AFFECTIVE MODELING ON SPOKEN DIALOGUE

By
Arodami Chorianopoulou

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
TECHNICAL UNIVERSITY OF CRETE
CHANIA, GREECE
JULY 2016

TECHNICAL UNIVERSITY OF CRETE
DEPARTMENT OF
ELECTRONICS AND COMPUTER ENGINEERING

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "**Affective Modeling on Spoken Dialogue**" by **Arodami Chorianopoulou** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: July 2016

Supervisor: _____
Associate. Prof. Polychronis Koutsakis

Readers: _____
Associate Prof. Alexandros Potamianos

_____
Prof. Evripidis Petrakis

TECHNICAL UNIVERSITY OF CRETE

Date: **July 2016**

Author: **Arodami Chorianopoulou**

Title: **Affective Modeling on Spoken Dialogue**

Department: **Electronics and Computer Engineering**

Degree: **M.Sc.** Convocation: **July** Year: **2016**

Permission is herewith granted to Technical University of Crete to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

To my sister, Elisabeth

Table of Contents

List of Tables

List of Figures

Words have a magical power. They can bring either the greatest happiness or deepest despair;
they can transfer knowledge from teacher to student;
words enable the orator to sway his audience and dictate its decisions.
Words are capable of arousing the strongest emotions and prompting all men's actions.

Sigmunt Freud

Abstract

Emotions are fundamental for human-human communication, impacting people's perception, communication and decision-making. These are expressed through speech, facial expressions, gestures and other non-verbal cues. Speech is the main channel of human communication, interpreting emotional and semantic cues. Affective computing and specifically emotion recognition, is the process of decoding communication signals. It aims to improve the human-computer interaction (HCI) in a cognitive level allowing computers to adapt to the users needs. Hence, speech emotion recognition suggests that vocal parameters reflect the affective state of a person. This assumption is supported by the fact that most affective states involve physiological reactions which in turn modify the process by which voice is produced. There are a number of potential applications for speech emotion recognition, including anger detection for Spoken Dialogue Systems (SDS) and emotional aids for people with autism.

Attention is a concept studied in cognitive psychology that refers to how a person actively processes information. Salience is the level to which something in the environment can catch and retain one's attention. While research on affective speech saliency is not extensive, salient information from audio and video has been investigated. It is argued that modeling the affective variation of speech can be approached by integrating acoustic parameters from various prosodic timescales, summarizing information from more localized (e.g. syllable-level) to more global prosodic phenomena (e.g. utterance-level).

In this thesis, speech prosody and related acoustic features, e.g., spectral and voice quality, are investigated for the task of emotion recognition. Features derived from the Amplitude and Frequency Modulation (AM-FM) model are also examined. Moreover, the contribution of different information levels is also addressed for the task of emotion recognition. Additionally, we investigate the affective salient information over time on spoken dialogue utterances using prosodic variations from different timescales of the speech signal, by weighting speech segments. The proposed models are evaluated on datasets of spontaneous speech.

For a human social and mental states are highly correlated. As a result affective speech is introduced on several areas of the computational community. For instance, people with Autism Spectrum Disorder (ASD) suffer from symptoms of anxiety and depression that significantly compromise their quality of life. Additionally, language in high-functioning autism is characterized by pragmatic and semantic deficits, and people with autism have a reduced tendency to integrate information. Motivated by these findings, we investigate the degree of engagament for children with ASD in interactions with their parents.

Περίληψη

Τα συναισθήματα είναι βασικά χαρακτηριστικά στην επικοινωνία μεταξύ ανθρώπων, επηρεάζοντας την αντίληψη, την επικοινωνία και την λήψη αποφάσεων. Όλα τα παραπάνω εκφράζονται μέσω της ομιλίας, των εκφράσεων προσώπου, των χειρονομιών και άλλων μη λεκτικών ενδείξεων. Η ομιλία είναι το βασικότερο μέσω επικοινωνίας μεταξύ των ανθρώπων, ερμηνεύοντας συναισθηματικές και γνωσιακές ενδείξεις. Η υπολογιστική αναγνώριση συναισθήματος είναι η διαδικασία με την οποία αποκωδικοποιούνται τέτοια σήματα επικοινωνίας. Σκοπός είναι να βελτιώσει την επικοινωνία μεταξύ ανθρώπου και υπολογιστή σε επίπεδο αντίληψης, επιτρέποντας στον υπολογιστεί να προσαρμοστεί στις ανάγκες ενός χρήστη. Ως εκ τούτου, η αναγνώριση συναισθήματος μέσω φωνής υποθέτει οτι φωνητικές παράμετροι κατοπτρίζουν την συναισθηματική κατάσταση ενός ανθρώπου. Αυτή η υπόθεση υποστηρίζεται και από το γεγονός οτι οι συναισθηματικές καταστάσεις εμπλέκουν ψυχολογικές αντιδράσεις, οι οποίες με τη σειρά τους αλλάζουν τη διαδικασία παραγωγής της φωνής. Υπάρχει ένα μεγάλο εύρος εφαρμογών για την αναγνώριση συναισθηματος από φωνή, συμπεριλαμβάνοντας την αναγνώριση θυμού για διαλογικά συστήματα και τη συναισθηματική υποστήριξη/βοήθεια για άτομα με αυτισμό.

Η προσοχή είναι μια έννοια που μελετάται στον κλάδο της γνωστικής ψυχολογίας και αναφέρεται στο πως ένας άνθρωπος ενεργά επεξεργάζεται την πληροφορία. Η σημαντικότητα είναι το επίπεδο στο οποίο κάτι από το περιβάλλον μπορεί να τραβήξει και να διατηρήσει την προσοχή ενός ανθρώπου. Ενώ ερευνητικά η συναισθηματική σημαντικότητα βάσει της φωνής δεν είναι εκτενής, η σημαντικότητα βάσει ήχου και εικόνας έχει ερευνηθεί. Υποστηρίζεται οτι η μοντελοποίηση της συναισθηματικής μεταβολής από τη φωνή μπορεί να προσεγγιθεί μέσω της ενσωμάτωσης ακουστικών παραμέτρων από διάφορα χρονικά πλαίσια της προσωδίας, συνοψίζοντας την πληροφορία από τοπικά (π.χ., συλλαβές) μέχρι πιο καθολικά φαινόμενα (π.χ. φράσεις).

Σε αυτήν την εργασία, η προσωδία καθώς και άλλα ακουστικά χαρακτηριστικά, όπως χαρακτηριστικά του φάσματος και της ποιότητας της φωνής, ερευνόνται για την αναγνώριση συναισθήματος. Χαρακτηριστικά τα οποία προέρχονται από το Amplitude and Frequency Modulation (AM-FM) μοντέλο επίσης εξετάζονται. Ακόμα, απευθύνεται στη συμμετοχή διαφορετικών επιπέδων πληροφορίας για την αναγνώριση συναισθήματος. Επιπλέον, μελετήσαμε τη συναισθηματική σημαντικότητα της πληροφορίας στο χρόνο σε διαλογικές φράσεις χρησιμοποιώντας προσωδιακές μεταβολές από διαφορετικά χρονικά πλαίσια του σήματος φωνής, ζυγίζοντας τα συγκεκριμένα πλαίσια. Τα προτενόμενα μοντέλα έχουν εκτιμηθεί σε σύνολα δεδομένων με αυθόρμητη ομιλία.

Η κοινωνική και διανοητική κατάσταση ενός ανθρώπου είναι άμεσα συνδεδεμένα.

Σαν αποτέλεσμα ο συναισθηματικός λόγος έχει εισαχθεί σε πολλές περιοχές της υπολογιστικής κοινώτητας. Για παράδειγμα, άνθρωποι με αυτισμό υποφέρουν από συμπτώματα άγχους και κατάθλιψης που διακυνδινεύουν αρκετά την καθημερινή ζωή τους. Επιπλέον, η γλώσσα σε υψηλά επίπεδα αυτισμού χαρακτηρίζεται από πραγματιστικές και σημασιολογικές διαταραχές και άνθρωποι με αυτισμό έχουν μειωμένη ικανότητα να αφομοιώσουν πληροφορίες. Έχοντας ως κίνητρο τα παραπάνω ευρήματα, ερευνήσαμε το επίπεδο της συμμετοχής παιδιών με αυτισμό σε αλληλεπιδράσεις με τους γονείς τους.

Acknowledgements

First and foremost, I would like to express my special appreciation and thanks to my advisor Prof. Alexandros Potamianos, for being a great mentor to me, for encouraging my research and for his valiable guidance and advice. I would also like to thank my committee members, Prof. Polychronis Koutsakis, for his support and help during those years and Prof. Euripidis Petrakis for serving as my committee.

I would also wish to thank Dr. Elias Iosif, who offered invaluable assistance and support. Another thank you is dedicated to the members and staff of the CVSP laboratory of the National Technical University of Athens.

Finally, I must express my very profound gratitude to my parents, my sister and to my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Chapter 1

Introduction

1.1 Affective Speech & Emotion Experience

Speech is one of the most natural communication forms between human beings, who also express their social and mental state via written language. Emotional speech implies that changes in the automatic nervous system indirectly alter speech. For instance, anger influences the vocal folds vibrations and the vocal track's shape. Subsequently, affects the acoustic characteristics of speech.

In psychology and philosophy, emotion is a subjective, conscious experience characterized primarily by psychophysiological expressions, biological reactions, and mental states. Moreover, it can be differentiated from a number of similar constructs. *Feelings* are subjective representations of emotions, private to the individual experiencing them. *Moods* are affective states that last for much longer durations than emotions and are usually less intense. *Affect* is a term used to describe the topics of emotion, feelings, and moods together, even though it is commonly used interchangeably with emotion.

The experience of emotion is a neurobiological process that emerges by psychological events. Emotions are determined by one of the oldest parts of our brain, the limbic system, including the amygdala, the hypothalamus, and the thalamus. Because they are primarily defined, the basic emotions are experienced almost the same way across cultures. It is argued that emotional stimulus produces changes in heart rate, respiration or sweating [17], while several psychological models relate emotion with behavior [61] and brain activity [18].

1.2 Affective Computing

Affective computing is the study of analyzing, recognizing and processing affective states. It combines engineering and computer science with psychology, cognitive science, neuroscience, ethics, and more. Affective systems use cues that humans use to perceive emotions, i.e., speech, written language, facial expressions, body posture, and gestures. Enabling systems to interpret speech for a more intuitive human machine

interaction suggests also understanding the transmitted affective and social aspects. Research by psychologists and neuroscientists has shown that emotion is highly related to decision-making. In Spoken Dialogue Systems (SDS) the analysis of speakers' emotion [20, 40, 56], age, gender [79] or personality [81] can significantly improve dialogue management strategies and improve the user experience.

1.2.1 Speech Emotion Recognition

The goal of automatic emotion recognition from speech is to recognize the speaker's emotional state from his voice. In order to employ a robust emotion detector, features from the speaker's speech signal able of describing the emotional content but independent of the speaker or the lexical content, have to be extracted. Research has shown that emotional reactions are strongly related to the pitch and energy of the speech. Speech produced in a state of fear, anger or joy becomes faster, louder, precisely expressed with a higher and wider pitch range. Other emotions such as tiredness, boredom or sadness, lead to slower, lower-pitched speech. Emotions are classified either on a discrete or a dimensional space. In the former approach, example emotions are happiness, sadness, anger, happiness and fear, while on the latter, arousal, valene and dominance dimensional spaces are the most significant. Specifically for spoken dialogue system applications, where speech is natural, additional features have been introduced describing speaker and dialogue characteristics, i.e., speaker gender, dialogue duration and the existence of speech overlap [40, 63].

1.3 Contributions

This thesis is focused on the task of speech emotion recognition on spoken dialogue, i.e., the analysis and development of affective models using the speaker's speech signal. We aim to recognize emotional states by analyzing the emotional content of sub-utterances. Hence, we weight the respective speech regions according to their emotional content and fuse the information over time in order to extract an utterance-level emotion decision.

Specifically, in order to recognize the amount of emotional information of spoken dialogue utterances, an affective salience model is proposed. It utilizes a regression model that combines features extracted from the acoustic signal and the posteriors of a segment-level classifier to obtain frame or segment-level ratings. The affective saliency model is trained using a minimum classification error (MCE) criterion that learns the weights by optimizing an objective loss function related to the classification error rate of the emotion recognition system. An information fusion model is also proposed, using the affective saliency scores to emphasize emotional segments over time. The fusion is employed either to weight the contribution of frame-level posteriors (late fusion) or features (early fusion) to the speech emotion classification decision. The models is evaluated for the task of anger detection on four call-center datasets of two languages, Greek and English.

Finally, a framework for engagement detection for typically developed (TD) and with Autism Spectrum Disorder (ASD) children is presented. Children with Autism Spectrum Disorder (ASD) face several difficulties in social communication. Hence,

analyzing social interaction can provide insight on their social and cognitive skills. Motivated by the assumption that ones degree of engagement is influenced in interactions with others, we used 66 videotaped sessions of children interacting with their parents in Greek. Features derived from both participants including acoustic, linguistic and dialogue act features are explored. The effect of visual cues is also investigated for engagement detection.

1.3.1 Publications

- Arodami Chorianopoulou, Polychronis Koutsakis and Alexandros Potamianos, "Emotion Recognition using Affective Saliency", in *INTERSPEECH*, 2016.
- Arodami Chorianopoulou, Efthymios Tzinis, Elias Iosif, Asimenia Papoulidi, Christina Papailiou and Alexandros Potamianos, "Engagement Detection on Children with Autism Spectrum Disorder", *ICASSP*, 2017.
- Jose Lopes, Arodami Chorianopoulou, Elisavet Palogiannidi, Helena Moniz, Alberto Abad, Katerina Louka, Elias Iosif and Alexandros Potamianos, "The SpeDial Datasets: Datasets for Spoken Dialogue Systems Analytics", 10th International Conference on Language Resources and Evaluation (LREC), 2016.
- Spiros Georgiladakis, Georgia Athanasopoulou, Raveesh Meena, Jose Lopes, Arodami Chorianopoulou, Elisavet Palogiannidi, Elias Iosif, Gabriel Skantze and Alexandros Potamianos, "Root-Cause Analysis of Miscommunication Hotspots in Spoken Dialogue Systems", in *INTERSPEECH*, 2016.

1.3.2 Thesis organization

This thesis is organized as follows: Chapter 2 presents the main ideas of speech production and perception, while the main models of the emotion representation are also introduced. In Chapter 3 the related research of the affective speech analysis is presented. Chapter 4 presents the affective salinecy model, while in Chapter 5 the fusion scenarios are analysed. In Chapter 6 we propose a framework for engagement detection, on children with autism while in Chapter 7 conclusions and future work are provided.

Chapter 2

Speech & Emotion Analysis

2.1 Introduction

This chapter introduces the main ideas of speech production and perception as speech is the most significant form of communication between individuals. Hence, speech carries information about the social and cognitive state of each speaker. Affective analysis can provide us with information about the social and mental state of a person while the relation of theoritical analysis with the computational community is also provided. The problem of sentiment analysis is the definition and formulation of the perceived emotions. However, the subjectivity of the perceived emotions are a significant issue throughout the research community, either psychologically or computationally. Hence, emotions are analyzed and investigated based on languages, speakers or even applications. Many theries of how emotions are perceived have been introduced over the years. Well established theories suggest that emotions are either discrete or lie on dimensional spaces. Both approaches are investigated using speech for social and mental evaluation tasks.

2.2 Physics of sound

Sound is a mechanical wave that is an oscillation of pressure transmitted through some medium (like air or water), composed of frequencies which are within the range of hearing. Sound that is perceptible by humans has frequencies from about 20 Hz to 20.000 Hz, although these limits are not definite. The speed of sound depends on the medium the waves pass through, and is a fundamental property of the material. In general, the speed of sound is proportional to the square root of the ratio of the elastic modulus (stiffness) of the medium to its density. Those physical properties and the speed of sound change with ambient conditions.

2.3 Speech Perception & Production

Speech is the vocalized form of human communication. It is based upon the syntactic combination of lexicals and names that are drawn from very large vocabularies. Each

spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units.

Speech perception refers to the processes by which humans are able to interpret and understand the sounds used in language [96]. Speakers show phonetic differences while producing the very same utterance, which occur at various linguistic levels and can be interpreted phonetically by several parameters such as voice quality, speech rate, loudness, fundamental frequency, breathing, articulatory behavior, etc. At the same time, listeners can vary in the way they exploit such cues for the purpose of speech perception and understanding. Figure 2.1 shows a human brain and the regions that are activated according to affective cues. Specifically, the amygdala are considered part of the limbic system and perform a primary role in the process of decision-making, and emotional reactions. When exreriencing fear, sensory stimuli reach the basolateral complexes of the amygdala, where they form associations with memories of the stimuli. There have been studies that show that damage to the amygdala can interfere with memory that is strengthened by emotion.

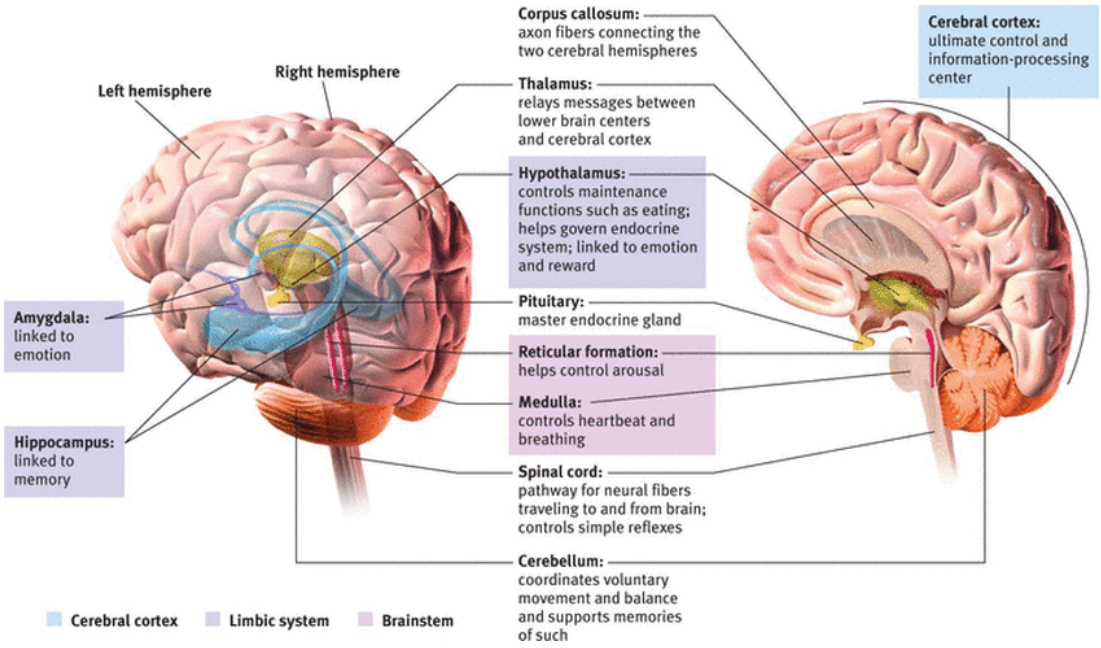

Figure 2.1: Human brain.

Speech production is the process by which thoughts are translated into speech. This process includes the selection of words, grammatical forms, and then the resulting sounds produced using the vocal mechanism. Speech is created by pressure provided by the lungs through the glottis in the larynx, which is then modified by the vocal tract into different vowels and consonants. The manner of articulation is the configuration and interaction of the articulators, i.e., speech organs such as the tongue, lips, and palate as shown in Figure 2.2, when making a speech sound. The physical structure of the articulators allows the production of many unique sounds, as sounds are produced in different areas, and with different muscles and breathing techniques. Difficulties in manner of articulation can contribute to speech difficulties and impediments. It is

suggested that infants are capable of making the entire spectrum of possible vowel and consonant sounds.

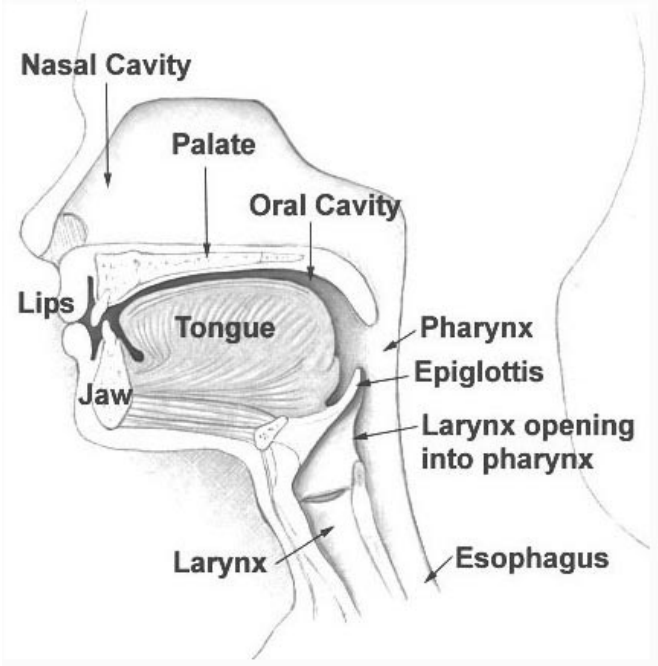

Figure 2.2: Human vocal tract.

2.4 Emotion Representation

The representations of emotions has mainly been researched from two viewpoints. The first is that emotions are discrete and fundamentally different constructs while the second asserts that emotions can be characterized on a dimensional basis in groupings. According to Paul Ekman [32], emotions are discrete, measurable, and physiologically distinct, while certain emotions appear to be universally recognized. His research findings led him to classify six basic emotions, namely anger, disgust, fear, happiness, sadness and surprise.

On the other hand, the multi dimensional analysis divides emotions into three dimensions known as valence, i.e., how negative or positive the experience was, arousal or activation, i.e., the extent of reaction to stimuli, and dominance, i.e., the disposition of an individual to assert control. The first two dimensions, arousal and valence, can be depicted on a 2D coordinate map. Robert Plutchik offers a three-dimensional model that is a hybrid of both basic-complex categories and dimensional theories [91], as shown in Figure 2.4. It arranges emotions in concentric circles where inner circles are more basic and outer circles more complex. Notably, outer circles are also formed by blending the inner circle emotions.

If discrete categorical labels are used, the emotional classs needs to be labeled. In general, there is a tradeoff between inter-evaluator agreement and description accuracy.

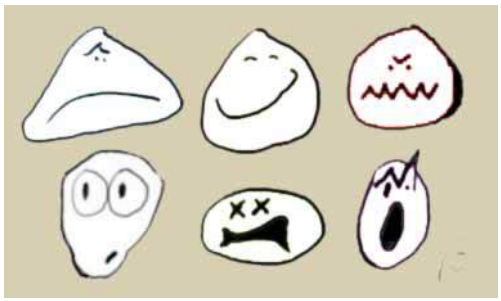

Figure 2.3: Basic emotions.

Figure 2.4: Plutchik's model

If the number of emotion categories is too extensive, the agreement between evaluators will be low. A simpler idea is to label the emotions as negative or non-negative. This approach makes the problem more concrete and specific. Moreover, the binary problem may refer to neutral and emotional classes. The underlying assumption is that expressive speech will differ from neutral speech in the feature space. Figure 2.5 demonstrates the activation-valence space and suggests that only emotion with high activation can be discriminated from neutral speech using only the fundamental frequency. Figure 2.7 presents the three-dimenional space, i.e., activation-valence-dominance.

Figure 2.5: Activation-valence space.

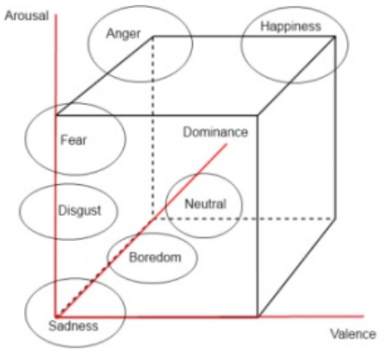

Figure 2.6: Activation-valence-dominance space.

Figure 2.7: Location of the emotional categories.

Another way is to determine the encoded emotion by the listener's reaction or answer. However, there is an explicit distinction between the encoding (speaker), the transmission and the representation (listener) of the emotion. These three distinctions in the models exist because expression and perception are two distinct and complex problems. The intended emotion encoded by the speaker may not necessarily match with the perceived emotion.

2.5 Affect & Engagement

Successful affective processing involves the determination of positive or negative valence and activation and the generation of affective experience. These spontaneous processes help guide behavior, especially in online interactions, such as social exchanges [16]. Additionally, the past years an interest in the role of emotions in academic settings has been grown, especially regarding the students' engagement and learning. In [70] the valence and activation of students was correlatted to their social behavior and engagement in group meetings and it was suggested that negative valence was highly correlated with a more lazy mood. In [87] students were asked to fill a questionare in order to analyze and evaluate their emotions. The Achievement Emotions Questionnaire (AEQ) contains several scales for representing a large number of emotions, such as hope, relief, anger, and anxiety, for students while studying or taking exams. The findings indicated that the analyzed emotions related to students' learning and performance.

2.5.1 Autism Spectrum Disorder (ASD)

Autism spectrum disorder (ASD) is a general term characterizing a group of complex disorders of brain development. These disorders are expressed in varying degrees, such as difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors. Autism could be counted a disorder of affective and social relations. One theory proposes that the social and communication deficits in autism are primarily affective [8]. More specifically, the affective theory states that in autism there is an innate inability to emotionally interact with other people. This theory was originally proposed by [54]. Figure 2.8 presents the position of the Affective Theory. However, autism is not an emotional response to trauma [9].

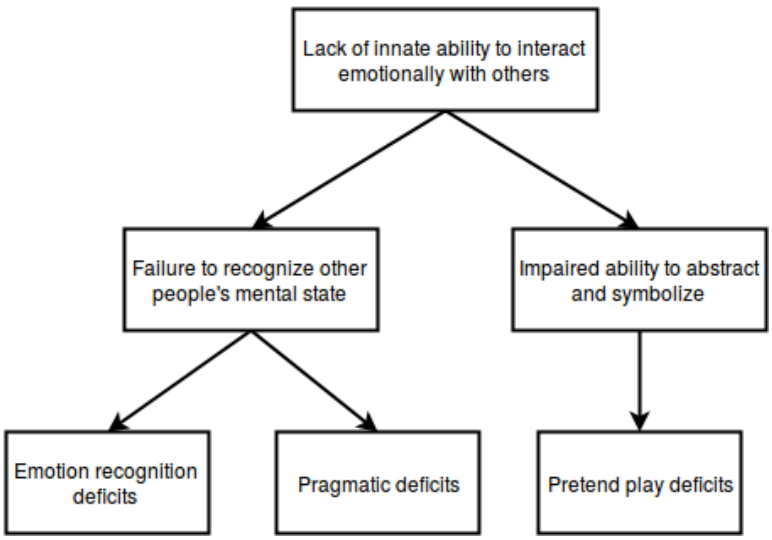

Figure 2.8: Position of Affective Theory in ASD.

Research reveals that ASD children have a mosaic of social and emotional skills

and show atypicalities in several areas of social life [113]. These atypicalities include the responsiveness to social signals, processing of faces, and the generation of negative emotional expressions [27]. However, individuals with ASD express and comprehend a full range of emotions [16]. Additionally, some ASD deficits in affective processing may be better attributed to cognitive or language factors [29].

Children with autism are continuously disadvantaged to access to the next stage of social competence and independent functioning, while most aspects of life depend on competence in interpersonal interaction [29]. As an unfortunate consequence of these impairments, children with autism spectrum disorders often exhibit limited autonomy and most remain reliant on caregivers as they enter adulthood.

2.6 Summary

In this chapter, we introduces the concepts of sound and speech production and their relation to the affective states and emotions. Moreover, we introduced the different senses of affect, i.e., moods and feelings, and the main models of describing emotions. Finally, we presented how affect can be correlated with mental and social deficits, such as Autism Spectrum Disorder.

Chapter 3

Affective Analysis

3.1 Introduction

Identifying speech features suitable to describe affective information is challenging. The standard approach in emotion recognition systems is to extract prosodic features, particularly pitch and energy [112, 62, 98]. In [69] Mel-Frequency Cepstral coefficients (MFCCs) have been used for training acoustic and phonetic tokens. However, randomly localized frequency perturbations have been found to influence spectral features. Hence, the amplitude and frequency modulation model (AM-FM) has been introduced in many tasks [3, 116], as it significantly affects speech recognition and perception. Features based on either part of the modulation model (AM or FM) are capable of providing acoustic information not captured by the linear-filter model. Compared to the MFCCs, AM-FM features model the structure of speech better due to the signal's decomposition. In [71] contextual features were proposed for spoken dialogue systems, including prosodic and discourse context.

Several machine learning techniques have been also explored for affective modeling. Support Vector Machines (SVM) [65], Hidden Markov Models (HMMs) [83], and Gaussian Mixture Models (GMMs) [21] are proposed for speech emotion recognition. In [59] the emotion recognition performance was compared using SVM, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) classifiers, while segment level approaches are also introduced to model the emotional aspects of the speech signal in [99]. In other paralinguistic tasks, e.g., cognitive load estimation, i-vectors have also been investigated [110].

In this Chapter we introduce the linear and non-linear models for speech production. We analyze the assumptions supporting both models and present several affective descriptors (see Section 3.3.1) based on both models.

3.2 Speech Production Models

3.2.1 Linear Model

According to the speech production theory, sound is the result of the air being push from the lungs through the glottis into the vocal tract. The vocal tract can be viewed

as a model that includes vocal cavities, which produce an oscillatory response of its characteristic frequency, *resonance frequency*. The tract amplifies the characteristic frequencies of the vocal cavities, hence the Fourier transform of the source waveform has dominant spectral peeks, *formants*, that correspond to the resonance frequencies [92].

The linear model

To model speech production the non-linear equations of accoustics are simplified to a one-dimensional linear form using the following assumptions [90, 93]:

- The air flow fills up the whole vocal tract uniformly.
- The air flow velocity is smaller than the sound velocity.
- The fluid cannot support shear stresses.

The *linear-filter model* assumes that the vocal tract consists of 4-5 cavities, each one modeled as a second-order linear filter. The impulse response of the filter is

$$r(n) = A\sigma^n \cos(2\pi(\frac{F}{F_s})n + \theta), n > 0 \tag{3.2.1}$$

where F and F_s is the resonance frequency of the cavity and the sampling frequency respectively, and $\sigma \in [0, 1]$ controls the energy dissipation rate.

According to the linear theory, the vocal tract can be thought of as a total of N linear resonances/filters with impulse response $w(n)$, given by:

$$w(n) = r_1(n) * r_2(n) \dots r_N(n) \tag{3.2.2}$$

where N is the number of speech formants. Additionally, the source-filter model assumes that the glottis and the vocal tract are not coupled, thus the source speech waveform is

$$s(n) = u(n) * w(n) \tag{3.2.3}$$

Linear Prediction Analysis

Although speech is a non-stationary process, features such as pitch period vary slow with time. Thus the parameter values of a linear model are assumed as constant values over a short frame of time. A mathematical framework for speech analysis over short-time windows is Linear Prediction. In linear prediction each sample is modeled as a linear combination of the previous samples:

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) \tag{3.2.4}$$

where $a_i, i = 1, 2, \ldots, p$ are the linear prediction coefficients (LPCs), which minimize the error:

$$E = \sum_{n=0}^{N} [s(n) - \sum_{i=1}^{p} a_i s(n-1)]^2 \tag{3.2.5}$$

where M is the number of samples in the short-time frame and p is the order of the linear predictor.

3.2.2 Non-linear Model

Strong evidence of a non-linear model are presented, in dissagreement with the linear speech production [102, 103]. The main principles are:

1. the air flow velocity in the vocal tract is not uniform but rather unstable.
2. the vortices modulate the air flow, amplifying certain frequencies and attenuating others [109].
3. the air flow is unstable during phonetic transitions, i.e. when the vocal tract's shape changes [107].

The interaction of the source with the vocal tract involves a *frequency modulation (FM)* component, meaning that the formant frequency is different during the different phases of the glottis. The FM is a result of the variations on the oscillation parameter. Instantaneous changes on the oscillation parameter can cause changes in the rate of decay of the amplitude envelope, *amplitude modulation (AM)*. Additionally, this interaction can effect the formant energy levels.

AM-FM Speech Modulation Model

The AM-FM speech modulation model presents each resonance signal as a combination of amplitude and frequency modulation [75, 77].

$$r(t) = a(t)\cos\underbrace{(2\pi(f_c t + f_m \int_0^t q(z)dz) + \phi(0))}_{\phi(t)} \tag{3.2.6}$$

where f_c is the formant frequency, $a(t)$ the time-varying amplitude signal, $q(t) \in [-1, 1]$ the frequency modulation signal, $f(t) = \frac{1}{2\pi}\phi(t) = f_c + f_m q(t)$ the instantaneous frequency defined as the normalized derivative of the phase $\phi(t)$ and f_m the maximum deviation of the instantaneous frequency from the formant frequency f_c, $(0 < f_m < f_c)$. The speech signal can now be represented as the sum of N resonance signals.

$$s(t) = \sum_{k=1}^{N} r_k(t) \tag{3.2.7}$$

where N is the number of formants. The AM-FM modulation model can describe the non-linear phenomena by decomposing each formant into amplitude envelope and instantaneous frequency signals.

Energy Separation Algorithm (ESA)

The Teager Energy Operator (TEO or Ψ operator) [51, 52] for a continuous time signal is:

$$\Psi_c[x(t)] = \dot{x}(t)^2 - x(t)\ddot{x}(t) \tag{3.2.8}$$

In discrete-time the energy operator is

$$\Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{3.2.9}$$

which is derived form the continuous-time operator by approximating derivatives with forward or backward operators.

The Energy Separation algorithm when applied to the AM-FM signal [76]

$$x(t) = a(t)\cos\underbrace{(2\pi(f_c t + f_m \int_0^t q(z)dz) + \phi(0))}_{\phi(t)} \tag{3.2.10}$$

can approximately estimate the squared product of amplitude $|a(t)|$ and frequency $f(t)$ signals, i.e. $\Psi_c[x(t)] \approx [a(t)f(t)]$, where $f(t) = \frac{d\phi}{dt}(t) = f_c + f_m q(t)$. The energy operator tracts the energy (per unit mass) of a linear oscillator $x(t) = A\cos(\omega_c(t) + \theta)$, when applied to it: $\Psi_c[A\cos(\omega_c(t) + \theta)] = (A\omega_c)^2$.

The ESA algorithm estimates the instantaneous frequency and amplitude of an AM-FM signal $s(t)$ as:

$$f(t) \approx \frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{s}(t)]}{\Psi[s(t)]}} \tag{3.2.11}$$

$$|a(t)| \approx \frac{\Psi[s(t)]}{\sqrt{\Psi[\dot{s}(t)]}} \tag{3.2.12}$$

Similar demodulation methods an be applied to discrete-time signals.

Multiband Demodulation Analysis (MDA)

Usually, the filters used in conjuction with TEO are Gabor filters. The impulse response and frequency response of a Gabor filter are:

$$h(t) = e^{(-\alpha^2 t^2)}\cos(2\pi v t) \tag{3.2.13}$$

$$H(f) = \frac{\sqrt{\pi}}{2\alpha}e^{-\frac{\pi^2(f-v)^2}{\alpha^2}} \tag{3.2.14}$$

where v is the central frequency of the filter chosen equal to the formant frequency and α is a bandwidth parameter. When the Ψ operator and the Gabor filtering are combined we get:

$$\Psi[x(t) * h(t)] = [x(t) * \frac{dh(t)}{dt}]^2 - x(t) * h(t)[x(t) * \frac{d^2h(t)}{dt^2}] \tag{3.2.15}$$

This process, called "Gabor ESA" is faster than the simple ESA and provides smoother instantaneous frequency times. An expression fot discrete-time signal would be

$$\Psi[x(n)] = \frac{x^2(n) - x(n-1)x(n+1)}{T^2} \tag{3.2.16}$$

where T is the sampling period of the signal $x(n)$.

3.3 Affective Descriptors

3.3.1 Speech Descriptors

Going over literature, a wide range of affective descriptors has been investigated. Table 3.1 summarizes the types and respective features.

Description	Features
Prosodic	Pitch, Energy, Duration, Zero crossing-rate
Short-term spectrum	MFCC, Spectral bands, Formant
Voice quality	Jitter, Shimmer, HNR, NHR

Table 3.1: List of affective descriptors.

The above list includes prosodic, spectral and voice quality features. Pitch is the relative highness or lowness of a tone and depends on the number of vibrations produced by the vocal cords. Algorithms for extracting the pitch frequency are mostly the autocorrelation function and methods using the wavelet transform. Additionally, speech energy can be exploited for emotion recognition because it is related to the arousal level of emotions [4, 98].

Vocal tract features suggest that the shape of the vocal tract is modified by the emotional state. Features that have been used to describe the shape of the vocal tract include formants, and coefficients derived from frequency transformations. One method to estimate the formants relies on the Linear Prediction Analysis which identifies the formants using coefficients (LPCs). Another widely used feature, is the energy of certain frequency bands. However, the Mel-frequency cepstral coefficients (MFCCs) provide a better representation of the signal than the frequency bands since they exploit the human auditory frequency response [112].

Voice quality features, such as jitter and shimmer measurement, depend on the knowledge of the length of the cycles of the speech waveform. Jitter (absolute) is the cycle-to-cycle variation of fundamental frequency, i.e. the average absolute difference between consecutive periods, while Shimmer (dB) is expressed as the variability of the peak-to-peak amplitude in decibels [38]. The contour of short-term acoustic features is affected by emotion states of anger, disgust, fear, joy and sadness. It is also a valuable feature for emotion recognition because they describe the temporal characteristics of an emotion.

Many studies focuses on a set of the basic emotions, namely joy, anger, fear, sadness, disgust and neutral. The behavior of five of the above emotional states is outlined in Table 3.2. However, many distinctions occur depending on the speaker's gender. The two vocal properties, i.e., intensity and pitch, differ on women and men. Pitch of a man's voice fall under low frequency, in contrary with woman's voice that has high pitch. Comparing the intensities of male vs female voice, female's voice has more frequency components compared to men, while women speak at one octave higher, too.

Anger is the emotion of the highest energy and pitch level, while disgust is expressed with a low mean pitch level, a low intensity level and a slower pitch rate than the neutral state does. Fear is correlated with high pitch level and raised intensity level. When sadness is expressed low levels of intensity and pitch are observed, while

	Pitch		Intensity		Timing	
	Mean	Range	Mean	Range	Speech rate	Duration
Anger	>>	>	>	>	<	<
Disgust	<	>	<		<<	
Fear	>>	>	>			<
Joy	>	>	>	>		<
Sadness	<	<	<	<	>	>

Table 3.2: Summary of the effects of several emotion states on selected acoustic features. >: increases, <: decreases. Double symbols indicate a change of increased predicted strength

the speech rate is generally slower than in neutral states. Joy shows similar outcomes as fear, however they differ on the pitch contour [111].

The AM-FM model suggests that the formant frequencies are not constant during a single pitch period, but they can vary around a center frequency. A wide range of features based on the AM-FM model have been introduced by the research community. In [30], the *Instantaneous Frequency Mean (IF-mean)*, *Mean Instantaneous Amplitude (IA-mean)* and *Frequency Modulation Percentages (FMPs)* are proposed. The short-time weighted mean of the instantaneous frequency signal $f_i(t)$, i.e., the Instantaneous Frequency Mean (IF-Mean), provides information about the speech formant structure taking advantage of the time resolution of the ESA. Transitional phenomena and instantaneous formant variations are mapped into those FM features. The Mean Instantaneous Amplitude (IA-Mean) features that are defined as the short-time mean of the instantaneous amplitude signal $|a_i(t)|$ for each speech resonance i. The IA-Mean features parametrize the resonance amplitudes and capture part of the nonlinear behavior of speech, e.g., the modulation pulses appearing within a single pitch period.

Frequency Modulation Percentages (FMPs) can partially capture the fluctuation of frequencies during a single pitch period and they are defined as:

$$FMP_i = \frac{B_i}{F_i} \tag{3.3.1}$$

where

$$F_i = \frac{\int_0^T f_i(t) a_i^2(t) dt}{\int_0^T a_i^2(t) dt} \tag{3.3.2}$$

$$B_i = \frac{\int_0^T [\dot{a}_i^2(t) + (f_i(t) - F_i)^2 a_i^2(t)] dt}{\int_0^T a_i^2(t) dt} \tag{3.3.3}$$

$i = 1, \ldots, n$ is the formant index and T is the time window length. F_i and B_i are called weighted mean frequency value and mean bandwidth of the formant i. So, before we can calculate the FMP_i for all i, we need to calculate the F_i and B_i for all i. And to do so, we need $a_i(t)$ and $f_i(t)$ for all i.

In [116] the *Variation of FM Component (TEO-FM-Var)* is introduced. The motivation for the TEO-FM-Var feature is to capture stress dependent information that may be present in changes within the FM component. Its processing is based on

the entire band although the final FM variations are computed around the restricted frequency band. The *Amplitude modulation cepstral coefficients (AMCCs)* [3] uses a smoothed nonlinear energy operator (SNEO) for amplitude modulation cepstral coefficients (AMCC) features. The advantage of NEO is that it uses only a few samples of the input signal to estimate the energy required to generate an AM-FM signal and separate it into amplitude and frequency components without imposing any stationary assumption as done by linear prediction or Fourier transform. In [108] *Amplitude modulation index (AMI)* provides a statistical analysis of amplitude modulations on bandpassed speech signals along the formant tracks. The AMI feature is computed for each pitch period and statistics are computed.

In [117], an alternative method for the estimation of the center frequencies fc of the Gabor filterbank, the iterative-ESA is proposed. This method implies the iterative application of ESA to the Gabor filtered signal and thus adjusting the center frequency of each filter after every iteration. The method is considered important since it reduces the importance of having good initial estimates of the center frequencies of the filterbank. The procedure started using center frequencies dictated by the mel-scale, updating each one of them after every iteration of the ESA, while keeping the bandwidth fixed. The algorithm is assumed to have converged when the center frequency of each filter does not change by more than 1% or reached a certain number of iterations.

3.3.2 Other Descriptors

Besides the speech affective process of the human brain, people tend to interpret several signals for analyzing the real life emotional aspects. Such cues that effect one's emotional state can be either lexical, visual or dialogue-level, while interacting with a group of people. In the first case scenario, where the affective information is expressed using lexical cues, the text-based affective analysis is orchestrated using affective lexica [26, 25]. Such lexica have been constructed for several languages and models using word or phrase-level emotional ratings.

When analyzing group interactions, several turn-taking and dialogue features can be employed. Interactions have been investigated for several tasks, including dialogue acts [55], conflict escalation [57] and sentiment analysis [19]. In [41], several affective, speech and text-based, lexical and semantic features are investigated for root-cause analysis and miscommunication hot-spot detection. Speaker identity and meeting type were investigated in [114] suggesting that both were highly correlated with one's involvement in interactions. In [37] two turn-level features, namely ASR confidence and word error rate (WER), were examined. WER were found to be increased when shouting was detected, while ASR confidence was decreasing in such conditions. Additionally, dialogue-level features were investigated, namely the total dialogue length, the number of turns and task success rate.

3.4 Affective Saliency

The saliency (or salience) is the relative state, widely used in the study of perception and cognition, that refer to any aspect of a stimulus that stands out from the rest.

Saliency detection is an attentional mechanism that uses learning in order to focus on perceptual and cognitive resources. Saliency typically arises from contrasts between items and their neighborhood and may be the result of emotional or cognitive factors.

In the domain of psychology, efforts have been made in modeling human attention. More specifically, in the area of computer vision, efforts have been made in modeling the bottom-up attentional mechanism [7]. One way is based on the spatial contrast analysis. For instance, in [48] a center-surround model is employed to define saliency across scales. Another way is based on the frequency domain, by using the amplitude spectrum to assign saliency, and in was first proposed in [47]. The phase spectrum was used instead in [44], while a system that uses both amplitude and phase information was proposed in [68].

Based on visual perception, an affective saliency map model is proposed in [6]. It considers psychological distance as well as the relative distribution of intensity, edge, color, and orientation. In [11], they focused on how visual distractors influence selection based on either the personal meaning (what a person knows about the distractor) or personal affect (how a person feels about the distractor). Faces are one of the most significant factors for effecting a person's visual stimuli, as they express emotion, intention and needs. Based on that assumption, the emotional facial expressions are investigated in [22], in order to understand the salient properties which trigger shifts of attention.

Several studies have also focused on the affecive properties of music covers. Philosophical and musicological analysis was used to determine such properties in [58], while perceptual analysis was investigated in [5]. Pitch and tonality [104], in addition to rhythm [60] from music have been explored, while the relation of such qualities, i.e., affect and acoustical cues, in music are addressed in [67].

3.5 Information Fusion

A need of improvement on the classsification performance has successfully introduced the concept of information fusion. Several alternative approaches have been proposed over the years for different tasks, including emotion recognition. Techniques of information fusion can be discriminated in three categories, connected to the classification process [95]. The first stage of fusion is on the *data level*, the second on *feature level* and the third on the *decision level*. The first two approaches are not extensively investigated. Mostly, heuristic methods have been used for feature-level fusion [10]. In [56] speech features derived from different timescales of the speech signal are fused and associated with different machine learning techniques. However, the proposed information fusion is employed on the decision-level.

Fusion of different modalities has been also investigated, based on the assumtion that a person's the affective state can be transmitted from different channels. Speech, text and visual information have been combined in [89] for sentiment analysis. In [115] facial and vocal expressions are investigated for emotion recognition, analyzing salient emotional features and cognitive models as well as multi-modal data fusion. A weighting scenario for audio-visual speech recognition is proposed in [42] using optimized weights to minimize the word error rate (WER). Acoustic and language information

are fused in [64]. Salience keywords were identified based on the most frequent words for a specific domain, while the different modalities were combined at the decision-level.

3.6 Summary

In this chapter we introduced the two models of speech production, namely the linear and non-linear model, as well as features derived from both of them. We presented affecive descriptors for speech emotion recognition mainly derived from the speech signal. Linguistic and dialogue features are also proposed. Last, we briefly described the concepts of affective saliency and information fusion that exploit the affective context and knowledge of speech.

Chapter 4

Affective Saliency Model

4.1 Introduction

In the recent years several work has focused on audio, video and text saliency [35]. In [34], audio-visual saliency is investigated for movie summarization. Audio saliency is assessed by quantifying multifrequency waveform modulations, while video saliency is estimated by spatiotemporal attention model driven by intensity, color and motion. In [53], audio saliency is applied on automatic acoustic scene classification of real life. Motivated by the human auditory system and its attention model, salient events of an audio clip are extracted in an unsupervised manner. A study of audio content analysis is presented in [74], in which an audio stream is segmented according to audio type or speaker identity. In [66], emotional vocal stimuli with varying degrees of acoustic cue saliency was used to create graded levels of stimulus-driven prosodic ambiguity.

Applying a discriminative procedure as Minimum Classification Error (MCE) training [50, 33] for information fusion over time has been investigated in the past for several tasks including automatic speech recognition and speaker recognition [72]. In [31] spectral distance features combined with a frame-level misclassification error have been investigated for information fusion over time using conditional random field classifiers. Such techniques are shown to reduce the classification error rate significantly and increase the discriminability among the different labels.

In this chapter we present an affective saliency model [24], that aggregates lower-level information in order to estimate their contribution of the utterance-level emotional perception. Affective saliency is estimated via a regression model that utilized features extracted from different timescales of the acoustic signal (e.g., F0) and the frame-level posterior probabilities. Thus, first a frame-level feature vector is constructed. It is assumed that each frame contains an expression of the emotion of the utterance it belongs to, and therefore it is given that same label. The resulting feature vector with the assumed frame-level labels is then given as input to train a frame-level classifier. The frame-level decisions of a given utterance are further combined in a weighting scheme, which emphasizes the most salient affective information over time. The regression parameters are trained iteratively by minimizing the classification error rate via Minimum Classification Error (MCE) training/ Generalized Probabilistic Descent (GPD). In our experiments, we used spoken dialogue call-center datasets and we focus on an anger

detection task (negative vs non-negative valence detection).

4.2 Affective Saliency Model

Let $X = \{x_1, \ldots, x_N\}$ be a frame vector of an utterance T, and C_i discrete affective labels, e.g. levels of anger vs. neutral, with $i = 1, \ldots, M$. The emotional content of an utterance T is computed over time by its corresponding frames and weighted according to the factor λ_j which indicates the affective saliency for frame j.

$$F(C_i|X) = \log P(C_i|X) = \frac{1}{N} \sum_{j=1}^{N} \lambda_j \log P(C_i|x_j) \tag{4.2.1}$$

where $P(C_i|x_j)$ are the frame-level posterior probabilities, while the weights λ_j are estimated via Minimum Classification Error (MCE). More specifically, given that the optimal weights are unknown, we train a regression model as:

$$\lambda_j = \sum_{k=1}^{K} a_k d_k \tag{4.2.2}$$

where a_k with $\sum_{k=1}^{K} a_k = 1$ the trainable weights and d_k the regression features, described in Section 4.2.2. The next step is to define the misclassification measure E, as shown below

$$E(X) = F(C_I|X) - F(C_C|X) \tag{4.2.3}$$

where C_I and C_C correspond to the incorrect and correct emotional classes, respectively. The loss function, which maps the misclassification error onto the interval $[0, 1]$ is a sigmoid function and it is defined as

$$l(X) = \frac{1}{1 + e^{-\gamma E(X)}}, \qquad \gamma > 1 \tag{4.2.4}$$

with γ representing the sigmoid scaling factor. The loss function approaches zero when $E(X) < 0$ and close to one otherwise. So by minimizing the loss function, the classification error is also minimized. The loss function $l(X)$ can be differentiated and optimized via an iterative gradient descent algorithm, by establishing the algorithmic convergence property [49]. The update equation of a specific unknown parameter w is

$$w^{'} = w - \epsilon \frac{1}{N_T} \sum_{\forall T} \frac{\partial l(X)}{\partial w} \tag{4.2.5}$$

where N_T is the total number of utterances T in the dataset, ϵ is a learning rate parameter used during the iterative MCE training and $\frac{\partial l(X)}{\partial w}$ the partial derivative of the loss function $l(X)$

$$\frac{\partial l(X)}{\partial w} = \frac{\partial l(X)}{\partial E(X)} \cdot \frac{\partial E(X)}{\partial \lambda_j} \cdot \frac{\partial \lambda_j}{\partial w} \tag{4.2.6}$$

4.2.1 Affective Classification Model

For the affective classification defined in (4.2.1), we found that the trainable parameters were more robust across datasets when computed on segment-level instead of frame-level. Hence, features were grouped in sets of 20 frames and statistics were computed over them. We use only 3 LLDs, namely energy, 1st Mel-Frequency Cepstral Coefficient (MFCC) and raw fundamental frequency (F0) and applied the following statistics: max, min, mean, median and standard deviation.

We selected only a set of those 3 LLDs after investigating the distribution of the class posterior probabilities of the trained classifier. When increasing the number of features of the affective classifier, the confidence for the segment-level decision was increasing favoring the majority class.

4.2.2 Regression Model

In this section we present the parameter estimation model and the saliency features d_k, as described in Eq. (4.2.2). Several features including features derived from the posterior probabilities and the acoustic signal were also evaluated as candidates for estimating affective saliency. We found that spectral flux and F0 extracted from different timescales of the speech signal, were robust across the different datasets. Specifically, we extracted spectral flux and F0 in a fixed window size of 200 ms and F0 in 30 ms with 10 ms update. Features extracted in 30 ms window size were further grouped in order to create segments and statistics were applied, namely max, min, mean, median, standard deviation. As an additional feature, we used the rate of unvoiced frames per segment using the Voice Activity Detector presented in [101].

4.3 Spoken Dialogue Datasets

Speech services have been constantly advancing the recent years, due to the growing need of telephone applications and industry. Despite recent progress in Spoken Dialogue System (SDS) technologies, there are a few spoken dialogue datasets containing interactions with real-users. For our experiments we used four spoken dialogue datasets from four call-centers in two languages. A brief description of the datasets is presented in Table 4.1.

	LEGO	CC	PB	MT
#non-negative	3309	1027	1095	1023
#negative	934	339	607	1106
#speakers	200	284	[1]	200
Language	English	English	Greek	Greek

Table 4.1: Dataset description.

[1] No information about the number of speakers was available for the phone banking dataset.

Bus Information - LEGO dataset

The LEGO dataset (a subset of the Let's Go dataset [97]) provides bus schedule information for buses during off-peak hours. The dataset was annotated in terms of events that could cause a hot-spot, while it was augmented with emotional labels, including anger and satisfaction. The anger-related labels (hot-anger detection) followed a 5-level scale: *friendly, neutral, slightly angry, angry, very angry.*

Incoming customer service calls (CC) dataset

The CC dataset of spontaneous speech was created by collecting user calls from a call center. For each call a wave file is available along with respective transcriptions. Each file (and the respective transcription) is annotated with one label: negative or non-negative.

Phone banking dataset

The phone banking dataset [1] consists of 1702 user utterances and their respective transcriptions (in Greek). The main functionality of this application was to provide information regarding bank services. The user utterances were annotated with respect to their emotional content (arousal/valence ratings and anger on a 5-scale scheme) and two personality dimensions, namely neuroticism and extraversion.

Movie ticketing dataset

The main functionality of the movie ticketing service is the retrieval of information about movies and showtimes followed by the booking of tickets. The dataset[2] [1, 73] cover two data types: 1) audio files, and 2) the respective transcriptions. In addition, each transcribed dialogue was annotated with respect to its emotional content. The dataset's annotations can be distinguished into two main categories: 1) annotations on dialogue level, and 2) annotations on utterance level. Each annotation aims to characterize either the emotional content, the personality of the caller, or dialogue characteristics. The emotional content was annotated with respect to arousal, valence and hot-anger. The labels used for anger annotation were discrete scores that lie in the $[1-5]$ interval capturing very angry user utterances (1) to friendly utterances (5)

4.4 Experimental Procedure

We conducted two types of experiments across all datasets: matched (training and testing on the same corpus) and cross-corpus. In the matched experiments, we divided each dataset in equally sized training, development and test sets, while for the cross-corpus experiments, we used (all the data of) three datasets for training and development and tested on the fourth. The development set was used for learning and optimizing the unknown parameters a_k of Eq. (4.2.2). Results on the trainable parameters are presented in Section 4.4.1.

[2]More information about how the dataset was created can be found at the Appendix B

4.4.1 Parameter Optimization

In this section, we present the results if the affective saliency model, i.e., the trainable parameters and the estimated salient weights. During MCE-training the a_k parameters were iteratively updated. In each iteration the average loss value was shown to decrease while the classification accuracy increased, as more misclassified utterances were corrected. Figure 4.1 shows the classification accuracy and the loss function values for four experimental datasets during training.

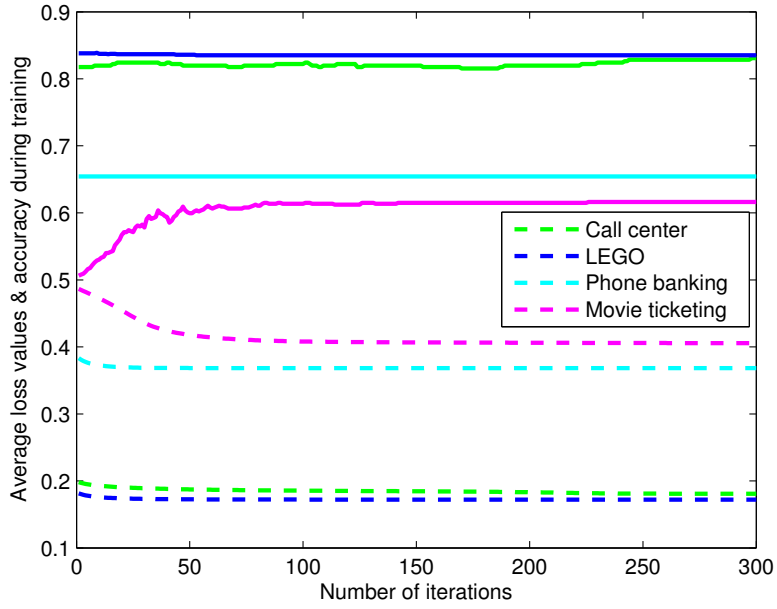

Figure 4.1: Classification accuracy and loss function values during training.

The optimal parameters are the ones that minimized the average loss function. The scaling factor γ of Eq. (4.2.4) and learning factor ϵ of Eq. (4.2.5) were set to $\gamma = 2$ and $\epsilon = 0.1$. We observed that after 300 iteration the GPD algorithm converges for the selected parameters γ and ϵ.

The parameters a_k were initially trained independently on each dataset to investigate the robustness of the proposed method. Results were pretty consistent across datasets. Finally we selected the median value across the datasets in order to construct a universal saliency model. The resulting weights for the $[0, 1]$ normalized features are presented in Tables 4.2 and 4.3.

Comparing the results on Tables 4.2 and 4.3 we observe that the trainable parameters a_k are more robust across datasets on the cross experiments than on the matched. This is justified as on the cross experiments the a_k are computed over three different datasets, decreasing the effect of the individual dataset's characteristics.

Figure 4.2 shows the speech signal and the frame-level pitch contour of the utterance "No, can I talk to a person?" with the weights λ_j computed according to Eq. (4.2.2). The weights are computed on segment-level and mapped to samples and/or

	F0 (30ms)					200ms		
	max	min	median	std	mean	Spec. Flux	F0	Unv. Rate
CC	0.143	0.099	0.243	-0.001	0.175	0.439	0.163	-0.264
LEGO	0.197	0.098	0.160	0.031	0.156	0.225	0.032	0.100
PB	0.224	0.086	0.155	0.040	0.156	0.191	0.015	0.129
MT	0.450	0.490	0.258	0.067	0.316	-1.688	0.616	0.488
median	0.211	0.099	0.202	0.036	0.166	0.208	0.098	0.114

Table 4.2: Estimated optimal parameters across all datasets for the matched experiments.

	F0 (30ms)					200ms		
	max	min	median	std	mean	Spec. Flux	F0	Unv. Rate
CC	0.273	0.157	0.205	0.064	0.205	-0.196	0.075	0.075
LEGO	0.218	0.116	0.173	0.027	0.170	0.114	0.029	0.148
PB	0.204	0.143	0.191	0.021	0.182	0.021	0.036	0.199
MT	0.220	0.130	0.182	0.037	0.180	0.081	0.018	0.148
median	0.219	0.137	0.187	0.032	0.181	0.051	0.033	0.174

Table 4.3: Estimated optimal parameters across all datasets for the cross experiments.

frames using linear interpolation. The weights' values vary across time and peaks are detected toward the end of the utterance where the word "person" is stressed (see also F0 contour). The saliency curve is very smooth since the saliency weights are computed on segment-level.

4.5 Conclusions

We have proposed an algorithm that utilizes a Minimum Classification Error (MCE) criterion in order to learn the most salient affective information over time. Hence, sub-utterance features were explored and used for training a regression model. The regression model uses features from different timescales and LLDs. Experiments on four different datasets of two different languages showed that the model's parameter optimization was robust across all datasets regardless the language.

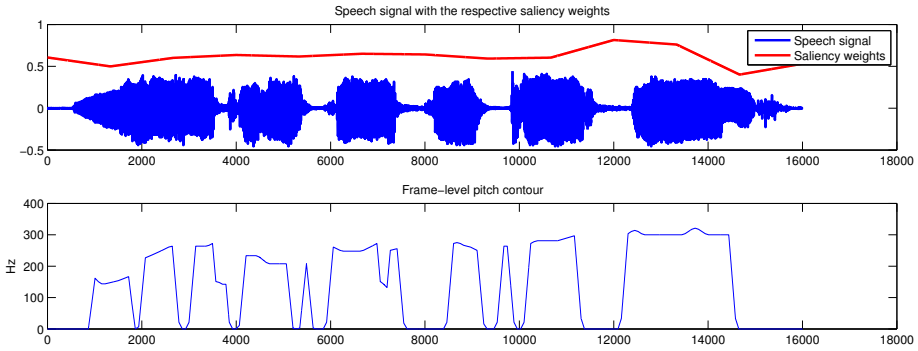

Figure 4.2: Utterance of the CC dataset with transcription: "No, can I talk to a person?". Estimated affective saliency (top) and fundamental frequency contour (bottom) is also shown.

Chapter 5

Fusion Over Time

5.1 Introduction

One of the main issues in affective classification is the level (phone, utterance) of information integration and decision fusion, as well as how information over different time-scales is fused over time. The most popular information fusion method for affective computing is feature-level fusion, where statistics of frame-level features (low-level descriptors) are estimated over a segment or for the whole utterance. In [95], a number of fusion methods are presented, while in [80] decision fusion over different modalities is presented.

Previous studies in this field have used thousands of paralinguistic features, mostly classified in three categories, namely prosodic, short-term spectral and voice quality [112]. As more studies are based on acted speech, where linguistic content and the produced emotions are simulated and controlled, prosodic features are the most widely used. Spectral features, especially MFCCs, were found to improve performance on many speech processing tasks, including emotion recognition.

In this Chapter, we present a model for information fusion over time that weights speech frames/segments based on their affective saliency. Figure 5.1 provides the system's description. The saliency weights are extracted by the affective saliency model presented in Chapter 4. This fusion is implemented following either an early (feature-level) or a late fusion scheme. We compare the early and late weighting scenarios with a baseline model, which fuses information by applying statistics over the frame-level features and achieve impoved performance.

5.2 Baseline Model

For our baseline model[1], we implemented an utterance-level with no weighting fusion scheme. More specifically, we computed statistics/functionals over an utterance's frames. Given a frame $j, 1 \leq j \leq N$, with feature value f_j the mean μ and standard

[1]Further experiments using the baseline model, larger feature sets and more datasets are provided in Appendix A.

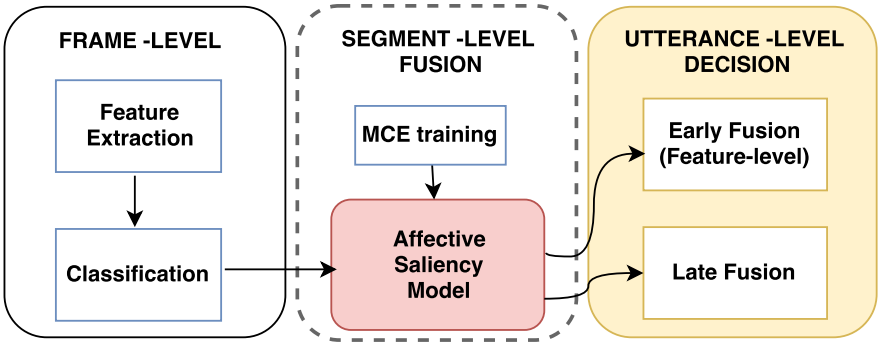

Figure 5.1: System architecture for the fusion scenarios using the affective saliency model.

deviation σ are:

$$\mu = \frac{\sum_{j=1}^{N} f_j}{N}, \qquad \sigma = \sqrt{\frac{\sum_{j=1}^{N} (f_j - \mu)^2}{N}} \tag{5.2.1}$$

The median value is estimated as the middle value of the sorted feature values f_j.

5.3 Early Fusion Model

The saliency weights are used to compute weighted statistics over the frames of an utterance, namely mean, standard deviation, max, min and median. Given a frame $j, 1 \leq j \leq N$, with feature value f_j and weight λ_j the weighted mean μ_w and standard deviation σ_w are:

$$\mu_w = \frac{\sum_{j=1}^{N} \lambda_j f_j}{\sum_{j=1}^{N} \lambda_j}, \qquad \sigma_w = \sqrt{\frac{\sum_{j=1}^{N} \lambda_j (f_j - \mu_w)^2}{\sum_{j=1}^{N} \lambda_j}} \tag{5.3.1}$$

The weighted median is estimated as feature values f_j that can appear multiple times, according to their weights λ_j.

5.4 Late Fusion Model

First we investigate a late fusion scheme for the utterance-level emotion decision. Specifically, we combine the computed weights λ_j as shown in Eq. (4.2.2) with the frame-level posterior probabilities of our affective classifier $P(C_i|x_j)$, as presented in Eq. (4.2.1). Then the utterance-level emotion decision is computed as:

$$C^* = \arg\max_{C_i} F(C_i|X) \tag{5.4.1}$$

where C_i, with $i = 1, \ldots, M$ the discrete affective labels.

5.5 Experimental Procedure

5.5.1 Affective Feature Extraction

A set of 33 frame-level features (low-level descriptors) and their deltas were extracted in a fixed window size of 30 ms with a 10 ms frame update, using the OpenSmile toolkit. The list of spectral and prosodic features used is given in Table 5.1.

Energy-related LLDs	Energy, Zero-Crossing Rate
Spectral LLDs	Energy 250-650Hz 1k-4kHz, Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity, MFCC 1-14, Roll Off Point 0.25, 0.50, 0.75, 0.90
Voicing realted LLDs	F0, Prob. of Voice, raw F0

Table 5.1: List of features

Regarding the baseline and early fusion scenarios the features in Table 5.1 were used along with their deltas. Similar to the saliency model (described in Chapter 4), features have been mapped into the [0,1] interval. In order to extract utterance-level features, the following functionals were applied: mean, standard deviation, median, max and min.

5.5.2 Experiments

For our experiments we used four spoken dialogue datasets from four call-centers in two languages: (1) bus information (LEGO, a subset of the Let'sGo dataset [97]), (2) US call center (CC) incoming customer service calls, (3) phone banking (PB) [1] and (4) movie ticketing (MT) [1, 73], presented in detail in Section 4.3.

We conducted two types of experiments across all datasets: matched (training and testing on the same corpus) and cross-corpus. In the matched experiments, we divided each dataset in equally sized training, development and test sets, while for the cross-corpus experiments, we used (all the data of) three datasets for training and development and tested on the fourth. Table 5.2 presents the average utterance duration per dataset, which as expected is an important factor for the model's performance.

	CC	LEGO	PB	MT
Average duration	1.85	1.67	4.17	1.43

Table 5.2: Average utterance duration in seconds per dataset.

Regarding the experimental procedure, the chance classifier assigns each test sample to the majority class. For our baseline experiments as well as the feature-level fusion an SVM classifier with polynomial kernel from the Weka toolkit is used [46]. We chose an SVM classifier due to its better performance compared to other classifiers tested. Additionally, a forward selection algorithm from the Weka toolkit was applied on the baseline system and the selected features were adapted on the early fusion scenario

as well. For the saliency model we chose a Naive Bayes classifier, in order to extract the class-posterior probabilities, and we present results before (pre-MCE) and after (post-MCE) MCE training.

5.5.3 Evaluation

Next, we present the unweighted average *(UA)* classification accuracy across all datasets and fusion scenarios for the matched and cross-corpus experiments.

	CC	LEGO	PB	MT	UA
Matched experiments					
pre-MCE	77.4	78.7	68.8	53.4	69.5
post-MCE	80.5	79.6	68.1	52.7	70.2
Cross-corpus experiments					
pre-MCE	81.4	79.0	65.6	58.0	71.0
post-MCE	81.6	79.5	66.0	58.2	71.4

Table 5.3: Late fusion: Classification accuracy (%) results for the matched and cross experiments.

In Table 5.3 the results for the late fusion scenario are presented for both the matched and cross experiments. The regression model (affective saliency weights) is initially trained independently by minimizing the average loss function on each dataset and further estimated across all datasets. Results are presented before (no weighting) and after MCE training. As we can see the MCE approach has better performance than the pre-MCE system when refering to the UA metric. When comparing each dataset's performance individually, for the cross-corpus post-MCE outperforms pre-MCE for all experiments, although the improvement is small.

	CC	LEGO	PB	MT	UA
Chance	73.4	79.4	64.2	52.7	67.4
Baseline	79.2	79.8	67.6	51.7	69.6
Early fusion	80.0	80.3	68.2	51.7	70.1

Table 5.4: Early fusion: Classification accuracy (%) results for the matched experiments.

	CC	LEGO	PB	MT	UA
Chance	75.2	77.9	64.3	51.9	67.3
Baseline	81.6	82.1	66.3	54.0	71.0
Early fusion	80.8	82.5	66.7	57.8	72.0

Table 5.5: Early fusion: Classification accuracy (%) results for the cross-corpus experiments.

In Table 5.4 the results of the early (feature-level) fusion are presented for the

matched experiments. For both the baseline and the fusion system, statistics are applied to frame-level LLDs in order to extract utterance-level features. However, for the feature-level fusion weighted statistics are used. The weights are computed according to the saliency model and mapped to frame-level using linear interpolation. We observe equal or better performance for each dataset individually, suggesting that the global nature of the affective saliency system is robust across the different datasets.

Table 5.5 shows the classification accuracy results for the early fusion scenario on the cross-corpus experiments. Here the affective model is computed on three datasets and tested on a fourth. We observe similar behavior with the results presented in Table 5.4, which suggests robustness across the different datasets. This is impressive given that our datasets are of different languages, sizes and SDS type.

Overall, we show improvement across all datasets using the affective saliency model either with the early or the late fusion fusion scenarios, suggesting that frame-level decisions can be fused more efficiently in order to characterize the utterance-level emotional content.

5.6 Conclusions

We investigated the automatic recognition of emotions in speech using an affective saliency model for fusing information over time. The proposed fusion algorithm exploits an affective saliency regression model to either weight frame-level posterior classification probabilities or frame-level features. We demonstrated that the proposed model can achieve modest performance improvement over the baseline. Our results suggest that MCE training increases the discriminability between emotional states, by enhancing the speech frames that carry the most salient information.

Chapter 6

Engagement Detection for Children with Autism Spectrum Disorder

6.1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that disturbs the ability for social engagement, i.e. the development of interpersonal sympathy and collaborative action [54, 39, 106]. Social engagement is based on motives for moving and responding to the physical and social environment. Periods of development represent the ability for more complex forms of social engagement which result from the combination of the two fundamental motives. At birth infants express a simple interest in others' expressions, while by the age of two months become more sensitive to the reciprocity of emotions and are able to recognize the others' communicativeness and its absence or appropriateness. During this period social engagement is characterized as interpersonal, since it does not refer to a topic in the environment; rather, it regards only emotions and intentions in the dyad. Around 3 months a typically developing infant often shifts attention to an object. The partner who seeks communication with the infant, may simultaneously look at the same object or follow the infant's gaze. This kind of communication may be characterized as converging interest, since the two partners attend to the same object, but the infant does not simultaneously pay attention to the other's intentions and feelings regarding that object. At 9 months infants show a more pronounced interest in exploring specific emotional reactions and relating them to external targets, a recognition of commands and prohibitions at 9 months [94, 105]. At this age an infant exhibits a new readiness to tune in with the intentions and interests of a partner in joint exploration and use of objects. This ability forms the basis for a creative imagination of roles, actions, and "tools" that are arbitrary or symbolic [105]. Children with ASD demonstrate different degrees of deficits in these aspects of social engagement. Thus, identifying impairments in the ability to respond to social cues revealing different aspects of social engagement may allow for distinction of young children with ASD from typically developing children and for early identification of this

disorder [28, 82]. An audio feature that conveys intentions and emotions in communication is prosody and child's response to this salient cue of communication may reveal her level of social engagement. However, relevant studies demonstrate that individuals with ASD do have more difficulty in perceiving at least some aspects of pragmatic/affective prosody [12, 13, 86, 88]. Verbal Response Latency (VRL) is another indicator of autism in children, and is defined as the time needed to respond during a conversation. VRLs can provide useful information about ones mental state, while long VRLs might occur when a complex conversation is performed [23]. ASD severity has been also analyzed in relation to vocal arousal and emotion dynamics [78, 15]. In [45], the degree of engagement on children with ASD was investigated using acoustic and duration features showing that vocal cues are highly related to engagement. This work is based on the assumption that vocal and language cues can model one's degree of engagement [14]. Motivated by this hypothesis, we experimented on a database consisting of sessions of typically developed (TD) and Autism Spectrum Disorder (ASD) children interacting with their parents under the supervision of a psychologist. The task that the subjects were asked to complete was for the parents to convince their child to play with a car, while the psychologist observed and participated when needed. We explored the role of acoustic-prosody, language and visual cues of both participants on assessing the child's engagement. Sentence complexity, i.e., the number of words per sentence and verbal use, also differ between TD and ASD children. Hence, we investigated these language cues for both participants as the partner/caregiver have an acting role on the sessions. Finally we experimented with visual cues based on the interactional role of the task, while we examined the effect of turn-taking cues, such as the parent's speech duration [43].

6.2 Experimental Dataset

6.2.1 Video Recordings

It was decided that a structured naturalistic procedure would be the most appropriate method for video recordings [2]. Recordings take place in the child's home and in familiar situations of everyday life. The recordings are characterized as structured because the introduction of certain situations by the psychologist did not leave the dyad complete freedom in play activities. The structured naturalistic method also ensured that all children would experience similar situations during the sampling period, thus, favoring comparisons. The set of toys included two different sized dolls, doll furniture, a tea set, a brush and a mirror, a school bus with little people in it, blocks, toy animals and a book. Mothers were asked to play with their child as they would normally do, introducing all the toys provided. Each session lasted approximately 45 minutes. A high quality video camera was used by an experienced psychologist so as to obtain high quality data for analysis. The structured naturalistic condition consists of the following situations: (a) familiarization (3 minutes), (b) still face condition (2 minutes): the mother and the child interact without toys for 30 seconds then the mother stays unresponsive for 30 seconds and the sequence is repeated once more, (c) play with toys provided by the psychologist (15 minutes), (d) mother pretends that she hurts herself and cries (1 minute), (e) play with toys provided by the psychologist (5 minutes), (f)

mother pretends that the doll is not eating (1 minute), and (g) play with toys provided by the psychologist (8 minutes).

Definition of car episode: on all four video recordings of each child, researchers located the points on the footage where the mother uttered the word car and defined a framework around the mother's utterance called episode. An episode begins when either the mother or the child first look or act at the car and ends when both the mother and the child shift their attention from the car. The mean duration of an episode was 4.86 minutes. The mean duration of each episode for the ASD group was 5.99 minutes and for the TD group was 3.72 minutes. This difference was not statistically significant ($t = 1.11$, $p = 0.283$). Microanalysis within an episode consists in noting the onset and offset of each manifested behavior from every category. This kind of analysis provides information on the duration that mother's and child's attention converge on the car, the initiator and the responder of the interaction as well as the type of ongoing interaction (e.g. solitary play, converging interest or joint attention).

	ASD	TD	ALL
#utterances	966	645	1611
#sessions	33	33	66
#children	9	8	17
#male	8	6	14
#female	1	2	3

Table 6.1: Dataset description.

Table 6.1 presents the dataset's characteristics, namely, the number of utterances, sessions and children.

6.2.2 Data Labeling

One expert annotator labeled the dataset using the ELAN software [100] and according to the following categories: transcription, gaze, action on object, action on partner and emotion. The partner in this case can be either the parent or the child. Interrater reliability was assessed with videotaped data from a random selection of 10% of the sessions. Cohen's kappa was 0.75 on average.

Using the aforementioned annotations, psychologists identified patterns for describing high-level categories of intention for the speaker:

1. *Solitary*: behavior used to learn and explore the environment
2. *Converging Interest*: two people express interest at the same object but they do not communicate between them about that
3. *Regulatory*: behavior used to influence the behavior of others
4. *Interpersonal*: joint attention revealing an interpersonal goal
5. *Interactional*: behavior used to develop social relationships and ease the process of interaction

Assuming that each of the above intention categories carries a variable degree of engagement, the intention/engagement annotation labels are presented in Table 6.2. Moreover, Table 6.2 shows which annotations are needed in order to identify the intention/engagement patterns for each class. The identification of the engagement patterns was applied on a time window, starting at the beginning of the mother's utterance until N seconds from its end. The labeling process was top-down, i.e., starting from the *Interactional* class to the *Solitary*. Ambiguity was observed for a small subset of utterances, as approximately 40 utterances were classified to two classes.

Intent	Engage	Gaze	Object	Partner	Emotion
Solitary	1	✓	✓		
Converging	2	✓	✓		
Regulatory	3	✓	✓	✓	
Interpersonal	4	✓	✓	✓	✓
Interactional	5,6,7,8	✓	✓	✓	

Table 6.2: Data annotations and intention/engagement labeling. Object: action on object, Partner: action on partner

In Table 6.3 and Figure 6.1 examples of the detected engagement patterns are presented. More specifically, in Table 6.3 two patterns are presented, along with the engagement and intention labels. Focusing on the second example, the mother says "The car, come.". During that time, the child is holding/inspecting the specific object, while in the following time frame the child looks at the mother and offers it to her. In Figure 6.1, the session engagement labels are presented over time. The timeline shows shifts on the child's degree of engagement starting with *Converging Interest* to *Interpersonal* and *Interactional*. By the time the highest degree of engagement is achieved, both participants are playing with the car.

Transcription	Gaze	Object	Partner	Intent	Engage
Do you want the car? What do you want?	LO LO LO	HI	MA	Converging	3
The car, come.	LPE	HI OG		Interactional	6

Table 6.3: Intent and engagement annotation examples; LO: looking object, LPE: looking partner's eyes/face, HI: holding/inspecting object, OG: offering/giving, MA: moving away.

Table 6.4 presents the number of utterances per intention class. No utterances are classified to the *Regulatory* class, however it is mentioned for completeness purposes.

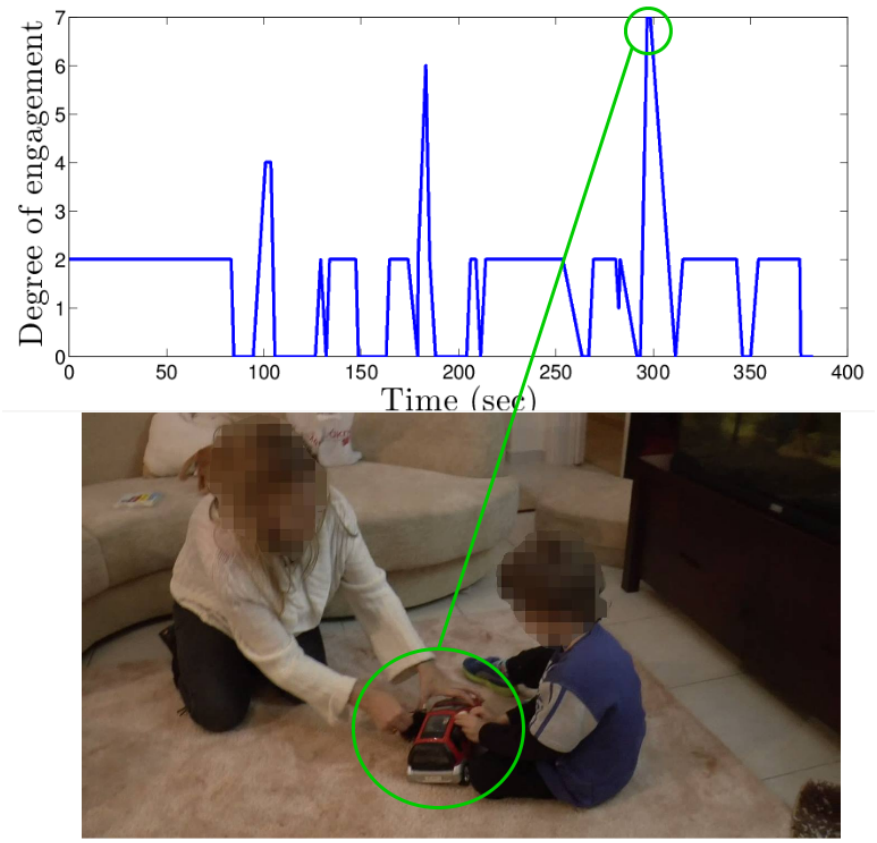

Figure 6.1: Degree of engagement labels over time for a session and example video frame for the highest engagement level.

6.3 Feature Extraction

In this section, we briefly describe various feature sets that are used for the automatic detection of engagement. A synopsis of the features is presented in Table 6.5. Linguistic features were extracted for both mother and child, while the rest feature sets, i.e., audio, video and affective text, were applied only on the mother's utterances.

6.3.1 Audio & duration features

In order to model the style and quality of speech a set of frame-level features (low-level descriptors, LLDs) were extracted in a fixed window size of 30 ms with a 10 ms frame update, using the OpenSmile toolkit [36]. The proposed feature set contains the following LLDs: energy, pitch, probability of voicing, harmonics to noice ratio (HNR) and the first ten LPC coefficients. In order to extract utterance-level features, the following functionals were applied: extremes, moments and percentiles.

Children with ASD tend to respond after a longer period of time compared to TD children. Hence, a voice activity detection (VAD) feature either for the child (interpreted as response) or for the mother (repeating herself) is employed. In both cases the feature is binary and activated only in the time window N used for extracting

Category	#utterances
Solitary	167
Converging Interest	510
Regulatory	0
Interpersonal	82
Interactional	69
No-engagement	784

Table 6.4: Number of utterances per intention category.

Audio - Duration	
Acoustic	Energy, Pitch, Probability of Voice, HNR, LPCs [1-10]
Duration	Utterance duration, VAD
Text	
Affective	Arousal, Valence, Dominance
Linguistic	#words, utterance repetition, #word repetition, #oov
Video	
Action-related	Gaze, action on object/partner

Table 6.5: List of features.

the engagement labels. Moreover, the mother's speech duration is used.

6.3.2 Text features

Linguistic

Based on the assumption that speech is altered when speaking to children with ASD, we created a set of lexical features on the transcribed utterances. These features include the number of words per utterance, a binary feature taking value 1 when the utterance is repeated and 0 otherwise, and the number of repeated words per utterance. Additionally, we observed that parents tend to use baby-talk (motherese) speech to describe sounds, for example the sound of a car. In order to recognize these words we compared our lexicon, consisting of 1.200 words, with a Greek vocabulary of approximately 300.000 words. Words that were not found in the vocabulary were annotated as out-of-vocabulary (OOV) and characterized as baby-speech.

Affective

The goal is to estimate the emotional content of the transcribed speaker utterances. A word w can be characterized regarding its affective content in a continuous space consisting of three dimensions, namely, valence, arousal, and dominance. In order to extract utterance-level ratings, the mean value of the ratings of the constituent words

is computed, using an affective lexicon. More details about the lexicon can be found in [85].

6.3.3 Action-related video features

As action-related features, we refer to the annotations regarding gaze and actions on objects/partner. Although these features were manually derived, they were included in the experimental features in order to investigate their role in engagement prediction. A detailed description of the annotations and their labels can be found in [2]. Information such as movements away or towards a person/object, gaze direction and symbolic or functional play are included.

6.4 Experimental Procedure

Our goal was to predict the child's engagement as a reaction to the mother's speech. The problem was posed as binary, i.e., *engagement vs. no-engagement*. The engagement classes were discriminated as follows: the degrees of engagement as presented in Table 6.2, i.e., *Solitary* to *Interactional*, are mapped to the *engagement* class while the utterances that did not match any of the engagement patterns are mapped to the *no-engagement* class.

Regarding the labeling procedure, the time window N during the pattern identification was a significant parameter. After preliminary experiments, we focused on time window $N = 1$ second. Another important factor was the duration of the mother's speech, based on the idea that the child's intention changes during that time.

For the experimental procedure, we adopted a leave-one-child-out scheme, i.e., testing on one child's utterances while training with the utterances of the rest children. The majority class classifier assigns each test sample to the majority class, while for the experiments an SVM classifier with polynomial kernel from the Weka toolkit [46] is used. We chose an SVM classifier due to its better performance compared to other classifiers tested. The classifiers were trained using the list of features presented in Table 6.5. Additionally, a forward selection algorithm was applied on the acoustic feature set and the selected features were adapted on the fusion scenario as well. As fusion we used the concatenation of the different features sets.

6.4.1 Evaluation

Next, we present the unweighted average classification accuracy *(UA)* and the unweighted average recall *(UR)* for all features sets as well as their fusion.

The results, as presented in Table 6.6, suggest that the action-related features are more successful at predicting the child's engagement outperforming the majority baseline and achieving 0.62 and 0.59 *UR* for the TD and ASD children respectively. The linguistic features, extracted from the mother's transcription, also achieve good performance for both TD and ASD children based on the unweighted recall metric (0.55 and 0.51 respectively).

Overall, we observe that the performance for TD children is better than for ASD, suggesting that the TD children behavior is easier to predict in these sessions.

	UA		UR	
	TD	**ASD**	**TD**	**ASD**
Majority class baseline	56.7	52.2	0.50	0.50
Parent's features				
Acoustic	47.6	47.1	0.46	0.50
Duration	56.6	46.8	0.44	0.47
Linguistic	**56.9**	50.7	**0.55**	**0.51**
Text Affective	50.4	46.3	0.49	0.50
Actions	**61.4**	**53.0**	**0.62**	**0.59**
Child's features				
Linguistic	49.2	44.7	**0.52**	0.48
Fusion				
All features	**63.3**	**53.9**	**0.64**	**0.57**

Table 6.6: Classification accuracy *(UA)* and unweighted recall *(UR)* results for the *engagement vs. no-engagement* task.

6.5 Discussion

In this section, we discuss the factors that seem to effect our system's performance according to the results presented in Table 6.6. Initially, in order to evaluate the human's perception of the child's engagement instead of employing automatic models, a subset of the dataset was annotated by two more annotators. The annotators' tasks were to predict whether the child is engaged or not by 1) only hearing the parent's utterance, and 2) only watching the parent's movements. Regarding the prediction, two labels were used: 1 when engagement was predicted and 0 otherwise. The inter annotator's agreement, according to the Cohen's coefficient, was computed and is presented in Table 6.7.

		TD		ASD	
Modality	Task	Agree	κ	Agree	κ
Audio	prediction	0.42	-0.24	0.51	-0.02
Video	prediction	0.65	0.29	0.56	0.10

Table 6.7: Inter-annotator's agreement wrt. engagement detection.

The κ values, presented in Table 6.7, suggest that the audio predictions achieve poor agreement, i.e., engagement prediction can not be estimated via audio only. However, the video prediction and assessment labeling can be interpreted as fair and moderate agreement respectively.

Our initial assumption was that the mother's prosody would be discriminative between the engagement classes. However, mothers have been found to speak motherese regardless of the child's degree of engagement. The low performance of the acoustic feature set can be also justified by the fact that the sessions are recorded between children and their parents, instead of a psychologist. We believe that psychologists are

inclined to use more strategic and less affective speech compared to parents.

The most significant factor on our analysis was whether the child was TD or ASD. As our results suggest, TD children are more responsive to the sessions than ASD children. Table 6.8 presents the Pearson correlation between the engagement labels and the VAD of child's/mother's speech and the repetition of the mother's utterance. The results suggest that the child's speech is more correlated to the engagement labels, although the children are from varying ages and the majority of them are non-verbal. Moreover, mother's speech in the time window N is not uncorrelated with the engagement labels as demonstrated on Table 6.8.

	Child VAD	Parent repetition	Parent VAD
TD	0.18	0.06	0.12
ASD	0.11	0.03	0.09

Table 6.8: Pearson correlation between engagement labels and features.

6.6 Conclusions

We investigated the engagement of TD and ASD children in sessions with their parents and focused on the utterance-level engagement vs. no-engagement classification task. We used feature sets from different modalities, namely audio, text and video, extracted mostly from the mother's utterances rather than the child's. Our results suggest that the child's engagement can be predicted by analyzing the mother's characteristics, but not with good accuracy. Prediction accuracy was higher to TD rather than ASD children. Although acoustic features were not expressive enough, movements/actions and lexical features from the mother's transcribed utterances were the most informative. Child speech was expected to be more correlated to the child's engagement level, however most of the children used non-verbal cues or reacted to the task's needs with movements and gazing rather than talking.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this chapter, we summarize the contributions of this thesis and we provide insights for future work. Motivated by the assumption that emotional changes alter the speech production process, we focused on the task of speech emotion recognition for spoken dialogue applications. This thesis presented a model for speech emotion recognition using affective saliency (*see Chapter 4*), while two fusion scenarios, early (feature-level) and late, were proposed (*see Chapter 5*). Several speech descriptors were examined and evaluated for the task, as well. Finally, we investigated the engagement for children with Autism Spectrum Disorder on interactions with their parents (*see Chapter 6*).

In this work, we emphazised on the affective feature extraction by examining several low-level descriptors (LLDs) for the task of speech emotion recognition. We investigated the performance of acoustic, spectral and voice quality descriptors, while we also examined features derived from the Amplitude and Frequency Modulation (AM-FM) model. The performance of the affective descriptors was evaluated on datasets of both acted and spontaneous speech. The concept of multilinguality was also explored by applying our models to datasets of different languages, i.e., English, Greek and German. The development of a Greek Spoken Dialogue dataset, namely the Movie Ticketing dataset, is included to the research work of this thesis.

One of the core models proposed in this thesis, the affective saliency model, aggregates lower-level information in order to estimate their contribution on the utterance-level emotional perception. The affective saliency is estimated using a regression model that utilizes several acoustic descriptors from different timescales combined with the frame-level posterior probabilities of an affective classifier. The regression model is trained using a Minimum Classification Error (MCE) criterior by optimizing an objective loss function. Trainable parameters were iteratively updated in order to minimize the classification error by establishing the algorithmic convergence property of a Generalized Probabilistic Descent (GPD) algorithm. The model's universality was also investigating by computing the parameter values across all datasets. Experiments on four different datasets of two different languages showed that the model's parameter optimization was robust regardless the language. Cross-corpus experiments, i.e., testing on one dataset and training with the rest three, demonstrated similar behavior.

Two fusion scenarios were then proposed for aggregating the affective saliency weigths to the utterance-level emotion decision. In the first scenario, an early fusion scheme is employed, in which frame-level descriptors are weighted according to the affective saliency. On the second scenario, a late fusion scheme is proposed. The saliency weights are combined with the frame-level posterior probabilities of an affective classifier. Both schemes are compared to baseline models, which decide the utterance-level emotion label with no weighting performed. Our results, for either weighting scheme, suggested that improvement can be achived compared to the baseline, while the MCE training improved the discriminability between the classes.

Finally, we examined the relation of affect and engagement in typically developed (TD) and with Autism Spectrum Disorder (ASD) children using sessions of interaction with their parents. Motivated by the fact that ones degree of engagement can be determined by all participants of an interaction, we examined both mother and child characteristics for engagement detection. Features derived from three modalities, namely speech, text and video are investigated and evaluated for the task. Prediction accuracy was higher to TD rather than ASD children, while video-related and lexical features from the mother's transcribed utterances were the most informative.

7.2 Future Work

In future work, it would be interesting to further explore the affective descriptors of the speech signal and investigate their contribution on the emotion decision. Regarding the affective saliency model, features and techniques will be examined for computing salient weights over time. Another interesting turn of the saliency model would be to adapt the affective salient weights into models based on other modalities, such as affective text analysis. A richer feature set, by applying a larger number of statistics and LLDs, and alternative machine learning algorithms will be evaluated for affective fusion. Further investigation and factor analysis concerning the improvement compared to the baseline system will be performed.

Regarding the engagement detection model, more features will be investigated and alternative machine learning algorithms will be evaluated. The contribution of each feature and its relation with the engagement labeling will be examined. Moreover, the factors concerning the participants of the study as well as the videotaped sessions need to be addressed. The age and severity of the ASD children may effect their social and mental deficits, subsequently effecting the system's performance. Finally, the action-related features as well as the transcribed utterances will be automatically extracted using video and speech recognizers.

Appendix A

Baseline Experiments

A.1 Introduction

In this Section, we present experiments using the baseline model and several low-level-descriptors (LLDs). Specifically, we used a set of 132 LLDs extracted using the OpenSmile toolkit and a set of statistics/functionals applied over them. We compared their performance with features derived from the AM-FM model and investigate their contribution on the emotion recognition task. For our experiments we used four datasets of three languages, English, German and French.

A.2 Feature Extraction

A.2.1 OpenSMILE feature set

Initially, we used the openSMILE toolkit on order to extract the 132 low-level descriptors (LLDs) from the InterSpeech 2012 configuration file. Then a set of functionals, including extrames, means, percentiles, peaks, and times were appleid over the frame-level LLDs in order to extract utterance-level features. The resulted baseline feature set (B_1) is comprised of 5757 features. The computed LLDs and the applied functionals are presented in detail in [81]. We used a feature selection algorithm using the Weka toolkit on the B_1 feature set, which resulted to the B_3 feature set.

A.2.2 Frequency Modulation Percentages (FMPs)

For this feature set, we computed the Frequency Modulation Percentages (FMPs), as presented in Section 3.3.1, which can partially capture the fluctuation of frequencies during a single pitch period. Functionals, implemented in matlab were applied on the LLDs, namely, average, maximum, minimum, median, standard deviation, mode, variance.

A.2.3 Fusion scenarios

In our first feature-level fusion scenario ($F_{FS(ALL)}$), we concatenated the B_1 feature set with the FMPs, creating a set comprised of all the features (ALL). Then, a forward

selection algorithm was applied for reducing the feature vector's dimensions. Additionally, in a second feature fusion ($F_{B3+FMPs}$) we concatenated the B_3 feature set, i.e., forward selection on the OpenSmile feature set, with the statistics applied on the FMPs.

A.3 Experimental Procedure

For training we adopted a leave-one-speaker-out scheme. The feature selection algorithm is applied in each training iteration and the final feature set is constructed by the intersection of the resulted feature sets. Regarding the classification algorithm, we experimented on 3 different classifiers, namely Naive Bayes, Support Vector Machines (SVM) and Random Forest. Finally, we evaluated our results with respect to weighted precision (A.1).

$$wPr = \frac{\sum_i^N Pr_i \cdot n_i}{\sum_{i=1}^N n_i} \quad Pr = \frac{tp}{tp + fp} \tag{A.1}$$

where tp, fp, fn and tn denote true-positive, false-positive, false-negative and true-negative samples for a class i. N is the number of categorical labels and n_i the true positive samples per class.

A.4 Experimental Datasets

For the evaluation task four databases were used, namely 1) Call Center (CC), 2) Berlin Database [84], 3) the SSPNet Personality Corpus [81] and 4) the Let's Go Data. The CC and LEGO datasets were presented in detail in Section 4.3. The Berlin dataset consists of 535 audio files containing 7 emotional labels by 10 speakers and 10 sentences. The SSPNet Corpus includes 640 audio clips of 10 seconds including personality scores assigned individually by 11 assessors. The scores are obtained from raw personality questionnaires and correspond to the following traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness.

Dataset	#Speakers	#Utterances	Language
Berlin	10	535	German
CallCenter	284	1366	English
Personality	332	640	French
LEGO	200	4243	English

Table A.1: Number of speakers per dataset.

Table A.1 presents the number of speakers and utterances and the language per dataset, while Table A.2 shows the feature vector's dimension for each feature set presented in Section A.2.

A.5 Evaluation & Results

Systems	Berlin	CallCenter	Personality				
			Agree	Consc	Extra	Neuro	Open
OpenSMILE B_1	5757	5757	5757	5757	5757	5757	5757
OpenSMILE F.S. B_3	203	142	21	15	23	26	12
FMPs	42	42	42	42	42	42	42
$F_{FS(ALL)}$	203	142	22	17	23	25	12
$F_{B3+FMPs}$	245	184	63	57	65	68	54

Table A.2: Dimensions per dataset and feature set.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.657	0.820	**0.365**	0.820	0.806
SVM	**0.814**	**0.837**	0.343	**0.849**	**0.839**
Random Forest	0.621	0.749	0.357	0.705	0.751

Table A.3: Weighted precision results for the Berlin Database.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.828	**0.861**	**0.778**	**0.856**	**0.862**
SVM	**0.839**	0.846	0.655	0.855	0.858
Random Forest	0.791	0.823	0.707	0.826	0.811

Table A.4: Weighted precision results for the Call Center Data.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.526	0.640	0.538	0.640	0.610
SVM	**0.564**	0.610	0.540	0.610	0.616
Random Forest	0.546	**0.647**	**0.543**	**0.647**	**0.637**

Table A.5: Weighted precision results for the Agreeableness dimension.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.540	0.551	0.552	0.551	0.587
SVM	0.617	0.617	**0.617**	0.617	0.617
Random Forest	**0.625**	**0.649**	0.563	**0.649**	**0.639**

Table A.6: Weighted precision results for the Conscientiousness dimension.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.539	0.561	0.455	0.561	0.560
SVM	0.553	0.623	**0.506**	0.623	0.607
Random Forest	**0.554**	**0.676**	0.493	**0.676**	**0.637**

Table A.7: Weighted precision results for the Extraversion dimension.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.548	0.618	0.499	0.618	0.618
SVM	0.507	**0.648**	0.529	**0.648**	**0.619**
Random Forest	**0.549**	0.635	**0.546**	0.635	0.615

Table A.8: Weighted precision results for the Neuroticism dimension.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.489	0.575	0.506	0.575	0.505
SVM	0.502	0.555	0.533	0.555	0.550
Random Forest	**0.566**	**0.579**	**0.555**	**0.579**	**0.565**

Table A.9: Weighted precision results for the Openness dimension.

	B_1	B_3	FMPs	$F_{FS(ALL)}$	$F_{B3+FMPs}$
Bayes	0.786	0.793	**0.748**	0.795	0.799
SVM	**0.834**	**0.805**	0.608	**0.806**	**0.819**
Random Forest	0.767	0.786	0.720	0.784	0.790

Table A.10: Weighted precision results for the Let's Go Data.

Appendix B

The Movie Ticketing Dataset

B.1 Introduction

The movie ticketing dataset [73] consists of 200 dialogues in Greek collected through a call center service for retrieving information about movies/show times and booking tickets. The dataset includes two data types for each dialogue: 1) audio recordings, and 2) the respective transcriptions. The annotation of dialogues was performed by an expert annotator, while the selected dialogues were balanced with respect to three factors: (i) gender of caller, (ii) call success, (iii) emotional content. To verify the quality of annotations, two additional annotators labeled a subset of the 60 dialogues from the original dataset for anger. The agreement between annotators found was 58% with 0.4 Kappa value - computed as the average pairwise agreement - according to the Fleiss coefficient, which can be interpreted as a moderate agreement.

B.2 Annotation Scheme

The first step prior to annotate the data was to manually transcribe the user utterances. The system prompts were also transcribed since no system logs were available, only the audio files from user and system turns. To annotate miscommunication, anger and satisfaction the speech transcription was presented to the annotator when performing the task, while he had access to the audio from the utterances too.

In miscommunication annotation the task was to evaluate if the system turn was problematic or not. Label 0 was used when system answer was not considered problematic, 1 when the system answer was problematic and 2 when the annotator could not decide from the context whether the system answer was problematic or not. During the annotation the annotator could see the whole dialogue.

The presence of anger, the satisfaction and the presence of repeated content in the utterances could be indicators that a miscommunication occurred. Along with the miscommunication annotation, the annotator had to listen to the utterance audio file and identify if anger was present, the degree of satisfaction of the user and if there is content repeated between the current user turn and the previous turn. The labels used for anger annotation were discrete scores that lie in the $[1 - 5]$ interval capturing very

angry user utterances (1) to friendly utterances (5). Satisfaction was annotated in a five point scale from 1 very unsatisfied to 5 very satisfied.

Moreover, 1 was used for user utterances in which repetition was observed and 0 otherwise. While listening to the dialogue the annotators were asked to be aware of gender. To annotate task success, the annotators should listen to the whole dialogue and verify that if the intention of the user was correctly answered by the system. The label 1 was used for successful dialogues and the 0 for unsuccessful dialogues.

B.3 Anger Detection on the MT dataset

The experimental results for the movie ticketing dataset are briefly presented.

B.3.1 Speech-based system

Here, the aim is to capture the speaker's emotional state using exclusively the speaker's speech signal. Hence, we utilize a set of low-level descriptors (LLDs) able to describe the emotional content. Such LLDs have been widely used and include prosody (pitch and energy), short-term spectral (Mel Frequency Cepstral Coeficients, MFCCs) and voice quality (Jitter) features. The LLDs were extracted in a fixed window size of 30 ms with a 10 ms frame update and were further exploited via the application of a set of functions, in order to map the speech contours to feature vectors. The following functions (statistics) computed at the utterance-level for each of the LLDs were used for the speech analysis: percentiles, extremes, moments, peaks.

B.3.2 Fusion of speech and text analysis

The main idea for the fusion of the two systems is motivated by the hypothesis that each system exhibits different types of errors. For example, cases of offensive language may be missed by the speech system, while cases of anger are likely to be missed by the text-based. In an attempt to improve the performance of the speech affective system, we employed a late fusion scheme. Specifically, the mean of the classification posterior probabilities of the two systems were used, while we classify to the class with the maximum posterior probability score.

B.3.3 Experiments and evaluation results

The goal is the detection of "angry" vs. "not angry" (i.e., 2-class classification problem) user utterances. For this purpose, the anger annotations were used. Specifically, the friendly and neutral labels were mapped to the "not angry" class, while the slightly angry, angry and very angry labels were mapped to the "angry" class. The evaluation was performed on the utterance level adopting the leave-one-dialogue-out process. The unweighted average recall (UAR) and the classification accuracy (CA) were used as evaluation metrics. The used feature set consisted of statistics over the first ten Mel-frequency cepstral coefficients (MFCCs) extracted via OpenSmile. In order to reduce the feature vector's dimensionality a forward selection algorithm was then applied using the WEKA toolkit, while a JRip classifier was used.

System	**UAR**	**CA (%)**
Speech	0.67	67
Text	0.61	59
Fusion of speech and text		
Mean of posterior probabilities	0.67	68

Table B.1: Movie ticketing dataset: "angry" vs. "not angry" classification.

The results of the affective analysis on the MT dataset are presented in Table B.1. Both systems exceed the performance of the majority-based classification regarded as naive baseline (0.5 UAR for binary problems and 59% CA). The best performance, with respect to CA, was obtained by the fusion of the speech- and text-based systems suggesting that the performance of the speech-based system can be (slightly) benefited by the incorporation of the text-based analysis.

Bibliography

[1] "Spedial project free data deliverable d2.1.," *SpeDial Project*, 2013–2015.

[2] "Deliverable WP3: Report on affective and cognitive modeling of TD and ASD children (months 1-9)," *BabyAffect Project*, 2015.

[3] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, and D. D. O'Shaughnessy, "Amplitude modulation features for emotion recognition from speech.," in *INTERSPEECH*, pp. 2420–2424, 2013.

[4] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog.," in *INTERSPEECH*, Citeseer, 2002.

[5] L.-L. Balkwill and W. F. Thompson, "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues," *Music perception: an interdisciplinary journal*, vol. 17, no. 1, pp. 43–64, 1999.

[6] S.-W. Ban, Y.-M. Jang, and M. Lee, "Affective saliency map considering psychological distance," *Neurocomputing*, vol. 74, no. 11, pp. 1916–1925, 2011.

[7] S.-W. Ban, I. Lee, and M. Lee, "Dynamic visual selective attention model," *Neurocomputing*, vol. 71, no. 4, pp. 853–856, 2008.

[8] S. Baron-Cohen, "Social and pragmatic deficits in autism: cognitive or affective?," *Journal of autism and developmental disorders*, vol. 18, no. 3, pp. 379–402, 1988.

[9] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind"?," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[10] J. C. Bezdek, J. Keller, R. Krisnapuram, and N. Pal, *Fuzzy models and algorithms for pattern recognition and image processing*, vol. 4. Springer Science & Business Media, 2006.

[11] A. T. Biggs, R. D. Kreager, B. S. Gibson, M. Villano, and C. R. Crowell, "Semantic and affective salience: The role of meaning and preference in attentional

capture and disengagement.," *Journal of experimental psychology: human perception and performance*, vol. 38, no. 2, p. 531, 2012.

[12] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist.," in *INTERSPEECH*, pp. 1043–1046, 2012.

[13] D. Bone, M. P. Black, A. Ramakrishna, R. Grossman, and S. Narayanan, "Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism," 2015.

[14] D. Bone, C.-C. Lee, T. Chaspari, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "Acoustic-prosodic, turn-taking, and language cues in child-psychologist interactions for varying social demand.," in *INTERSPEECH*, pp. 2400–2404, 2013.

[15] D. Bone, C.-C. Lee, A. Potamianos, and S. S. Narayanan, "An investigation of vocal arousal dynamics in child-psychologist interactions using synchrony measures and a conversation-based model.," in *INTERSPEECH*, pp. 218–222, 2014.

[16] M. Braverman, D. Fein, D. Lucci, and L. Waterhouse, "Affect comprehension in children with pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 19, no. 2, pp. 301–316, 1989.

[17] R. Buck, "What is this thing called subjective experience? reflections on the neuropsychology of qualia.," *Neuropsychology*, vol. 7, no. 4, p. 490, 1993.

[18] R. Buck, "The biological affects: a typology.," *Psychological review*, vol. 106, no. 2, p. 301, 1999.

[19] F. Burkhardt, M. Van Ballegooy, K.-P. Engelbrecht, T. Polzehl, and J. Stegmann, "Emotion detection in dialog systems: applications, strategies and challenges," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6, IEEE, 2009.

[20] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pp. 110–127, 2012.

[21] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 582–596, 2009.

[22] M. G. Calvo and L. Nummenmaa, "Detection of emotional faces: salient physical features guide effective visual search.," *Journal of Experimental Psychology: General*, vol. 137, no. 3, p. 471, 2008.

[23] T. Chaspari, D. Bone, J. Gibson, C.-C. Lee, and S. Narayanan, "Using physiology and language cues for modeling verbal response latencies of children with ASD," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 3702–3706, IEEE, 2013.

[24] A. Chorianopoulou, P. Koutsakis, and A. Potamianos, "Emotion recognition using affective saliency," in *INTERSPEECH (to be presented)*, 2016.

[25] G. L. Clore and A. Ortony, "The semantics of the affective lexicon," in *Cognitive perspectives on emotion and motivation*, pp. 367–397, Springer, 1988.

[26] G. L. Clore, A. Ortony, and M. A. Foss, "The psychological foundations of the affective lexicon.," *Journal of personality and social psychology*, vol. 53, no. 4, p. 751, 1987.

[27] G. Dawson, L. Carver, A. N. Meltzoff, H. Panagiotides, J. McPartland, and S. J. Webb, "Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development," *Child development*, vol. 73, no. 3, pp. 700–717, 2002.

[28] G. Dawson, K. Toth, R. Abbott, J. Osterling, J. Munson, A. Estes, and J. Liaw, "Early social attention impairments in autism: social orienting, joint attention, and attention to distress.," *Developmental psychology*, vol. 40, no. 2, p. 271, 2004.

[29] M. Dawson, I. Soulières, M. A. Gernsbacher, and L. Mottron, "The level and nature of autistic intelligence," *Psychological science*, vol. 18, no. 8, pp. 657–662, 2007.

[30] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust am-fm features for speech recognition," *Signal Processing Letters, IEEE*, vol. 12, no. 9, pp. 621–624, 2005.

[31] S. Dimopoulos, A. Potamianos, E.-F. Lussier, and C.-H. Lee, "Multiple time resolution analysis of speech signal using mce training with application to speech recognition," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3801–3804, IEEE, 2009.

[32] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, pp. 45–60, 1999.

[33] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden markov modeling," *Information Theory, IEEE Transactions on*, vol. 35, no. 5, pp. 1001–1013, 1989.

[34] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, and Y. Avrithis, "Movie summarization based on audiovisual saliency detection," in *2008 15th IEEE International Conference on Image Processing*, pp. 2528–2531, IEEE, 2008.

[35] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3553–3556, IEEE, 2009.

[36] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, ACM, 2010.

[37] A. Fandrianto and M. Eskenazi, "Prosodic entrainment in an information-driven dialog system.," in *INTERSPEECH*, pp. 342–345, 2012.

[38] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition.," in *INTERSPEECH*, pp. 778–781, 2007.

[39] U. Frith, *Autism and Asperger syndrome*. Cambridge University Press, 1991.

[40] D. Galanis, S. Karabetsos, M. Koutsombogera, H. Papageorgiou, A. Esposito, and M.-T. Riviello, "Classification of emotional speech units in call centre interactions," in *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pp. 403–406, IEEE, 2013.

[41] S. Georgiladakis, G. Athanasopoulou, R. Meena, J. Lopes, A. Chorianopoulou, E. Palogiannidi, E. Iosif, G. Skantze, and A. Potamianos, "Root-cause analysis of miscommunication hotspots in spoken dialogue systems," in *INTERSPEECH (to be presented)*, 2016.

[42] H. Glotin, D. Vergyr, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1, pp. 173–176, IEEE, 2001.

[43] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[44] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on*, pp. 1–8, IEEE, 2008.

[45] R. Gupta, C.-C. Lee, D. Bone, A. Rozga, S. Lee, and S. Narayanan, "Acoustical analysis of engagement behavior in children.," in *WOCCI*, pp. 25–31, 2012.

[46] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[47] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[48] L. Itti, C. Koch, E. Niebur, *et al.*, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[49] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 257–265, 1997.

[50] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification [pattern recognition]," *Signal Processing, IEEE Transactions on*, vol. 40, no. 12, pp. 3043–3054, 1992.

[51] J. F. Kaiser, "On a simple algorithm to calculate theenergy'of a signal," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 381–384, IEEE, 1990.

[52] J. F. Kaiser, "On teager's energy algorithm and its generalization to continuous signals," in *Proc. 4th IEEE digital signal processing workshop*, 1990.

[53] O. Kalinli, S. Sundaram, and S. S. Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing.," in *MMSP*, vol. 9, pp. 5–7, 2009.

[54] L. Kanner *et al.*, *Autistic disturbances of affective contact.* publisher not identified, 1943.

[55] T. Kawahara, K. Sumi, Z.-Q. Chang, and K. Takanashi, "Detection of hot spots in poster conversations based on reactive tokens of audience.," in *INTERSPEECH*, pp. 3042–3045, Citeseer, 2010.

[56] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pp. 48–51, IEEE, 2007.

[57] S. Kim, S. H. Yella, and F. Valente, "Automatic detection of conflict escalation in spoken conversations.," in *INTERSPEECH*, pp. 1167–1170, 2012.

[58] P. Kivy, *New essays on musical understanding.* Oxford University Press, 2001.

[59] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals.," in *INTERSPEECH*, Citeseer, 2003.

[60] E. W. Large and J. F. Kolen, "Resonance and the perception of musical meter," *Connection science*, vol. 6, no. 2-3, pp. 177–208, 1994.

[61] J. LeDoux, *The emotional brain: The mysterious underpinnings of emotional life.* Simon and Schuster, 1998.

[62] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pp. 240–243, IEEE, 2001.

[63] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[64] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition.," in *INTERSPEECH*, 2002.

[65] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes.," in *Interspeech*, pp. 205–211, 2004.

[66] D. I. Leitman, D. H. Wolf, J. D. Ragland, P. Laukka, J. Loughead, J. N. Valdez, D. C. Javitt, B. Turetsky, and R. Gur, "It's not what you say, but how you say it: a reciprocal temporo-frontal network for affective prosody," *Frontiers in human neuroscience*, vol. 4, p. 19, 2010.

[67] M. Leman, V. Vermeulen, L. Voogdt, and D. Moelants, "Using audio features to model the affective response to music," in *Proceedings of the International Symposium on Musical Acoustics*, 2004.

[68] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 996–1010, 2013.

[69] M. Li, "Automatic recognition of speaker physical load using posterior probability based features from acoustic and phonetic tokens," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[70] L. Linnenbrink-Garcia, T. K. Rogat, and K. L. Koskey, "Affect and engagement during small group instruction," *Contemporary Educational Psychology*, vol. 36, no. 1, pp. 13–24, 2011.

[71] J. Liscombe, G. Riccardi, and D. Hakkani-Tur, "Using context to improve emotion detection in spoken dialog systems," 2005.

[72] C.-S. Liu, C.-H. Lee, W. Chou, B.-H. Juang, and A. E. Rosenberg, "A study on minimum error discriminative training for speaker recognition," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 637–648, 1995.

[73] J. Lopes, A. Chorianopoulou, E. Palogiannidi, H. Moniz, A. Abad, K. Louka, E. Iosif, and A. Potamianos, "The spedial datasets: Datasets for spoken dialogue systems analytics,"

[74] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, 2002.

[75] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE transactions on signal processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[76] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on signal processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[77] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pp. 421–424, IEEE, 1991.

[78] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. Tal, and O. Golan, "Emotion in the speech of children with autism spectrum conditions: prosody and everything else.," in *WOCCI*, pp. 17–24, 2012.

[79] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features.," in *INTERSPEECH*, pp. 2818–2821, Citeseer, 2010.

[80] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 2462–2465, IEEE, 2010.

[81] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *Affective Computing, IEEE Transactions on*, vol. 3, no. 3, pp. 273–284, 2012.

[82] P. Mundy and F. Acra, "Joint attention, social engagement, and the development of social competence," *The development of social engagement neurobiological perspectives*, pp. 81–117, 2006.

[83] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[84] A. Paeschke and W. F. Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[85] E. Palogiannidi, P. Koutsakis, E. Iosif, and A. Potamianos, "Affective lexicon creation for the Greek language," in *10th International Conference on Language Resources and Evaluation (LREC), At Portorož (Slovenia)*, 2016.

[86] R. Paul, L. D. Shriberg, J. McSweeny, D. Cicchetti, A. Klin, and F. Volkmar, "Brief report: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders," *Journal of Autism and Developmental Disorders*, vol. 35, no. 6, pp. 861–869, 2005.

[87] R. Pekrun, T. Goetz, A. C. Frenzel, P. Barchfeld, and R. P. Perry, "Measuring emotions in students' learning and performance: The achievement emotions questionnaire (aeq)," *Contemporary educational psychology*, vol. 36, no. 1, pp. 36–48, 2011.

[88] S. Peppé, J. McCann, F. Gibbon, A. O'Hare, and M. Rutherford, "Receptive and expressive prosodic ability in children with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 4, pp. 1015–1028, 2007.

[89] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis.," in *ACL (1)*, pp. 973–982, 2013.

[90] A. D. Pierce *et al.*, *Acoustics: an introduction to its physical principles and applications*, vol. 20. McGraw-Hill New York, 1981.

[91] R. Plutchik, "A general psychoevolutionary theory of emotion," *Theories of emotion*, vol. 1, pp. 3–31, 1980.

[92] A. Potamianos, *Speech processing applications using an AM-FM modulation model*. PhD thesis, Harvard University Cambridge, Massachusetts, 1995.

[93] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice Hall, 1978.

[94] V. Reddy and M. Barrett, "Prelinguistic communication," *The development of language*, pp. 25–50, 1999.

[95] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," *Computing and Information systems*, vol. 7, no. 1, pp. 1–10, 2000.

[96] K. R. Scherer, "A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology.," in *INTERSPEECH*, pp. 379–382, 2000.

[97] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system.," in *LREC*, pp. 3369–3373, 2012.

[98] B. Schuller, M. Lang, and G. Rigoll, "Automatic emotion recognition by the speech signal," *Institute for Human-Machine-Communication, Technical University of Munich*, vol. 80290, 2002.

[99] M. Shami and W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," *Speech Communication*, vol. 49, no. 3, pp. 201–212, 2007.

[100] H. Sloetjes and P. Wittenburg, "Annotation by category – ELAN and ISO DCR," in *6th International Conference on Language Resources and Evaluation (LREC)*, 2008.

[101] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 798–807, 2010.

[102] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[103] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech production and speech modelling*, pp. 241–261, Springer, 1990.

[104] E. Terhardt, "Pitch, consonance, and harmony," *The Journal of the Acoustical Society of America*, vol. 55, no. 5, pp. 1061–1069, 1974.

[105] C. Trevarthen, "Infant semiosis1," *Origins of semiosis: Sign evolution in nature and culture*, vol. 116, p. 219, 1994.

[106] C. Trevarthen, "Autism as a neurodevelopmental disorder affecting communication and learning in early childhood: prenatal origins, post-natal course and effective educational support," *Prostaglandins, Leukotrienes and Essential Fatty Acids*, vol. 63, no. 1, pp. 41–46, 2000.

[107] D. J. Tritton, *Physical fluid dynamics*. Springer Science & Business Media, 2012.

[108] P. Tsiakoulis and A. Potamianos, "Statistical analysis of amplitude modulation in speech signals using an am-fm model," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3981–3984, IEEE, 2009.

[109] J. Van den Berg, J. Zantema, and P. Doornenbal Jr, "On the air resistance and the bernoulli effect of the human larynx," *The journal of the acoustical society of America*, vol. 29, no. 5, pp. 626–631, 1957.

[110] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework.," in *INTERSPEECH*, pp. 751–755, 2014.

[111] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[112] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, pp. I–593, IEEE, 2004.

[113] J. L. Wilbarger, D. N. McIntosh, and P. Winkielman, "Startle modulation in autism: positive affective stimuli enhance startle response," *Neuropsychologia*, vol. 47, no. 5, pp. 1323–1331, 2009.

[114] B. Wrede and E. Shriberg, "Relationship between dialogue acts and hot spots in meetings," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pp. 180–185, IEEE, 2003.

[115] C.-H. Wu, J.-C. Lin, W.-L. Wei, and K.-C. Cheng, "Emotion recognition from multi-modal information," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pp. 1–8, IEEE, 2013.

[116] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 201–216, 2001.

[117] A. Zlatintsi and P. Maragos, "Am-fm modulation features for music instrument signal analysis and recognition," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pp. 2035–2039, IEEE, 2012.