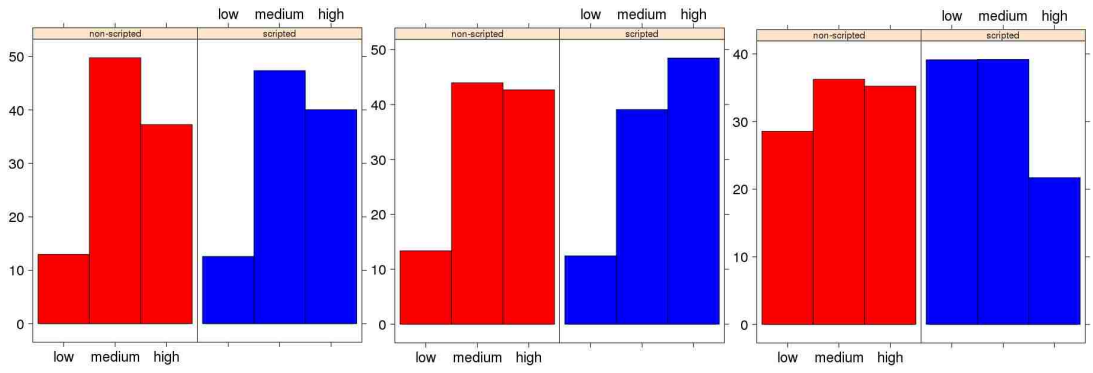
(c) Power (d) Valence

Figure 5.2: Word-Level Discrete Emotion Distribution on the Spontaneous AVEC2012 Database

5.2.2 Distribution of Emotions in Acted Dialogues

Figure 5.3 illustrates the distribution of the transformed discrete utterance-level emotion annotations on the IEMOCAP database. Recall that when collecting the acted IEMOCAP database, there were two types of acting: non-scripted and scripted acting (see Section 3.3.2). In Figure 5.3, the bars in red represent utterances collected by non-scripted acting, and the bars in blue represent utterances collected by scripted acting. The three columns from left to right in each graph represents the *low*, *medium*, and *high* categories of each emotion dimension, respectively. The y axis in the figure is percentage of total data instances.

Compared to the emotion distributions on the spontaneous AVEC2012 database (Figure 5.2), the emotion categories are more balanced on the acted IEMOCAP database. This indicates the advantage of collecting emotion databases by acting, which is that the data is more balanced (Zeng et al., 2009). An interesting observation is that there are fewer utterances with low Arousal or low Power compared to utterances with medium and high Arousal or Power. This reflects the fact that during collection of the IEMOCAP database, the acting scenarios were biased towards more active and dominant situations (e.g., an intense argument at customer service). Distributions of emotion annotation for utterances collected by scripted or non-scripted acting are approximately the same, which is possibly due to similar scenario design in both cases. The difference between the distributions of emotion annotation on the AVEC2012 database and the IEMOCAP database indicates spontaneous and acted dialogues are different in terms of speaker's emotional status.



(a) Arousal (b) Power (c) Valence

Figure 5.3: Utterance-Level Emotion Distribution on the Acted IEMOCAP Database

5.3 Distribution of DIS-NVs

In Section 5.2 we studied the differences between spontaneous and acted dialogue in terms of speaker's emotional states. In this section, we study the differences between spontaneous and acted dialogue in terms of DIS-NVs in speech. As suggested by Trouvain (2014), DIS-NVs are more common in spontaneous and unscripted dialogue. This is because actors are trained to be fluent, and DIS-NVs are often not included in scripts for collecting scripted dialogue. Therefore, we expect fewer utterances with DIS-NVs in the IEMOCAP database of acted dialogue than in the AVEC2012 database of spontaneous dialogue.

To compare the occurrences of DIS-NV in spontaneous and acted dialogue, we report the percentage of utterances containing each type of DIS-NV in both databases in Table 5.1. "FP" represents filled pause, "FL" represents filler, "ST" represents stutter, "LA" represents laughter, "AB" represents audible breath. As we can see, utterances with filled pause, laughter, and audible breath are less frequent in the acted IEMOCAP database than in the spontaneous AVEC2012 database. This is consistent with previous findings. However, fillers and stutters are more frequent in the non-scripted utterances of the IEMOCAP database. This indicates that compared to scripted acting, non-scripted acting is more similar to spontaneous dialogue in terms of the number of disfluencies in utterances.

To study the differences in DIS-NV distribution between spontaneous and acted dialogue in detail, in addition to the overall frequency, we also analyze the distributions of DIS-NV features for each emotion dimension of the spontaneous AVEC2012 database and the acted IEMOCAP database in Section 5.3.2.

Table 5.1: Percentages of Utterances with DIS-NV in Spontaneous and Acted Dialogue

Databases	FP(%)	FL(%)	ST(%)	LA(%)	AB(%)
AVEC2012	**32.0**	14.7	9.4	**11.9**	**2.7**
IEMOCAP (non-scripted)	14.9	**33.0**	**10.1**	2.4	0.8
IEMOCAP (scripted)	7.8	15.9	2.9	0.9	0.4

5.3.1 Additional DIS-NVs

The DIS-NVs we annotated are only a subset of all the DIS-NVs occurring in speech. Here we study the influence of including other types of common DIS-NV. More

specifically, we annotated speech repairs (SR, when speaker corrects him/herself), turn-taking times (TT, silent pause at the beginning of a turn), and prolongations (PL, prolonged uttering of a syllable) as additional DIS-NVs in the IEMOCAP database. Percentage of utterances containing these additional DIS-NVs are shown in Table 5.2. As we can see, compared to FP, FL, and ST shown in Table 5.1, SR and PL are less frequent in the IEMOCAP database.

We also conduct 10-fold cross validation experiments with an SVM model (C-SVC with RBF kernel) to compare performance of our original DIS-NV set containing 5 DIS-NVs (FP, FL, ST, LA, AB) and the expanded DIS-NV set including the three additional DIS-NVs (FP, FL, ST, LA, AB, SR, TT, PL) on the IEMOCAP database. We report the results (F1-measures) in Table 5.3. As we can see, adding these additional DIS-NVs does not improve emotion recognition performance. The original DIS-NV set of 5 DIS-NVs consistently outperforms the expanded DIS-NV set of 8 DIS-NVs.[1] Therefore, in later experiments, we continue to use the original DIS-NV set of 5 DIS-NVs.

Table 5.2: Percentages of Utterance with Additional DIS-NV in IEMOCAP Database

Databases	SR(%)	TT(%)	PL(%)
IEMOCAP (non-scripted)	**3.4**	27.9	**3.8**
IEMOCAP (scripted)	1.0	**35.1**	1.3

Table 5.3: Using Additional DIS-NVs for Emotion Recognition on IEMOCAP Database

Models	Arousal(%)	Power(%)	Valence(%)	Mean(%)
Original DIS-NV set	**36.3**	**40.7**	**32.8**	**36.6**
Expanded DIS-NV set	35.6	38.0	29.9	34.5

5.3.2 Distribution of DIS-NVs in Spontaneous and Acted Dialogues

In this section, we examine the distributions of filled pause and laughter as example DIS-NVs to study the differences between spontaneous and acted dialogue. As described in Table 4.1 of Section 4.2.3, filled pause and laughter have the highest individual effectiveness for emotion recognition in spontaneous dialogue. To study

[1] Arousal: $p = 0.0013$, Power: $p = 0.0068$, Valence: $p = 0.0024$

the distribution differences in more detail, we plot DIS-NVs in the non-scripted and scripted IEMOCAP database separately. Because annotation of the Expectancy emotion dimension is missing for the IEMOCAP database, here we only plot filled pause and laughter distributions on the Arousal, Power, and Valence dimensions.

To compare the distribution on different types of dialogue, we plot smoothed density graphs with lines representing the AVEC2012 database, the non-scripted subset of the IEMOCAP database, and the scripted subset of the IEMOCAP database, respectively in Figures 5.4 and 5.5. In these figures, the x axis represents the percentage of the total duration of an utterance being a DIS-NV, the y axis represents the percentage of utterances having the value on the x axis. We limit the maximum value of y axis to 0.8 to zoom in to the utterances containing DIS-NVs. Histograms of all the DIS-NV distributions on both databases can be found in Section C.1 of Appendix C.

5.3.2.1 Distribution of Filled Pauses in Spontaneous and Acted Dialogues

As shown in Figure 5.4, the blue (scripted IEMOCAP) and green (non-scripted IEMOCAP) lines stop before the x axis reaches 0.68, while the pink line (AVEC2012) reaches 1.0 on the x axis. Recall that in these figures the x axis represents the percentage of the total duration of an utterance being filled pauses. This shows that there are no utterances that are more than 70% filled pause in the acted IEMOCAP database, while in the spontaneous AVEC2012 database there are utterances that are entirely a filled pause. This reflects the fact that during data collection, the IEMOCAP database used professional actors, who are trained to have fewer disfluencies than the general public participating in the AVEC2012 data collection.

Filled pauses on the non-scripted and scripted IEMOCAP dialogue have similar distributions in Figure 5.4, except that there are fewer filled pauses in the scripted acting dialogue than in the non-scripted acting dialogue or the spontaneous dialogue (the blue line reaches 0 on y axis earlier than the green line in all graphs of Figure 5.4). Although Busso et al. (2008) argue that non-scripted acting is similar to spontaneous dialogue, as we can see, there remain fundamental differences in filled pause distributions between spontaneous (the red lines) and non-scripted acting (the green lines) dialogue.

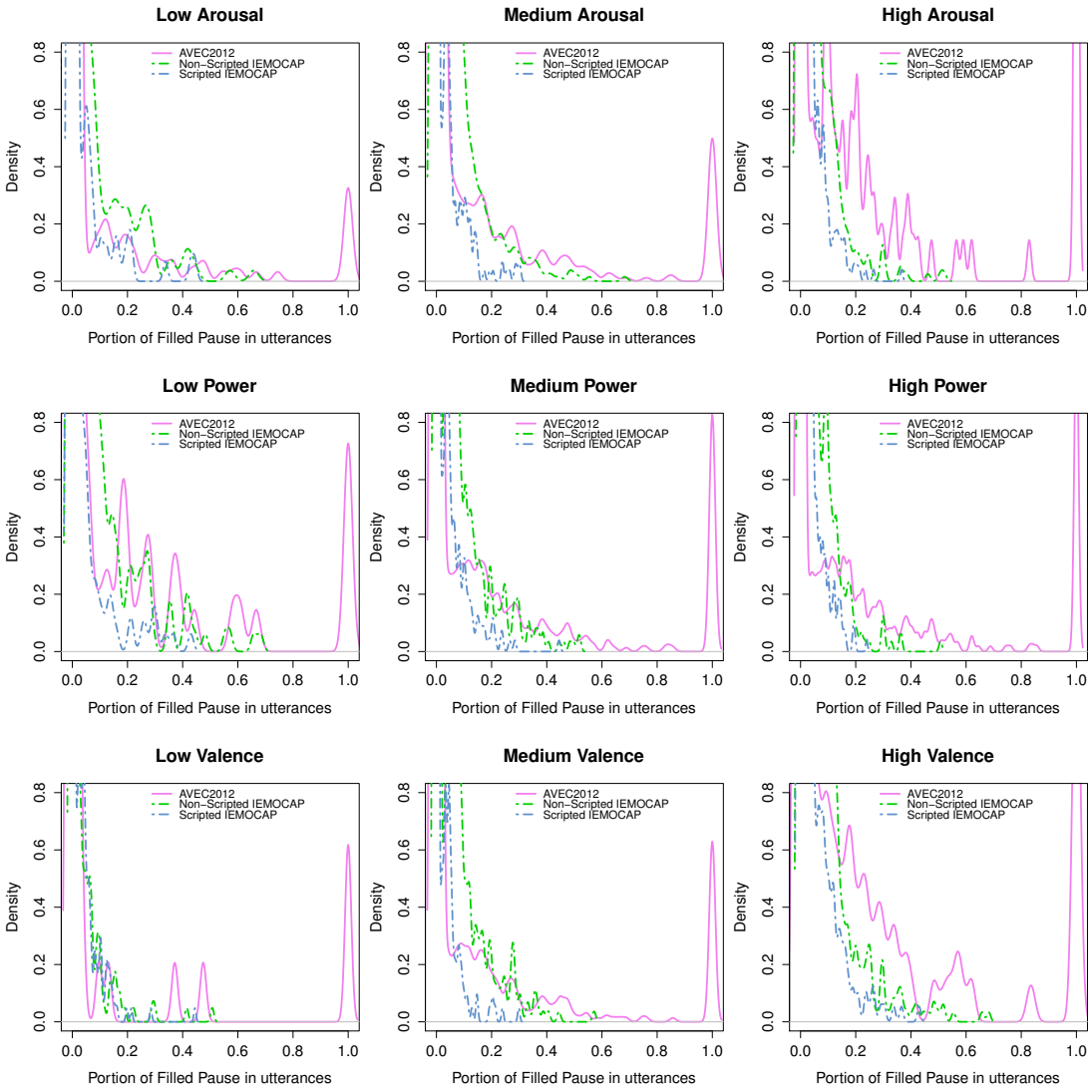


Figure 5.4: Filled Pause Distribution on Arousal, Power, Valence in Utterances

5.3.2.2 Distribution of Laughter in Spontaneous and Acted Dialogues

As shown in the Low Valence graph in Figure 5.5, the spontaneous AVEC2012 utterances annotated as having low Valence (negative emotion) can contain laughter, while none of the acted IEMOCAP utterances annotated as having low Valence contain laughter (the green and blue lines have x=0). When we investigated specific utterances in the AVEC2012 database which contain laughter and are annotated with low Valence, we identified various types of laughter other than the joyful laughter of amusement, which is consistent with Cognitive Science findings that laughter is a complex behaviour in spontaneous dialogue (Glenn, 2003; Szameitat et al., 2009b; Wildgruber et al., 2013). For example, in training session No.2, where the speaker

talked to Spike, the rude and offensive virtual agent, at turn No.42 she was annoyed by the agent accusing her of violating the data collection rule and said *"It wasn't a question. It was an indirect speech act. (LAUGH) Shut up! Why don't you go... go stuff your head in a box and give me Prudence instead."* (Prudence is the calm and neutral virtual agent). The speaker showed taunting laughter (Szameitat et al., 2009a) here which expresses a negative and aggressive emotional state in order to humiliate her conversational partner. Another example is in training session No.18, where the speaker talked to Obadiah, the pessimistic and depressive virtual agent. At turn No.101 the speaker realized how much he missed his family who lived somewhere else and said *"Oh God I'm not cheery. (LAUGH) Em... No can't be cheery all the time. Actually che... people that are cheery all the time kinda irritate me."* In this case laughter signalled embarrassment and submissiveness (Adelswärd, 1989).

As shown in Figure 5.5, laughter on the non-scripted and scripted IEMOCAP dialogues have similar distributions, except that there is less laughter in the scripted acting dialogues than in the non-scripted acting or the spontaneous dialogues. In contrast to the spontaneous AVEC2012 database, there is no laughter in utterances with low Valence in the acted IEMOCAP database. This indicates that variations of laughter types may be overlooked during the acting for collecting the IEMOCAP database.

Note that in this study we use the laughter annotation provided by the AVEC2012 and the IEMOCAP database, which does not differentiate types of laughter and only has a binary annotation of laughter being present or absent. This simplification may limit the effectiveness of the laughter features for emotion recognition. It will be interesting to collect more detailed laughter annotations in the future and study their relationship with emotions in spoken dialogue.

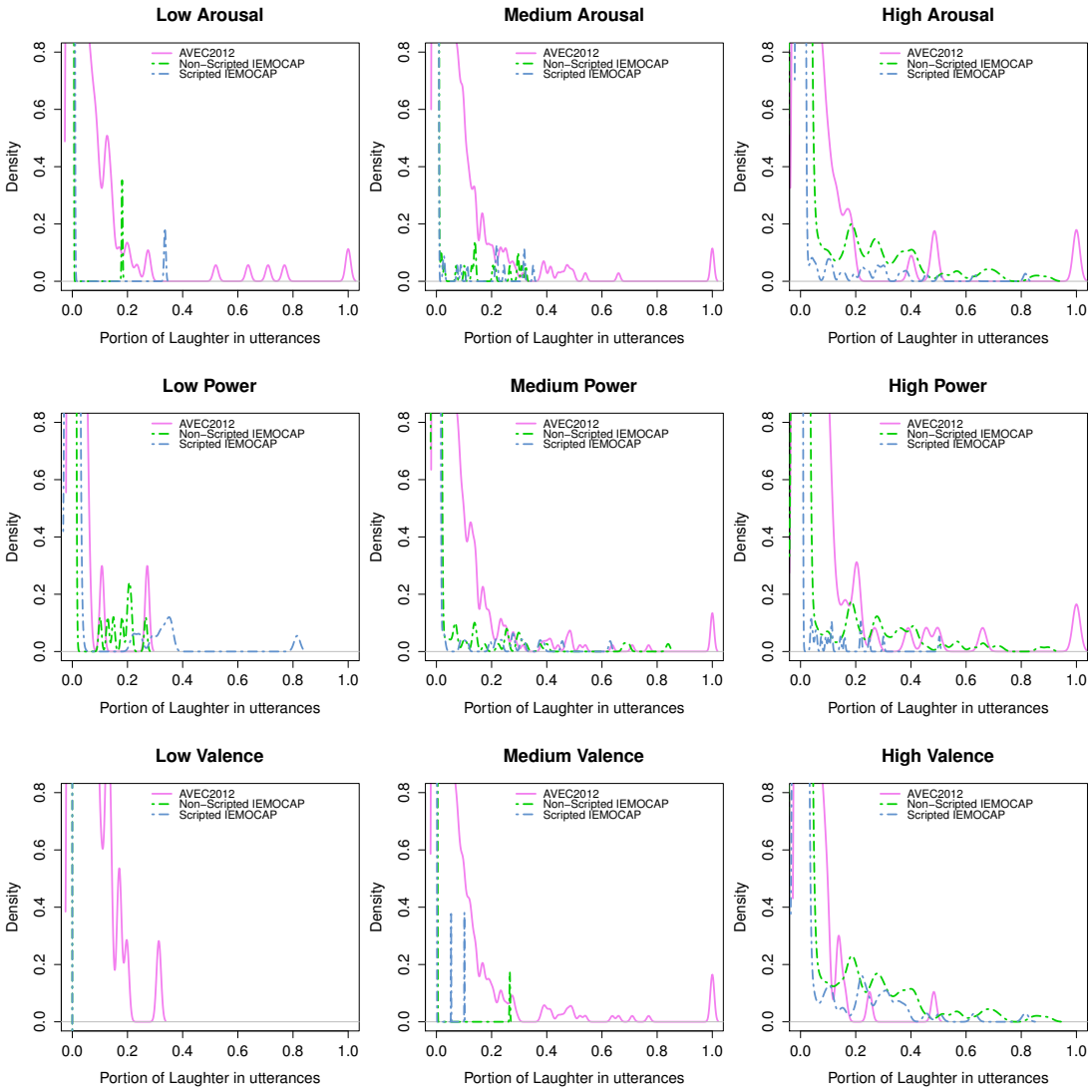


Figure 5.5: Laughter Distribution on Arousal, Power, Valence in Utterances

5.4 Distribution of Acoustic Features

Besides the distribution of speaker's emotions and DIS-NVs in speech, in this section, we study the acoustic differences between the spontaneous and acted dialogue by studying the distribution of the Global Prosodic features of Bone et al. (2014) on both databases. Recall that the Global Prosodic features are the median of log pitch, intensity, and voice quality (HF500) over the utterance (see Section 3.4.1.3). This small set of utterance-level acoustic features describes the major emotion-related acoustic characteristics of the speech signal and were shown to be predictive of emotions in cross-corpora studies. We study the scripted and non-scripted IEMOCAP utterances separately to have a more detailed understanding. Because the Expectancy dimension

annotations are missing in the IEMOCAP database, we only compare the distributions on the Arousal, Power, and Valence dimensions between the two databases in this section. Distributions of GP features on the Expectancy dimension in the AVEC2012 database and smoothed density plots of all GP features on both databases can be found in Section C.2 of Appendix C.

5.4.1 Distribution of Log Pitch

As shown in Table 5.4, the distribution of median log pitch is more skewed and has smaller standard deviation in spontaneous dialogue than in non-scripted and scripted dialogue. These observations indicate that the distribution of median log pitch values has more variation in acted dialogue than in spontaneous dialogue.

Table 5.4: Distribution of Log Pitch on AVEC2012 and IEMOCAP Databases

Databases	Mean	Standard Deviation	Skewness
AVEC2012	0.412	**0.172**	**-0.360**
IEMOCAP (non-scripted)	0.487	0.195	-0.190
IEMOCAP (scripted)	0.480	0.199	-0.191

5.4.2 Distribution of Intensity

As shown in Table 5.5, the distribution of median intensity is more skewed and has smaller standard deviation in spontaneous dialogue than in non-scripted or scripted acted dialogue. This indicates that the loudness of speech has wider range in acted dialogue than in spontaneous dialogue.

Table 5.5: Distribution of Intensity on AVEC2012 and IEMOCAP Databases

Databases	Mean	Standard Deviation	Skewness
AVEC2012	0.581	**0.095**	**-0.647**
IEMOCAP (non-scripted)	0.445	0.134	0.073
IEMOCAP (scripted)	0.455	0.134	0.232

5.4.3 Distribution of Voice Quality

Recall that HF500 is calculated as the ratio between the total energy above 500Hz and the lower-frequency energy in an utterance. As shown in Table 5.6, similar to intensity, the distribution of HF500 is more skewed and has smaller standard deviation in spontaneous dialogue than in non-scripted or scripted acted dialogue. This indicates that there is more voice quality variation in acted dialogue than in spontaneous dialogue.

Table 5.6: Distribution of Voice Quality on AVEC2012 and IEMOCAP Databases

Databases	Mean	Standard Deviation	Skewness
AVEC2012	0.570	**0.395**	**0.750**
IEMOCAP (non-scripted)	0.542	0.506	0.325
IEMOCAP (scripted)	0.400	0.463	0.747

5.5 Experiment 5: Influence of Dialogue Type on Effectiveness of DIS-NV Features

In previous sections, we illustrated differences of DIS-NV distribution and acoustic variations in spontaneous and acted dialogue. Our statistical analyses show that compared to spontaneous dialogue, in acted dialogue, there are fewer DIS-NVs and more acoustic variation. In this section, we conduct emotion recognition experiments on both the spontaneous AVEC2012 and the acted IEMOCAP databases to study how these differences between spontaneous and acted dialogues influence the effectiveness of features for emotion recognition in spoken dialogue.

5.5.1 Methodology

To study the influence of dialogue type on the effectiveness of DIS-NV features, we perform emotion recognition on both the spontaneous AVEC2012 database and the acted IEMOCAP database, and compare the experimental results (Tian et al., 2015a). The Support Vector Machine model described in Section 3.5.1 is used for building all the emotion recognizers. As discussed in Section 3.1.6, unlike the regression experiments in Chapter 4, we perform 10-fold cross-validation experiments

for classification experiments, and report the weighted F-measures to avoid the imbalanced data issue. We evaluate significance of performance differences using the paired Permutation test (Menke and Martinez, 2004) with 100,000 randomisations.

Five types of benchmark acoustic and lexical features are extracted on both databases to compare with our DIS-NV features, namely LLD (see Section 3.4.1.1) acoustic features, eGeMAPS acoustic features (see Section 3.4.1.2), GP acoustic features (see Section 3.4.1.3), non-sparse PMI lexical features (see Section 3.4.2.1), and CSA lexical features (see Section 3.4.2.2).

5.5.2 Results and Discussion

Results of the unimodal emotion recognition models on the AVEC2012 database are shown in Table 5.7. Results of the unimodal emotion recognition models on the IEMOCAP database are shown in Table 5.8. "Mean" represents the arithmetic mean of the results on all emotion dimensions. Note that the IEMOCAP database did not provide Expectancy annotations, thus the results on the Expectancy dimension is missing for the IEMOCAP database. We include a baseline model which predicts the majority class.

Consistent with our results in Experiment 1 in Section 4.3.1, DIS-NV features have the best performance in predicting the Expectancy emotion dimension on the spontaneous AVEC2012 database,[2] and achieved the best overall performance compared to other acoustic and lexical features.[3] However, as we expected, the DIS-NV features are less effective on the acted IEMOCAP database.

We also observe that acoustic features are more predictive of emotions than lexical features on the acted IEMOCAP database, while lexical features are more predictive than acoustic features on the spontaneous AVEC2012 database. This is consistent with our finding in Section 5.4 that acted emotions are more acoustically exaggerated. Our results indicate that effectiveness of features depends largely on the specific emotion recognition task, especially the type of dialogue.

[2]DIS-NV vs. IS10LLD on Expectancy: $p = 0.0089$

[3]DIS-NV vs. CSA on Arousal: $p = 0.0138$, on Expectancy $p = 0.0001$, on Power $p = 0.0140$, on Valence $p = 0.0421$ (not significant)

Table 5.7: Unimodal Emotion Recognition with SVM on the Spontaneous AVEC2012 Database

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	51.6	55.6	66.4	58.8	58.1
AVEC-LLD	52.4	60.8	67.5	59.2	60.0
IS10-LLD	52.9	60.8	67.6	59.2	60.1
eGeMAPS	56.9	60.1	73.4	66.8	64.3
GP	56.3	60.0	72.4	66.8	63.9
DIS-NV	55.9	**61.4**	**74.7**	66.8	**64.7**
PMI	55.7	60.7	73.0	66.8	64.0
CSA	**57.5**	59.8	73.0	**67.1**	64.4

Table 5.8: Unimodal Emotion Recognition with SVM on the Acted IEMOCAP Database

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	31.7	#	28.7	27.0	29.1
LLD	**65.2**	#	**53.8**	**53.5**	**57.5**
eGeMAPS	60.9	#	52.2	49.4	54.1
GP	57.0	#	49.7	41.5	49.4
DIS-NV	36.3	#	40.7	32.8	*36.6*
PMI	47.8	#	48.1	32.9	42.9
CSA	47.0	#	47.2	29.5	41.2

5.5.3 Summary

In Experiment 5, we showed that the type of dialogue has significant influence on effectiveness of features. Our DIS-NV features are predictive of emotions in spontaneous dialogue, but are less predictive of emotions in acted dialogue due to fewer occurrences of DIS-NVs in acted speech. Results also showed that acoustic features have better performance than lexical features for emotion recognition in acted dialogue due to exaggerated acting. Our findings indicate that it is import to study the data, especially the type of dialogue, before performing emotion recognition on spoken dialogue.

5.6 Experiment 6: Using Deep Learning for Unimodal Emotion Recognition

As discussed in Section 3.1.4, previous work has identified the Long Short-Term Memory Recurrent Neural Network (LSTM) model as highly predictive of emotions because of its ability to model long-range context. Our Experiment 3 also verifies that emotion recognition can benefit from including temporal context. However, our emotion recognition experiments so far all used the non-contextual SVM model. Therefore, in Experiment 6, we study the gain of using the deep, contextual LSTM model instead of the shallow, non-contextual SVM model. We build unimodal LSTM emotion recognition models on both databases (Tian et al., 2015b).

5.6.1 Methodology

Performance of unimodal LSTM models on the AVEC2012 and IEMOCAP databases is reported in Tables 5.9 and 5.10. "Mean" represents the arithmetic mean of the results on the four emotion dimensions. LSTM models with a single hidden layer are used when building these unimodal models and the number of memory cells was selected based on cross-validation experiments.[4] We include a baseline model which predicts the majority class. Similar to Experiment 5, we conduct 10-fold cross-validation experiments on both databases and report the weighted F-measures, and evaluate significance of performance differences using the paired Permutation test (Menke and Martinez, 2004) with 100,000 randomisations.

5.6.2 Results and Discussion

As shown in Table 5.9, consistent with our previous findings, DIS-NV features are predictive of emotions in spontaneous dialogue, especially for predicting the Expectancy emotion dimension,[5] whether used with the SVM model or the LSTM model. The CSA lexical features benefit more from using the contextual LSTM model relative to the non-contextual SVM model, and achieve the best overall performance on the AVEC2012 database.[6] Compared to the performance of the SVM models shown in Table 5.7, the LSTM models achieve improved performance on all emotion dimensions

[4]The number of memory cells for each model is: LLD = 32, CSA = 16, GP = DIS-NV = PMI = 8.

[5]DIS-NV vs. PMI on Expectancy: $p << 0.0001$

[6]CSA vs. DIS-NV on Arousal: $p << 0.0001$, on Expectancy: $p = 0.00001$, on Power: $p = 0.0141$, on Valence: $p << 0.0001$

using each feature set. This indicates the effectiveness of the deep, contextual LSTM model for emotion recognition in spoken dialogue.

As discussed in Section 3.4.1.1, for a fair comparison, we also extract the InterSpeech 2010 LLD set (IS10-LLD) in addition to the AVEC2012 baseline LLD set (AVEC-LLD) on the AVEC2012 database. As shown in Tables 5.7 and 5.9, IS10-LLD set yields similar or improved performance for the SVM model[7] and the LSTM model[8] compared to the AVEC-LLD set. Therefore, in the AVEC2012 experiments in the Chapter 6, we report the results using the IS10-LLD feature set instead of the AVEC-LLD feature set.

As shown in Table 5.10, consistent with our previous findings, effectiveness of features varies when the type of dialogue is different. Compared to the performance of the SVM models shown in Table 5.8, the LSTM models achieve improved performance on all emotion dimensions using the smaller GP, DIS-NV, PMI, and CSA feature sets on the IEMOCAP database. However, performance of the LSTM model using the LLD and the eGeMAPS feature set is worse than the SVM models. This may be because the LLD and the eGeMAPS feature sets have higher dimensionality (1582 for LLD, 88 for eGeMAPS), which results in more complex LSTM models that have more parameters to optimize during the training. There are fewer training instances in the IEMOCAP database than in the AVEC2012 database (approximately 10,000 for IEMOCAP, approximately 50,000 for AVEC2012), which may limit the optimization of the LSTM models using the LLD and the eGeMAPS feature set. This issue of insufficient training data is consistent with previous emotion recognition studies using deep learning models which we discussed in Section 3.1.4.

Another reason for the LSTM model not bringing large performance gains may be that the IEMOCAP database annotated data at the utterance level, while the AVEC2012 database annotated data at the word level. The long-range temporal context the LSTM model attempts to incorporate may not be helpful or necessary when the time scale of the data instances is as long as an utterance. We also observed that the knowledge-inspired features perform better than the statistical LLD features on both spontaneous and acted dialogue in most cases, which is consistent with previous emotion recognition studies (Bone et al., 2014; Eyben et al., 2015b; Savran et al., 2012).

[7]For SVM models AVEC-LLD vs. IS10-LLD on Arousal: $p = 0.0169$, on Expectancy $p = 0.3608$, on Power $p = 0.1261$, on Valence $p = 0.2079$

[8]For LSTM models AVEC-LLD vs. IS10-LLD on Arousal: $p = 0.0007$, on Expectancy $p = 0.0015$, on Power $p = 0.0068$, on Valence $p = 0.0030$

Table 5.9: Unimodal Emotion Recognition with LSTM on the Spontaneous AVEC2012 Database

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	51.6	55.6	66.4	58.8	58.1
AVEC-LLD	56.5	61.6	72.1	66.4	64.2
IS10-LLD	57.1	61.4	72.7	67.1	64.6
eGeMAPS	56.2	60.3	72.6	66.8	64.0
GP	56.0	60.3	72.4	66.8	63.9
DIS-NV	56.2	**65.9**	72.8	67.3	*65.5*
PMI	56.0	62.7	72.3	66.7	64.4
CSA	**58.1**	61.7	**75.2**	**70.2**	**66.3**

Table 5.10: Unimodal Emotion Recognition with LSTM on the Acted IEMOCAP Database

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	31.7	#	28.7	27.0	29.1
LLD	53.7	#	46.2	38.6	46.2
eGeMAPS	**60.1**	#	**52.2**	**46.6**	**53.0**
GP	58.0	#	50.6	41.8	50.1
DIS-NV	41.6	#	37.8	34.0	37.8
PMI	48.8	#	48.7	32.9	43.5
CSA	50.0	#	48.1	44.5	47.5

5.6.3 Summary

In Experiment 6, we verified that, compared to the flat and non-contextual SVM model, the deep and contextual LSTM model typically yields better performance for emotion recognition. However, the effectiveness of the complex LSTM model may be limited by the amount of training data available. Considering the small size of emotion databases, using knowledge-inspired features with a neural network model will result in better recognition performance than having the neural network learn more abstract feature representations from noisy statistical features. This leads to our Experiment 8 in Section 6.2.2 where we combine knowledge-inspired features in our LSTM model for multimodal emotion recognition. Our results show that the feature representations automatically learned by deep learning models are still less

effective than knowledge-inspired features, which indicates that there is room for improvement for the ability of deep learning models to learn relevant features for emotion recognition.

5.7 Discussion

In this chapter, we discussed the differences between spontaneous and acted dialogue. We found that there are typically more DIS-NVs in spontaneous dialogue than in acted dialogue. We also found complex variations in non-verbal vocalisations in spontaneous dialogue which were overlooked when designing data collection by acting. According to Global Prosodic feature distributions, loudness and voice quality distribution in acted dialogue have wider value ranges than in spontaneous dialogue, and there is more pitch variation in acted dialogue than in spontaneous dialogue. Dialogue collected by non-scripted acting shares similarities with spontaneous dialogue, while there are fundamental differences between scripted acted dialogue and spontaneous dialogue.

Our cross-corpora experiments showed that DIS-NV features are less predictive of emotions in acted dialogue because there are fewer DIS-NVs in acted dialogue compared to spontaneous dialogue. However, this chapter only considered unimodal emotion recognition models. Recall that in Experiment 2 in Chapter 4, we found that DIS-NVs contain emotion-related information beyond the acoustic characteristics and lexical content in spontaneous dialogue. Thus, incorporating the DIS-NV features with benchmark acoustic and lexical features may improve the performance of emotion recognition in acted dialogue as well. Thus, we will study the performance of multimodal emotion recognition models incorporating the DIS-NV features on both spontaneous and acted dialogue in the Chapter 6.

We also investigated the gain of using deep, contextual LSTM models compared to using shallow, non-contextual SVM model for emotion recognition. Our results show that although the LSTM model achieves better performance than the SVM model in most cases, optimization of the complex LSTM model may be limited by the small amount of training data available. Therefore, it may be better to use the knowledge-inspired features with the LSTM model which have lower dimensionality and thus fewer parameters to optimize. In the future, to further study the benefit of using LSTM models for emotion recognition, we would also like to compare the performance of a LSTM model using embedding layers learned from the data-driven low-level features with a LSTM model using knowledge-inspired directly features. In

terms of the gain of including temporal contexts, we find that the IEMOCAP database which annotated emotions at the utterance-level benefits less from using the LSTM model than the AVEC2012 database which annotated emotions at the word-level. This indicates that the LSTM's ability to model long range temporal context may be more useful to emotion recognition tasks at a small time scale (e.g., frame or word level) instead of at a large time scale (e.g., utterance or conversation level).

Chapter 6

Multimodal Emotion Recognition with Hierarchical Fusion

The Solarians have given up something mankind has had for a million years; something worth more than (...) everything; because it's something that made everything possible (...) The tribe, sir. Cooperation between individuals.

— Isaac Asimov, *The Naked Sun (1956)*

In previous chapters, most of our experiments investigated unimodal emotion recognition. However, consistent with human studies on emotion recognition, including information from multiple modalities when building emotion recognition models typically yields better performance. Therefore, in this chapter, we study the performance of multimodal emotion recognition using DIS-NV features and benchmark acoustic and lexical features. In Experiment 2 of Chapter 4, we found that simple feature concatenation may not be enough to boost the gains of modality fusion. Thus, we are motivated to identify a better fusion strategy for multimodal emotion recognition. More specifically, we propose a HierarchicaL (HL) fusion strategy, which combines information in a knowledge-inspired hierarchical structure. In HL fusion, features that describe data at lower levels of abstraction (e.g., statistical features) or smaller time scales (e.g., frame-level features) are used at lower levels of the hierarchy. In this chapter, we study the effectiveness of the proposed HL fusion for multimodal emotion recognition using acoustic and lexical features compared with the benchmark Feature-Level (FL) fusion and Decision-Level (DL) fusion. We are aware that information beyond the acoustic and lexical modalities can contribute to emotion recognition as well. In fact, later in Chapter 7, we will investigate multimodal emotion recognition using other modalities in addition to the acoustic and lexical modality.

However, in this chapter, by multimodal model we refer to emotion recognition models using acoustic and lexical information.

6.1 Multimodal Emotion Recognition

As discussed in Section 3.1.5, humans convey and perceive emotions through all communicative modalities. We have better emotion recognition performance when given information from multiple modalities (Zeng et al., 2009). In addition, consistent with the human studies, multimodal emotion recognition models typically outperform unimodal emotion recognition models (D'Mello and Kory, 2012). There are two types of fusion strategy used in current multimodal emotion recognition models: Feature-Level (FL) fusion (or "early fusion") and Decision-Level (DL) fusion (or "late fusion"). In this section, we first describe the FL and DL fusion strategies and identify their limitations. We then propose the HL fusion strategy and discuss how it may improve multimodal emotion recognition performance compared to FL and DL fusion.

6.1.1 Feature-Level (FL) Fusion and Decision-Level (DL) Fusion

In FL fusion, feature sets from different modalities are concatenated before performing recognition, as shown in Figure 6.1 (e.g., Nazari et al. (2015)). In some studies, feature engineering is first applied to the concatenated feature set or to a unimodal feature set (e.g., Gievska et al. (2015)). However, it is hard to apply knowledge on intra-modality differences in FL multimodal models. In contrast, DL fusion applies a rule-based decision model (e.g., Wu and Liang (2011)) or a machine learning model (e.g., Pei et al. (2015)) over the predictions given by each unimodal model, as shown in Figure 6.2. Previous studies comparing these two fusion strategies show that DL fusion typically outperforms FL fusion (He et al., 2015a; Soleymani et al., 2012a). However, detailed information about features within each modality (i.e., inter-modality differences) is lost in the final decision model of DL fusion and interactions between features from different modalities are not modelled. This results in DL fusion having worse performance than FL fusion in cases where the features used are already highly predictive of emotions on their own (Huang et al., 2015b; Jin et al., 2015; He et al., 2015a).

Compared to unimodal models, the improvement given by multimodal models using FL and DL fusion is often limited (D'Mello and Kory, 2012). The reason may

be that both FL and DL fusion incorporate modalities at the same level. However, as discussed in Section 3.1.3, different features may be extracted at different time scales (e.g., frame-level vs. utterance-level) or have different levels of abstraction (e.g., statistical vs. knowledge-inspired).

The cognitive process during human dialogue is often defined as a four-level structure (Benotti, 2009). For example, the communication model proposed by Clark (1996) classifies communication into four steps: attention, identification, understanding and consideration. At the attention step, the listener becomes aware that his/her conversational partner is speaking. At the identification step, the listener perceives the acoustic variations and recognizes the content of the speech (s)he hears. After identifying the content, in the understanding step, the listener analyses the meaning of the sentences (s)he just recognized. At the final consideration step, the meaning conveyed in the perceived speech evokes specific reactions and verbal/emotional responses of the listener based on his/her personal memories, knowledge, or goals. Similarly, when perceiving emotions in dialogue, humans make use of information received at different cognitive steps at different time steps under influence of both long-term and short-term contexts (Grandjean et al., 2008). As suggested by the Appraisal emotion theory (Ortony et al., 1990), emotions are responses to the perceived relationship between the environment and a person's internal goals. Thus, in Clark's (1996) model, the acoustic features are perceived earlier at the identification step, the lexical features are perceived later at the understanding step, and emotions are produced and updated at the consideration step.

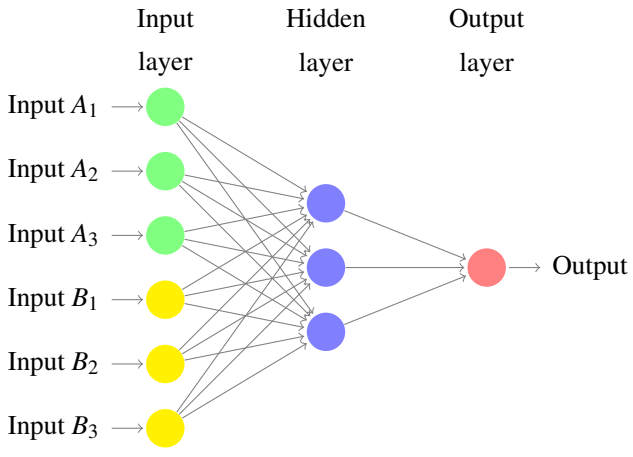


Figure 6.1: An Example of Feature-Level (FL) Fusion

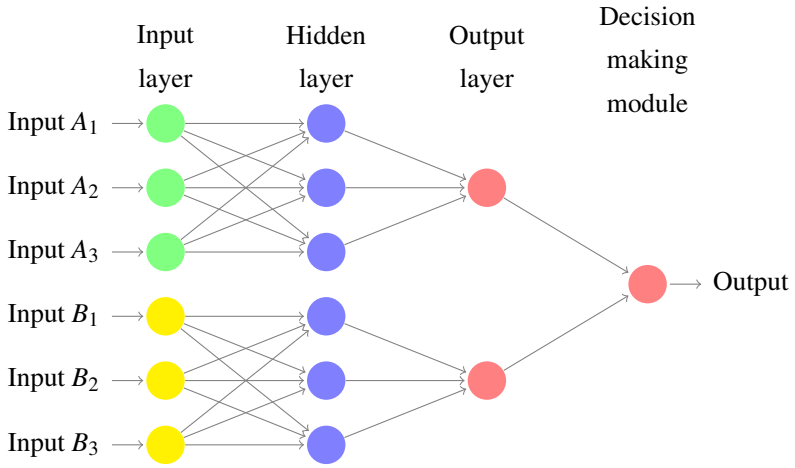


Figure 6.2: An Example of Decision-Level (DL) Fusion

6.1.2 HierarchicaL (HL) Fusion

To address the limitations of FL and DL fusion, we propose the HierarchicaL (HL) fusion strategy, which incorporates features that are extracted at a larger time scale (e.g., utterance-level) or are more abstract (e.g., knowledge-inspired) at higher levels of its hierarchical structure, as shown in Figure 6.3. Compared to FL fusion, HL fusion incorporates prior knowledge about different modalities by using features at different layers based on the time scale at which the features are extracted or the levels of abstraction of the features (intra-modality differences). Compared to DL fusion, HL fusion preserves more detailed information of features in each unimodal model (inter-modality differences) and the recognition model has access to individual features through the network structure when making the final decision. Because the knowledge-inspired hierarchy of HL fusion is able to model both inter- and intra-modality differences, our hypothesis is that multimodal models using the HL fusion strategy will have better performance than multimodal models using FL or DL fusion strategies.

Combining information in a hierarchical manner has been shown to improve the gain of information fusion in previous studies on using meta-data for model adaptation (e.g., Kuznetsov et al. (2016)). However, to the best of our knowledge, the only previous work using a similar hierarchical approach for multimodal emotion recognition is by Chen and Jin (2015). In their work, features from the audio, visual, and physiological modalities were used to recognize frame level continuous Arousal and Valence values in French dialogue. However, their hierarchical model differentiates between modalities, but does not take into account differences between

features within a single modality. In the work of Kim et al. (2015) which performed emotion recognition with frame level statistical acoustic features, a logistic regression model incorporating features derived from prosody, spectral envelope, and glottal information in a hierarchical structure outperformed a logistic regression model using all acoustic features at the input level. This indicates that a hierarchy capturing differences both between and within modalities is desirable for multimodal emotion recognition. The motivation for the hierarchical fusion of Chen and Jin (2015) is to address the fact that signals from different modalities change asynchronously. Performance of the hierarchical fusion of Chen and Jin (2015) outperformed FL fusion, but performed worse than DL fusion. The reason that the hierarchical fusion of Chen and Jin (2015) has limited performance may be that different feature sets have different levels of abstraction over the data. Compared to Chen and Jin (2015), our HL model has a knowledge-inspired structure that incorporates both inter- and intra-modality differences. The hierarchy of our HL model is motivated both by the temporal characteristics (the time scale at which the features are extracted) and the levels of abstraction of the features. In Experiment 6, we show that our HL fusion can outperform both FL and DL fusion.

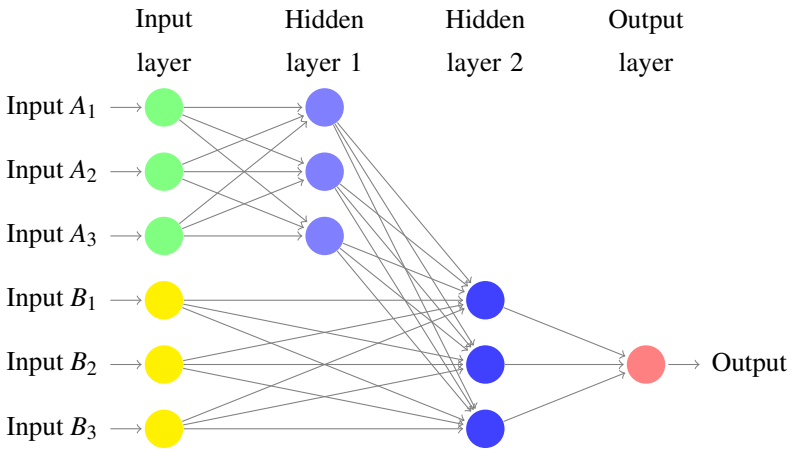


Figure 6.3: An Example of Proposed HierarchicaL (HL) Fusion

6.2 Multimodal Emotion Recognition with DIS-NV Features and HL Fusion

As discussed in Section 6.1.2, we propose HL fusion to overcome limitations of FL and DL fusion in multimodal emotion recognition. In this section, we study the

performance of multimodal emotion recognition using the proposed DIS-NV features and HL fusion strategy. The DIS-NV features are combined with the five benchmark acoustic and lexical feature sets used in Experiments 5 and 6 in Chapter 5. We compare the performance of multimodal emotion recognition models using HL fusion with those using FL and DL fusion on both the spontaneous AVEC2012 database and the acted IEMOCAP database. Consistent with the experimental settings in Chapter 5, we conduct 10-fold cross-validation experiments and report weighted F-measures. We also include a baseline model which always predicts the majority class.

6.2.1 Experiment 7: Multimodal Emotion Recognition with HL Fusion and All Features

To study the effectiveness of HL fusion, in Experiment 7, we build multimodal emotion recognition models using FL, DL, and HL fusion strategies and combine the DIS-NV features with all five benchmark feature sets (LLD, eGeMAPS, GP, PMI, and CSA) on both the spontaneous AVEC2012 database and the acted IEMOCAP database.

6.2.1.1 Methodology

We build an LSTM model with three hidden layers for the FL and HL fusion models in Experiment 7. The number of memory cells in each layer ($h_{bottom} = 32$, $h_{middle} = 16$, $h_{top} = 8$) is optimized by cross-validation experiments. The FL fusion model uses all of the features as inputs to the bottom hidden layer.

For the HL fusion model, the LLD and eGeMAPS features are used as inputs to the bottom hidden layer, the GP and DIS-NV features are added at the middle hidden layer, and the PMI and CSA features are added at the top hidden layer, as shown in Figure 6.4. The LLD and eGeMAPS features are input at the bottom level because in terms of the time scale at which the features are extracted, they are frame-level features while the other features are utterance-level features. The PMI and CSA features are used at the top level because in terms of the abstraction level, they encode prior information about emotional states based on lexical identity and thus have a higher level of abstraction. Note that the number of neurons shown in this figure is only an indication and is not the number of neurons used in the actual models. In the actual models, the number of input neurons equals the number of features used, and the number of neurons in the hidden layers are optimized by cross-validation experiments.

For DL fusion, predictions given by unimodal models are used as inputs to

another LSTM model which has one hidden layer with 16 memory cells (optimized by cross-validation experiments). We also tested using rule-based decision models (e.g., selecting the class with highest probability). However, these models performed worse than DL fusion using a LSTM model as the decision model, thus we focus on the latter.

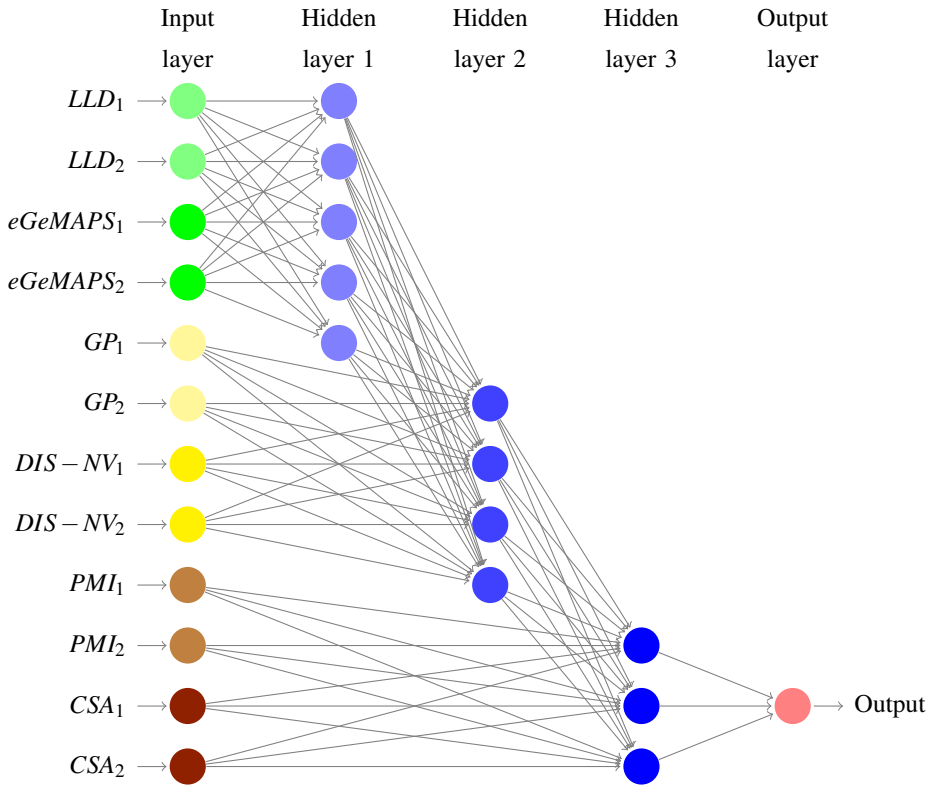


Figure 6.4: Structure of HL Model using All Features

6.2.1.2 Results and Discussion

Results of 10-fold cross-validation experiments of multimodal emotion recognition using all feature sets on both databases are shown in Tables 6.1 and 6.2.[1] As we can see, HL fusion outperforms FL fusion on all emotion dimensions on both databases. In addition, HL fusion outperforms DL fusion on most emotion dimensions on both databases. The only exception is the Valence dimension of the IEMOCAP database.[2] As shown in Table 6.2, the large LLD feature set is not highly predictive of Valence for the IEMOCAP database. As the negative influence of the LLD feature set is

[1] In Table 6.1, LLD, FL, DL, HL results are slightly different from Tian et al. (2016). This is because here we use the InterSpeech 2010 LLD feature set for the AVEC2012 experiments, instead of the AVEC2012 baseline LLD feature set used in Tian et al. (2016).

[2] $p << 0.0001$ in all cases, except for $p = 0.0001$ for HL vs. DL on Valence for IEMOCAP

larger for the HL model than for the DL model, the HL model performs worse than the DL model in this particular case. Moreover, there are fewer training instances in the IEMOCAP database (approximately 10,000) than in the AVEC2012 database (approximately 50,000). This limits the performance of the HL model, which has more parameters to fit than the DL model.

Table 6.1: AVEC2012 Multimodal Emotion Recognition with All Features

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Unimodal LSTM Models					
Baseline	51.6	55.6	66.4	58.8	58.1
LLD	57.1	61.4	72.7	67.1	64.6
eGeMAPS	56.2	60.3	72.6	66.8	64.0
GP	56.0	60.3	72.4	66.8	63.9
DIS-NV	56.2	**65.9**	72.8	67.3	65.5
PMI	56.0	62.7	72.3	66.7	64.4
CSA	**58.1**	61.7	**75.2**	**70.2**	**66.3**
Multimodal LSTM Models Using All Features					
FL	56.6	63.7	72.5	68.2	65.3
DL	58.7	65.3	73.4	69.2	66.7
HL	**59.4**	**67.9**	**73.7**	**70.9**	**68.0**

Table 6.2: IEMOCAP Multimodal Emotion Recognition with All Features

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Unimodal LSTM Models					
Baseline	31.7	#	28.7	27.0	29.1
LLD	53.7	#	46.2	38.6	46.2
eGeMAPS	**60.1**	#	**52.2**	**46.6**	**53.0**
GP	58.0	#	50.6	41.8	50.1
DIS-NV	41.6	#	37.8	34.0	37.8
PMI	48.8	#	48.7	32.9	43.5
CSA	50.0	#	48.1	44.5	47.5
Multimodal LSTM Models Using All Features					
FL	53.4	#	48.7	37.1	46.4
DL	52.4	#	50.3	**47.4**	50.0
HL	**57.3**	#	**51.1**	*45.4*	**51.3**

6.2.1.3 Summary

Experiment 7 on multimodal emotion recognition using all extracted features showed that the HierarchicaL (HL) fusion strategy we proposed typically achieves better performance than the FL and DL fusion strategies. This verified that modelling both inter- and intra-modality differences can improve the performance of emotion recognition in both spontaneous and acted dialogue. However, our results also show that when the amount of training data is limited, including noisy statistical features with high dimensionality may have a negative effect on the performance of multimodal emotion recognition models.

6.2.2 Experiment 8: Multimodal Emotion Recognition with HL Fusion and Knowledge-Inspired Features

In Experiment 7, we found that the performance of the multimodal emotion recognition model may be limited by the issue of insufficient training data. In Experiment 8, we study whether or not we can improve the performance by building multimodal models using only knowledge-inspired feature sets. In addition, to better understand the influence of incorporating prior knowledge when designing the recognition model, we investigate building HL models with randomly grouped features instead of designing

the hierarchy based on prior knowledge about the features.

6.2.2.1 Methodology

When building the multimodal models using a subset of all features, we remove the PMI features for experiments on both databases. This is because both the PMI and CSA features describe relations between lexical content and emotions. However, the PMI features are likely to over-fit the database that the emotion recognition experiment is performed on. Thus, they are less robust than the CSA features which are non-domain-specific. For the AVEC2012 database, we remove both the LLD and eGeMAPS features because they are relatively high in dimensionality yet low in effectiveness. For the IEMOCAP database, because the eGeMAPS features are highly effective in this specific case, we only remove the LLD features. We simplify the HL and FL models by removing the input neurons connected to the removed features. Structures of the HL models using only knowledge-inspired features are shown in Figures 6.5 and 6.6. For DL fusion, predictions given by unimodal models are used as inputs to another LSTM model which has one hidden layer with 8 memory cells (optimized by cross-validation experiments).

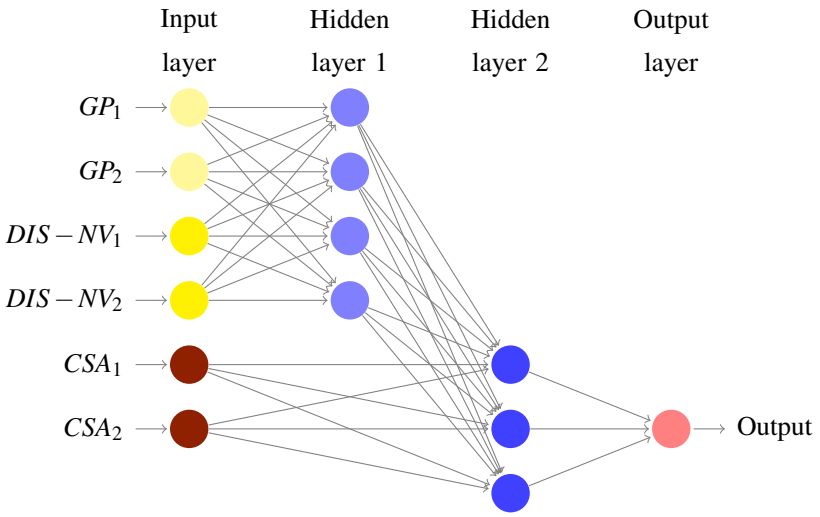


Figure 6.5: Structure of the HL Model for the AVEC2012 Database using GP, DIS-NV and CSA Features

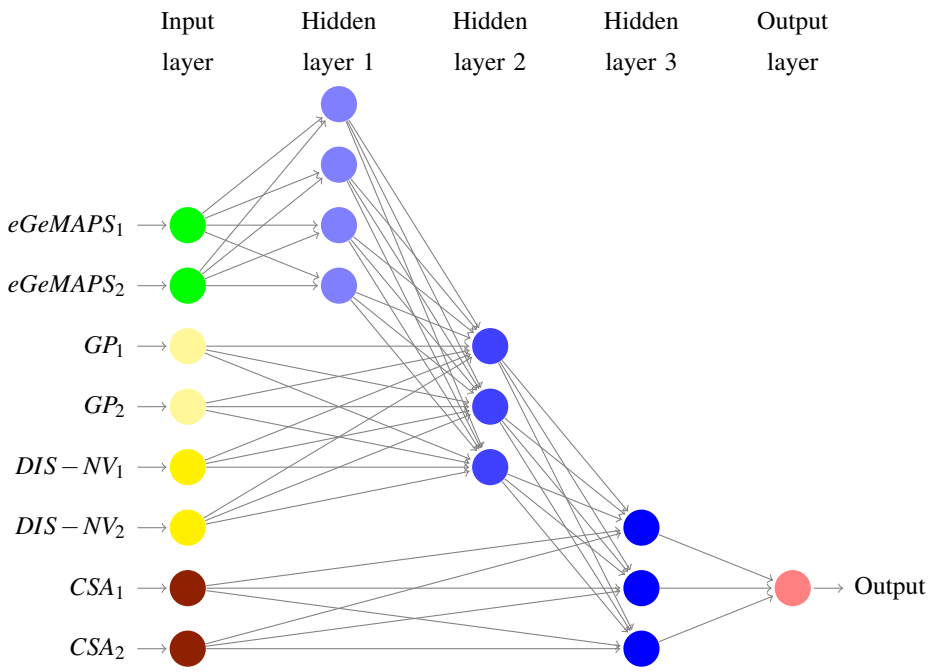


Figure 6.6: Structure of the HL Model for the IEMOCAP Database using eGeMAPS, GP, DIS-NV and CSA Features

6.2.2.2 Results and Discussion

Results of the multimodal models using only knowledge-inspired features are reported in Tables 6.3 and 6.4. As we can see, performance of HL and FL models increases when using only smaller, knowledge-inspired feature sets. Performance of DL models decreases compared to using all features. This is because the decision making module of the DL model only has access to final outputs of each unimodal model. The lack of data issue has smaller influence on the DL decision making module which has a simpler neural network structure than on the FL or HL model. Thus, after removing feature sets, the decision making module of the DL model does not benefit from the simplified model structure as much as the FL or HL model, while having less input information. When using only knowledge-inspired features, HL fusion outperforms both FL and DL fusion on predicting all emotion dimensions on both databases. The HL model also achieves the best performance on all emotion dimensions on both databases compared to all unimodal and multimodal emotion recognition models we built so far.[3] Compared to the hierarchical fusion model of Chen and Jin (2015), our HL model incorporates prior knowledge on differences both between and within modalities. Such knowledge-inspired structure of our HL model is able to model both

[3] $p << 0.0001$ in all cases

inter- and intra-modality differences. Our HL model achieves results better than both FL and DL fusion, while the hierarchical fusion model of Chen and Jin (2015) only outperformed the FL fusion. This verifies the efficacy of our HL model on multimodal emotion recognition in spoken dialogue.

To better understand the benefits of incorporating features in a knowledge-inspired structure, we investigate grouping features randomly instead of basing on prior knowledge when used in the HL model. We use the same LSTM model for the HL model with random feature grouping. As shown in Tables 6.3 and 6.4, the random-grouping HL models ("HL(Random)" in the tables) have significantly worse performance than the original HL models using a knowledge-inspired structure for predicting emotion dimensions on both databases.[4] The random-grouping HL model has similar overall (mean) performance with the FL and DL model. This indicates that compared to using features at the same level, incorporating features in a hierarchical structure only brings improvement when prior knowledge on the features is considered when designing the hierarchical structure.

Table 6.3: AVEC2012 Multimodal Emotion Recognition with GP, DIS-NV, and CSA Features

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Unimodal LSTM Models					
Baseline	51.6	55.6	66.4	58.8	58.1
LLD	57.1	61.4	72.7	67.1	64.6
eGeMAPS	56.2	60.3	72.6	66.8	64.0
GP	56.0	60.3	72.4	66.8	63.9
DIS-NV	56.2	**65.9**	72.8	67.3	65.5
PMI	56.0	62.7	72.3	66.7	64.4
CSA	**58.1**	61.7	**75.2**	**70.2**	**66.3**
GP + DIS-NV + CSA					
FL	60.1	68.1	74.8	71.7	68.7
DL	56.6	63.3	73.5	68.0	65.3
HL	**61.8**	**69.2**	**76.2**	**72.4**	**69.9**
HL(Random)	60.8	68.2	75.0	71.5	68.9

[4] $p << 0.0001$ in all cases

Table 6.4: IEMOCAP Multimodal Emotion Recognition with eGeMAPS, GP, DIS-NV, and CSA Features

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Unimodal LSTM Models					
Baseline	31.7	#	28.7	27.0	29.1
LLD	53.7	#	46.2	38.6	46.2
eGeMAPS	**60.1**	#	**52.2**	**46.6**	**53.0**
GP	58.0	#	50.6	41.8	50.1
DIS-NV	41.6	#	37.8	34.0	37.8
PMI	48.8	#	48.7	32.9	43.5
CSA	50.0	#	48.1	44.5	47.5
eGeMAPS + GP + DIS-NV + CSA					
FL	55.2	#	50.8	47.2	51.1
DL	51.6	#	49.7	46.8	49.3
HL	**61.7**	#	**52.8**	**51.2**	**55.3**
HL(Random)	57.9	#	51.5	43.9	51.1

6.2.2.3 Summary

In Experiment 8, we showed that with a limited amount of training data, using smaller, knowledge-inspired feature sets improves performance of multimodal emotion recognition. Consistent with Experiment 7, the proposed HL fusion outperforms FL and DL fusion on predicting all emotion dimensions of both spontaneous and acted data, which verifies the efficacy of the proposed HL fusion. Our results also indicate that in order to achieve better performance, it is important to incorporate prior knowledge of features when designing the structure of an emotion recognition model.

6.3 Discussion

In this chapter, we proposed the HierarchicaL (HL) fusion strategy for multimodal emotion recognition. The HL model incorporates different feature sets in a knowledge-inspired hierarchical structure. We compared HL fusion with the widely used Feature-Level (FL) and Decision-Level (DL) fusion strategies. Experiments on two emotion databases of different types of dialogue show that HL fusion consistently outperforms FL and DL fusion. The HL model achieves state-of-the-art

performance for recognizing emotions in spoken dialogue compared to other unimodal and multimodal models using acoustic and lexical information. While developed for emotion recognition in spoken dialogue, the proposed HL fusion could in principle, be applied to other multimodal recognition tasks, as it allows us to incorporate specific knowledge of feature abstraction and the time scale at which the features are extracted.

Note that lack of training data may limit the performance of the HL emotion recognition model. We addressed this issue by using only knowledge-inspired features in the multimodal model and achieved improved performance. However, as discussed in Section 3.1.4, semi-supervised and unsupervised methods may also be helpful for addressing the issue of lack of labelled data. Therefore, in the future, we plan to study how we can improve performance of our emotion recognition models further by applying such approaches. Moreover, in this study, due to time constraints we did not discuss possible improvements given by modifications to the network architectures. For example, using state-of-the-art techniques such as pre-training the LSTM models, or decision-level combination of the unimodal models, the FL model, and the HL model. It is possible that there is room for performance improvements by using more sophisticated neural network architectures. However, this is not the focus of our research here and will be left for future study.

In our research so far, we have found that it is beneficial to include prior knowledge on human emotions for emotion recognition in spoken dialogue, whether in the features extracted or in the recognition model structure. We are interested to study whether this finding can be generalized to emotion-related tasks other than emotion recognition in spoken dialogue. Thus, in the next chapter, we will apply the proposed DIS-NV features and HL fusion to predict movie-induced emotions and study their efficacy for this task.

Chapter 7

Generalizing the Proposed Approaches to Other Emotion-Related Tasks

All knowledge is one. When a light brightens and illuminates a corner of a room, it adds to the general illumination of the entire room.

— Isaac Asimov, *Atom: Journey Across the Subatomic Cosmos (1991)*

In previous chapters, we proposed the DIS-NV features and the HL fusion strategy for emotion recognition in spoken dialogue. Our experiments verified the efficacy of the proposed approaches on emotion recognition in both spontaneous and acted dialogue. In this chapter, we explore how other emotion-related tasks can benefit from the DIS-NV features and HL fusion. In particular, we collaborated with Michal Muszynski, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel from the University of Geneva, and applied the DIS-NV features and the HL fusion to predict the emotional responses of movie audiences induced by affective movie content. (Tian et al., 2017).

Recognizing movie-induced emotions is a challenging task in current Affective Computing studies. Previous work has focused on using audiovisual movie content to predict movie-induced emotions. However, it is unknown whether or not information beyond the audiovisual content, such as dialogue cues and aesthetic highlights[1], will contribute to recognizing movie-induced emotions. Moreover, the relationship between the audience's perceptions of the affective movie content

[1]Aesthetic highlight refers to significant movie moment judging by its aesthetic importance and artistic value

(perceived emotions) and the emotions evoked in the audience (induced emotions) remains unclear. Previous work hypothesized the perceived and induced emotions to be consistent (Hanjalic and Xu, 2005). However, this may not always be the case. For example, in a thriller where the audience is aware of an upcoming threat to the movie characters, the audience may feel anxious (induced emotion) while watching a happy scene (perceived emotion). We identify three perspectives of emotions in movies, as shown in Figure 7.1: the audience's perspective, the actor's perspective, and the director's perspective. Movie audiences interpret the movie content and perceive the emotions it conveys (the perceived emotions). This then induces emotional responses which the audience feels (the induced emotions). Movie actors express emotions based on their interpretation of the script and may experience emotions themselves during acting (the expressed emotions). Movie directors create scripts with expectations of what emotions they intend the movie to induce in the audiences (the intended emotions). Here we focus on the audience perspective of movie emotions.

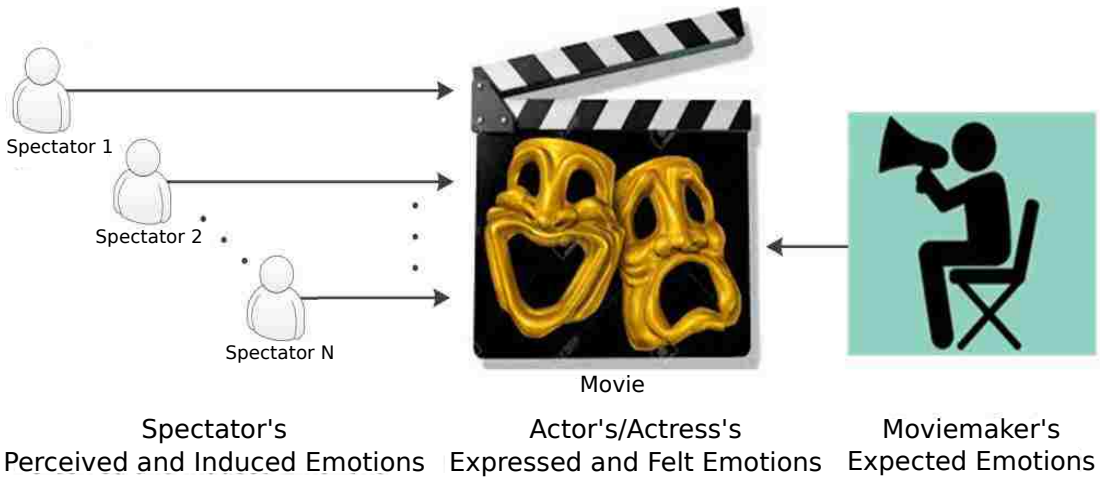


Figure 7.1: The Three Perspectives of Emotions in Movies

In this chapter, we study whether or not our DIS-NV features and HL fusion model are effective for recognizing movie-induced emotions, as well as the relationship between emotions perceived from and induced by movies. First, we transcribe the movies of the LIRIS-ACCEDE database (Baveye et al., 2015b) and label DIS-NVs and movie highlights. We then apply the HL fusion strategy to build multimodal models for recognizing movie-induced emotions. Second, we annotate emotions perceived from the movies in a crowd-sourced manner to study the relationship between perceived and induced emotions of movie audience. Note that the induced emotion annotations provided by the LIRIS-ACCEDE database are annotated on the Arousal and Valence

dimensions, while the perceived emotion annotations we collected are annotated on the Arousal, Power, and Valence dimensions.

7.1 Applying DIS-NV Features and HL Fusion to Affective Content Analysis

Recently, increased attention has been paid to recognizing emotions in spectators induced by affective content, motivated by potential applications, such as emotion-based content delivery (Hanjalic, 2006) or video indexing and summarization (Arifin and Cheung, 2008). However, recognizing the emotions induced by affective content remains a challenging task, with only weak to moderate correlations achieved between automatic predictions and human annotations of emotions (Dellandréa et al., 2016). This limits the efficacy of affective content analysis in related applications.

State-of-the-art studies on recognizing induced emotions have focused on extracting features from the audiovisual content of the stimuli. However, lexical content, such as movie dialogue or lyrics of songs may also influence the emotional response of the audience. For example, movie dialogue has been shown to be effective for recognizing violence in movies (Gninkoun and Soleymani, 2011). Moreover, because of the suggested positive correlation between perceived and induced emotions (Hanjalic and Xu, 2005), cues of perceived emotions in movies may be used for the recognition of induced emotions as well. Thus, we are motivated to study the efficacy of our DIS-NV features for recognizing movie-induced emotions. Beyond feature effectiveness, how the features are modelled also influences recognition performance. Thus, we study the impact of temporal context (history) on the recognition model and the application of our HL fusion strategies for combining multimodal information beyond the acoustic and lexical modalities.

To conduct our experiments, we chose the recently collected LIRIS-ACCEDE database (Baveye et al., 2015b) which contains continuous Arousal and Valence annotations of movie-induced emotions. Note that unlike the AVEC2012 database, the Power and Expectancy dimensions are not annotated in the LIRIS-ACCEDE database. This database has been widely used in state-of-the-art studies on movie-induced emotions, including benchmark challenges such as MediaEval2016 (Dellandréa et al., 2016). To study the relationship between emotions perceived from and induced by movies, we collect crowd-sourced annotations on perceived Arousal, Valence, and

Power of the movie dialogue. We add manual transcripts of the LIRIS-ACCEDE movies, as well as expert annotations of DIS-NV in dialogues and aesthetic highlights (Kostoulas et al., 2015a).

7.1.1 Review of Affective Content Analysis

The field of affective content analysis studies the relationship between information conveyed by the stimuli and emotional responses it evoked in the spectator. It remains a challenging task where only limited performance has been achieved for predicting induced emotions using movie-based features (Baveye et al., 2017). Here we briefly review the state-of-the-art of affective content analysis. In particular, we investigate previous work on induced emotion recognition on the LIRIS-ACCEDE database to identify limitations that we can improve.

The LIRIS-ACCEDE database was collected and released to provide resources for researchers to collaborate on affective content analysis (Baveye et al., 2015b). Here we focus on the continuous subset of the LIRIS-ACCEDE database, which contains 30 full movies, totalling 442 minutes (Baveye et al., 2015a). The 30 movies included in the continuous LIRIS-ACCEDE database are short independent movies ($<$ 20 minutes) shared under the Creative Commons licenses with diverse content and genres. During data collection, 10 participants watched each movie once and annotated continuous Arousal and Valence scores (value range [−1,1]) of the emotions they felt during watching (movie-induced emotions). The means of scores given by the participants over each second of the movie were used as the gold-standard annotations. A follow-up study screened these 30 movies to another 13 participants wearing sensors and collected physiological and behavioural measurements of the audiences during the movie (Li et al., 2015).

Table 7.1 provides an overview of the state-of-the-art for recognizing induced Arousal (A) and Valence (V) of movie audiences using the LIRIS-ACCEDE database. Note that annotations of the Power and Expectancy dimensions are missing in the LIRIS-ACCEDE database. Regression models in previous work include Support Vector Regression (SVR) (Boser et al., 1992), Long Short-Term Memory Recurrent Neural Networks (LSTM) (Hochreiter and Schmidhuber, 1997), Partial Least Squares (PLS) (Abdi, 2003), and Convolutional Neural Networks (CNN) (Cireşan et al., 2010).

As we can see, previous work on the LIRIS-ACCEDE database predicted movie-induced emotions with various regression models. The Pearson Correlation

Table 7.1: Mean Squared Error (MSE) and Pearson's Correlation Coefficients (CC) Reported in Previous Work

Model	A-mse	A-cc	V-mse	V-cc
Multimodal Models				
AudioVisual SVR (Baveye et al., 2015b)	0.326	0.242	0.343	**0.221**
AudioVisual SVR (Anastasia and Leontios, 2016)	#	**0.265**	#	0.110
AudioVisual SVR (Chen and Jin, 2016)	**0.120**	0.236	**0.099**	0.142
AudioVisual LSTM (Ma et al., 2016)	0.124	0.054	0.102	0.017
Unimodal Models				
Audio PLS (Jan et al., 2016)	0.129	-0.072	0.141	-0.062
Visual CNN (Baveye et al., 2015a)	**0.021**	0.152	**0.027**	0.197
Visual SVR (Baveye et al., 2015a)	0.022	**0.337**	0.034	**0.296**
Visual SVR (Liu et al., 2016b)	0.126	0.056	0.106	0.019

Coefficient (CC) is the most commonly reported evaluation metric. Mean Squared Error (MSE) is sometimes reported in addition (e.g., Baveye et al. (2015b)). Only weak or moderate correlations have been achieved in state-of-the-art studies,[2] which shows that recognizing induced emotions of movie audiences is a challenging task. Note that different studies have different experiment protocols, such as data pre-processing and training-testing partitions. Thus, their results may not be directly comparable. Previous work has focused on using features extracted from audiovisual movie content (e.g., Anastasia and Leontios (2016); Chen and Jin (2016)). However, lexical information from the movie dialogue is overlooked, even though it has proved to be important in other emotion recognition studies (Poria et al., 2017). Moreover, the usefulness of knowledge-inspired affective cues in movies, such as aesthetic highlights (Li et al., 2015), has not been explored for predicting movie-induced emotions. Another important limitation of current studies using movie content features to predict movie-induced emotions is that they overlooked the audience reactions, while emotion induction is heavily influenced by individual variances.

Many previous studies have examined unimodal models for induced emotion recognition (e.g., Jan et al. (2016); Liu et al. (2016b)). In fact, Baveye et al. (2015a) built a SVR model using only visual features and achieved best reported CC for this task. However, combining multimodal information has improved performance for

[2]The best reported CC for Arousal is 0.337, for Valence is 0.296 (Baveye et al., 2015a)

a number of other emotion recognition tasks (e.g., Tian et al. (2016)). Thus, we are motivated to study modality fusion strategies that may benefit induced emotion recognition. In addition, the LSTM model has low performance for predicting movie-induced emotions (Ma et al., 2016),[3] yet it has achieved leading performance in various emotion recognition tasks due to its ability to model temporal context (e.g., Brady et al. (2016)). Ma et al. (2016) predict movie-induced emotions at an interval of 10 seconds, which already contains temporal context. This may limit the gain when using a LSTM model to include more history. However, the suitable amount of history to include for predicting movie-induced emotions is unclear. In Section 7.2, we address the above limitations in the state-of-the-art of movie-induced emotion recognition by studying the effectiveness of features beyond audiovisual movie content, and by testing the gain of including history and combining multimodal information.

7.1.2 Transcription and DIS-NV Annotations of LIRIS-ACCEDE

Here we provide protocols for collecting the extended annotations of the continuous LIRIS-ACCEDE database. We choose the 8 English movies listed in Table 7.2 because they contain relatively more dialogue. Moreover, these movies are in the double-reality art form, where the lead characters switch between two worlds. This mirrors the activity of movie-watching where the reality and the movie world together create double-reality experience for the movie audience. Thus, the audiences may empathize more with the movie characters which is a particularly interesting scenario for understanding perceived and induced emotions. In total, we annotated 118 minutes of movies containing 870 utterances.

The movie transcription and affective cue annotation was conducted by two expert annotators. To increase the annotation speed, audio recordings of the movie were first passed through the IBM Watson Speech-to-Text service,[4] which provides automatic speech transcription with word timings. This auto-generated transcript was then manually corrected and annotated by the annotators in parallel, each annotating five movies. To evaluate the annotation agreement, *First Bite* and *Spaceman* were annotated by both annotators and we computed the Normalized Damerau-Levenshtein (NDL) distances (Bard, 2007) of their transcripts, as well as the Pearson Correlation Coefficient (CC) of the word timings.

NDL distance is a widely used measurement of the distance between two strings.

[3] CC for Arousal is 0.054, for Valence is 0.017 (Ma et al., 2016)

[4] `https://www.ibm.com/watson/developercloud/speech-to-text.html`

Table 7.2: Statistic of Selected LIRIS-ACCEDE Movies

Movie	Genre	Utterance Count	Mean Utterance Duration (s)	Total Duration (s)
After the Rain	Drama	77	3.000	231.000
First Bite	Romance	54	2.056	111.024
Nuclear Family	Comedy	147	2.694	396.018
Payload	Adventure	121	2.488	301.048
Spaceman	Adventure	133	2.489	331.037
Superhero	Drama	161	2.832	455.952
Tears of Steel	Adventure	79	2.165	171.035
The Secret Number	Drama	98	2.724	266.952

It is computed as the minimum number of operations required to transform one string to the other, divided by the length of the longer string of the pair. NDL distance of 0 indicates that the two strings are identical. Thus, values closer to 0 show stronger annotation agreement. We find that 94.8% of the words transcribed are identical for the two annotators, with average NDL distance of 0.049. Considering the average length of words is 4 characters in the compared transcript, an average NDL distance of 0.049 means for every five words there is less than one character difference. CC for the word and utterance timings of the transcript is reported in Table 7.3. As we can see, the utterance and word timings annotated by the two annotators are strongly correlated. This verifies that the two annotators strongly agreed on movie transcription. We also report the concordance correlation coefficient (CCC) (Lawrence and Lin, 1989) which takes into account both the correlation and the value shift. CCC between two example variables x and y is calculated as below:

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} \tag{7.1}$$

Where μ_x, μ_y are the means of the two variables, σ_x^2, σ_y^2 are the corresponding variances, ρ_{xy} is the CC between the two variables. CCC is a widely used measurement for agreement on continuous values. As we can see, the utterance and word timings annotated by the two annotators are strongly correlated. This verifies that the two annotators strongly agreed on movie transcription.

The same annotators also annotated DIS-NVs in movie dialogue.[5] To test the

[5]All remaining words other than the 5 DIS-NVs are labelled as “General lexicon”

Table 7.3: Movie Transcription and DIS-NV Label Agreement

Labels	StartTime (CC)	EndTime (CC)	StartTime (CCC)	EndTime (CCC)
Utterance	0.998	0.998	0.997	0.998
Word	0.999	0.999	0.999	0.999
General lexicon	0.989	0.989	0.988	0.988
Filled pause	0.625	0.625	0.560	0.549
Filler	0.920	0.920	0.744	0.744
Stutter	0.916	0.916	0.835	0.836
Laughter	0.635	0.635	*0.369*	*0.369*
Audible breath	0.766	0.764	0.620	0.637

agreement, we divide the annotations into six subsets based on the DIS-NV labels and compute both CC and CCC of the start and end timings of words in each subset. As shown in Table 7.3, although the annotation agreement on DIS-NV labels is lower compared to movie transcription, the annotations remain strongly correlated. Note that some types of DIS-NV are rare in movie dialogue. Thus, a small variance in the number of instances can result in large differences in the CC and CCC. For example, laughter has the lowest agreement in Table 7.3. However, the only difference between the two annotators is that one annotator labelled two giggles at the beginning of *First Bite* while the other annotator did not label them, yet there are only seven laughs in total in the compared movie dialogue.

7.2 Experiment 9: Recognizing Movie-Induced Emotions with DIS-NV Features and HL Fusion

In Experiment 9, we discuss our unimodal and multimodal experiments on recognizing emotions induced by movies. In our unimodal experiments, we first study the influence of temporal context by building LSTM models with different time steps, then compare the effectiveness of different features, including our DIS-NV features, for movie-induced emotion recognition. In our multimodal experiments, we study the gain from combining multimodal information with the HL fusion strategy compared to FL and DL fusion.

7.2.1 Methodology

The original Arousal and Valence annotations on the LIRIS-ACCEDE database are provided for each second of the movie. To include a suitable amount of data for feature extraction, we use a 5 second sliding window with a 4 second overlap between neighbouring windows to compute all features. The average Arousal and Valence scores over each window are used as the gold-standard induced emotion annotations. We also remove the end credits of each movie because participants started to remove the wearable sensors at this point, which introduced outliers in the signals. This results in 7103 data instances in total.

7.2.1.1 Movie Based Features

Similar to previous work (e.g., Ma et al. (2016)), we extract features from the audiovisual movie content with OpenSMILE (Eyben et al., 2010b). For each sliding window, we extract 1582 InterSpeech 2010 Paralinguistic Challenge Low-Level Descriptor audio features (Schuller et al., 2010b) and 1793 visual features. The latter are histograms of Local Binary Pattern, HSV (hue, saturation, and value), and optical flow of each image region (Eyben et al., 2016). These are standard benchmark features used in various emotion recognition tasks (Poria et al., 2017). To reduce feature dimensionality, we apply the ReliefF algorithm (Robnik-Šikonja and Kononenko, 2003) and rank the individual effectiveness of features by performing regression with 20 nearest neighbours. We select the top 100 audio features and the top 100 visual features for Arousal and Valence respectively in order to maintain similar feature set sizes between different feature sets. We conduct ReliefF feature ranking on the remaining 22 movies of the continuous LIRIS-ACCEDE database outside the 8 movies we perform recognition experiments on. This allows us to incorporate in-domain knowledge and avoid including test data during feature selection.

Besides the data-driven audiovisual features, we additionally extract three knowledge-inspired feature sets. These include DIS-NV features (see Section 7.1.2), CSA lexical features computed from the movie transcript (see Section 3.4.2.2), and aesthetic movie highlights over each window. Note that different from Section 7.1.2, here we extract six DIS-NV features (including General lexicon), which are computed as the total duration of each type of DIS-NV in each sliding window divided by the window length of 5 seconds.

The aesthetic movie highlights correspond to critical movie moments defined

by experts in terms of art form and content (Kostoulas et al., 2015a). They are knowledge-inspired cues and are more abstract than the audiovisual movie content. The motivation of using aesthetic highlights to predict movie induced emotions is that moments with high aesthetic value can evoke emotions in the audience as a result of simulation and empathy. Moreover, aesthetic movie highlights may contain art techniques designed intentionally to arousal strong emotional responses in the audience. We record occurrences of six aesthetic highlights in each window:

- Form highlights:
  - Spectacular: e.g., technical choices, special effects.
  - Subtle: e.g., use of the camera, lightening, music.
- Content highlights:
  - Character: e.g., character's emotions (the expressed emotions in Figure 7.1), responses to increasingly dramatic events.
  - Dialogue: e.g., clarifying the action's motivation, showing tensions among the characters.
  - Theme: e.g., unusual close-up, development of the urban theme.
- Any type of highlight above has occurred.

The knowledge-inspired features are more sparse than the audiovisual features. These dialogue cues and highlights are infrequent events in the movie. Thus, the majority of the knowledge-inspired feature values are zero vectors.

7.2.1.2 Audience Reaction Based Features

To compare with movie based features, we include two audience reaction based feature sets, namely physiological features and behavioural features. We use a third order low-pass Butterworth filter with cut-off frequency at 0.3Hz to filter physiological and behavioural signals of the movie audience before feature extraction. This is a widely used filtering technique in Affective Neuroscience to reduce noise (Li et al., 2015). The physiological features are 273 statistics over the sliding window based on the original measurement of the electrodermal activity of the audience and its first and second derivatives (Li et al., 2015). The behavioural features are 273 statistics over the sliding window based on the original measurement of signals collected from acceleration

sensors attached to the audience's hands and its first and second derivatives (Kostoulas et al., 2015b). Note that these physiological and behavioural measurements are collected from a different group of participants than those whose induced emotions were annotated as the gold-standard we are predicting in the unimodal and multimodal experiments.

7.2.1.3 Recognition Models

Unlike Experiments 6, 7 and 8, in Experiment 9 we build LSTM models for recognizing movie-induced emotions using the Keras library (Chollet, 2015). Compared to PyBrain, Keras has more advanced functions for training a LSTM regression model, such as drop-out. In Experiment 9, we train the LSTM models using RMSprop with a learning rate of 0.0001 and the MSE evaluation metric.[6] All LSTM models have three hidden layers (number of neurons in each hidden layer: $h_1 = 64$, $h_2 = 32$, $h_3 = 16$). To prevent over-fitting, we use 0.5 drop-out rate in h_1 and set the maximum training iteration to 50 epochs with an early stopping tolerance of 10 epochs.

For multimodal experiments, we again compared our HL fusion with FL and DL fusion. In our multimodal models, for FL fusion, all features are used at the input layer of the LSTM model. For DL fusion, predictions of unimodal LSTM models are input to another LSTM model. For HL fusion, input neurons of low-level features are connected to h_1, while input neurons of high-level features are connected to h_2 directly. We build multimodal models combining all features, as well as multimodal models using only movie based features. For the HL model using all features, the physiological and behavioural features are used at the higher layer because they are measurements of audience's reactions (see Figure 7.2). For the HL model using movie based features, the audiovisual features are used at the higher layer because we include in-domain knowledge during feature selection (see Figure 7.3).

We perform leave-one-movie-out cross-validation for the unimodal and multimodal recognition experiments and report unweighted average of MSE, absolute CC, and absolute CCC. We evaluate the significance of performance differences using two-sample Wilcoxon tests with $p < 0.05$ as being significant. We compare Arousal and Valence predictions for pairs of models that have the closest performance in each experiment (e.g., Visual vs. DIS-NV on Arousal in Table 7.5), and find that all of them are significantly different with $p << 0.0001$, except for Lexical vs. Highlight on

[6]The initializers in the recurrent layer are the same as the LSTM model settings in Section 3.5.2

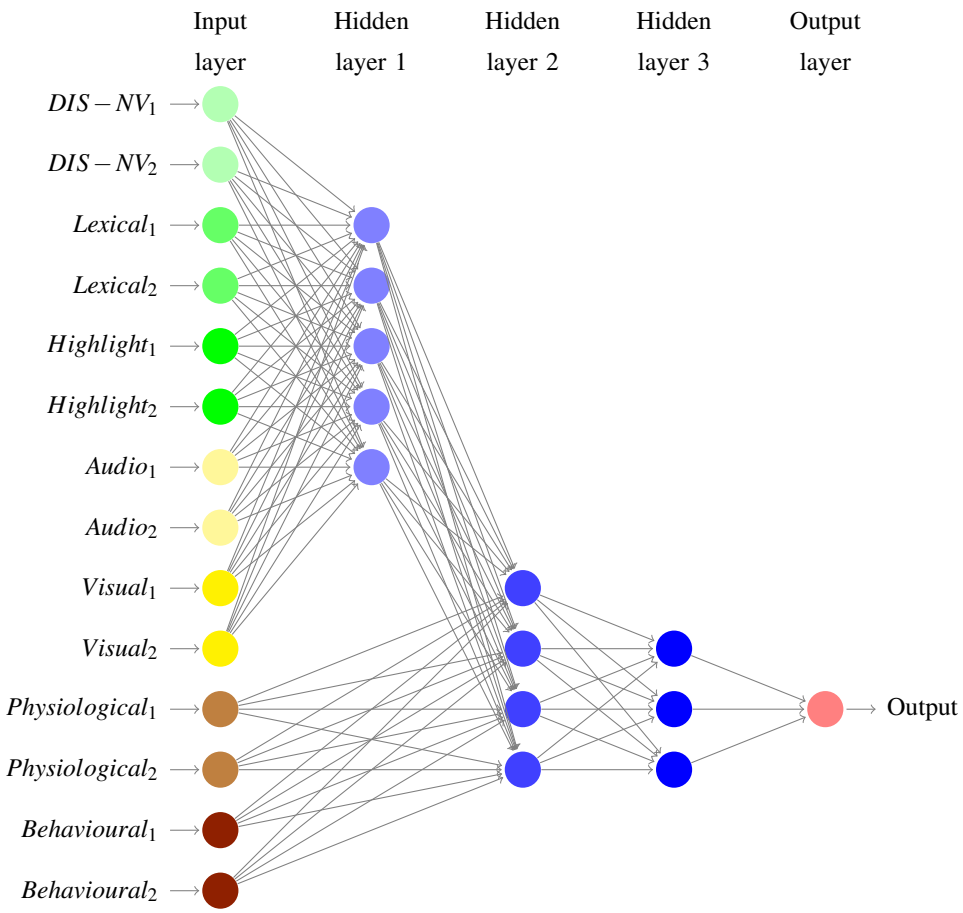


Figure 7.2: Structure of HL Model using All Features

Valence in Table 7.5 which has $p = 0.424$.[7] Note that because of different experiment protocols, such as the use of the overlapping window, our results are not directly comparable with previous work reported in Table 7.1.

[7] Lexical vs. Highlight on Arousal has $p << 0.0001$. The second closest pair on Valence (Audio vs. DIS-NV) has $p << 0.0001$

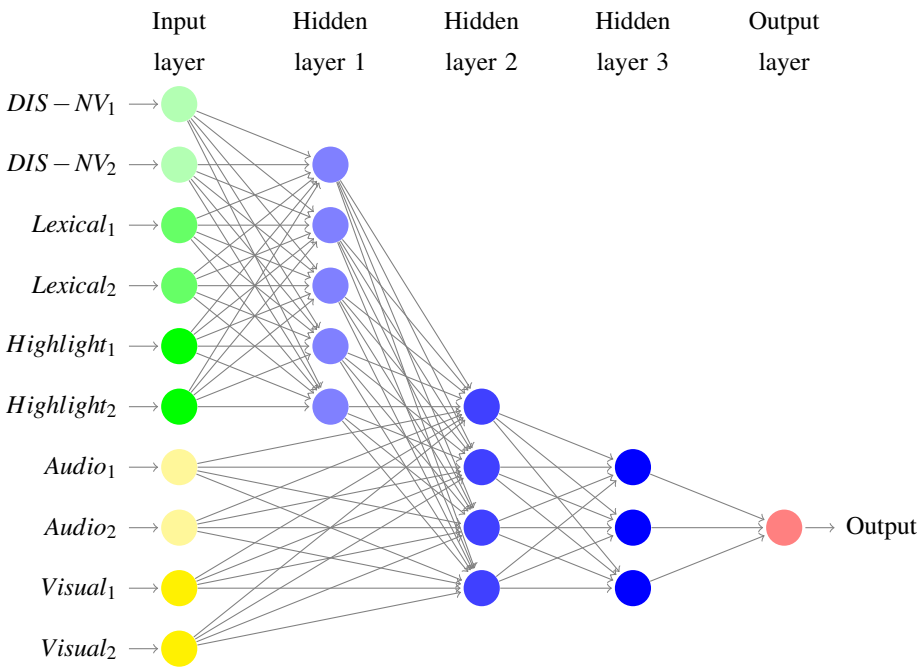


Figure 7.3: Structure of HL Model using Movie Based Features

7.2.2 Influence of Temporal Context on Induced Emotion

The original induced emotion annotation provided by the continuous LIRIS-ACCEDE database is at every single second, where the average absolute difference between adjacent emotion annotations of Arousal is 0.006 and of Valence is 0.005. This is extremely small considering the annotation value range is [−1,1]. Previous work has shown that human emotions are context dependent and typically do not change rapidly over a small time interval (Poria et al., 2017). However, the amount of history needed varies for different tasks. The suitable amount of temporal context for predicting induced movie emotions remains unknown. Similarly, in Experiment 6 of Chapter 5, we find that compared to the AVEC2012 database which annotated emotions at word level, the IEMOCAP database which annotated emotions at utterance level benefits less from using the LSTM model to incorporate temporal context.

We attempt to identify a suitable amount of history for recognizing induced emotions by testing the LSTM model using physiological features with different time steps. We use physiological features because they are direct representatives of the audience's induced responses (Kostoulas et al., 2015b). Results of our experiment on influence of context length is shown in Table 7.4 and Figure 7.4. Note that for MSE a smaller number means better performance while for CC and CCC a bigger number means better performance. Numbers in bold are the best performance. As we can see,

Table 7.4: Influence of Context on Movie-Induced Emotions

Time Step	A-mse	A-cc	A-ccc	V-mse	V-cc	V-ccc
1	0.053	**0.225**	**0.058**	0.066	0.401	0.075
2	0.055	0.201	0.050	0.070	0.411	**0.076**
3	**0.047**	0.190	0.044	**0.066**	**0.432**	0.072
4	0.056	0.205	0.051	0.070	0.389	0.065
5	0.051	0.190	0.048	0.074	0.379	0.061

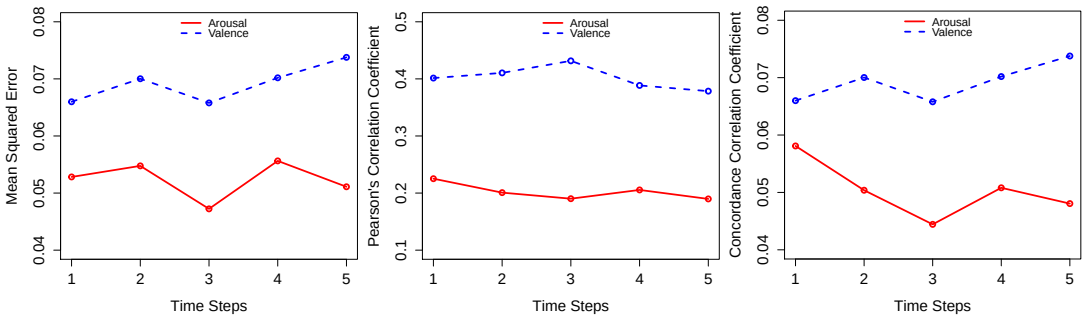


Figure 7.4: LSTM Model with Different Time Steps

overall a time step of 3 gives better recognition performance than shorter or longer time steps.[8] Recall that our feature vectors are extracted over a 5 second sliding window with 4 seconds overlap. With 3 history feature vectors the model will have 9 seconds of context (including the current window). A duration of 9 seconds is typically longer than a word, while this may be shorter than some long utterances. This explains why the word-level AVEC2012 database benefited more from using the LSTM model than the utterance-level IEMOCAP database in Experiment 6 of Chapter 5. In later induced emotion recognition experiments in this Chapter, all our LSTM models will use a time step of 3. Note that in this experiment we have a limited collection of different movie genres. Thus, our finding may not generalize to other movie genres, such as horror.

7.2.3 Recognizing Movie-Induced Emotions with Unimodal Models

In Table 7.5, we report results of our unimodal experiments on recognizing movie-induced emotions. Numbers in bold indicate the best performance in the experiment. As we can see, the physiological features achieved the best performance

[8] In Table 7.4, time step of 3 has three numbers in bold, time step of 1 has two numbers in bold, time step of 2 has one number in bold.

Table 7.5: Unimodal Movie-Induced Emotion Recognition

Model	A-mse	A-cc	A-ccc	V-mse	V-cc	V-ccc
LSTM						
Audience Reaction Features						
Physiological	**0.047**	**0.190**	0.044	0.066	**0.432**	**0.072**
Behavioural	0.049	0.183	**0.082**	**0.064**	0.129	0.054
Movie Based Features						
Audio	0.054	**0.218**	**0.055**	0.069	0.134	**0.033**
Visual	0.060	0.126	0.018	0.090	**0.152**	0.025
Lexical	0.050	0.085	0.029	0.071	0.060	0.014
DIS-NV	0.049	0.124	0.010	**0.069**	0.115	0.011
Highlight	**0.049**	0.153	0.042	0.070	0.056	0.006

for predicting induced Valence. Note that the physiological features are based on measurements of 13 participants that are different from the 10 participants whose induced emotions are being predicted. This indicates that people share similarities in how and what emotions are induced by the same movie. The behavioural features are less predictive than the physiological features for Valence. This indicates that hand movements of the audience may be caused by various factors besides induced emotions, and contain more noise compared to the electrodermal measure.

The audio features achieved the best CC for predicting induced Arousal. This suggests that including in-domain knowledge can benefit induced emotion recognition. The DIS-NV features achieved better MSE than other movie-based features. This indicates the effectiveness of our DIS-NV features for recognizing movie-induced emotions when there is presence of dialogue. Knowledge-inspired features based on affective cues (lexical, DIS-NV, and highlight features) achieved better MSE predicting induced Arousal and Valence than data-driven features based on audiovisual movie content (audio and visual features). This different behaviour of CC and MSE shows that an evaluation metric combining correlation and error may be better for evaluating induced emotion recognition performance. Thus, we also report the Concordance Correlation Coefficient (CCC, see Section 7.1.2) (Lawrence and Lin, 1989).

To study the difference of feature predictiveness in more detail, we plot unimodal predictions on *Superhero* which has the most dialogue (Figures 7.5 and 7.6) and on *First Bite* which has the least dialogue (Figures 7.7 and 7.8) as shown in Table 7.2. As

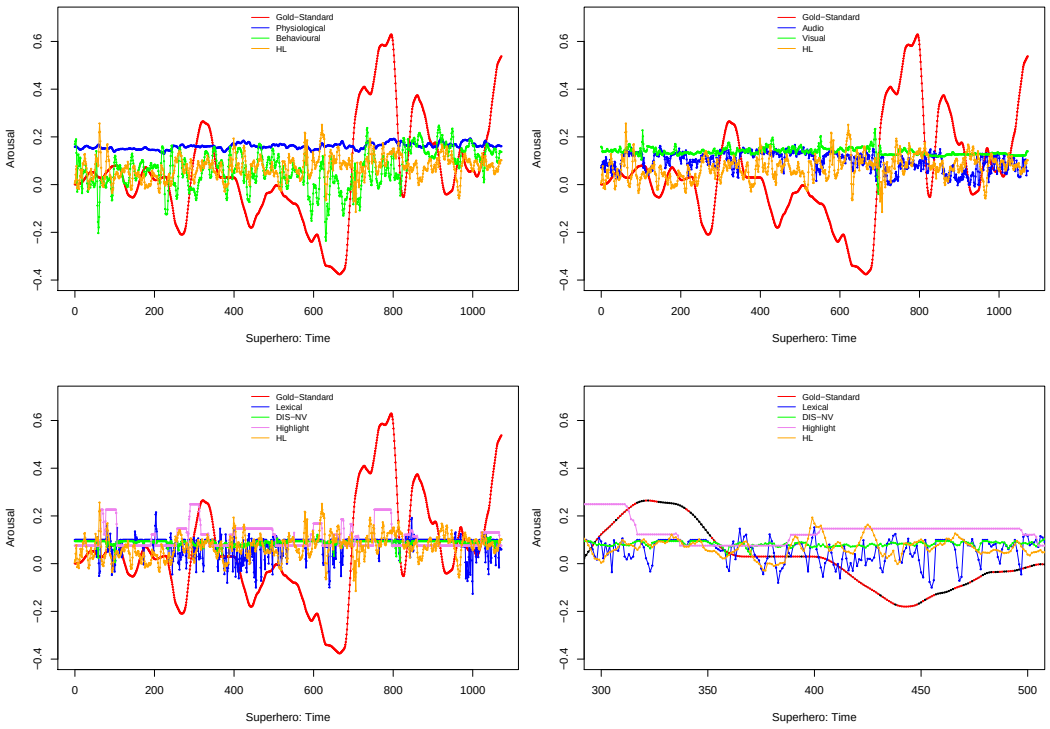


Figure 7.5: Arousal Predictions using Different Features on “Superhero”

we can see, when there is no dialogue (red dots of the gold-standard Arousal or Valence lines in the figures), the lexical and DIS-NV features predict mean values. Thus they work better for *Superhero* which has the most dialogue. The audiovisual predictions are flatter than the physiological, behavioural, or knowledge-inspired features, which may be due to noise in the audiovisual movie content. These findings are consistent with our observations in Experiment 1 of Chapter 4.

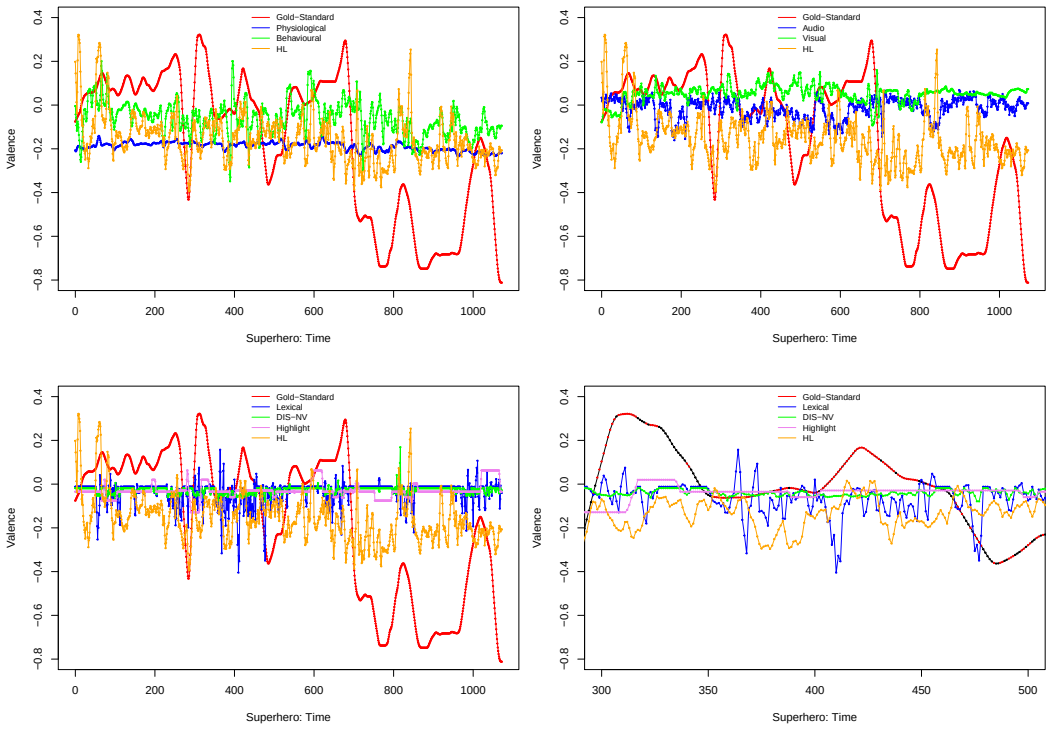


Figure 7.6: Valence Predictions using Different Features on “Superhero”

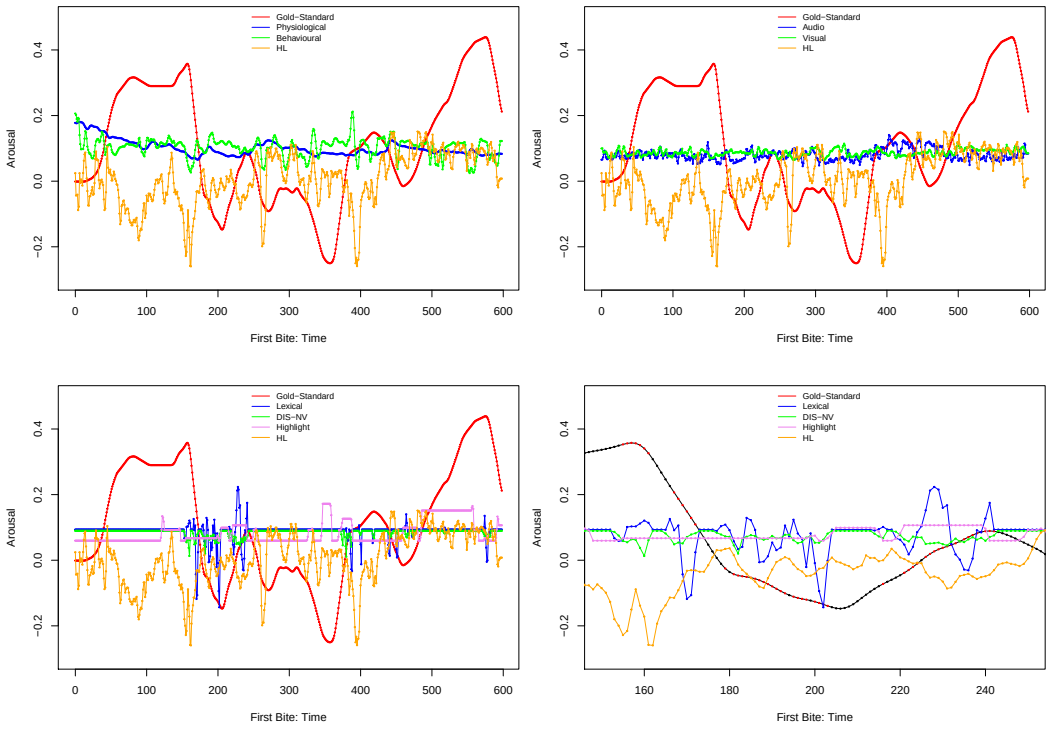


Figure 7.7: Arousal Predictions using Different Features on “First Bite”

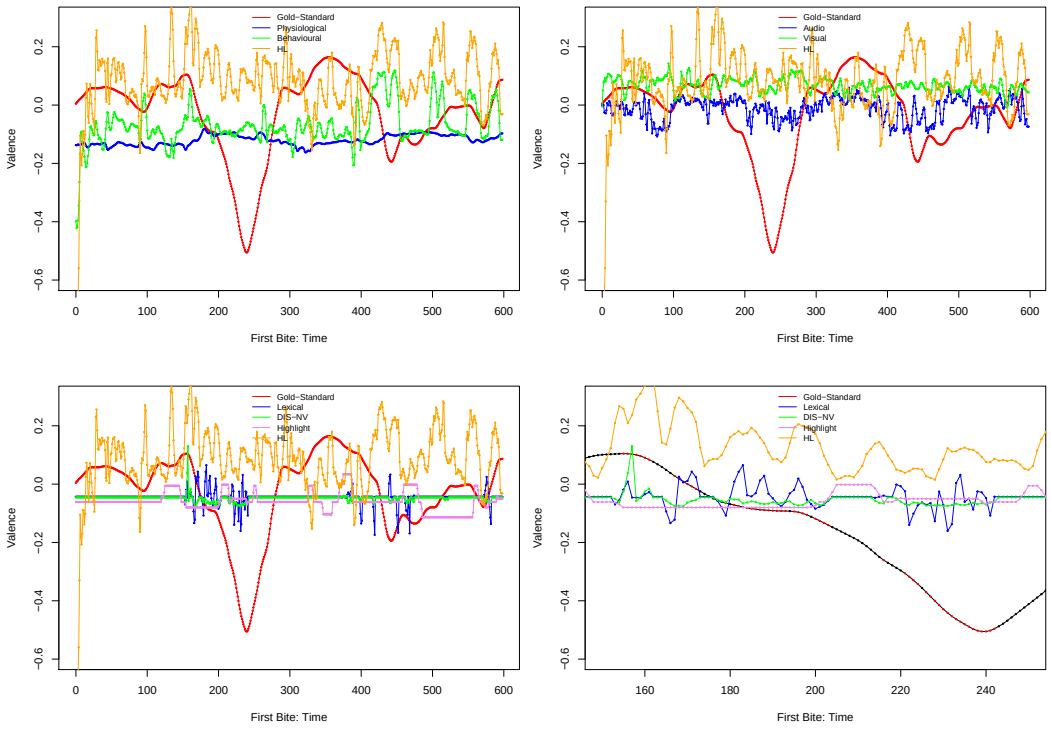


Figure 7.8: Valence Predictions using Different Features on “First Bite”

7.2.4 Recognizing Movie-Induced Emotions with Multimodal Models

In Table 7.6, we report results of our multimodal experiment on recognizing movie-induced emotions. We build multimodal models using all features as well as multimodal models using only movie based features.

For multimodal models using all features, the FL model has the best CC for predicting Arousal, while the HL model has the best CC for predicting Valence. Recall that the physiological and behavioural features are used at a higher layer than all other features in HL fusion. In the unimodal experiment (Table 7.5), the audio features have the best CC for predicting Arousal, while the physiological features have the best CC for predicting Valence. The audio features have larger influence in FL fusion than in HL or DL fusion, which results in better CC for predicting Arousal using FL fusion. The DL model has the best MSE for predicting induced Arousal and Valence. This may be related to the fact that DL fusion is not influenced by feature dimension, thus the DL model can benefit more from the smaller DIS-NV and Highlight feature sets which have strong MSE performance (Table 7.5). Compared to unimodal models, the multimodal models achieved better performance on predicting Arousal, but not

Valence. This may be because of the issue of insufficient training data.

Multimodal models using only movie based features have significantly worse performance than multimodal models using all features. This indicates that emotion recognition models can benefit from combining information from all possible modalities. For multimodal models using movie based features, the DL model has the best MSE performance and best CC of Valence, while the FL model has the best CC of Arousal. This is because the audio features, which achieved best unimodal CC for predicting induced Arousal (Table 7.5), have larger influence in FL fusion than in DL fusion. HL fusion achieved the best CCC scores for both Arousal and Valence. However, HL fusion does not give better CC or MSE scores than FL or DL fusion in this particular case. The reason may be that we extract all features using overlapping windows and we reduce noise in the data-driven audiovisual features by performing feature selection. Thus, the difference between movie based feature sets in terms of the time scale at which the features are extracted or the level of abstraction of the features is not as large as in our experiments in Chapter 6. This limits the gain of combining features at different levels using HL fusion. Similar to multimodal models using all features, multimodal models using movie based features did not outperform unimodal models in all cases, which may be because of insufficient training data again.

In Figures 7.5, 7.6, 7.7, and 7.8 we plot the predictions given by the HL model using all features in addition to the unimodal predictions. As we can see, compared to the models using knowledge-inspired features, the HL models avoided straight-line predictions when there is no presence of movie dialogue. Compared to the models using audiovisual features, the overall shape of the HL model predictions is less flat and better captures the raises and falls of the gold-standard predictions. This illustrates how the HL multimodal model typically yields better performance than the unimodal models.

It is difficult to directly compare our unimodal and multimodal models with state-of-the-art performance shown in Table 7.1 of Section 7.1.1 because of different experimental settings, such as the time-interval that emotions are predicted on. However, our experiments indicate that performance improvements can be achieved by including history information, and by incorporating knowledge-inspired affective cues in addition to the audiovisual movie content for recognizing movie-induced emotions.

Table 7.6: Multimodal Movie-Induced Emotion Recognition

Model	A-mse	A-cc	A-ccc	V-mse	V-cc	V-ccc
LSTM						
With All Features						
FL	0.057	**0.271**	0.080	0.071	0.107	0.039
DL	**0.044**	0.189	0.021	**0.070**	0.163	0.038
HL	0.074	0.159	**0.093**	0.095	**0.227**	**0.117**
With Movie Features						
FL	0.054	**0.216**	0.038	0.069	0.118	0.039
DL	**0.044**	0.182	0.012	**0.057**	**0.178**	0.032
HL	0.073	0.157	**0.111**	0.075	0.083	**0.050**

7.2.5 Summary

In Experiment 9, we studied the effectiveness of our DIS-NV features and HL fusion for recognizing emotions induced by movies. Our experiments show that when there is dialogue in the movie, DIS-NV features are predictive of the movie-induced emotions. Our HL fusion gave better performance than the FL and DL fusion when evaluated with CCC. However, because the features in this experiment are extracted at similar time scales and have similar levels of abstraction, the gain of using HL fusion is not as large as in Chapter 6. In the future, it may be beneficial to use the movie highlight features as a prior to learn more flexible time windows for building the multimodal model.

7.3 Perceived and Induced Emotions

To improve on state-of-the-art performance of emotion recognition, beyond identifying effective features and modelling approaches, another important aspect is to collect more natural emotion databases. However, as discussed in Section 5.2, the natural and spontaneous emotion databases are often biased with majority of the data being neutral or having mild emotions. To address this issue, current studies have investigated using affective stimuli to induce certain emotions. For example, in the AVEC2012 database, virtual agents with different personality types are used to induce target emotions in the participants. When selecting stimuli to induce emotions, it is often assumed that emotions perceived from the affective content (**perceived emotions**) are consistent

with emotions induced by the stimuli in the spectators (**induced emotions**). However, we found that this is not always the case in the AVEC2012 database. For example, the virtual agent Poppy is designed to be cheerful and to induce positive emotions with high Valence in the participants. However, some of the participants were actually annoyed by Poppy and the conversation induced negative emotions with low Valence instead. Tan (2013) also suggested that in movie making, the emotions induced in the audiences are not always consistent with the emotions intended by the director.

Perceived and induced emotions are often not distinguished in affective research. In fact, only a handful of previous studies have addressed the differences between the perceived emotions of affective content and the induced emotions of the spectator, mainly in music emotion research (Kallinen and Ravaja, 2006). However, the empirical study of Gabrielsson (2001) has shown that emotions perceived from music are not always consistent with the emotional responses induced by the music in the audience. This suggests the necessity of distinguishing perceived and induced emotions. In this chapter, we discuss the relationship between perceived and induced emotions under the context of movie audience.

In comparison to music, movies convey complex information through multiple modalities. Tan (1995) argued that emotions perceived from a movie can influence the emotions induced in the audience by evoking empathy, which suggests a positive correlation between perceived and induced emotions. However, Baveye et al. (2017) argued that emotions intended by the directors may not always be consistent with emotions induced in the audience, although they did not discuss perceived emotions. To the best of our knowledge, there has been no previous work formally addressing the relationship between perceived and induced emotions of movie audiences. Therefore, we are motivated to bridge this gap by performing statistical analyses of the emotions perceived from movie content and emotions induced in movie audiences. This will provide a foundation for understanding how affective content induces emotions in audiences, and how to use movie content information to predict movie-induced emotions.

7.3.1 Perceived vs. Induced Emotions

When experiencing affective content, such as listening to music or watching a movie, we perceive emotions conveyed by the affective content from characteristics of the stimuli, such as tempi and pitch of music (Gabrielsson, 2001). On the contrary, induced

emotions of a spectator evoked by the stimuli are related to personal experience and individual preferences (Plantinga, 2012). For example, Gabrielsson (2001) reported that a song perceived as happy induced stronger depression in a subject who is already in a depressed mood. Moreover, the work of Matthews et al. (1990) indicates that perceived emotions are more objective than induced emotions, and annotators typically have stronger agreement over perceived emotions than induced emotions (Song et al., 2016b). Previous work on affective content analysis does not always distinguish between perceived and induced emotions. Although Knautz and Stock (2011) have found consistencies between perceived and induced emotions, music emotion research has identified fundamental differences between perceived and induced emotions (e.g., Gabrielsson (2001); Tarvainen et al. (2014)). Kallinen and Ravaja (2006) have also suggested that induced emotions can have more intensive Arousal and less intensive Valence ratings compared to perceived emotions of the same stimuli.

Compared to music emotions, there has only been limited work studying the relationship between perceived and induced emotions of movie audiences. Hanjalic and Xu (2005) hypothesized positive correlations between perceived and induced emotions of movie audiences because perceived emotions can be used to estimate a spectator's affective reactions. However, current research on the cinematic art form suggests that emotion induction is a complex and delicate mechanism which is not always consistent with the expectations of the film maker and the perception of the objective characteristics of the movie content (Tan, 2013). To the best of our knowledge, there has been no previous work formally studying how perceived and induced emotions of movie audiences are related. To bridge the gap between perceived and induced emotions of movie, we extend the LIRIS-ACCEDE database by annotating perceived emotions of the movie content in a crowd-sourced manner, and analyse the relationship between perceived and induced movie emotions in Section 7.3.

7.3.2 Annotating Perceived Movie Emotions

Emotion annotation is more subjective compared to movie transcription in Section 7.1.2. Previous work has suggested that to achieve reliable emotion annotations, it is desirable to have more than 6 annotators (Busso et al., 2008). To collect a large amount of annotations in a time and cost efficient manner, we annotate perceived emotions of movie audiences using Amazon Mechanical Turk,[9]

[9] https://requester.mturk.com/

a crowd-sourced annotation platform. We segment movies into utterance clips using manual transcription of utterance timings (see Section 7.1.2) and collect at least 10 annotations from different annotators for each clip. The annotators were instructed to rate the emotions expressed by movie characters on the Arousal (A), Power (P), and Valence (V) dimensions with 1 to 9 integer scores. Note that unlike the AVEC2012 database, the Expectancy dimension is not annotated here. We also provide explanations of each emotion dimension and meaning of different scores. Each Human Intelligence Task (HIT) contains clips of 5 continuous utterances from the same movie in their original order to provide movie context to the annotators. Each utterance appears at each of the five video windows in different HITs to reduce bias. The HITs are in random order and we kept track of previous annotators of each movie to prevent an utterance being annotated by the same annotator more than once. Annotators were only allowed to annotate a clip after it finished playing, and could only submit after annotating all clips. We published 1809 HITs and 129 annotators with various cultural and educational backgrounds participated. An example of the annotation interface is shown in Figure 7.9.

The 1 to 9 scores collected from the crowd-sourced annotation are normalized to $[-1,1]$[10] to be consistent with the induced emotion annotation. We compute means of the annotations from multiple annotators collected on each utterance of the movie dialogue as the perceived emotion annotation, resulting in utterance-level Arousal, Power, and Valence annotations of perceived emotion of movie audiences.

7.3.3 Experiment 10: Perceived and Induced Emotions

In Experiment 10, we study the relationship between emotions perceived from and induced by movies. Note that the induced emotions are annotated at each second, while the perceived emotions are annotated at utterance-level which is generally longer than one second. Thus, we align the annotations by computing mean values of induced Arousal and Valence scores over each movie utterance as the utterance-level induced emotion annotation. We then calculate the CC between perceived and induced emotions for each movie independently. We computed the weighted average of CC over all 8 movies and report the results in Table 7.7. In the first row, "Per" represents perceived emotions, "Ind" represents induced emotions. To evaluate the practical significance of CC, following the model of Cohen (1988), we interpret absolute CC

[10] "1" = −1, "2" = −0.75, "3" = −0.5, "4" = −0.25, "5" = 0, "6" = 0.25, "7" = 0.5, "8" = 0.75, "9" = 1

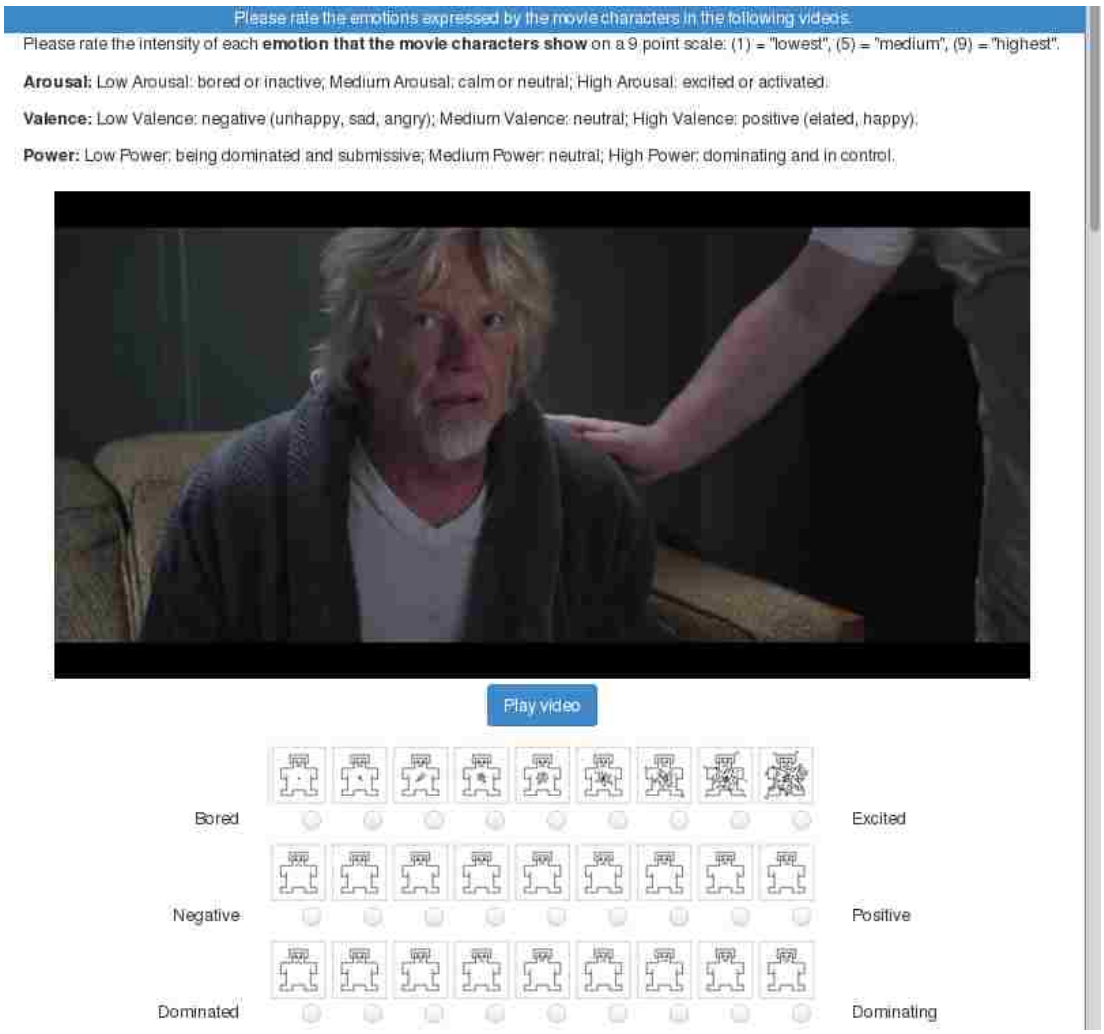


Figure 7.9: Example of Amazon Mechanical Turk Annotation Interface

values at around 0.1, 0.3, and 0.5 as reflecting the effect size of small, medium, and large magnitude respectively (coloured as yellow, blue, and red in Table 7.7).

As we can see, perceived Arousal, Power, and Valence are highly positively correlated with each other, while induced Arousal and Valence are moderately negatively correlated with each other. This may be related to perceived emotion annotation being a more objective task. The negative correlation between induced Arousal and Valence is consistent with the work of Warriner et al. (2013) which found a CC of −0.185 between crowd-sourced annotations of induced Arousal and Valence collected for nearly 14,000 English lemmas. This suggests that induced negative emotions may have stronger Arousal than induced positive emotions. However, no definitive conclusions can be made because of the small absolute CC value. Induced Valence and perceived emotions have moderately positive correlations, while induced

Table 7.7: CC Between Perceived and Induced Emotions

Emotion	Per-A	Per-V	Per-P	Ind-A
Per-V	0.538	#	#	#
Per-P	0.652	0.471	#	#
Ind-A	-0.095	-0.366	-0.170	#
Ind-V	0.243	0.345	0.307	-0.388

Arousal and perceived emotions are weakly or moderately negatively correlated. In particular, perceived Arousal and induced Arousal are only weakly negatively correlated. This inconsistency between perceived and induced emotions indicates fundamental differences between emotions perceived from and induced by movies. Emotion induction is a complex process. Various factors other than the emotions the movie content conveys can influence what emotional response is induced in the audiences. The assumption that perceived and induced emotions are consistent is not accurate and researchers need to take extra caution when designing experiments for affective content analysis research and when collecting emotion databases by inducing. For example, pilot studies should be conducted to test whether or not a chosen stimuli can induce the target emotion in a specific person.

In Table 7.8, we study the individual differences on induced and perceived emotion annotations by calculating the average standard deviations of the annotations on each movie. In the first row, "Per" represents perceived emotions, "Ind" represents induced emotions. For induced emotions, we use the original annotations per second. For perceived emotions, we use the annotations per utterance. At an emotion annotation step t (a second or an utterance of a movie), we compute the standard deviation over all N annotators as below:

$$Std_t = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Annotation_{t_{rater_i}} - \overline{Annotation_t})^2} \tag{7.2}$$

We report the average of the standard deviations $\overline{Std_t}$ for all emotion annotation steps of a movie. As we can see, the average standard deviation for perceived emotions is larger than that for induced emotions for all movies. This may be due to the use of crowd-sourced annotation for perceived emotion annotation. The perceived emotion annotation was given by 129 untrained annotators from various cultural and educational backgrounds, while the induced emotion annotation was given by 10 trained annotators who are undergraduate to recently graduated master students from

Table 7.8: Standard Deviations of Induced and Perceived Emotion Annotations of Multiple Annotators

Movie	Per-A	Per-V	Per-P	Ind-A	Ind-V
After the Rain	0.433	0.389	0.374	0.340	0.230
First Bite	0.404	0.328	0.353	0.239	0.196
Nuclear Family	0.432	0.425	0.462	0.307	0.319
Payload	0.445	0.425	0.421	0.306	0.208
Spaceman	0.390	0.364	0.365	0.294	0.222
Superhero	0.387	0.456	0.444	0.307	0.253
Tears of Steel	0.462	0.398	0.430	0.302	0.278
The Secret Number	0.439	0.365	0.390	0.264	0.247

France. Therefore, these 10 trained annotators share more similarities in emotion induction and agree more in their annotations.

7.3.4 Summary

In Experiment 10, we collected crowd-sourced annotations of emotions perceived from movies and studied the relationship between emotions perceived from and induced by movies. Our statistical analysis showed that perceived and induced emotions are not always positively correlated. Thus, when collecting emotion databases by inducing, instead of assuming that the emotions perceived from the stimuli will be consistent with the emotions induced in the participants, pilot studies should be done first to verify whether or not the stimuli can induce the target emotions.

7.4 Discussion

In this chapter, we applied our DIS-NV features and HL fusion to recognizing movie-induced emotions. Our unimodal experiment indicates that the DIS-NV features are effective for recognizing movie-induced emotions when there is dialogue, and the amount of temporal context suitable for movie-induced emotion recognition is 9 seconds. Our multimodal experiment indicates that improved performance can be achieved by combining knowledge-inspired affective cues with audiovisual movie content. Our HL fusion is shown to be more predictive than FL and DL fusion for recognizing movie-induced emotions when using an evaluation metric combining

both correlation and value error. Our experiments indicate that although designed specifically for emotion recognition in spoken dialogue, other emotion-related tasks may benefit from our proposed DIS-NV features and HL fusion strategy as well. Note that our induced emotion recognition experiments here are preliminary. Improved performance may be achieved in the future by optimizing feature representations and model structures.

Similar to our experiments in Chapters 5 and 6, the small amount of labelled data available for affective content analysis limits performance of movie-induced emotion recognition significantly. Inspired by audiovisual features benefitting from including in-domain knowledge, it may be beneficial to apply transfer learning for predicting movie-induced emotions. In addition to using semi-supervised and unsupervised algorithms to address the issue of insufficient training data, it is also important to collect more natural emotion databases. To collect emotion databases of target emotions, affective contents, such as movie or music, are often used to induce emotions in the participants. To understand the relationship between the emotions perceived from affective content and the emotions induced in the spectator, we collect crowd-sourced annotations of perceived emotions on the continuous LIRIS-ACCEDE database. We find that emotions perceived from and induced by movies are not always positively correlated. When selecting stimuli for emotion induction, there is much more to be considered than simply assuming that the perceived emotions of the stimuli will be consistent with the emotions induced in spectators.

Chapter 8

Towards an Emotional Human-Computer Interaction System

The robopsychologist continued: "Here is what we're going to do."

— Isaac Asimov, *I, Robot (1950)*

In this chapter, we discuss the contributions and limitations of this thesis. We also discuss possible future directions for applying the proposed emotion recognition model to a Human-Computer Interaction (HCI) system to improve emotional interaction quality.

8.1 Conclusion

In this thesis, we explored approaches to improve state-of-the-art performance of emotion recognition in spoken dialogue using acoustic and lexical features. In this section, we first summarize the contributions of our work. We then discuss the limitations and possible future directions to address them.

8.1.1 Major Findings and Contributions

Emotion recognition has been a focus in Affective Computing since the establishment of the field. We are especially interested in recognizing emotions in spoken dialogue because of its potential to be applied to Human-Computer Interaction systems and improve the interaction quality. Emotion recognition in spoken dialogue is a challenging task and state-of-the-art performance leaves considerable room for improvement. In this thesis, we proposed two approaches for improving

the state-of-the-art of emotion recognition in spoken dialogue, namely the use of DISfluency and Non-verbal Vocalisation (DIS-NV) features, and the HierarchicaL (HL) fusion strategy. The knowledge-inspired DIS-NV features describe occurrences of filled pause, filler, stutter, laughter, and audible breath in utterances. The HL fusion strategy combines information from different modalities in a hierarchical structure which incorporates features that are more abstract or are extracted from longer time intervals at higher layers of the model. We conducted experiments on the AVEC2012 database of spontaneous dialogue and the IEMOCAP database of acted dialogue to study the performance of the proposed approaches. Emotions are defined as vectors in the Arousal-Expectancy-Power-Valence multi-dimensional space in this work.

Our results show that the DIS-NV features are predictive of emotions in spontaneous dialogue when used alone. The DIS-NV features achieved the best-reported result on the Expectancy emotion dimension, which describes levels of uncertainty. This is consistent with Psycholinguistic studies that suggest that disfluencies are indicators of speaker uncertainty. Among different types of DIS-NV features, filled pauses and laughter are the most predictive of emotions in terms of individual effectiveness. Incorporating the DIS-NV features with other benchmark acoustic and lexical features yields improved performance for emotion recognition on both spontaneous and acted dialogue. This verifies that DIS-NVs contain additional information that is predictive of emotions beyond the acoustic characteristics or the lexical content of speech. Conventional acoustic and lexical features used in state-of-the-art research on recognizing emotions in spoken dialogue have focused on speech in isolation, while specific characteristics of spoken dialogue compared to other forms of speech (e.g., monologue) are often overlooked. The effectiveness of the proposed DIS-NV features suggests that awareness of dialogue-specific cues in speech can benefit emotion recognition in spoken dialogue.

Our cross-corpora experiments illustrate the fundamental differences between spontaneous and acted dialogue, and how these differences influence the effectiveness of features and models for emotion recognition. In general, there are fewer DIS-NVs in acted dialogue, which limits the effectiveness of DIS-NV features for emotion recognition in acted dialogue. Acted dialogue is also acoustically exaggerated compared to the spontaneous dialogue, which results in acoustic features being more effective than lexical features for emotion recognition in acted dialogue. Our results also verified that models that are able to include contextual information and automatically learn more abstract feature representations typically yield better

performance than non-contextual models which use the extracted features directly. However, the lack of annotated training data in emotion databases may limit the performance of emotion recognition models.

To increase the gain from combining different modalities, we proposed a HierarchicaL (HL) fusion strategy. Compared to the state-of-the-art Feature-Level (FL) and Decision-Level (DL) fusion strategies which combine modalities at the same level, our HL fusion is able to model both inter and intra modality differences. The knowledge-inspired structure of HL fusion captures the differences between various feature representations on two aspects: the level of abstraction over the data, and the time scale at which the features are extracted. Our experiments on both spontaneous and acted dialogue showed improved results of multimodal emotion recognition using HL fusion compared to using FL or DL fusion.

To study how the efficacy of the proposed approaches can generalize to other emotion related tasks, we applied our DIS-NV features and the HL fusion strategy to the problem of recognizing movie-induced emotions of audiences. The experiments indicated that our emotion recognition model is predictive of movie-induced emotions. The HL fusion strategy outperformed FL and DL fusion, which suggests that other multimodal recognition tasks can benefit from HL fusion as well.

Our work contributes to the Affective Computing community by improving the state-of-the-art of emotion recognition in spoken dialogue. We identified effective DIS-NV features, and studied the relationship between DIS-NVs and emotions in dialogue. This contributes to the Psycholinguistic understanding of emotions in dialogue as well. We also studied how different aspects of the data can influence performance of emotion recognition approaches. Our experiments suggest that features and models should be chosen carefully to fit the context of a specific emotion recognition task. The HL fusion strategy we proposed was shown to be an effective modality fusion strategy for multimodal recognition tasks beyond emotion recognition in spoken dialogue.

8.1.2 Limitations

In the proposed emotion recognition model, for the lexical modality, we use basic features derived from affective dictionaries. However, sentiment analysis from text itself is a growing field. Thus, in the future we are interested in experimenting with an advanced text sentiment analysis classifier for predicting emotions in spoken dialogue,

e.g., the Stanford CoreNLP tool (Manning et al., 2014). We expect that our emotion recognition model will benefit from using a more sophisticated sentiment analysis method in the lexical modality, especially on the Valence dimension.

When building the emotion recognition model, we down-sampled frame-level features to the longer time interval (e.g., taking the mean of frame-level LLD features over a word as the LLD features for this word). This results in detailed information from shorter time intervals being lost in the recognition model, which may reduce the performance of emotion recognition. Thus, we plan on studying whether or not performance improvement can be achieved by building a recognition model capable of preserving frame-level information with asynchronous feature processing. For example, using the Phased LSTM model (Neil et al., 2016) instead of the regular LSTM model. We are also interested in identifying the optimum time interval for annotating and recognizing each emotion dimension.

Our experiments in Chapter 7 indicate that different emotion dimensions may be correlated. Thus, the emotion recognition model may benefit from using a multi-task learning algorithm which predicts all emotion dimensions simultaneously. The issue of insufficient training data encountered in this work and current emotion recognition research highlights the need to collect more natural emotion databases and to develop semi-supervised and unsupervised emotion recognition models. Current emotion recognition studies have focused on building a robust model. However, for HCI systems, an emotion recognition model being able to adapt to an individual user is preferred. This stresses the need for semi-supervised and unsupervised emotion recognition models that learn from and adapt to incremental data collected from an individual user over the interaction.

Our work and most state-of-the-art emotion recognition studies rely on intrinsic metrics. Although most current studies report results that are statistically significantly different, the improvement is often limited, with small absolute values. This indicates that emotion recognition is a challenging task. Because our long-term goal is to improve the interaction quality of HCI systems, we are curious to learn if the performance improvements shown in the intrinsic evaluation of different emotion recognition approaches are noticeable to human users when applied to a HCI system. Thus, in the future we would like to combine our emotion recognition models with emotion interaction models and emotion synthesis models, and perform extrinsic evaluation methods to study whether or not the efficacy of the proposed approaches can improve the emotional interaction quality and benefit fully automatic HCI systems.

Note that in a fully automatic HCI system it is important to provide real-time responses to the user. However, we did not test the real-time factor of our emotion recognition model discussed in this thesis. Thus, in the future we also plan to study the trade-off between better emotion recognition performance and shorter response delays when applying our emotion recognition model to a fully automatic HCI system. Under the HCI setting, it is also possible to view emotions as a hidden variable of the system response. Thus, we are interested in exploring learning an end-to-end interaction strategy in the future instead of requiring emotion labels to train the emotion recognition model.

8.2 Emotional Interaction in Human-Computer Interaction Systems

Similar to most studies on automatic emotion recognition, our work relies on intrinsic measures to evaluate different approaches (e.g., correlation coefficients or classification accuracies). This leads to an open question in emotion recognition: when the emotion recognition model is applied to a HCI system, will the performance improvements shown in intrinsic tests translate to improvements in emotional interaction quality (e.g., higher engagement and satisfaction of the user)? We plan to work on this question in the future by integrating our emotion recognition model with a working HCI system and performing extrinsic evaluation measures.

To apply our emotion recognition model to a HCI system and realize emotional interaction, we will need an emotion interaction model, which can decide suitable responses to the recognized emotions of the user, as well as an emotion synthesis model, which can generate expressive feedback. In this section, we review state-of-the-art approaches on emotion modelling and emotion synthesis.

8.2.1 Emotion Modelling

Emotional interaction is a complex process in human communication (Zhao et al., 2014). Emotion modelling remains an open question in Affective Computing. Most earlier emotion interaction models in HCI systems were rule-based. For example, rules written by pedagogy experts were used in the tutoring system of Wiemer-Hastings et al. (1998). Such rule-based models can generate suitable responses and work fairly well in a more restricted scenario, but the lack of flexibility may bore the user

quickly. The rules are brittle and it is impossible to cover all possible occasions. This is in fact a common issue in rule-based Artificial Intelligence systems. In contrast, more recent studies build machine learning models that learn from multiple dialogue corpora to establish a pool of detailed and flexible response rules (e.g., Shawar and Atwell (2005)). However, the model requires large amounts of transcribed dialogue to learn sophisticated rules. Another common strategy is simple mimicking of the user's behaviour (e.g., McGettigan et al. (2015)). Psychological studies have shown that humans mimic their conversational partner's behaviour (e.g., facial expressions) during interaction to strengthen the social bond (Chartrand and Bargh, 1999). Besides lack of flexibility, rule-based emotion interaction models also have the issue of ignoring long-term context. In other words, a rule-based model generates feedback based on the current emotional states of the user, but does not consider long-term factors such as dialogue history or the personality of the user. However, such contextual information has been shown to be important for improving the interaction quality of HCI systems (Zhang et al., 2016a).

Beyond rule-based emotion interaction models, there has been increasing interest in developing more flexible and context-aware emotion interaction models in the current Affective Computing community. Marsella et al. (2010) provided an overview of current emotion modelling approaches. As shown in Figure 8.1, many of the state-of-the-art emotion models are based on Ortony's appraisal theory of emotion (Ortony et al., 1990). An example of the appraisal-based emotion models is the EMA (EMotion and Adaptation) model of Marsella and Gratch (2009). As shown in Figure 8.2, the EMA model studies the appraisal dynamics generated by a person's causal interpretation of the environment. This single-level appraisal model is able to describe naturalistic emotional situations at multiple temporal and cognitive levels. There are also hybrid emotion models, such as the ALMA (A Layered Model of Affect) emotion interaction framework by Gebhard (2005). In the ALMA model, emotional interaction is described with a three layer model and emotions are defined as vectors in the Arousal-Power-Valence emotional space. The ALMA model uses the current emotional states of the user as the short-term factors, moods of the user as the mid-term factors, and personality traits of the user as the long-term factors. The ALMA model is widely used for developing virtual agents (e.g., Schroder et al. (2012); Jia et al. (2014); Jain and Asawa (2016)), Non-Player Characters in games (e.g., Lim et al. (2012)), and social robots (e.g., Magnenat-Thalmann and Zhang (2014)) in current HCI research. Considering that our emotion recognition model is based on the appraisal emotion

theory, it is natural to combine our emotion recognition model with an appraisal-based emotion interaction model.

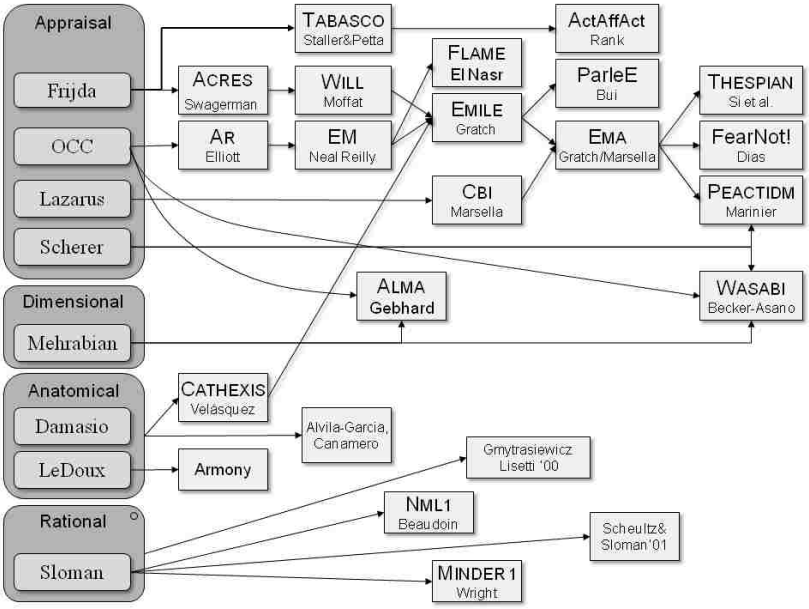


Figure 8.1: An Overview of Computational Models of Emotions (Marsella et al., 2010)

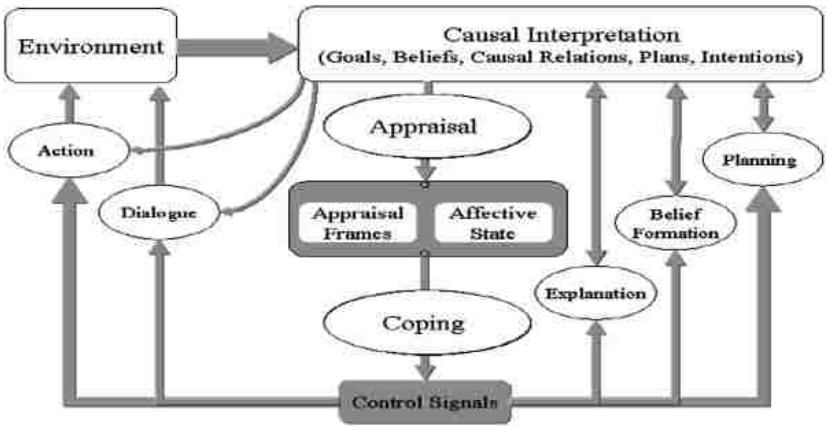


Figure 8.2: The EMA Emotion Model (Marsella and Gratch, 2009)

8.2.2 Emotion Synthesis

Humans are sensitive to expressions of emotions. Moreover, currently, how to generate natural emotion expressions that are believable to humans remains an open question. Current research on synthesizing emotional feedback is conducted mainly on the audio modality (expressive speech synthesis) and the visual modality (facial expression synthesis, gesture and movement synthesis). For the audio modality,

most state-of-the-art expressive speech synthesis approaches have focused on prosody modelling. Parameter mapping rules based on Paralinguistic studies were used in earlier systems (e.g., Schröder and Trouvain (2003)). In recent studies, machine learning algorithms that automatically learn the parameter mapping have gained increasing attention. For example, Kobayashi (2015) used a Hidden Markov Model to enhance prosodic variances and to transform the synthesized neutral speech to expressive speech. Besides parameter mapping, unit selection methods that generate speech by selecting and concatenating short segments of speech recordings are also applied to expressive speech synthesis (e.g., Zhang et al. (2015)).

For the visual modality, to synthesize facial expressions in virtual agents, machine learning algorithms are used either to learn facial landmark movement trajectories (e.g., Cao et al. (2013)) or to model muscle movements (e.g., Yu et al. (2014)) from human facial expression databases. To synthesize gestures and movements in virtual agents or robots, contextual models such as Hidden Markov Models used by Bozkurt et al. (2015) are applied to learn the relations between the intensity and energy of movements and perceived emotions. There are empirical studies identifying relations between specific movements and emotions as well (e.g., Dael et al. (2013); Novikova et al. (2015)).

State-of-the-art HCI systems typically perform emotion synthesis on multiple modalities to achieve more believable emotion expressions. In a dialogue system, emotions are mostly conveyed through the audio modality. However, emotion masking of the lexical contents of the synthesized speech is also important for a natural and believable emotional response. In addition, the balance between naturalness and expressiveness remains an open question in emotion synthesis.

8.3 Summary

The major finding of this thesis is that including prior knowledge on emotions, either in the feature representation or in the model structure, is beneficial for automatic emotion recognition. The context of the emotion recognition task is vital for building an effective recognition model. In particular, for emotion recognition in spoken dialogue, instead of analysing each utterance in isolation as a common speech emotion recognition task, it is important to take dialogue-related knowledge into account.

Our work contributes to the Affective Computing community by identifying effective approaches for emotion recognition in spoken dialogue. Moreover, our

work contributes to the Psycholinguistic understanding of emotions by exploring the relationship between dialogue characteristics and speaker's emotions in different types of spoken dialogue. In the future, our emotion recognition model has the potential to be integrated into HCI systems and improve the interaction quality.

Appendix A

Review of Existing Emotion Databases

To support our discussion in Section 3.1.2 on current emotion databases, in the following tables we review the most recent or widely used emotion databases. In the "Modality" columns, "a" represents the database includes audio recordings, "v" represents the database includes video recordings, "p" represents the database includes physiological signal recordings. In the "Emotion" columns, "Big-6" represents the Big-6 categorical emotions are used for emotion annotation (Ekman et al., 1987). In the "Type" columns, "a" represents data collection by acting, "i" represents data collection by inducing, "n" represents data collection by recording natural and spontaneous behaviours.

Table A.a. Recent Multimodal Emotion Databases

Database	Modality	Emotion	Type	Language	Size	Description
AVEC2013/14 (Valstar et al., 2013)	a, v	Depression, Arousal, Valence	a, i	German	292 participants	read passage with particular emotion or recall emotional memories
LIRIS-ACCEDE (Baveye et al., 2013)	a, v	Big-6	a	English	9800 clips	1518 annotators labeling movie clips
DEAP (Koelstra et al., 2012)	v, p	other dimensional	i	null	32 participants	participants watching videos
GEMEP (Bänziger et al., 2012)	a, v	other categorical	a	French	10 participants	acting emotional scenarios
Belfast Induced (Sneddon et al., 2012)	a, v	other categorical	i, n	English	~250 participants	watch films or finish tasks
MAHNOB-HCI (Soleymani et al., 2012a)	a, v, p	other categorical, other dimensional	i	English	27 participants	watch emotional clips
HUMANIE (Douglas-Cowie et al., 2011)	a, v	other categorical, other dimensional	a, i, n	multiple	unknown	add emotion annotations to several existing databases
SAL (McKeown et al., 2010) (SEMAINE, AVEC2011/12)	a, v	other categorical, other dimensional	n	English	24 participants	Human talk to WOZ agents with personality bias
IEMOCAP (Busso et al., 2008)	a, v	arousal, valence, Big-6, other categorical	a	English	10 participants	acting emotional scenarios
Vera am Mittag (VAM) (Grimm et al., 2008)	a, v	other dimensional	n	German	12 hours	TV talk show
AMI (AMIDA) (Hain et al., 2006)	a, v	dialogue act	n	English	100 hours	Role-play meeting recordings
eNTERFACE'05 (Martin et al., 2006)	a, v	Big-6	a	English	42 participants	participants act emotional stories
EmoTV (Abrilian et al., 2005)	a, v	other dimensional	n	French	48 participants	TV talk show
Belfast naturalistic (Douglas-Cowie et al., 2003)	a, v	other categorical, other dimensional	n	English	31 male, 94 female	10–60s clips taken from TV chat shows, current affairs interviews

Table A.b. Audio-Only Emotion Databases

Database	Emotion	Type	Language	Size	Description
AIBO (InterSpeech'09) (Schuller et al., 2009)	other categorical	n	German	51 kids	kids play with the AIBO dog robot
EMOVO (Giovannella et al., 2009)	Big-6	a	Italian	14 participants	read sentences in particular emotions
UT-SCOPE (Varadarajan et al., 2006)	other categorical	n	English	30 participants	people interact with ASR system
SYMPAFLY (Batliner et al., 2003)	other categorical	n	German	110 dialogues	Naive users book flights using machine dialogue system
Berlin (EMO-DB) (Kienast and Sendlmeier, 2000)	other categorical	a	German	10 participants	participants asked to read passages written with appropriate emotion

Table A.c. Video-Only Emotion Databases

Database	Emotion	Type	Size	Description
FaceWarehouse (Cao et al., 2014)	other categorical	a	150 participants	recording and motion capture of 3D facial expressions
FEEDB (Szwoch, 2013)	other categorical	a	50 participants	motion capture and recording of making facial expressions
DISFA (Mavadati et al., 2013)	other dimensional	i	27 participants	recording facial expressions while participant watching video clips
CASME (Yan et al., 2013)	other categorical	i	35 participants	recording facial micro-expressions while participant watching video clips
GAPED (Dan-Glauser and Scherer, 2011)	other dimensional	i	60 participants	photos
ICT-3DRFE (Stratou et al., 2011)	other categorical	a	23 participants	3D models and images of facial expressions with illumination variance control
USTC-NVIE (Wang et al., 2010)	Big-6	a, i	215 participants	showing images to participants or ask them to act
Extended Cohn-Kanade (CK+) (Lucey et al., 2010)	Big-6	a, n	210 participants	photos and videos
MUG (Aifanti et al., 2010)	Big-6	a, i	86 participants	watching videos or posing
YorkDDT (Warren et al., 2009)	other categorical	a	9 participants	participants giving emotional or unemotional truthful descriptions or lies

Table A.d. Other Unimodal Emotion Databases

Database	Modality	Emotion	Type	Language	Size	Description
Fleureau et al. (2012)	physiological	Arousal, Valence	i	null	10 participants	participants watching various videos
Liu et al. (2010)	EEG signal	Arousal, Valence	i	null	10 participants	participants listen to emotional sound clips
Lv et al. (2008)	keystroke pressure	Big-6	i	null	50 participants	participants listen to emotional stories while typing
Chanel et al. (2006)	physiological	Arousal, Valence	i	null	4 participants	participants watching various images
WordNet-Affect (Strapparava and Valitutti, 2004)	text	other categorical	n	English	100 million words	words and emotions correlation

Appendix B

Lists of Low-Level Descriptors (LLDs) and Functionals Used for Acoustic Feature Extraction

In Section 3.4.1 we introduced the LLD acoustic features and the eGeMAPS acoustic features, which are benchmark feature sets used in state-of-the-art emotion recognition research. Here we provide the list of LLDs and functionals used for extracting these features. The LLDs and functionals shared between the AVEC2012 baseline LLD set and the IS10 LLD set are in bold in the tables.

B.1 AVEC2012 Baseline LLD Set and InterSpeech 2010 LLD Set

We extracted two LLD acoustic feature sets: the AVEC2012 baseline LLD set, and the InterSpeech 2010 LLD set. Below are lists of LLDs and functionals used for extracting these features. The items in bold are shared between these two feature sets. In Table B.1 and Table B.3, "F2F" is short for "frame-to-frame". F0 is sub-harmonic summation followed by Viterbi smoothing. F2F jitter is pitch period length deviations. F2F shimmer is amplitude deviations between pitch periods. In Table B.2 and Table B.4, "STD" is short for standard deviation. "LA" is short for linear approximation. In Table B.3, the loudness is normalized intensity raised to a power of 0.3. In Table B.4, "IQR" is short for inter-quartile range.

Table B.1: 31 LLDs used for the AVEC2012 LLD set (Schuller et al., 2012)

Energy & spectral (25)	
loudness (auditory model based)	skewness
energy in bands 250Hz-650Hz and 1kHz-4kHz	harmonicity
25%, 50%, 75%, and 90% spectral roll-off points	entropy
spectral flux	MFCC 1-10
zero crossing rate	kurtosis
variance	psychoacousitc sharpness
Voicing related (6)	
probability of voicing	F0
logarithmic Harmonics-to-Noise Ratio	**F2F jitter**
differential F2F jitter	**F2F shimmer**

Table B.2: 42 functionals used for the AVEC2012 LLD set (Schuller et al., 2012)

Statistical functionals (38)	
(positive) **arithmetic mean**	**arithmetic STD**
outlier-robust min value (1% percentile)	skewness
outlier-robust max value (99% percentile)	kurtosis
outlier robust signal range (1%-99% percentile)	quartiles
percentage of frames contour above 25%, 50%, 90%	flatness
percentage of frames contour is rising	inter-quartile ranges
max, min, mean, STD of segment length	LPC 1-5
mean, STD of rising and falling slopes	LP gain
mean, STD of inter maxima distances	root quadratic mean
amplitude mean of maxima and minima	amplitude range of maxima
Regression functionals (4)	
quadratic regression coefficient a	**linear error of LA**
linear regression slope	**quadratic error of LA**

Table B.3: 38 LLDs used for the InterSpeech 2010 LLD set (Eyben et al., 2010a)

Energy & spectral (32)	
8 line spectral pair frequencies (from 8 LPC coefficients)	loudness
logarithmic power of Mel-frequency bands 0-7	MFCC 0-14
Voicing related (6)	
envelope of the smoothed F0 contour	**F2F jitter**
voicing probability of the final F0 candidate	**F2F shimmer**
differential F2F jitter	smoothed F0 contour

Table B.4: 21 functionals used for the InterSpeech 2010 LLD set (Eyben et al., 2010a)

Statistical functionals (17)	
absolute frame position of max	**arithmetic mean**
absolute frame position of min	**arithmetic STD**
first quartile (25%, 50%, and 75% percentile)	skewness (3rd order moment)
percentage of time the signal is above 75% and 90%	kurtosis (4th order moment)
outlier-robust min value (1% percentile)	IQR quartile2-quartile1
outlier-robust max value (99% percentile)	IQR quartile3-quartile2
outlier robust signal range (1%-99% percentile)	IQR quartile3-quartile1

Regression functionals (4)	
linear error of LA	slope (m) of LA
quadratic error of LA	offset (t) of LA

B.2 Expanded Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)

Compared to the AVEC2012 baseline LLD features and the InterSpeech 2010 LLD features, the eGeMAPS feature set contains a small number of LLDs motivated by Psycholinguistic studies. The LLDs and functionals used for extracting the eGeMAPS feature set are listed below. In Table B.5, loudness is estimation of perceived signal intensity from an auditory spectrum, spectral flux is the difference of the spectra of two consecutive frames, pitch is logarithmic F0 on a semitone frequency scale starting at 27.5 Hz, "RoE" is short for ratio of energy, "LRS" is short for Linear Regression Slope.

Table B.5: 25 LLDs used for the eGeMAPS feature set (Eyben et al., 2015b)

Energy & spectral (15)	
LRS of the log power spectrum of 0-500Hz	loudness
LRS of the log power spectrum of 500-1500Hz	spectral flux
ratio of summed energy: 50Hz-1kHz, 1kHz-5kHz	MFCC 1-4
ratio of strongest energy peak in 0-2kHz to 2kHz-5kHz	RoE: F1 to F0
RoE: first F0 harmonic to second F0 harmonic	RoE: F2 to F0
RoE: first F0 harmonic to the highest F3 harmonic	RoE: F3 to F0
Voicing related (10)	
deviations in individual consecutive F0 period lengths	pitch
peak amplitude difference of consecutive F0 periods	centre frequency: F1, F2, F3
RoE: harmonic components to noise-like components	bandwidth: F1, F2, F3

Table B.6: 13 Functionals used for the eGeMAPS feature set (Eyben et al., 2015b)

Functionals applied to the voiced segments (8)	
the range of 20th to 80th percentile	20th percentile
mean and standard deviation of rising slopes	50th percentile
mean and standard deviation of falling slopes	80th percentile
Functionals applied to the unvoiced segments (5)	
Hammarberg Index	mean of Alpha Ratio
spectral slopes from 0-500 Hz and 500-1500Hz	mean of spectral flux

Appendix C

Descriptive Statistics of the DIS-NVs and the Global Prosodic Features

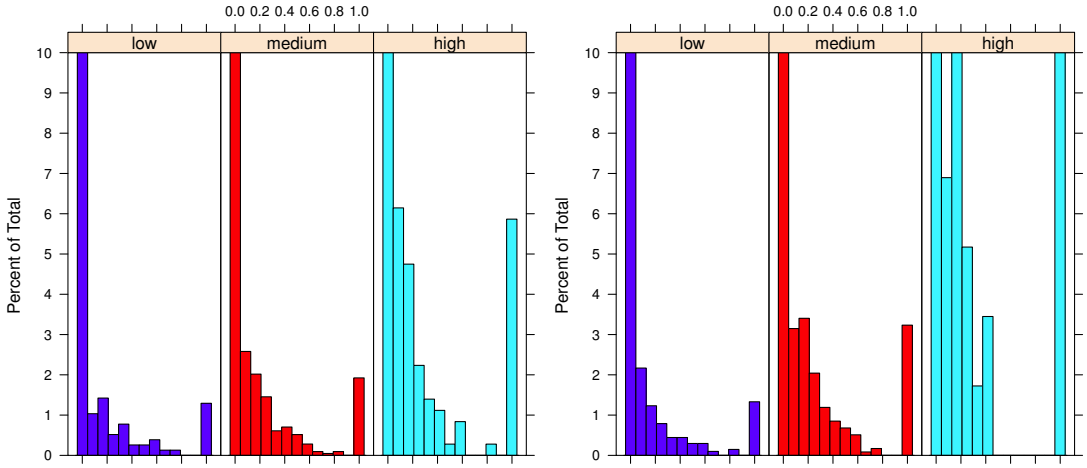
To study the differences between spontaneous and acted dialogue, we examined distributions of DIS-NVs and Global Prosodic (GP) features of Bone et al. (2014) on each emotion dimension for each databases in Chapter 5. This appendix contains additional distribution figures.

C.1 Distribution of DIS-NVs

In these histograms of DIS-NV distribution, dark blue, red, and light blue represent the category *low*, *medium*, and *high* respectively for each emotion dimension. The x axis represents the percentage of the total duration of a utterance being a DIS-NV, the y axis represents the percentage of *utterances* with the same value on the x axis. We set the maximum value of y axis to 10 (10% of total utterances) and only plot the utterances that contains this type of DIS-NV in order to show the distribution more clearly.

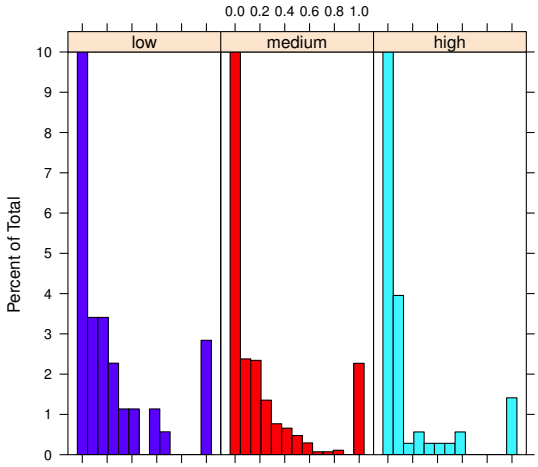
As shown in the Expectancy graph of Figure C.2, there is no laughter in utterances with high Expectancy in spontaneous dialogue. This is consistent with the empirical knowledge that laughter is more likely to occur as an unexpected affective burst (Scott, 2013).

C.1.1 Spontaneous AVEC2012 Database

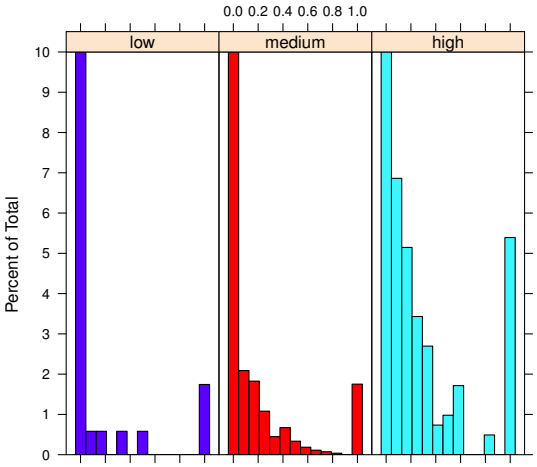


(a) Arousal

(b) Expectancy

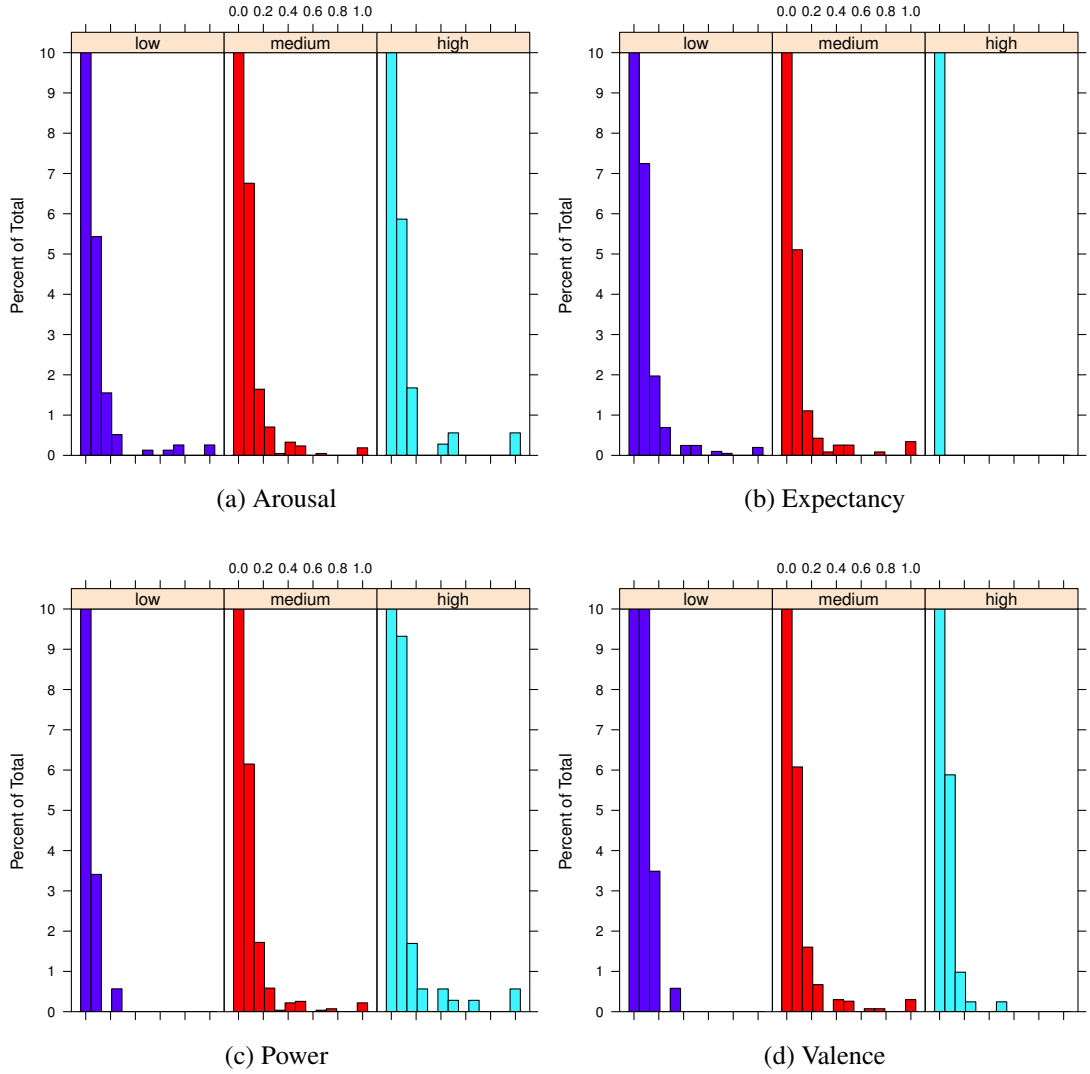


(c) Power



(d) Valence

Figure C.1: Filled Pause Distribution in AVEC2012 Utterances



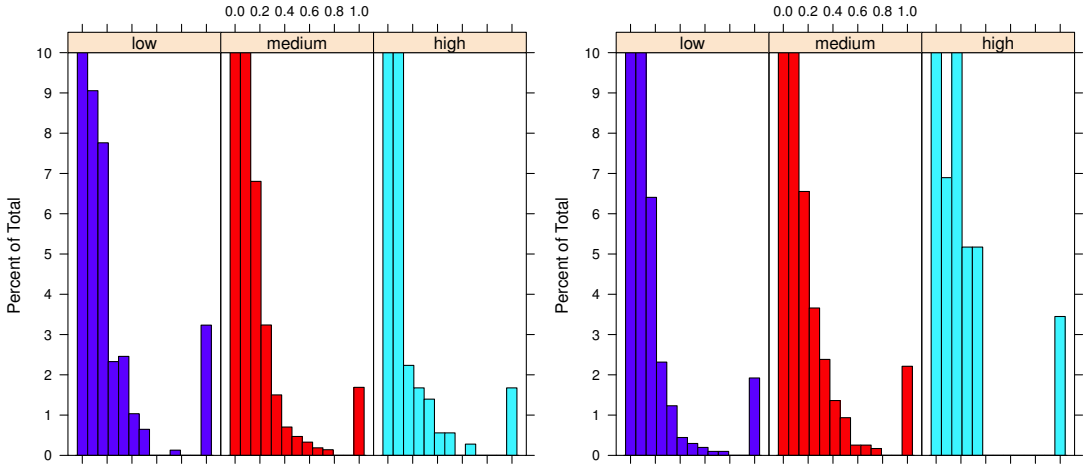
(a) Arousal

(b) Expectancy

(c) Power

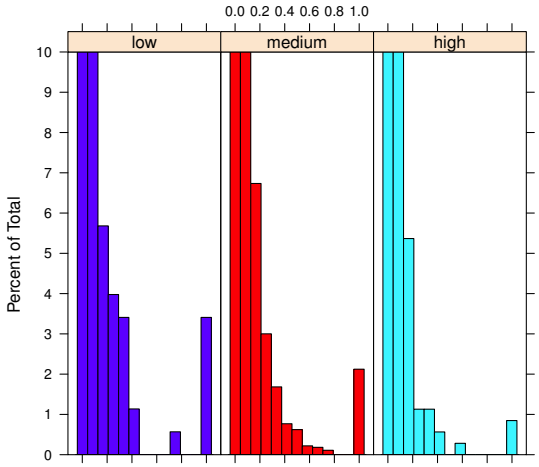
(d) Valence

Figure C.2: Laughter Distribution in AVEC2012 Utterances

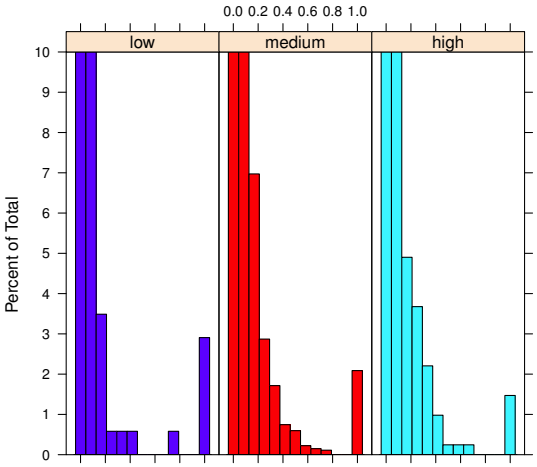


(a) Arousal

(b) Expectancy

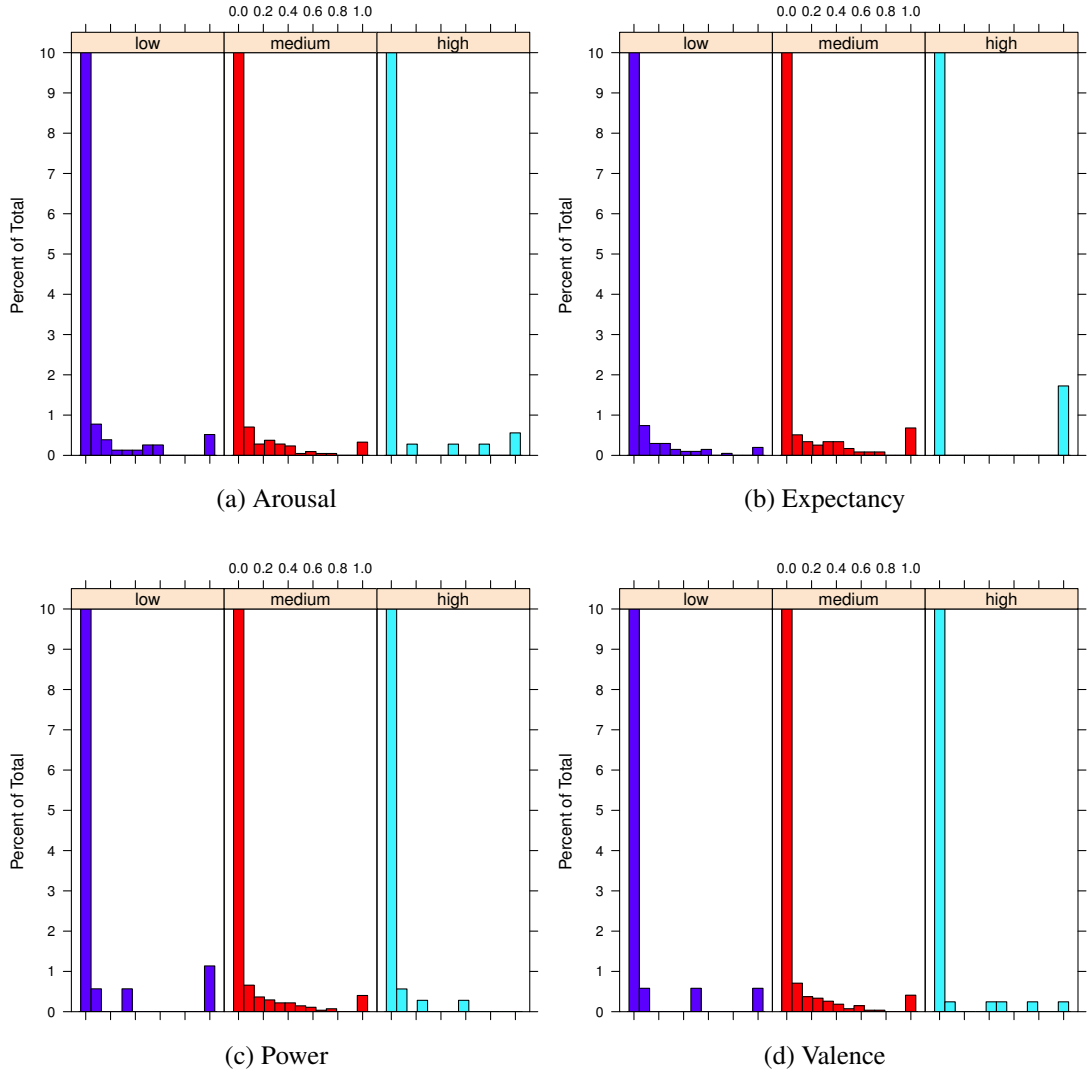


(c) Power



(d) Valence

Figure C.3: Filler Distribution in AVEC2012 Utterances



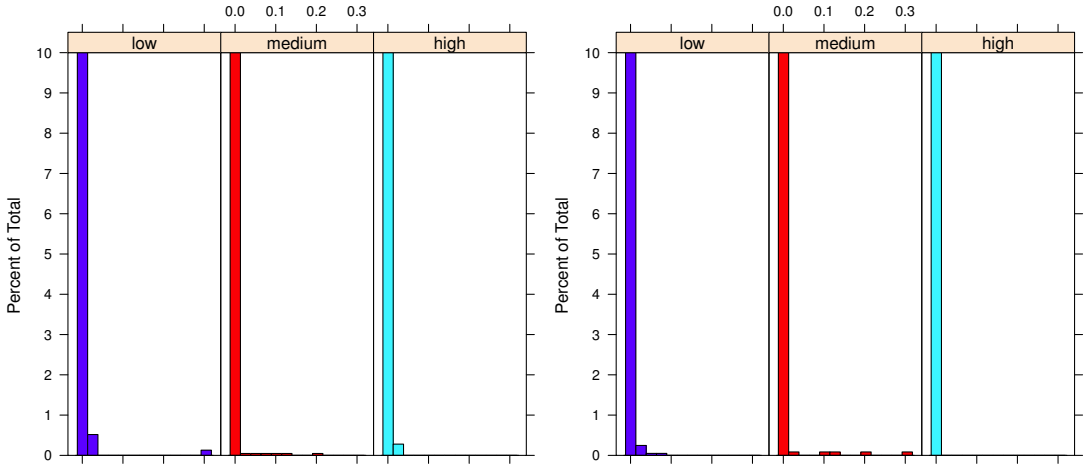
(a) Arousal

(b) Expectancy

(c) Power

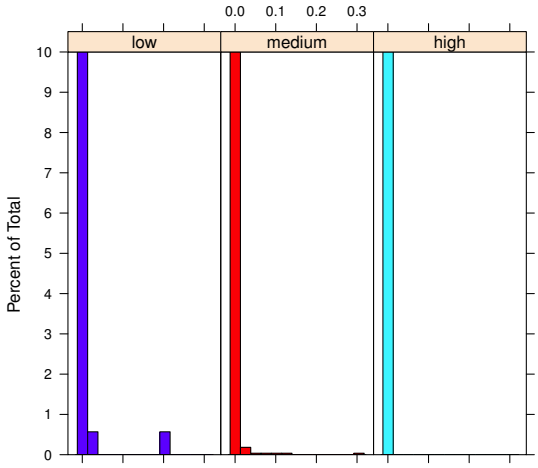
(d) Valence

Figure C.4: Stutter Distribution in AVEC2012 Utterances

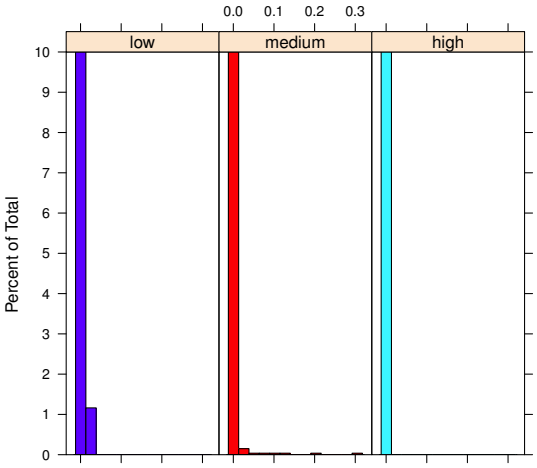


(a) Arousal

(b) Expectancy



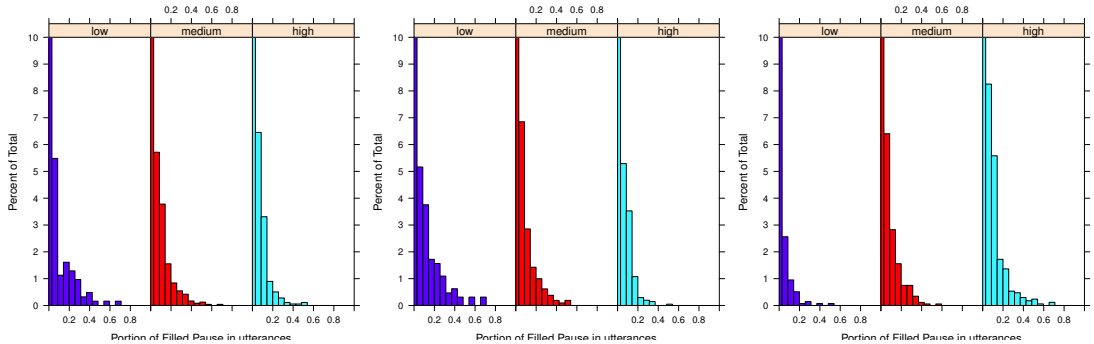
(c) Power



(d) Valence

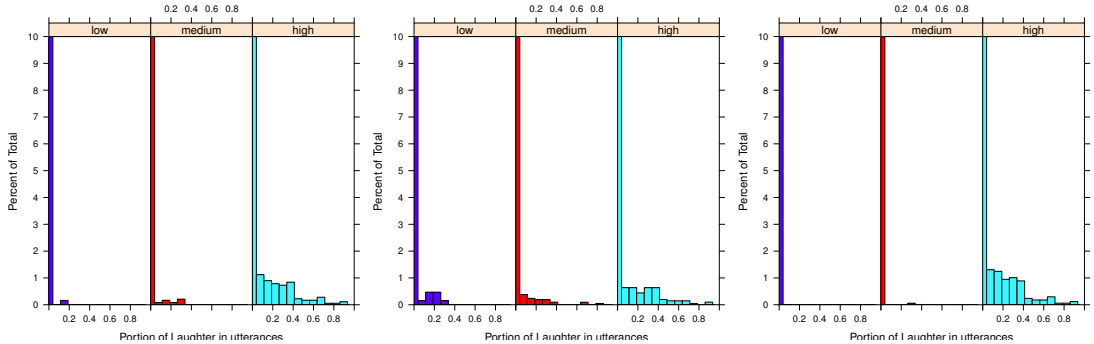
Figure C.5: Audible Breath Distribution in AVEC2012 Utterances

C.1.2 Non-Scripted Subset of the IEMOCAP Database



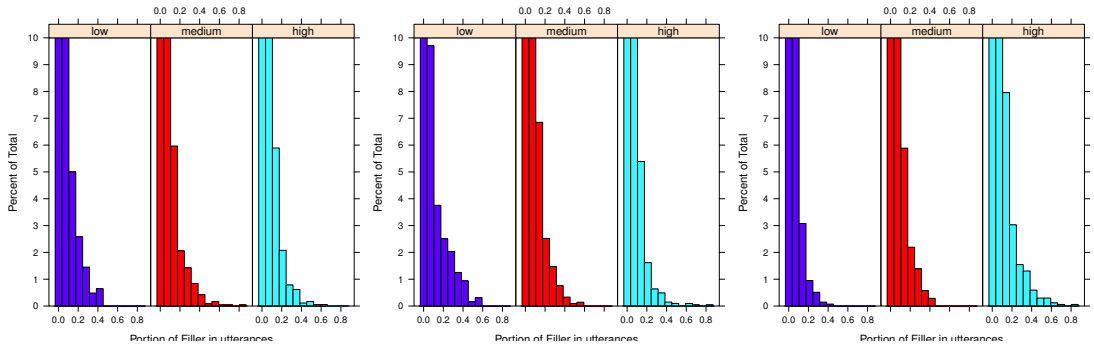
(a) Arousal (b) Power (c) Valence

Figure C.6: Filled Pause Distribution in Non-Scripted IEMOCAP Utterances



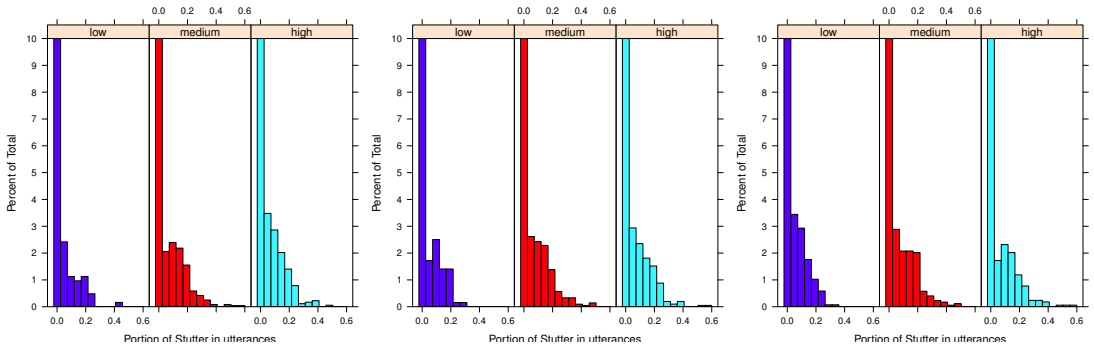
(a) Arousal (b) Power (c) Valence

Figure C.7: Laughter Distribution in Non-Scripted IEMOCAP Utterances



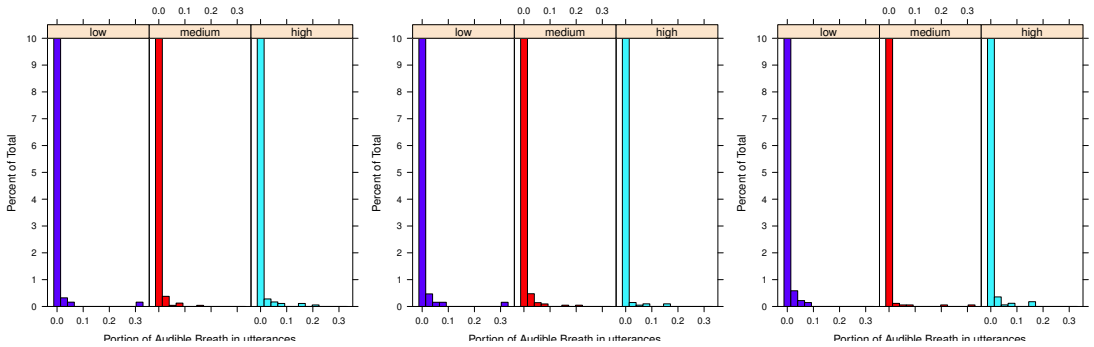
(a) Arousal (b) Power (c) Valence

Figure C.8: Filler Distribution in Non-Scripted IEMOCAP Utterances



(a) Arousal (b) Power (c) Valence

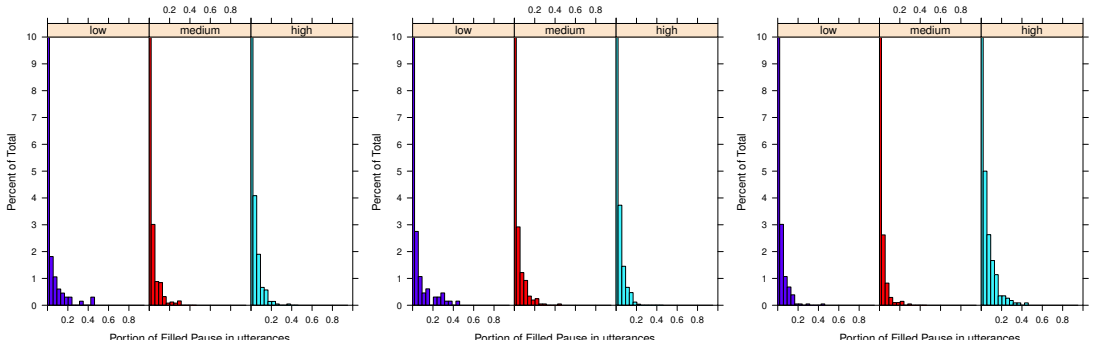
Figure C.9: Stutter Distribution in Non-Scripted IEMOCAP Utterances



(a) Arousal (b) Power (c) Valence

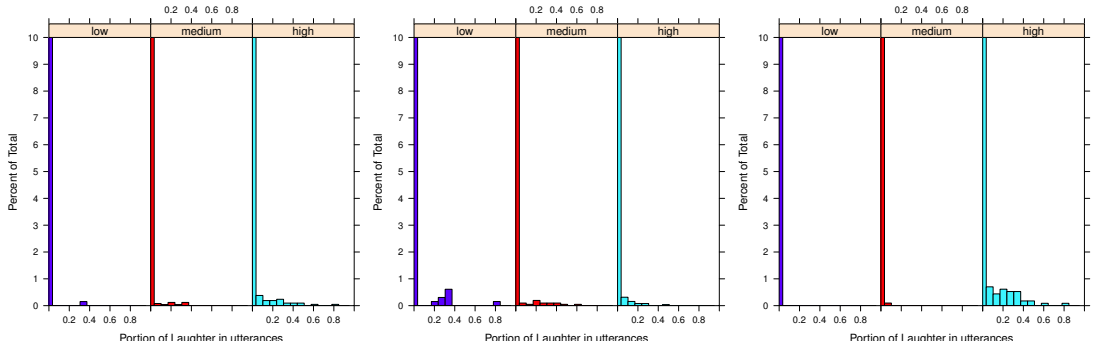
Figure C.10: Audible Breath Distribution in Non-Scripted IEMOCAP Utterances

C.1.3 Scripted Subset of the IEMOCAP Database



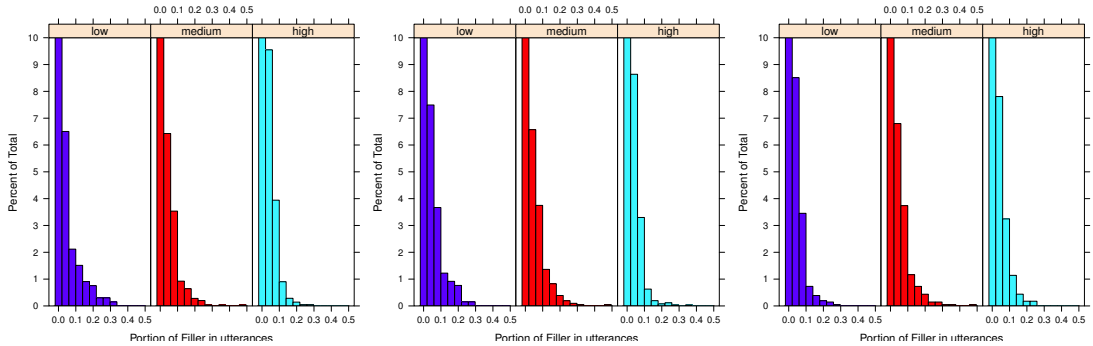
(a) Arousal (b) Power (c) Valence

Figure C.11: Filled Pause Distribution in Scripted IEMOCAP Utterances



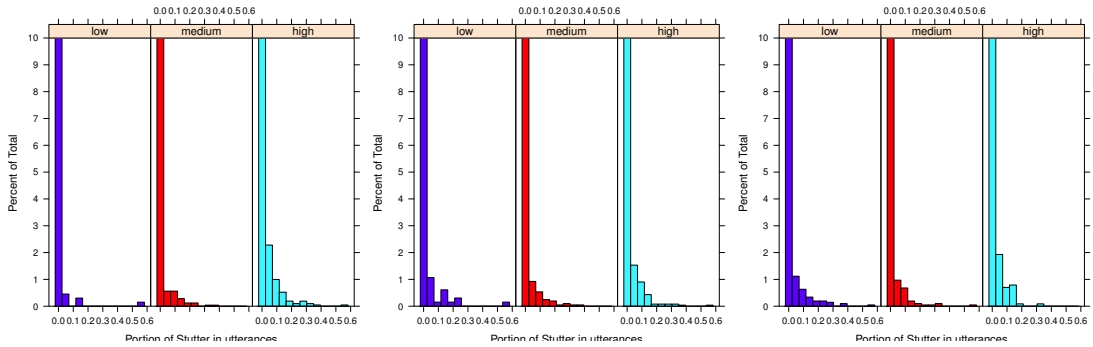
(a) Arousal (b) Power (c) Valence

Figure C.12: Laughter Distribution in Scripted IEMOCAP Utterances



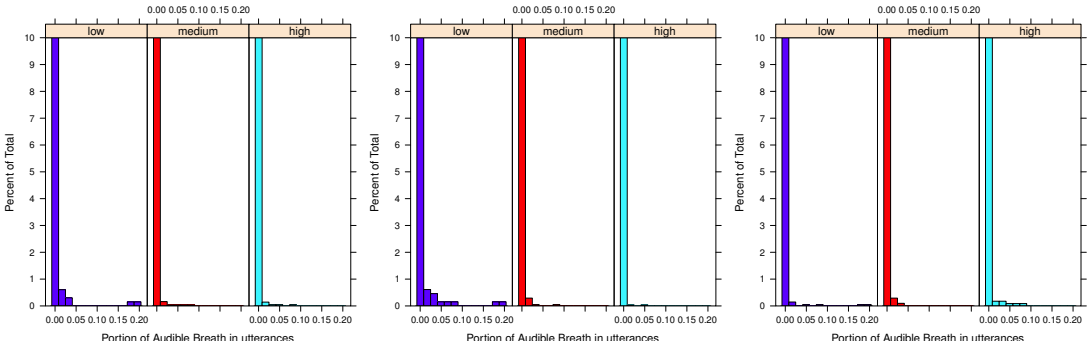
(a) Arousal (b) Power (c) Valence

Figure C.13: Filler Distribution in Scripted IEMOCAP Utterances



(a) Arousal (b) Power (c) Valence

Figure C.14: Stutter Distribution in Scripted IEMOCAP Utterances



(a) Arousal (b) Power (c) Valence

Figure C.15: Audible Breath Distribution in Scripted IEMOCAP Utterances

C.2 Distribution of Global Prosodic (GP) Features

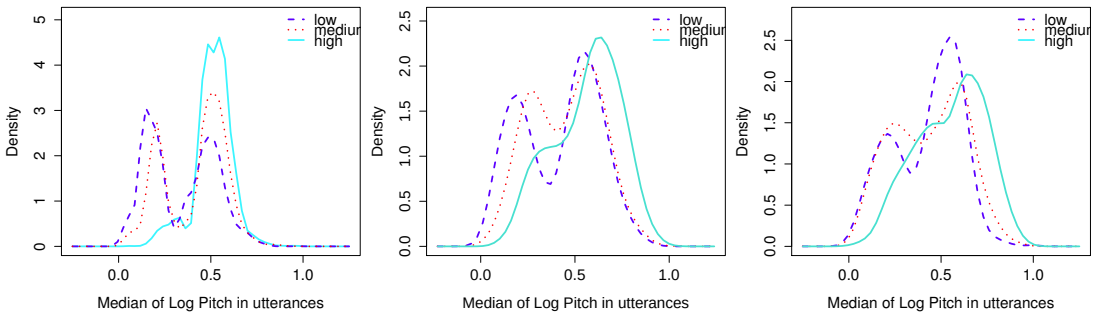
Here we plot the smoothed density graphs of the Global Prosodic (GP) features of Bone et al. (2014) over utterances of the spontaneous AVEC2012 database and the acted IEMOCAP database. Dark blue, red, and light blue represent *low*, *medium*, and *high* Expectancy, respectively.

C.2.1 Distribution of GP Features on the Arousal, Power, and Valence Dimension

Note that the Expectancy dimension was not annotated on the IEMOCAP database. Therefore, here we compared the distribution of GP features in both databases on the Arousal, Power, and Valence dimensions.

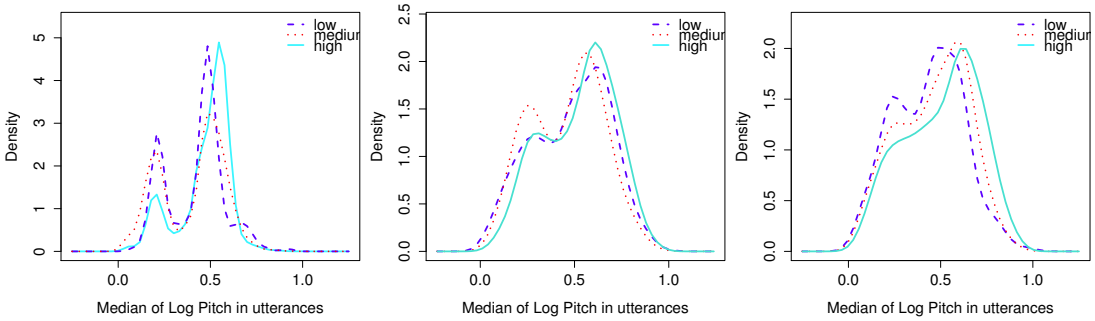
C.2.1.1 Distribution of Log Pitch

In Figures C.16, C.17, and C.18, we plot the distributions of the median log pitch over utterances of both databases. As we can see, the log pitch distribution on the AVEC2012 database (left-most graphs in each figure) is more grouped than the distribution on the IEMOCAP database (middle and right-most graphs in each figure), which indicates the utterances in the AVEC2012 database have less variation in log pitch values than utterances in the IEMOCAP database. For log pitch distribution on non-scripted IEMOCAP utterances (middle graphs in each figure), such grouping effect is less obvious, while in the scripted IEMOCAP utterances (right-most graphs in each figure) this grouping is hardly visible.



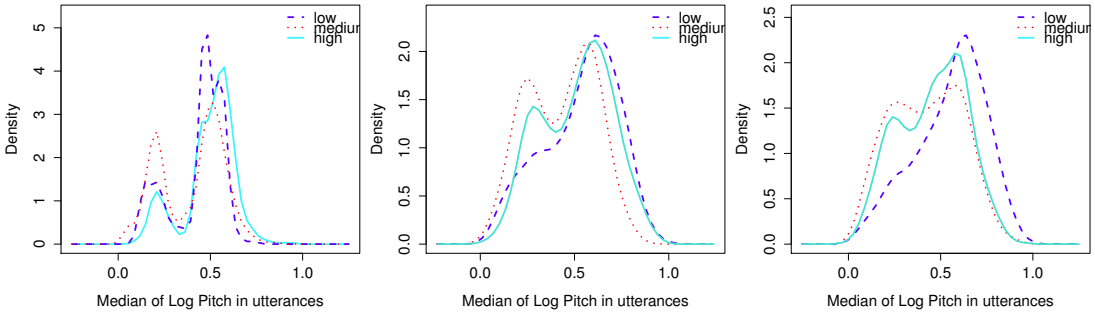
(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.16: Log Pitch Distribution on Arousal



(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.17: Log Pitch Distribution on Power

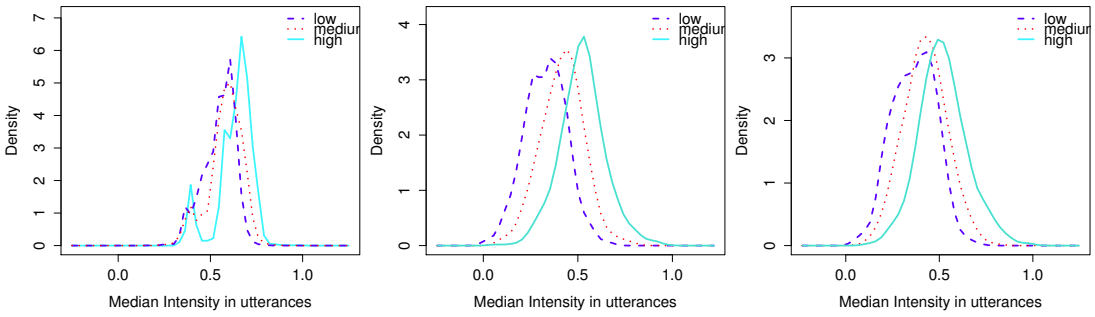


(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.18: Log Pitch Distribution on Valence

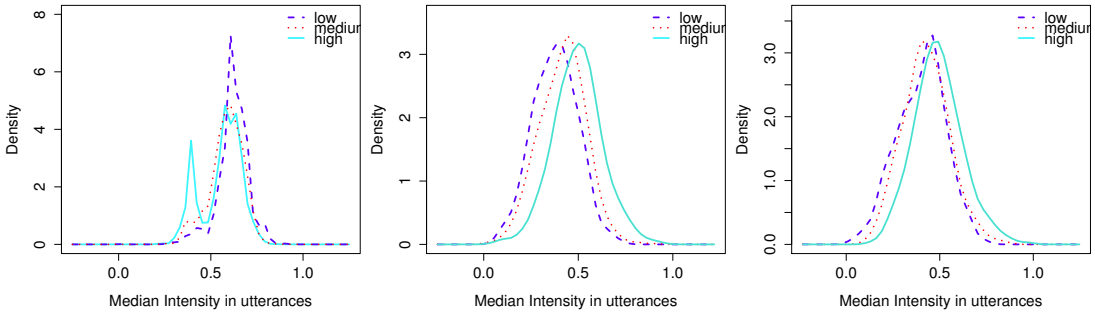
C.2.1.2 Distribution of Intensity

In Figures C.19, C.20, and C.21, we plot the distribution of the median intensity over utterances of both databases.



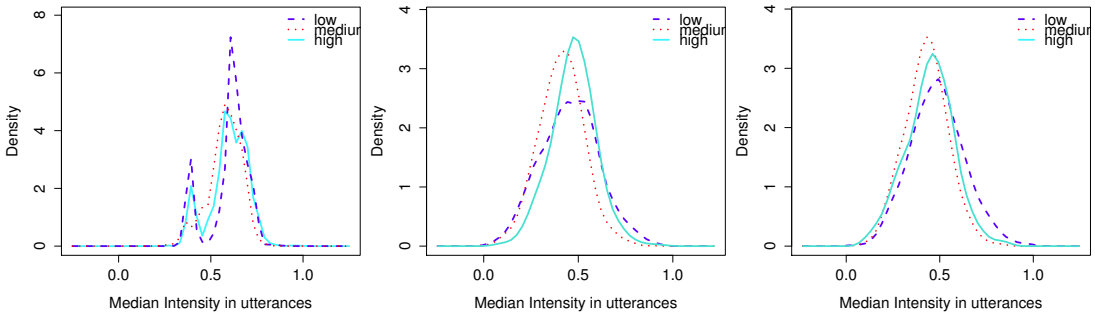
(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.19: Intensity Distribution on Arousal



(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.20: Intensity Distribution on Power

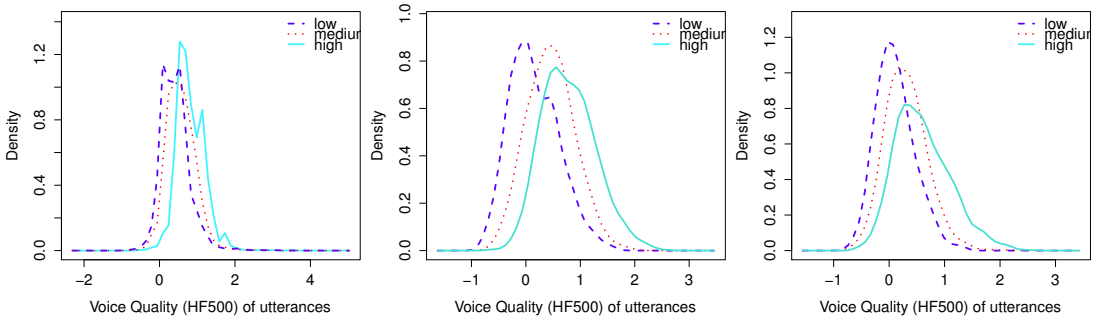


(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.21: Intensity Distribution on Valence

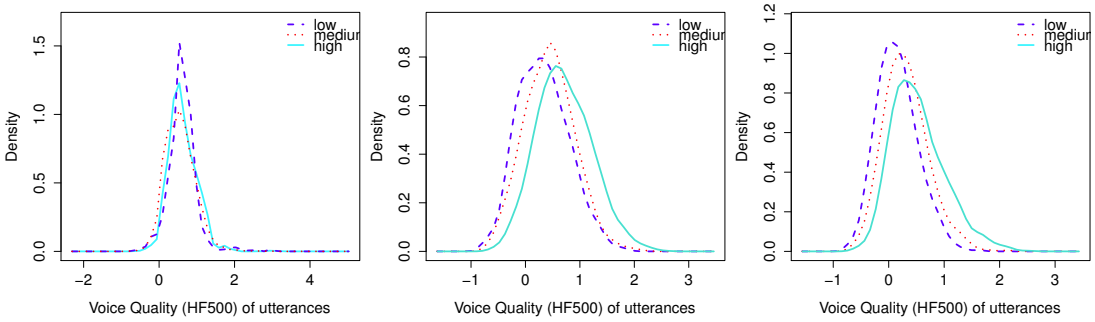
C.2.1.3 Distribution of Voice Quality

In Figures C.22, C.23, and C.24, we plot the distribution of the Voice Quality (HF500) over utterances of both databases.



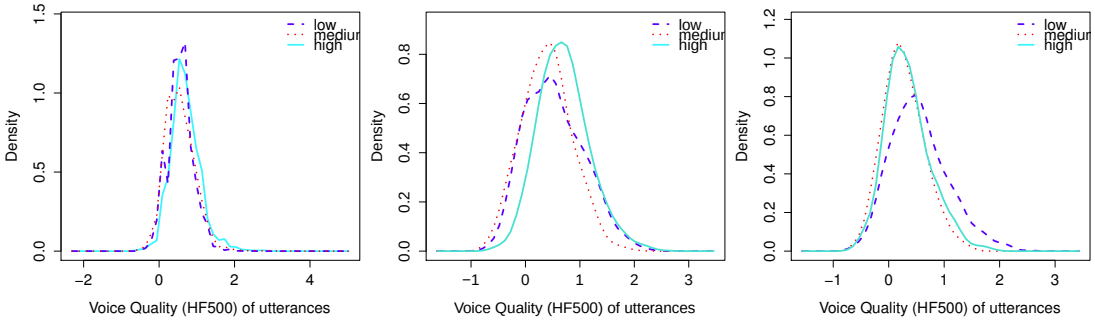
(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.22: Voice Quality (HF500) Distribution on Arousal



(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.23: Voice Quality (HF500) Distribution on Power

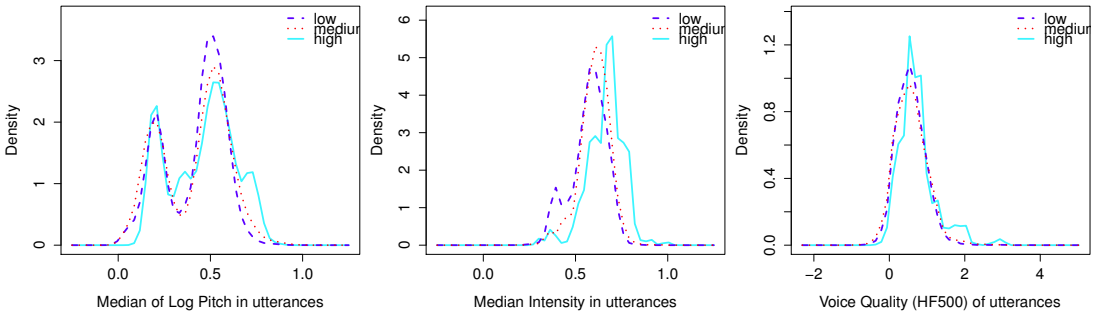


(a) AVEC2012 (b) Non-Scripted IEMOCAP (c) Scripted IEMOCAP

Figure C.24: Voice Quality (HF500) Distribution on Valence

C.2.2 Distribution of GP Features on the Expectancy Dimension

Figure C.25 contains smoothed density graphs of GP feature distributions on the Expectancy dimension in the AVEC2012 database.



(a) Log Pitch (b) Intensity (c) Voice Quality (HF500)

Figure C.25: GP Distributions on Expectancy Dimension in AVEC2012 Utterances

Bibliography

Abdelwahab, M. and Busso, C. (2015). Supervised domain adaptation for emotion recognition from speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5058–5062. IEEE.

Abdi, H. (2003). Partial least square regression (PLS regression). *Encyclopedia for research methods for the social sciences*, pages 792–795.

Abrilian, S., Devillers, L., Buisine, S., and Martin, J.-C. (2005). EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*.

Adelswärd, V. (1989). Laughter and dialogue: The social significance of laughter in institutional discourse. *Nordic Journal of Linguistics*, 12(02):107–136.

Aifanti, N., Papachristou, C., and Delopoulos, A. (2010). The MUG facial expression database. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pages 1–4. IEEE.

Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256.

Anastasia, T. and Leontios, H. (2016). AUTH-SGP in MediaEval 2016 emotional impact of movies task. In *Proceedings of the MediaEval 2016 Multimedia Benchmark Workshop*.

Arifin, S. and Cheung, P. Y. (2008). Affective level video segmentation by utilizing the pleasure-arousal-dominance information. *IEEE Transactions on Multimedia*, 10(7):1325–1341.

Audhkhasi, K., Kandhway, K., Deshmukh, O. D., and Verma, A. (2009). Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4857–4860. IEEE.

Averill, J. R. (1980). A constructivist view of emotion. In *Theories of emotion*, pages 305–339. Elsevier.

Aytar, Y., Vondrick, C., and Torralba, A. (2016). SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900.

Bachorowski, J.-A., Smoski, M. J., and Owren, M. J. (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, 110(3):1581–1597.

Baltrusaitis, T., Banda, N., and Robinson, P. (2013). Dimensional affect recognition using continuous conditional random fields. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Banda, N., Engelbrecht, A., and Robinson, P. (2015). Continuous emotion recognition using a particle swarm optimized NARX neural network. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 380–386. IEEE.

Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161.

Barczewska, K. and Igras, M. (2013). Detection of disfluencies in speech signal. *Challenges of Modern Technology*, 4.

Bard, G. V. (2007). Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric. In *ACSW2007*, volume 68, pages 117–124. Australian Computer Society, Inc.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., and Haas, J. (2003). User states, user strategies, and system performance: how to match the one with the other. In *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*.

Baveye, Y., Bettinelli, J.-N., Dellandréa, E., Chen, L., and Chamaret, C. (2013). A large video data base for computational models of induced emotion. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 13–18. IEEE.

Baveye, Y., Chamaret, C., Dellandréa, E., and Chen, L. (2017). Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affective Computing*.

Baveye, Y., Dellandréa, E., Chamaret, C., and Chen, L. (2015a). Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 77–83. IEEE.

Baveye, Y., Dellandrea, E., Chamaret, C., and Chen, L. (2015b). LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55.

Benotti, L. (2009). Clarification potential of instructions. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 196–205. Association for Computational Linguistics.

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems.

Bone, D., Lee, C., and Narayanan, S. (2014). Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. *Affective Computing, IEEE Transactions on*, 5(2):201–213.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. pages 144–152.

Bozkurt, E., Erzin, E., and Yemez, Y. (2015). Affect-expressive hand gestures synthesis and animation. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., and Huang, T. S. (2016). Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *AVEC2016*, pages 97–104. ACM.

Braunschweiler, N. and Chen, L. (2013). Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. In *Proceedings of the 8th ISCA Speech Synthesis Workshop*, pages 1–6.

Brueckner, R. and Schuller, B. (2015). Be at odds? Deep and hierarchical neural networks for classification and regression of conflict in speech. In *Conflict and Multimodal Communication*, pages 403–429. Springer.

Burmania, A., Abdelwahab, M., and Busso, C. (2016). Trade-off between quality and quantity of emotional annotations to characterize expressive behaviors. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5190–5194. IEEE.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Camargo, J. E. and González, F. A. (2016). Multimodal latent topic analysis for image collection summarization. *Information Sciences*, 328:270–287.

Cao, C., Weng, Y., Lin, S., and Zhou, K. (2013). 3D shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41.

Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K. (2014). FaceWarehouse: a 3D facial expression database for visual computing. *Visualization and Computer Graphics, IEEE Transactions on*, 20(3):413–425.

Cardinal, P., Dehak, N., Koerich, A. L., Alam, J., and Boucher, P. (2015). ETS system for AV+EC 2015 challenge. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 17–23. ACM.

Chanel, G., Kronegg, J., Grandjean, D., and Pun, T. (2006). Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals. In *Multimedia content representation, classification and security*, pages 530–537. Springer.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Chao, L., Tao, J., Yang, M., Li, Y., and Wen, Z. (2015). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM.

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

Chen, S., Gao, Z., and Wang, S. (2016). Emotion recognition from peripheral physiological signals enhanced by EEG. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2827–2831. IEEE.

Chen, S. and Jin, Q. (2015). Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56. ACM.

Chen, S. and Jin, Q. (2016). RUC at MediaEval 2016 emotional impact of movies task: Fusion of multimodal features. In *Proceedings of the MediaEval 2016 Multimedia Benchmark Workshop*.

Chollet, F. (2015). Keras. `https://github.com/fchollet/keras`.

Chong, C. S., Kim, J., and Davis, C. (2015). Exploring acoustic differences between Cantonese (tonal) and English (non-tonal) spoken expressions of emotions. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Cireşan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, 22(12):3207–3220.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge, 2 edition.

Cornelius, R. R. (2000). Theoretical approaches to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Crystal, D. (1976). *Prosodic systems and intonation in English*, volume 1. CUP Archive.

Dael, N., Goudbeek, M., and Scherer, K. (2013). Perceived gesture dynamics in nonverbal expression of emotion. *Perception*, 42(6):642–657.

Dan-Glauser, E. S. and Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior research methods*, 43(2):468–477.

De Silva, L. C., Miyasato, T., and Nakatsu, R. (1997). Facial emotion recognition using multi-modal information. In *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, volume 1, pages 397–401. IEEE.

Dellandréa, E., Chen, L., Baveye, Y., Sjöberg, M., Chamaret, C., and Lyon, E. (2016). The MediaEval 2016 emotional impact of movies task. In *Proceedings of the MediaEval 2016 Multimedia Benchmark Workshop*.

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072.

Dhall, A. et al. (2012). Collecting large, richly annotated facial-expression databases from movies.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. (2017). From individual to group-level emotion recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 524–528. ACM.

D'Mello, S. and Kory, J. (2012). Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 31–38. ACM.

D'mello, S. K. and Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):43.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1):33–60.

Douglas-Cowie, E., Cox, C., Martin, J.-C., Devillers, L., Cowie, R., Sneddon, I., McRorie, M., Pelachaud, C., Peters, C., Lowry, O., and Batliner, A. (2011). The HUMAINE database. In *Emotion-Oriented Systems*, pages 243–284. Springer.

Dupont, S., Çakmak, H., Curran, W., Dutoit, T., Hofmann, J., McKeown, G., Pietquin, O., Platt, T., Ruch, W., and Urbain, J. (2016). Laughter research: a review of the ilhaire project. In *Toward Robotic Socially Believable Behaving Systems-Volume I*, pages 147–181. Springer.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., and Scherer, K. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

Eyben, F., Huber, B., Marchi, E., Schuller, D., and Schuller, B. (2015a). Real-time robust recognition of speakers' emotions and characteristics on mobile platforms. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 778–780. IEEE.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., et al. (2015b). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*.

Eyben, F., Weninger, F., Wöllmer, M., and Schuller, B. (2016). open-Source Media Interpretation by Large feature-space Extraction.

Eyben, F., Woellmer, M., and Schuller, B. (2010a). the Munich open speech and music interpretation by large space extraction toolkit. *IEEE Netw.*, 24(2):36–41.

Eyben, F., Wöllmer, M., and Schuller, B. (2010b). OpenSMILE: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM.

Finlayson, I. R. (2014). *Testing the roles of disfluency and rate of speech in the coordination of conversation*. PhD thesis, QUEEN MARGARET UNIVERSITY.

Fleureau, J., Guillotel, P., and Huynh-Thu, Q. (2012). Physiological-based affect event detector for entertainment video applications. *Affective Computing, IEEE Transactions on*, 3(3):379–385.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

Forbes-Riley, K. and Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9):1115–1136.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709.

Fung, P., Bertero, D., Wan, Y., Dey, A., Chan, R. H. Y., Siddique, F. B., Yang, Y., Wu, C.-S., and Lin, R. (2016). Towards empathetic human-robot interactions. *arXiv preprint arXiv:1605.04072*.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(1_suppl):123–147.

Gao, L., Qi, L., and Guan, L. (2016). Information fusion based on kernel entropy component analysis in discriminative canonical correlation space with application to audio emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2817–2821. IEEE.

Gebhard, P. (2005). ALMA: a layered model of affect. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 29–36. ACM.

Gievska, S., Koroveshovski, K., and Tagasovska, N. (2015). Bimodal feature-based fusion for real-time emotion recognition in a mobile context. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 401–407. IEEE.

Giovannella, C., Conflitti, D., Santoboni, R., and Paoloni, A. (2009). Transmission of vocal emotion: do we have to care about the listener? the case of the Italian speech corpus EMOVO. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE.

Glenn, P. (2003). *Laughter in interaction (Vol. 18)*. Cambridge University Press.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Gninkoun, G. and Soleymani, M. (2011). Automatic violence scenes detection: A multi-modal approach.

Gordon, S. L. (1990). Social structural effects on emotions. *Research agendas in the sociology of emotions*, pages 145–79.

Grandjean, D., Sander, D., and Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition*, 17(2):484–495.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865–868. IEEE.

Gu, Y., Postma, E., and Lin, H.-X. (2015). Vocal emotion recognition with log-Gabor filters. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 25–31. ACM.

Hain, T., Burget, L., Dines, J., Garau, G., Karafit, M., Lincoln, M., and Wan, V. (2006). The AMI meeting transcription system. In *Proc. NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Worskhop*, page 12.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

Han, J., Zhang, Z., Ringeval, F., and Schuller, B. (2017). Reconstruction-error-based learning for continuous emotion recognition in speech. In *42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*.

Hanjalic, A. (2006). Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100.

Hanjalic, A. and Xu, L.-Q. (2005). Affective video content representation and modeling. *IEEE transactions on Multimedia*, 7(1):143–154.

He, L., Jiang, D., and Sahli, H. (2015a). Multimodal depression recognition with dynamic visual and audio cues. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 260–266. IEEE.

He, L., Jiang, D., Yang, L., Pei, E., Wu, P., and Sahli, H. (2015b). Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huang, K.-Y., Lin, J.-K., Chiu, Y.-H., and Wu, C.-H. (2015a). Affective structure modeling of speech using probabilistic context free grammar for emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5286–5290. IEEE.

Huang, Z., Dang, T., Cummins, N., Stasak, B., Le, P., Sethu, V., and Epps, J. (2015b). An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 41–48. ACM.

Huang, Z. and Epps, J. (2016). Detecting the instant of emotion change from speech using a martingale framework. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5195–5199. IEEE.

Jaimes, A. and Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1):116–134.

Jain, S. and Asawa, K. (2016). Programming an expressive autonomous agent. *Expert Systems with Applications*, 43:131–141.

Jan, A., Gaus, Y. F. A., Zhang, F., and Meng, H. (2016). BUL in MediaEval 2016 emotional impact of movies task. In *Proceedings of the MediaEval 2016 Multimedia Benchmark Workshop*.

Jia, J., Wu, Z., Zhang, S., Meng, H. M., and Cai, L. (2014). Head and facial gestures synthesis using PAD model for an expressive talking avatar. *Multimedia Tools and Applications*, 73(1):439–461.

Jin, Q., Li, C., Chen, S., and Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4749–4753. IEEE.

Juslin, P. N. and Scherer, K. R. (2005). Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, pages 65–135.

Kächele, M., Thiam, P., Palm, G., Schwenker, F., and Schels, M. (2015). Ensemble methods for continuous affect recognition: multi-modality, temporality, and challenges. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 9–16. ACM.

Kallinen, K. and Ravaja, N. (2006). Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, 10(2):191–213.

Kaya, H., Eyben, F., Salah, A. A., and Schuller, B. (2014). CCA based feature selection with application to continuous depression recognition from acoustic speech features. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3729–3733. IEEE.

Kennedy, L. and Ellis, D. (2004). Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop*, pages 118–121.

Kienast, M. and Sendlmeier, W. F. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Kim, M. J., Yoo, J., Kim, Y., and Kim, H. (2015). Speech emotion classification using tree-structured sparse logistic regression. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Knautz, K. and Stock, W. G. (2011). Collective indexing of emotions in videos. *Journal of Documentation*, 67(6):975–994.

Knox, M. T. and Mirghafori, N. (2007). Automatic laughter detection using neural networks. In *INTERSPEECH*, pages 2973–2976.

Kobayashi, T. (2015). Prosody control and variation enhancement techniques for hmm-based expressive speech synthesis. In *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, pages 203–213. Springer.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. (2012). DEAP: A database for emotion analysis; using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1):18–31.

Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, T. (2015a). Dynamic time warping of multimodal signals for detecting highlights in movies. In *INTERPERSONAL2015*, pages 35–40. ACM.

Kostoulas, T., Chanel, G., Muszynski, M., Lombardo, P., and Pun, T. (2015b). Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals. In *QoMEX2015*, pages 1–6. IEEE.

Krikke, T. F. and Truong, K. P. (2013). Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech.

Kuznetsov, V., Liao, H., Mohri, M., Riley, M., and Roark, B. (2016). Learning n-gram language models from uncertain data. *Proceedings of Interspeech 2016*.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003). Emotion recognition by speech signals. In *INTERSPEECH*. Citeseer.

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Lee, J. and Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Lefter, I., Nefs, H. T., Jonker, C. M., and Rothkrantz, L. J. (2015). Cross-corpus analysis for acoustic recognition of negative interactions. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 132–138. IEEE.

Levenson, R. W. (1992). Autonomic nervous system differences among emotions. *Psychological science*, 3(1):23–27.

Li, T., Baveye, Y., Chamaret, C., Dellandréa, E., and Chen, L. (2015). Continuous arousal self-assessments validation using real-time physiological responses. In *ASM2015*, pages 39–44. ACM.

Lickley, R. J. (2015). Fluency and disfluency. *The handbook of speech production*, page 445.

Lim, M. Y., Dias, J., Aylett, R., and Paiva, A. (2012). Creating adaptive affective autonomous NPCs. *Autonomous Agents and Multi-Agent Systems*, 24(2):287–311.

Liu, Y., Feng, X., and Zhou, Z. (2016a). Multimodal video classification with stacked contractive autoencoders. *Signal Processing*, 120:761–766.

Liu, Y., Gu, Z., Zhang, Y., and Liu, Y. (2016b). Mining emotional features of movies. In *Proceedings of the MediaEval 2016 Multimedia Benchmark Workshop*.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1526–1540.

Liu, Y., Sourina, O., and Nguyen, M. K. (2010). Real-time EEG-based human emotion recognition and visualization. In *Cyberworlds (CW), 2010 International Conference on*, pages 262–269. IEEE.

Lotfian, R. and Busso, C. (2015). Emotion recognition using synthetic speech as neutral reference. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4759–4763. IEEE.

Lotfian, R. and Busso, C. (2016). Practical considerations on the use of preference learning for ranking emotional speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5205–5209. IEEE.

Lubis, N., Sakti, S., Neubig, G., Toda, T., Purwarianti, A., and Nakamura, S. (2016). Emotion and its triggers in human spoken dialogue: Recognition and analysis. In *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 103–110. Springer.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE.

Lv, H.-R., Lin, Z.-L., Yin, W.-J., and Dong, J. (2008). Emotion recognition based on pressure sensor keyboards. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 1089–1092. IEEE.

Ma, Y., Ye, Z., and Xu, M. (2016). THU-HCSI at MediaEval 2016: Emotional impact of movies task. In *Proceedings of the MediaEval 2016 Multimedia Benchmark Workshop*.

Magnenat-Thalmann, N. and Zhang, Z. (2014). Assistive social robots for people with special needs. In *Contemporary Computing and Informatics (IC3I), 2014 International Conference on*, pages 1374–1380. IEEE.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Marsella, S., Gratch, J., and Petta, P. (2010). Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, 11(1):21–46.

Marsella, S. C. and Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The eNTERFACE audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 8–8. IEEE.

Matthews, G., Jones, D. M., and Chamberlain, A. G. (1990). Refining the measurement of mood: The UWIST mood adjective checklist. *British journal of psychology*, 81(1):17–42.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). DISFA: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160.

McGettigan, C., Walsh, E., Jessop, R., Agnew, Z., Sauter, D., Warren, J., and Scott, S. (2015). Individual differences in laughter perception reveal roles for mentalizing and sensorimotor systems in the evaluation of emotional authenticity. *Cerebral cortex (New York, NY: 1991)*, 25(1):246–257.

McKeown, G., Valstar, M. F., Cowie, R., and Pantic, M. (2010). The SEMAINE corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084. IEEE.

Menke, J. and Martinez, T. R. (2004). Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 2, pages 1331–1335. IEEE.

Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *Affective Computing, IEEE Transactions on*, 3(2):184–198.

Monkaresi, H., Hussain, M. S., and Calvo, R. A. (2012). Classification of affects using head movement, skin color features and physiological signals. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2664–2669. IEEE.

Moore, J., Tian, L., and Lai, C. (2014). Word-level emotion recognition using high-level features. In *Computational Linguistics and Intelligent Text Processing*, pages 17–31. Springer.

Nazari, Z., Lucas, G., and Gratch, J. (2015). Multimodal approach for automatic recognition of machiavellianism. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 215–221. IEEE.

Neil, D., Pfeiffer, M., and Liu, S.-C. (2016). Phased LSTM: Accelerating recurrent network training for long or event-based sequences. In *Advances in Neural Information Processing Systems*, pages 3882–3890.

Nenkova, A. (2013). Multimodal affect prediction using audio, lexical and video indicators. Invited talk in the University of Edinburgh.

Niedenthal, P. M. and Brauer, M. (2012). Social functionality of human emotion. *Annual review of psychology*, 63:259–285.

Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., et al. (2013). Laugh-aware virtual agent and its impact on user amusement. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 619–626. International Foundation for Autonomous Agents and Multiagent Systems.

Novikova, J., Ren, G., and Watts, L. (2015). It's not the way you look, it's how you move: Validating a general scheme for robot affective behaviour. In *Human-Computer Interaction*, pages 239–258. Springer.

Ortony, A., Clore, G. L., and Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological review*, 97(3):315.

O'Shaughnessy, D. and Gabrea, M. (2000). Automatic identification of filled pauses in spontaneous speech. In *Electrical and Computer Engineering, 2000 Canadian Conference on*, volume 2, pages 620–624. IEEE.

Ozkan, D., Scherer, S., and Morency, L.-P. (2012). Step-wise emotion recognition using concatenated-HMM. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 477–484. ACM.

Palogiannidi, E., Iosif, E., Koutsakis, P., and Potamianos, A. (2015). Valence, arousal and dominance estimation for English, German, Greek, Portuguese and Spanish lexica using semantic models. In *Proceedings of Interspeech*, pages 1527–1531.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *BMVC*, volume 1, page 6.

Pecune, F., Mancini, M., Biancardi, B., Varni, G., Ding, Y., and Pelachaud, C. (2015). Laughing with a virtual agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1817–1818. International Foundation for Autonomous Agents and Multiagent Systems.

Pei, E., Yang, L., Jiang, D., and Sahli, H. (2015). Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 208–214. IEEE.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., and Booth, R. (2007). Operator's manual: Linguistic inquiry and word count. *LIWC2007. Austin, TX: LIWC*.

Petridis, S. and Pantic, M. (2008). Audiovisual discrimination between laughter and speech. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 5117–5120. IEEE.

Petrushin, V. A. (2000). Emotion recognition in speech signal: experimental study, development, and application. In *Sixth International Conference on Spoken Language Processing*.

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

Pini, S., Ben-Ahmed, O., Cornia, M., Baraldi, L., Cucchiara, R., and Huet, B. (2017). Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In *19th ACM International Conference on Multimodal Interaction*.

Plantinga, C. (2012). Art moods and human moods in narrative cinema. *New Literary History*, 43(3):455–475.

Pokorny, F. B., Graf, F., Pernkopf, F., and Schuller, B. W. (2015). Detection of negative emotions in speech signals using bags-of-audio-words. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 879–884. IEEE.

Popescu, A., Broekens, J., and van Someren, M. (2014). GAMYGDALA: An emotion engine for games. *Affective Computing, IEEE Transactions on*, 5(1):32–44.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 1:34.

Raghunathan, T. E., Rosenthal, R., and Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, 1(2):178.

Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., and Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmi, M., and Pantic, M. (2017). AVEC 2017–Real-life depression, and affect recognition workshop and challenge.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., and Setter, J. (1998). Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association*, 28:83–94.

Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69.

Rothgänger, H., Hauser, G., Cappellini, A. C., and Guidotti, A. (1998). Analysis of laughter and speech sounds in Italian and German students. *Naturwissenschaften*, 85(8):394–402.

Ruinskiy, D. and Lavner, Y. (2007). An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals. *IEEE transactions on audio, speech, and language processing*, 15(3):838–850.

Russell, J. A., Bachorowski, J.-A., and Fernández-Dols, J.-M. (2003). Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1):329–349.

Sagha, H., Coutinho, E., and Schuller, B. (2015). Exploring the importance of individual differences to the automatic estimation of emotions induced by music. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 57–63. ACM.

Salamin, H., Polychroniou, A., and Vinciarelli, A. (2013). Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 4282–4287. IEEE.

Sauter, D. A., Eisner, F., Calder, A. J., and Scott, S. K. (2010a). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly journal of experimental psychology*, 63(11):2251–2272.

Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010b). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6):2408–2412.

Savran, A., Cao, H., Shah, M., Nenkova, A., and Verma, R. (2012). Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 485–492. ACM.

Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Rückstieß, T., and Schmidhuber, J. (2010). PyBrain. *Journal of Machine Learning Research*.

Schröder, M. (2003). Experimental study of affect bursts. *Speech communication*, 40(1):99–116.

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Ter Maat, M., McKeown, G., Pammi, S., Pantic, M., et al. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183.

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315. Citeseer.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2011). The INTERSPEECH 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., 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, B., Valster, M., Eyben, F., Cowie, R., and Pantic, M. (2012). AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010a). Cross-corpus acoustic emotion recognition: variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B. W., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., Narayanan, S. S., et al. (2010b). The INTERSPEECH 2010 paralinguistic challenge. In *Interspeech*, volume 2010, pages 2795–2798.

Scott, S. (2013). Laughter-the ordinary and the extraordinary. *Psychologist*, 26(4):264–268.

Shawar, B. A. and Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics*, 10(4):489–516.

Shi, X. (2016). Automatic Detection of Disfluencies in Spoken Dialogues. Master's thesis, School of Informatics, the University of Edinburgh, United Kingdom.

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Ninth European Conference on Speech Communication and Technology*.

Sneddon, I., McRorie, M., McKeown, G., and Hanratty, J. (2012). The Belfast induced natural emotion database. *Affective Computing, IEEE Transactions on*, 3(1):32–41.

Snedecor, G. W. and Cochran, W. G. (1989). Statistical methods (8th edition). *Ames: Iowa State Univ. Press Iowa*.

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012a). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55.

Soleymani, M., Pantic, M., and Pun, T. (2012b). Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223.

Song, P., Ou, S., Zheng, W., Jin, Y., and Zhao, L. (2016a). Speech emotion recognition using transfer non-negative matrix factorization. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5180–5184. IEEE.

Song, Y., Dixon, S., Pearce, M. T., and Halpern, A. R. (2016b). Perceived and induced emotion responses to popular music. *Music Perception: An Interdisciplinary Journal*, 33(4):472–492.

Spelman, E. (1989). Anger and insubordination. *Ann Garry and Marilyn Pearsall. Boston: Unwin Hyman*, 72.

Stolar, M. N., Lech, M., and Allen, N. B. (2015). Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 987–991. IEEE.

Stouten, F. and Martens, J.-P. (2003). A feature-based filled pause detection system for Dutch. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 309–314. IEEE.

Strapparava, C. and Valitutti, A. (2004). WordNet Affect: an affective extension of WordNet. In *LREC*, volume 4, pages 1083–1086.

Stratou, G., Ghosh, A., Debevec, P., and Morency, L. (2011). Effect of illumination on automatic expression recognition: a novel 3D relightable facial database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 611–618. IEEE.

Sun, F., Harwath, D., and Glass, J. (2016). Look, listen, and decode: Multimodal speech recognition with images. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 573–578. IEEE.

Szameitat, D. P., Alter, K., Szameitat, A. J., Darwin, C. J., Wildgruber, D., Dietrich, S., and Sterr, A. (2009a). Differentiation of emotions in laughter at the behavioral level. *Emotion*, 9(3):397.

Szameitat, D. P., Alter, K., Szameitat, A. J., Wildgruber, D., Sterr, A., and Darwin, C. J. (2009b). Acoustic profiles of distinct emotional expressions in laughter. *The Journal of the Acoustical Society of America*, 126(1):354–366.

Szwoch, M. (2013). FEEDB: a multimodal database of facial expressions and emotions. In *Human System Interaction (HSI), 2013 The 6th International Conference on*, pages 524–531. IEEE.

Tan, E. S. (2013). *Emotion and the structure of narrative film: Film as an emotion machine*. Routledge.

Tan, E. S.-H. (1995). Film-induced affect as a witness emotion. *Poetics*, 23(1-2):7–32.

Tarvainen, J., Sjöberg, M., Westman, S., Laaksonen, J., and Oittinen, P. (2014). Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments. *IEEE Transactions on Multimedia*, 16(8):2085–2098.

Teigen, K. H. (2008). Is a sigh just a sigh? Sighs as emotional signals and responses to a difficult task. *Scandinavian journal of Psychology*, 49(1):49–57.

Tian, L. (2013). Emotion Recognition using Feature-Level Fused Audio-Visual Data. Master's thesis, School of Informatics, the University of Edinburgh, United Kingdom.

Tian, L., Lai, C., and Moore, J. (2015a). Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations. In *Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech*.

Tian, L., Moore, J., and Lai, C. (2016). Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. *Proceedings of Speech Language Technology 2016*.

Tian, L., Moore, J. D., and Lai, C. (2015b). Emotion recognition in spontaneous and acted dialogues. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 698–704. IEEE.

Tian, L., Muszynski, M., Lai, C., Moore, J. D., Kostoulas, T., Lombardo, P., Pun, T., and Chanel, G. (2017). Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same? In *Proceedings of 7th International Conference on Affective Computing and Intelligent Interaction (ACII'17)*, pages 28–35, San Antonio, Texas, USA. IEEE.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Zafeiriou, S., et al. (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)*, pages 5200–5204. IEEE.

Trouvain, J. (2014). Laughing, breathing clicking: The prosody of nonverbal vocalisations. *Proceedings of Speech Prosody (SP7), Dublin*, pages 598–602.

Trouvain, J. and Truong, K. P. (2012). Comparing non-verbal vocalisations in conversational speech corpora. *European Language Resources Association (ELRA)*.

Truong, K. P. and Van Leeuwen, D. A. (2007). Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of AVEC'16, co-located with ACM MM 2016*, Amsterdam, The Netherlands. ACM.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM.

van der Maaten, L. (2012). Audio-visual emotion challenge 2012: a simple approach. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 473–476. ACM.

Varadarajan, V., Hansen, J. H., and Ayako, I. (2006). UT-SCOPE – a corpus for speech under cognitive/physical task stress and emotion. In *Proc. of LREC Workshop en Corpora for Research on Emotion and Affect, Genoa*, pages 72–75.

Vidrascu, L. and Devillers, L. (2005). Detection of real-life emotions in call centers. In *INTERSPEECH2005*, pages 1841–1844.

Vielzeuf, V., Pateux, S., and Jurie, F. (2017). Temporal multimodal fusion for video emotion classification in the wild. *arXiv preprint arXiv:1709.07200*.

Vlasenko, B. and Wendemuth, A. (2015). Annotator's agreement and spontaneous emotion classification performance. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences*, 9(12):585–594.

Wang, C. (2016). Automatic Detection of Non-Verbal Vocalisations in Spoken Dialogues. Master's thesis, School of Informatics, the University of Edinburgh, United Kingdom.

Wang, J.-C., Wang, H.-M., and Lanckriet, G. (2015). A histogram density modeling approach to music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 698–702. IEEE.

Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., and Wang, X. (2010). A natural visible and infrared facial expression database for expression recognition and emotion inference. *Multimedia, IEEE Transactions on*, 12(7):682–691.

Warren, G., Schertler, E., and Bull, P. (2009). Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1):59–69.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207.

Wei, J., Pei, E., Jiang, D., Sahli, H., Xie, L., and Fu, Z. (2014). Multimodal continuous affect recognition based on LSTM and multiple kernel learning. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE.

Wiemer-Hastings, P., Graesser, A. C., Harter, D., Group, T. R., et al. (1998). The foundations and architecture of autotutor. In *International Conference on Intelligent Tutoring Systems*, pages 334–343. Springer.

Wildgruber, D., Szameitat, D. P., Ethofer, T., Brück, C., Alter, K., Grodd, W., and Kreifelts, B. (2013). Different types of laughter modulate connectivity within distinct parts of the laughter perception network. *PloS one*, 8(5):e63441.

Williams, C. E. and Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2013). LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163.

Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., and Narayanan, S. S. (2010). Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In *INTERSPEECH*, pages 2362–2365.

Wöllmer, M., Metallinou, A., Katsamanis, N., Schuller, B., and Narayanan, S. (2012). Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4157–4160. IEEE.

Wu, C.-H. and Liang, W.-B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21.

Wu, C.-H., Liang, W.-B., Cheng, K.-C., and Lin, J.-C. (2015). Hierarchical modeling of temporal course in emotional expression for speech emotion recognition. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 810–814. IEEE.

Xia, R. and Liu, Y. (2015). Leveraging valence and activation information via multi-task learning for categorical emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5301–5305. IEEE.

Xu, J., Broekens, J., Hindriks, K., and Neerincx, M. A. (2014). Effects of bodily mood expression of a robotic teacher on students. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 2614–2620. IEEE.

Xu, X., Deng, J., Zheng, W., Zhao, L., and Schuller, B. (2015). Dimensionality reduction for speech emotion features by multiscale kernels. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):25.

Yan, W.-J., Wu, Q., Liu, Y.-J., Wang, S.-J., and Fu, X. (2013). CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE.

Yang, N., Yuan, J., Zhou, Y., Demirkol, I., Duan, Z., Heinzelman, W., and Sturge-Apple, M. (2017). Enhanced multiclass svm with thresholding fusion for speech-based emotion classification. *International Journal of Speech Technology*, 20(1):27–41.

Yannakakis, G. N., Cowie, R., and Busso, C. (2017). The ordinal nature of emotions. In *Int. Conference on Affective Computing and Intelligent Interaction*.

Yu, D., Seide, F., and Li, G. (2012). Conversational speech transcription using context-dependent deep neural networks. In *ICML*.

Yu, H., Garrod, O., Jack, R., and Schyns, P. (2014). Realistic facial animation generation based on facial expression mapping. In *Fifth International Conference on Graphic and Image Processing*, pages 906903–906903. International Society for Optics and Photonics.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zayats, V., Ostendorf, M., and Hajishirzi, H. (2016). Disfluency detection using a bidirectional LSTM. *arXiv preprint arXiv:1604.03209*.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.

Zhang, J., Thalmann, N. M., and Zheng, J. (2016a). Combining memory and emotion with dialog on social companion: A review. In *Proceedings of the 29th International Conference on Computer Animation and Social Agents*, pages 1–9. ACM.

Zhang, R., Lou, X., and Wu, Q. (2015). Duration refinement for hybrid speech synthesis system using random forest. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 792–796. IEEE.

Zhang, Z., Ringeval, F., Dong, B., Coutinho, E., Marchi, E., et al. (2016b). Enhanced semi-supervised learning for multimodal emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5185–5189. IEEE.

Zhao, R., Papangelis, A., and Cassell, J. (2014). Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*, pages 514–527. Springer International Publishing.

Zheng, W., Yu, J., and Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 827–831. IEEE.