Imperial College London
Department of Computing

Machine Learning Techniques for Holistic Computational Paralinguistics

Yue Zhang

August 2018

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing and the Diploma of Imperial College London.
This thesis is entirely my own work, and, except where otherwise indicated,
describes my own research.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

Analysing the voice behind the words represents a central aspect of human communication and is thus the key to intelligent machines such as social robots and virtual assistants. In the realm of Computational Paralinguistics (CP), research has aimed at more natural human-computer interaction by enabling machines to recognise various speaker characteristics, social signals and other non-verbal cues. Shaped by the standard methodology in machine learning, research in CP has devised single-task learning systems for recognising dozens of speaker attributes. However, human-like machine perception demands the *holistic* analysis of paralinguistic phenomena, learning how these are interrelated and using this knowledge to gain a better understanding of the whole context. Thus, in this thesis, two major challenges in current research are addressed: (i) Overcoming the bottleneck of data annotation to meet the urgency for obtaining labelled data; (ii) Exploiting task relatedness in learning multiple tasks at the same time.

To tackle the first problem, efficient data annotation methods based on active learning and semi-supervised learning are proposed, with the aim to leverage the vast amount of unlabelled data for training while reducing manual labelling effort. In determining *whether or not* and *how many* human annotators are required in regard to label ambiguity, a generic cooperative learning framework is introduced, using measures of model confidence for human-machine arbitration and an early-stopping criterion based on interrater agreement. Techniques of this kind are particularly important for subjective tasks that do not come with ground truth labels. The effectiveness of the proposed algorithms is demonstrated on the tasks of speech emotion recognition and prosody estimation.

To enable holistic analysis in meeting the second thesis objective, it is important to know which related tasks may mutually benefit from joint learning and thus to avoid negative transfer. Building upon feature relevance analysis and non-metric dimensional scaling, a data-driven approach is proposed to construct a paralinguistic atlas, in which task similarities are transformed into distance measures in a two-dimensional space. On this basis, various multi-task learning methods are applied to related tasks to improve generalisation performance by means of data aggregation, feature relevance analysis, auxiliary labels, and shared representation learning.

In particular, two approaches have been conceived to handle the prevalence of single-label databases, which represents a main obstacle for holistic analysis. One possibility to exploit task interdependencies is to generate multi-label datasets. This is achieved with the proposed cross-task labelling method, which automatically completes the missing target labels, e. g., when performing joint learning of deception detection and sincerity classification. Alternatively, the single-label datasets at hand can be fed into a multi-task deep neural network, which learns different tasks in parallel through shared hidden layers and separate output layers. This method also effectively addresses the 'universality' aspect of the holistic concept, i. e., recognising various paralinguistic patterns such as different emotion expressions in a simultaneous way.

Acknowledgements

I would like to thank my supervisor Prof. Björn Schuller for inspiring me to carry out research into the fascinating field of Computational Paralinguistics and for the opportunity to study at Imperial College London. I wish to express my sincere gratitude to Prof. Duncan Gillies and Prof. Roger Moore for being my viva voce examiners.

Particularly, I would like to thank my colleagues and collaborators Andrea Michi, Yifan Liu, Yuxiang Zhou, Jie Shen, Caijiao Quan, Maximilian Schmitt, Florian Hönig, Harry Strömfelt and many others for the fruitful discussions and joint efforts. Thanks to all my friends for their support and the memorable time.

Furthermore, I would like to thank Ms Teresa Ng for her invaluable help in day-to-day business, and Ms Amani El-Kholy for her academic advices.

My research was supported by funding from the European Union's Framework Programme for Research and Innovation HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA) and the European Union's Seventh Framework Programme under the ERC Starting Grant No. 338164 (iHEARu).

Foremost, I thank my parents for their constant support in various aspects of life, and my husband Felix for his unconditional love, guidance, encouragement and companionship throughout the time of my PhD.

In loving memory of my grandparents.

Notation

Scalars

a	Agreement level
b	Bias
d	Distance / dissimilarity
r	Pearson's correlation coefficient
Δw	Window shift
$\hat{y}_i$	Prediction of instance i
y_i	Label of instance i
B	Batch size
C	Complexity constant
D	Number of frequency bands/ bins
F	Number of frame-level features
F_0	Fundamental frequency
G	Number of functionals
H	Number of hidden layers
J	Loss
L	Number of training instances
M	Number of MFCCs
N	Number of test instances
O	Number of output nodes

P	Number of classes
Q	Merit
R	Number of annotators
S	Number of tasks
T	Number of frames
W	Number of samples in window
Z	Confidence
α, β	Lagrange multipliers
ε	Tolerance
φ	Phase
γ	Spectral compression exponent
η	Learning rate
ζ	Slack variable
μ	Momentum
ρ	Spearman's rank correlation coefficient
ξ	Exponent of error norm
$\perp$	Undefined label

Indexing

f	Feature index
o	Annotator (oracle) index
s	Task index
t	Frame index
q	Training step
τ	Sample index

Vectors

$\mathbf{b}$	Bias vector
$\mathbf{m}$	Mel-scale spectrum
$\tilde{\mathbf{m}}$	MFCCs
$\mathbf{v}$	Sample window
$\mathbf{w}$	Weight vector
$\mathbf{x}_i$	Features of instance i
$\mathbf{z}$	DFT (complex)

Matrices

$\mathbf{A}$	Linear activation matrix
$\mathbf{H}$	Hidden activation matrix
$\mathbf{W}$	Weight matrix
$\mathbf{X}$	Feature matrix

Graphs, Sets and Spaces

$\mathcal{D}$	Directed acyclic graph
$\mathcal{E}$	Edges
$\mathcal{J}$	Subset of selected instances
$\mathcal{L}$	Training dataset
$\mathcal{R}$	Pool of annotators
$\mathcal{T}$	Test set
$\mathcal{U}$	Unlabelled dataset
$\mathcal{V}$	Vertices
$\mathcal{X}$	Feature space
$\mathcal{Y}$	Label space

Functions

h	Model
k	Kernel
u	Activation function
x	Audio signal
ψ	Non-linear transformation

List of Acronyms

AL	Active Learning
CC	Correlation Coefficient
CDCC	Cross-Domain Correlation Coefficient
CFS	Correlation-Based Feature Selection
CL	Cooperative Learning
ComParE	Computational Paralinguistic Challenge
CTL	Cross-Task Labelling
DAL	Dynamic Active Learning
DCL	Dynamic Cooperative Learning
DNN	Deep Neural Network
LLD	Low-Level Descriptor
LOSGO	Leave-One-Speaker-Group-Out
LOSO	Leave-One-Speaker-Out
MFCC	Mel-Frequency Cepstral Coefficient
MTL	Multi-Task Learning
NMDS	Non-Metric Dimensional Scaling
PL	Passive Learning
SGD	Stochastic Gradient Descent
SHL	Shared Hidden Layers

SMO Sequential Minimal Optimisation

SSL Semi-Supervised Learning

STL Single-Task Learning

SVM Support Vector Machine

SVR Support Vector Regression

UAR Unweighted Average Recall

Contents

Abstract **iii**

1 Introduction **1**
1.1 Motivation and Objectives 1
1.2 Problem Statement 3
1.3 Contributions 4
1.4 Thesis Outline 5

I Automatic Analysis of Speaker Characteristics **7**

2 Computational Paralinguistics: Overview **9**
2.1 Bridging the Non-Verbal Gap Between Human and Machine 9
2.2 Typology of Paralinguistic Phenomena 12
2.3 Databases 15
2.4 Benchmarks 23

3 Machine Analysis of Speaker Characteristics: Background **29**
3.1 Processing Chain 29
3.2 Feature Extraction 32
3.3 Machine Learning Models 36
3.4 Machine Learning Tasks 45

II Efficient Methods for Data Annotation **51**

4 Dynamic Active Learning **53**
4.1 The Sub-optimality of Majority Voting in Data Annotation 53
4.2 Dynamic Query Strategies 56
4.3 Experiments 57
4.4 Results and Discussion 59

5 A Generic Human-Machine Annotation Framework **61**
5.1 Learning From Limited Labelled Data (But Unlimited Unlabelled Data) 62
5.2 Dynamic Cooperative Learning Framework 64
5.3 Experiments 68
5.4 Results and Discussion 70

III Holistic Speaker Analysis **75**

6 A Guide to the Paralinguistic Universe **77**
6.1 The Whole and Its Parts 77
6.2 Measures of Task Relatedness 79
6.3 Non-Metric Dimensional Scaling 80
6.4 The Paralinguistic Atlas 81

7 Drowsiness Detection **85**
7.1 Alcohol Intoxication and Sleep Deprivation 85
7.2 Feature Relevance Analysis 86
7.3 Experiments 87
7.4 Results and Discussion 89

8 Language Proficiency Assessment **93**
8.1 Non-native Prosody and Native Language 93
8.2 Auxiliary Attributes 95
8.3 Experiments 95
8.4 Results and Discussion 97

9 Two Sides of the Same Coin? **99**
9.1 Deception and Sincerity 99
9.2 Learning Schemes for Related Tasks 102
9.3 Experiments 105
9.4 Results and Discussion 106

10 Holistic Emotion Recognition **109**
10.1 Breaking Down the Walls Between Emotion Representations 109
10.2 Multi-Task Shared-Hidden-Layer Nets 111
10.3 Experiments 113
10.4 Results and Discussion 115

11 Conclusions and Outlook **117**
11.1 Summary of Thesis Achievements 117
11.2 Towards Holistic Machine Perception 119

Chapter 1

Introduction

The most intelligible factor in language is not the word itself, but tone, strength, modulation, tempo with which a sequence of words is spoken — in brief, the music behind the words, the passion behind the music, the person behind the passion; everything, in other words, that cannot be written.

Friedrich Nietzsche

1.1 Motivation and Objectives

The fascination of Artificial Intelligence (AI) is more than a practical one; we form a deeper attachment to human-like AI that interacts with us in a personal way. In the science fiction film "Her", the personalised operating system named Samantha distinguishes herself in her ability to engage in conversation through the usage of non-verbal communication signals. With recent technology advances, audio, speech and language processing systems have attracted major attention due to a multitude of possible applications, such as computer-aided learning tools, safety and security appliances, as well as digital health solutions. In recent years, unprecedented developments of deep learning for acoustic and language modelling have realised the first of their kind smart devices for home automation (Amazon Echo, Google Home) and virtual assistants (Apple Siri, Google Assistant, Microsoft Cortana) in Internet of Things (IoT)-inspired environments. The functioning core of these voice-operated systems is a speech recognition and natural language processing engine, which enables computers to parse the verbal content and thereby support a wide range of user queries, e. g., performing phone actions, handling device settings, searching the Internet, scheduling events and reminders, navigating areas, and giving recommendations. Moreover, rule-based strategies for dialogue management allow contemporary virtual assistants (VAs) to sustain rudimentary conversation with human interlocutors using a

repertoire of predefined utterances. Further, with continuing use, commercial software agents return results customised to the users' individual language usages, searches, and preferences. However, despite their many practical advantages, there remains a significant gap between current computational systems and human-like machine intelligence.

One major shortcoming is that today's conversational agents cannot yet manage non-verbal communication with humans through sending and receiving wordless cues such as vocal tones. In other words, they can at best understand *what is said*, but are mostly unaware of *how the speaker says the words*, i. e., the paralanguage (from Greek para "alongside" language) in human communication that may modify, reinforce or nuance meaning, and convey information about the speaker. Due to the lack of intelligent audio analysis, the existing personal assistants are severely limited in their capability to recognise and adapt to individual speaker characteristics such as the emotional state, resulting in 'social impairment' and affective insensitivity. Another issue is that without the ability to decode the non-verbal cues carried in the tone of voice, VAs are prone to misinterpret ambiguous expression (e. g., irony) and often encounter difficulties in comprehending the actual intentions of users.

In speech processing, machine analysis (and synthesis) of *the voice behind the words* has emerged as an increasingly important field termed Computational Paralinguistics (CP) [185]. Beyond the traditional fields of automatic speech recognition and natural language processing, which are concerned with the processing and understanding of verbal content, research into paralinguistic speech phenomena aims at leveraging a fundamental meta-communication mechanism for intelligent machine listening. For over a decade, researchers in this field have worked towards AI that is aware of paralinguistic properties, inter alia prosody, pitch, volume, and intonation. These acoustic features modulate the speech signal in aspects of speaker states (e. g., affect, health condition) and traits (e. g., age, gender, personality, cultural background), social signals (e. g., laughter, sigh), as well as illocutionary force [10], i. e., the intended meaning of an utterance. Chapter 2 will expatiate on the paradigm of CP.

As a core machine perception area, computer audition makes use of data-driven machine learning methods. In paralinguistic research, tremendous effort has been undertaken in collecting and annotating task-specific databases, which serve to train models that can make predictions on unseen instances. Using these data resources, the INTERSPEECH Computational Paralinguistic Challenges (ComParE) set the benchmark for state-of-the-art recognition performances on a wide range of classification and regression tasks. The baseline computation is carried out in an elaborated workflow, starting from audio preprocessing to automatic feature extraction, to model training and evaluation. The analytical framework is embedded in a free software environment, using tools such as openSMILE [62] for automatic feature extraction from speech signals, and Weka 3 [86] for data mining. This foundation work will serve as a starting point for the thesis.

1.2 Problem Statement

The whole is greater than the sum of its parts.

ARISTOTLE

From a methodological point of view, today's systems are designed to handle separate recognition tasks, using standard methods of machine learning. Labelled data are of paramount importance for Computational Paralinguistics, like any other field intersecting with machine learning. Over the past decade, a few dozen of mostly small databases labelled in single paralinguistic dimensions have been created. However, in anticipation of the plethora of paralinguistic phenomena that are yet to be addressed, there is an urgent need to quicken the pace of research development. The main drawback is that data annotation is commonly performed by human experts, incurring the cost of manual labour and high time consumption. This is compounded by the subjectivity in determining the gold standard labels (i. e., the 'pseudo-truth' given a set of possibly noisy annotations) in regard to many paralinguistic tasks such as emotion or interest recognition. For tasks of this kind, multiple annotators are usually required to reduce personal bias by establishing an interrater agreement. Upon entering the "Big Data" era, it has become infeasible to label the massive amounts of data by hand. To satisfy the thirst for data, machine labelling is crucial to make data annotation more efficient.

Another shortcoming in this context is that paralinguistic phenomena are considered in isolation, yet, strong interdependencies exist, particularly between various speaker characteristics. This hypothesis is corroborated by evidences amassed in social sciences. Taking the example of emotion expressions [243], these are clearly shaped by the speaker's age, gender, personality, cultural background, and many other factors; an introverted person would probably not distinctly display elation, thus an emotion recogniser would need to adjust its sensitivity. Or, another example, a person who has caught a cold may sound older than he/ she is because of the hoarse voice. All these interdependencies cannot be modelled in the conventional single-task approach. Figuratively speaking, the existing, mostly desultory studies have been aimed at discovering more pieces of a puzzle symbolising a speaker's profile. However, little work has been done to examine the interconnections between the individual aspects [193], in the same way as humans perceive and analyse a speaker as a whole person. Still, before this can be achieved on a larger scale, richly labelled data need to be acquired, which brings us back to the problem of data paucity as mentioned before.

In this thesis, holism (from Greek holos "all, whole, entire") is construed as the concept that speaker characteristics are interrelated as parts of a whole, and as such, they need to be analysed in a holistic context. To quote Aristotle, "*the whole is greater than the sum of its parts*". In this spirit, the central objective of this thesis is to make machine analysis of speaker characteristics more human-like through two approaches. First, more recognition

tasks will be tackled systematically, in order to cover the manifold facets of paralinguistic phenomena. Especially, in a way that the human knowledge of speaker attributes is constantly expanding and evolving, the learning systems need to acquire new training data as efficiently as possible by means of self-learning and human-machine collaboration. Second, the interplays and synergies between paralinguistic phenomena will be examined and leveraged for holistic analysis. By learning multiple tasks at the same time and in relation to each other, large-scale data aggregation will be enabled to improve generalisation performance. In verifying the gains through holistic analysis, the thesis will lend impulses to advance CP to the next level.

1.3 Contributions

As a matter of fact, the main driving force for CP research is the paralinguistic Challenge series, paving the way for the automatic recognition of various speaker characteristics. A summary of the Interspeech Challenges 2009–2013 (with different Challenge titles until then) is given in the survey article by the author and her colleagues [206], which marks the beginning of the ComParE Challenges in retrospect. This thesis provides a complete overview of the Challenge series, including the new recognition tasks addressed by the author and her colleagues in the follow-up ComParE 2014–2017 ([199, 202, 204, 205]).

For efficient data annotation, a novel approach to active learning (dynamic active learning) is presented, which adapts the number of human annotators to the difficulty of labelling a data point such as to economically allocate human resources. To this end, an agreement-based early stopping is enabled, improving on the conventional majority voting rule. The variants of this technique were first described in the works [252, 253, 254] by the author and her colleagues.

A further boost in labelling efficiency is achieved through a novel cooperative learning algorithm, which avoids human annotation completely if the machine-predicted labels are sufficiently reliable. To capture prediction uncertainty, deep learning based confidence measures are used. In contrast to the previous approach that is limited to classification tasks [262], the new cooperative learning algorithm can also be effectively applied to regression problems, e. g., prosody estimation as studied in this thesis.

A main progress towards holistic analysis is made by exploring the paralinguistic universe, encompassing the bulk of the hitherto studied speaker attributes. As a highlight of this thesis, in a data-driven approach based on non-metric dimensional scaling, the ComParE tasks are projected onto a paralinguistic atlas, employing different distance measures to model the task relations in a two-dimensional space. The maps are oriented in different ways such as to help interpret the topological space and the task relations. Based on this guide, several clusters of the tasks are more closely examined as candidates for multi-task learning.

An effective way to determine the predictive features in cross-corpus experiments is given by feature relevance analysis. Using the selected features, data from different domains such as alcohol intoxication and sleepiness can be combined to train a joint model for general drowsiness detection. The study [261] by the author and her colleagues shows the advantage of a drowsiness classifier for safety-related applications, where the specific cause of the drowsy condition is unknown or less relevant in some circumstances.

Furthermore, a multi-task learning scheme based on auxiliary attributes is proposed, appending the labels of related tasks as additional features to the target label. This approach is evaluated for language proficiency assessment of non-native English speakers (cf. [256]), showing that the regression analysis of non-native prosody can be improved by including the knowledge of the native language of the speakers.

In most cases, auxiliary labels are not available since the existing paralinguistic databases have usually been created for single prediction tasks. To generate multi-label data, a novel cross-task labelling technique is proposed. To this end, single-task classifiers are trained using the labelled data of each database and then used to predict the missing target labels in all others. This method was first applied to the joint analysis of emotion, personality and voice likeability in the work by the author and her colleagues [258]. In this thesis, the effectiveness of this approach is demonstrated for deception and sincerity classification (cf. [257]), improving the accuracies on both tasks.

An alternative approach to label completion is the usage of multi-task shared-hidden-layer deep neural networks, as investigated by the author and her colleagues in the work [259]. In the proposed multi-task network topology, different tasks share the same features in the hidden layers, but are associated with separate output layers. This is far more efficient than using a set of single-task networks, since the input-to-hidden and hidden-to-hidden connections have to be computed only once for each input vector, and the number of parameters in each output layer is small. Given the diversity of emotion representations such as categorical and dimensional descriptors, a compact model particularly lends itself to holistic emotion recognition.

1.4 Thesis Outline

The thesis is structured as follows. Part I provides a comprehensive insight into the realm of Computational Paralinguistics. Chapter 2 overviews the paralinguistic speech databases, research Challenges and benchmark accuracies for a variety of recognition tasks. The machine learning background for paralinguistic analysis is expounded in Chapter 3, providing a detailed explanation of the state-of-the-art analytical systems and learning algorithms. Part II introduces novel techniques for efficient data annotation, with the aim to reduce human labelling work to a

minimum while retaining the same reliability of labels. To this end, dynamic query strategies in active learning settings are proposed in Chapter 4. On top, a human-machine annotation framework is introduced in Chapter 5, combining semi-supervised learning with dynamic active learning to enable dynamic cooperative learning. In Part III, the interrelations between speaker characteristics are analysed in a holistic context. Chapter 6 presents a paralinguistic atlas, which comprises different maps as a guidance for finding related tasks. The ensuing chapters provide in-depth analysis of the related tasks, including detection of alcohol intoxication and sleepiness (Chapter 7), native language identification and degree of nativeness estimation (Chapter 8), deception and sincerity classification (Chapter 9), and universal emotion recognition (Chapter 10). To exploit the interdependencies between these tasks, different multi-task learning techniques are applied to improve generalisation performance. The thesis concludes with a summary of achievements and an outlook for promising research avenues towards holistic machine perception.

Part I

Automatic Analysis of Speaker Characteristics

Chapter 2

Computational Paralinguistics: Overview

The best listeners listen between the lines.

Nina Malkin

This Chapter provides an overview of the research field of Computational Paralinguistics, which centres on the meta-information accompanying language as conveyed by the tones of voice. Paralinguistic speech phenomena encompass a plethora of speaker characteristics, ranging from transient speaker states to permanent traits. Besides the time scale, other taxonomic dimensions such as objectivity – subjectivity have been devised to codify paralinguistic patterns. Accordingly, the paralinguistic realm can be represented as a two-dimensional space spanned by the axes 'time' and 'label ambiguity'. The state of the art in this emerging field is presented in light of the ComParE series, which sets the benchmark for the performance on a large variety of prediction tasks. In addition, a description of the available resources, benchmark datasets and evaluation campaigns is provided. To start with, the related fields of Affective Computing (AC) and Social Signal Processing (SSP) are coined to draw some lines of demarcation.

2.1 Bridging the Non-Verbal Gap Between Human and Machine

Within the paradigm of human-like computing, machine learning research has aimed at more natural human-machine communication by endowing machines with perceptual and analytical abilities to understand human non-verbal behaviour. The findings of Albert Mehrabian on the relative importance of verbal and non-verbal messages on communicating feelings and attitudes

(i. e., like-dislike) have led to the often misquoted rule "total feeling = 7% verbal feeling + 38% vocal feeling + 55% facial feeling" [136]. Although this rule cannot be quantitatively generalised, it brings out the impact of paralinguistic and kinesic communication elements beyond the literal meaning of words.

The first endeavours towards emotionally intelligent machines have been undertaken in the field of AC that "relates to, arises from, or deliberately influences emotion or other affective phenomena" [153]. Affective computers are aimed at improving human affective experience by perceiving and responding to human emotion e. g., by generating emotion expressions (though it does not imply that the affective AI would possess its own emotions [220]). Coupled with new wearable technologies, many of the applications envisioned in Picard's pioneering work for computer-assisted learning, perceptual information retrieval, arts and entertainment, as well as human health and interaction have already been realised and are encroaching ever more into everyday life. Extending beyond emotional information processing, the field of SSP [234, 236] centres on the *social* (from Latin socialis "allied") *demeanour* such as dominance and rapport in interactions between humans or between human and machine. Spanning different areas of machine perception, SSP envisages multi-modal processing for the automatic detection of non-verbal cues, including *auditory cues* as well as *visual cues* such as facial expressions, physical appearance, mutual gaze, interpersonal distance, gestures and posture. According to its definition, SSP pays special attention to the social interaction between senders and receivers, who exchange behavioural signals in a conscious effort.

As an integral part of machine perception, computer audition or machine listening focusses on processing sound, music and speech signals. Speech serves as the principal mode of human communication. Drawing on observational learning and accumulated background knowledge, humans naturally develop the ability to identify speaker characteristics from acoustic and linguistic features, exposing it as being crucial in acquiring salience from the social environment in order to achieve adaptive behaviour for survival. As a transmission medium, the tone of voice makes for powerful social signals by underpinning what is said, expressing one's sentiment and feelings, as well as signalling comprehension, (dis)agreement, and intentions. On the part of an attentive listener, the voice timbre reveals a whole range of speaker attributes, encompassing temporary states and stable/ inherent traits. Through listening between the lines, humans are able to characterise a speaker by building and gradually refining a specific model. For example, if someone receives a telephone call from an unknown caller, he/ she will picture the speaker's profile, probably starting with the person's gender and age, then mental and physical state (e. g., mood, health condition), side-activities such as eating and running, and eventually forming a mental image of the speaker's physical appearance, personal background and personality traits. The decoding of non-verbal messages on the one hand, and the analysis of (subconsciously) exposed personal characteristics on the other hand, form an integral and central part of human social intelligence – and the nucleus of paralinguistic research.

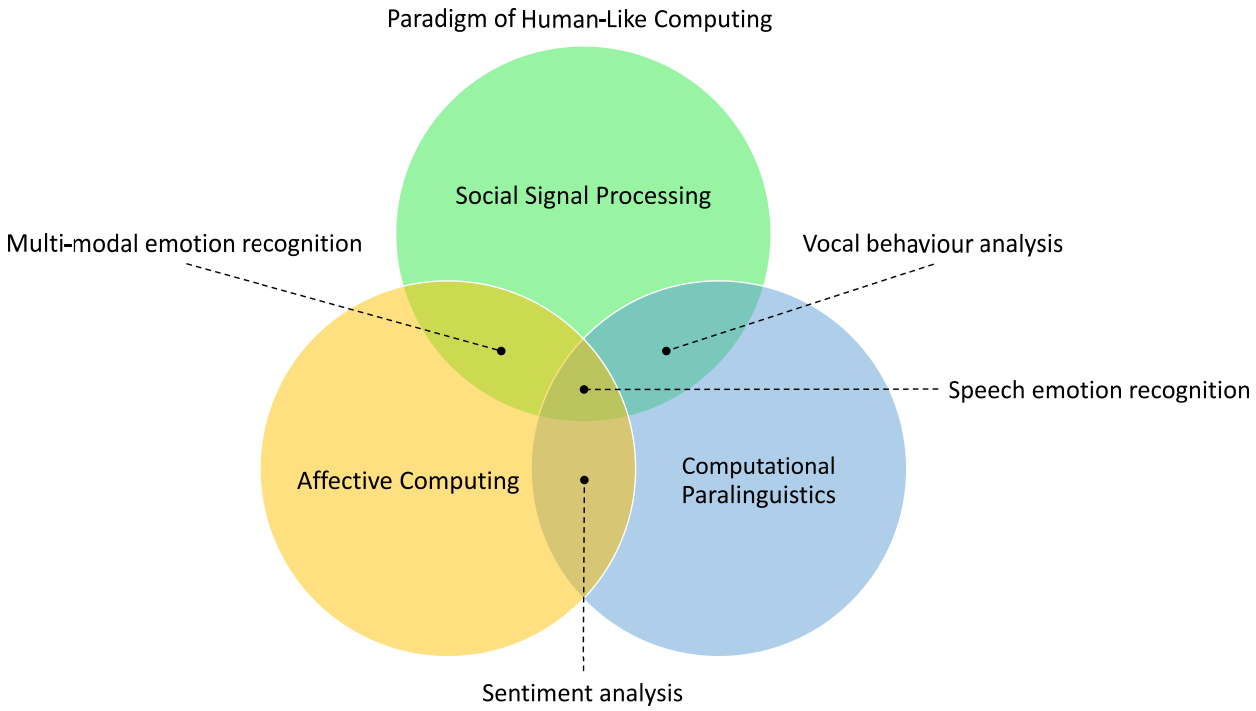


Figure 2.1: A Venn diagram showing the disciplines of Human-Like Computing.

The term 'paralanguage' can be traced back to the works of the American linguists Archibald Hill [93] and George Trager in 1958 [226]. The study of paralanguage, i. e., paralinguistics, began to flourish in the 1960s with roots in semiotics and was shaped by several foundation works [4, 37, 227]. In a broader sense, paralanguage was used synonymously with non-verbal elements in conversation according to Hill [93], subsuming kinesics (body language) [19]. In a narrow sense, the subject area of paralinguistics is confined to "variations in tone of voice", following the definition by Crystal [38], excluding the visual components of non-verbal communication such as kinesics, proxemics, and haptics. With recent developments in human-computer interaction, the traditional field of paralinguistics has seen a resurgence in popularity, entering the new era of Computational Paralinguistics [185]. The term 'computational' refers to the automatic analysis and synthesis of paralinguistic phenomena, which modulate natural language in spoken (speech) or written modality. Due to the substantially different methodologies, paralinguistic synthesis is excluded from the scope of this thesis.

Figure 2.1 illustrates a conceptual approach to coining the related fields of AC and SSP in demarcation to CP within the paradigm of human-like computing. AC lays the focus on human emotions whereas SSP covers social signals in general, thus the intersection of these fields is construed as multi-modal processing of emotional information. Because CP has a core interest in a person's emotional state, but is confined to speech and/ or text processing, sentiment analysis (i. e., opinion mining from text for valence detection [151]) is the field of application where SSP and the intersection of CP and AC are disjoint. Having the auditory modality in common, CP meets SSP in terms of voice analysis – as regards social signals. Here, it is important to note that

speaker characteristics (e. g., personality traits) can be intentionally displayed as social signals, but they can also be subconsciously (or unwillingly) revealed, hence not all paralinguistic tasks come under this intersection. On these grounds, the centrepiece shared by all three realms on the basis of the definitions by Picard [153], Vinciarelli et al. [234], Schuller and Batliner [185] is speech emotion recognition, which highlights the importance of speech in expressing emotions.

2.2 Typology of Paralinguistic Phenomena

The structure of paralinguistic patterns can be represented in various ways. Several interpretive scales have been proposed in [185] to codify paralinguistic phenomena, including time variance (short-term vs long-term) and labelling scheme (objective vs subjective). Formally, according to the Oxford English Dictionary[1], a 'trait' is defined as "a distinguishing quality or a genetically determined characteristic, typically one belonging to a person", whereas a 'state' describes "a particular condition that someone is in at a specific time". In terms of label ambiguity, a distinction is made between ground truth (which can be validated by objective facts or measurement) and gold standard (which is assigned by human annotators and thus depends on their subjective appraisal). Temporal dynamics and task subjectivity have an immediate impact on data annotation, feature extraction, and statistical modelling. The annotation of changing, subjective patterns generally requires more human effort in interrater-agreement procedures, as will be expounded in Part II of the thesis. For feature extraction, the appropriate segmentation level (frames, words, utterances) needs to be chosen based on the time granularity. In statistical modelling, the recognition of medium- and long-term phenomena can make effective use of smoothing predictions within intervals where the speaker attribute is expected to be time-invariant [245].

Figure 2.2 depicts the paralinguistic realm explored in light of the ComParE instalments, which have been held in conjunction with the annual Interspeech conference since 2009. The first Interspeech 2009 Emotion Challenge (IS09EC) [188, 192] targeted spontaneous emotion recognition cast in a binary (idle vs negative) and a five-way (anger, emphatic, neutral, positive, and rest) classification task. The follow-up Interspeech 2010 Paralinguistic Challenge (IS10PC) [189, 196], evaluated the continuous-valued level of interest ([-1,+1]) and two inherent attributes, i. e., age (child, youth, adult, senior) and gender (female, male, children). In the ensuing Interspeech 2011 Speaker State Challenge (IS11SSC) [191, 200], intoxication (above or below 0.5 per mill blood alcohol concentration) and sleepiness (above or below 7.5 on the Karolinska sleepiness scale) had to be automatically detected. Next, in the Interspeech 2012 Speaker Trait Challenge (IS12STC) [195, 203], personality (openness, conscientiousness, extraversion, agreeableness, neuroticism), likability, and speech intelligibility were addressed in

[1] https://www.oxforddictionaries.com/

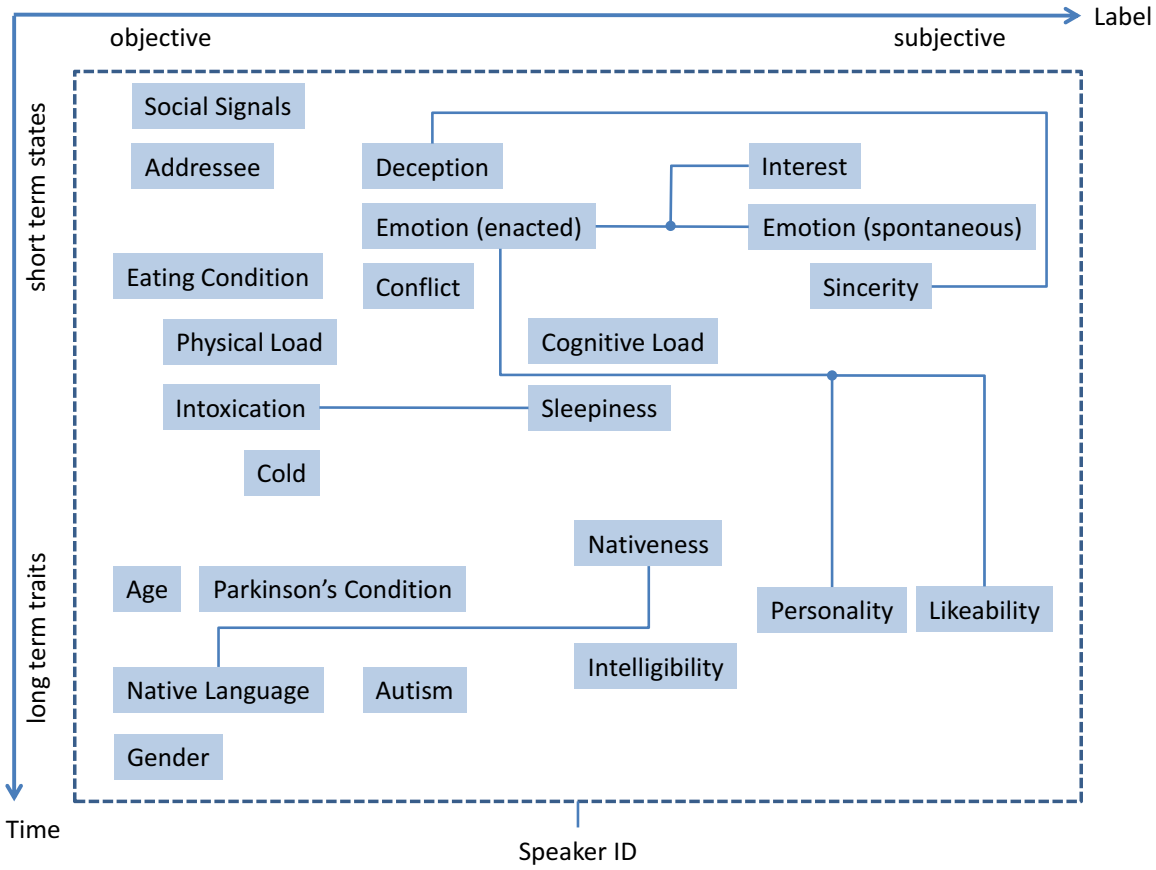


Figure 2.2: Computational Paralinguistics Challenge tasks from 2009–2017; Interdependencies between speaker attributes exploited for multi-task learning are indicated through connectors.

binary classification tasks (above or below average). Since the 2013 instalment, the Challenge series has been consistently named to ComParE, subsuming all paralinguistic tasks under one umbrella. As the only temporal localisation task to date, the ComParE 2013 featured the frame-wise detection of social signals such as laughter and fillers. Besides, the occurrence of conflict in political debates had to be classified as 'high' or 'low'. Moreover, the task of emotion recognition was revisited, with the difference that the emotional states were played by professional actors. In addition, atypical speech patterns due to pervasive developmental disorders (autism) were investigated. In the ComParE 2014, the level of cognitive load (in terms of working memory) and physical load (based on measures of heart rate and skin conductivity) were classified. The ComParE 2015 provided two regression tasks, featuring the degree of nativeness (i. e., non-native English prosody) and Parkinson's condition on the unified Parkinson's disease rating scale (UPDRS). Moreover, in the Eating Condition Sub-Challenge, six different food types, which the subjects were eating while talking at the same time, were distinguished. The ComParE 2016 posed new challenges in detecting deceptive vs non-deceptive speech, estimating the degree of sincerity, and identifying the native language out of eleven L1 classes of English L2 speakers. The ComParE 2017 introduced new baseline approaches on new classification tasks. In the Addressee Sub-Challenge, it had to be determined whether speech produced by an adult was

directed towards another adult or a child; in the Cold Sub-Challenge, speakers having a cold had to be told apart from healthy speakers.

Figure 2.2 illustrates one interpretive approach to allocating the ComParE tasks in the taxonomic space, spanned by the 'time' and 'label' axes. To establish the systematic framework, let us first examine the key points. In the upper left corner, social signals in terms of non-verbal vocalisations such as laughter or fillers often occur in fractions of a second, wherefore frame-wise localisation is required. At the other end of the time scale, gender is commonly regarded as a time-invariant, biological attribute. Farther to the right are the increasingly subjective phenomena, ranging from diagnosed, (self-)assessed, observed, to perceived characteristics. For distinction, the level of interest represents a speaker attribute from the perspective of external observers, whereas the perceived voice likeability clearly depends on individual judgement. In the centre, sleepiness was evaluated by the subjects (self-assessment) and additionally by experimental assistants using a standard psychoanalytic questionnaire, and thus represents medium-term, ascertainable conditions. Other physical conditions, which can usually be determined through biological markers such as heart rate or blood alcohol concentration, include physical load, intoxication and health condition. These measurable attributes are placed in the medium-term, objective area. In some cases, the ground truth is given by the experimental condition that induced a specific behaviour such as eating or deception. The speaker characteristics not mentioned above are interspersed within the paralinguistic realm *in interrelation with each other*, where the speaker's ID relates to all other attributes. Some attributes such as speaking styles can be construed as short-, medium- or long-term since their coordinate on the time axis highly depends on the concrete task. For instance, sincerity can be regarded as a consistent personal quality or it can refer to a transient state of mind (e. g., when apologising). It remains to mention that most of the speaker characteristics, especially affect-related, can either be enacted or spontaneous, where the label ambiguity in the latter case is arguably higher.

The connecting lines in Figure 2.2 indicate the intertwined speaker characteristics, of which the interdependencies will be leveraged for holistic analysis in the Chapters of Part III, and in the works published by the author and her collaborators, which are referenced in passing here. In Chapter 7, the coherence between the medium-term speaker states, i. e., alcohol intoxication and sleepiness, is utilised for drowsiness detection. Chapter 8 sheds light on the interrelation between the degree of nativeness and the native language of English learners, which are jointly analysed to improve the recognition accuracy in these tasks. Chapter 9 tests the hypothesis that insincerity is a concomitant of deceptive speech, and reversely, sincerity indicates non-deception. The multi-task learning technique presented in Chapter 9 has also been applied in [258] to the triad of personality, emotion and likeability. Chapter 10 deals with holistic affect recognition by unifying multiple emotion representations in model training. In the work [260], various speaker traits such as age, gender, likeability and personality are automatically recognised to determine the speaker ID in the application of speaker diarisation (i. e., assessing who speaks when).

2.3 Databases

This section provides a general description of the benchmark datasets used in the ComParE, which will be jointly used to derive the paralinguistic atlas presented in Chapter 6. Moreover, the majority of the databases will be used in the studies within the scope of this thesis. Further details will be supplied in the respective Chapters. The audio files of the databases are, if not otherwise noted, encoded and stored in 16-bit PCM WAV file format at a sampling rate of 16 kHz. In the ComParE Challenges, data partitioning has been performed so as to achieve speaker-independence and data stratification according to the available meta-information such as demographics. At the end of this section, the core information is précised in tabular form (Table 2.1).

2.3.1 FAU Aibo Emotion Corpus

The FAU (Friedrich Alexander University) "Aibo Emotion Corpus" (AEC) [216] was used in the IS09EC. It contains spontaneous, emotionally coloured speech in German. The recordings (sampling rate 48 kHz down-sampled to 16 kHz, quantisation 16 bit) were collected from 51 children (age 10-13, 21 male, 30 female) at two schools, interacting with the Sony's pet robot Aibo that was programmed to perform specific actions. To elicit specific emotions, the children were led to believe that Aibo was following their orders, where the robot was actually controlled by an unseen human operator (Wizard of Oz experiment) and sometimes disobeyed them on purpose. In total, the database comprises 18 216 instances, totalling 9.2 hours of speech data. Each instance corresponds to a manually defined chunk that consists of multiple words according to the syntactic-prosodic criteria. The annotation of emotional states was undertaken on a per word basis. The rationale is that many chunks consists of short commands with long pauses in between, during which the children might experience emotional changes, in waiting for Aibo's reaction. Five experts (advanced students of linguistics) annotated the data, totalling 48 401 words, using a given set of 11 emotional states: neutral (39 169), angry (84), touchy (225), reprimanding (310), emphatic (2,528), surprise (0), joyful (101), helpless (3), motherese (1 260), bored (11), and other (3). The numbers put in parentheses refer to the frequencies of words falling under the respective category by a majority decision of the labellers, i. e., if three or more labellers agreed, the label was assigned to the word. Consequently, there remain 4 707 words for which a majority vote could not be established. A heuristic approach was applied to map these labels from the word-level to the chunk-level. For the emotion classification task, the rare categories were subsumed under the 5 umbrella terms, i. e., Anger (including angry, touchy, reprimanding), Emphatic, Neutral, Positive (covering motherese and joyful), and Rest. Additionally, a binary classification task was given by dichotomising the 11 emotion classes to **NEG**(ative) states with negative valence and **IDL**(e) states.

2.3.2 aGender Corpus

The "aGender" Corpus [28] was created as a benchmark database for age and gender recognition. It contains 47 hours of speech data (sampling rate 8 kHz, quantisation 8 bit) collected from 945 subjects via telephone calls in mixed indoor and outdoor environment. The subjects were prompted by an automated interactive voice response system to repeat sentences or speak freely (totalling 65 364 utterances). In the Age Sub-Challenge (IS10PC), the subjects were classified into four age groups: **C**hild (7-14), **Y**outh (15-24), **A**dult(25-54), and **S**enior (55-80). As to gender classification, children (131) were considered separately from **F**emale (410) and **M**ale (404) subjects. The Cartesian product of both attributes (with exception of children), i. e., $\{C\} \cup (\{Y, A, S\} \times \{M, F\}) = \{(C), (Y, M), (Y, F), (A, M), (A, F), (S, M), (S, F)\}$, results in seven age/ gender tuples in a pseudo multi-task learning scheme. To obtain speaker-independent partitions in a roughly 40 %/ 30 %/ 30 % ratio, 25 speakers from each of the seven classes were randomly assigned to the test set, while the utterances of the other 770 speakers were divided into training and development set.

2.3.3 TUM AVIC Corpus

The TUM (Technical University of Munich) "Audiovisual Interest Corpus" (AVIC) [187] was set in the scenario that an experimenter led subjects through a commercial. The subjects' role was to naturally interact with the presenter by listening to the product presentation and ask questions of interest. The participants were specially requested not to pretend to be generally interested as a courtesy to the presenter, but to display genuine interest or disinterest. For the Affect Sub-Challenge (IS10PC), two hours of lapel microphone recordings (44.1 kHz, 16 bit) from 21 subjects (11 male, 10 female) were used. All subjects were non-native, yet advanced English speakers. The speech material was segmented into 3 880 turns, which were annotated with respect to the level of interest (LoI) by four psychologists using a five-point ordinal scale, designating disinterest, indifference, neutrality, interest, and curiosity. For the regression task, gold-standard labels were obtained by averaging the ratings and mapping the mean value to the interval [-1,1].

2.3.4 Alcohol Language Corpus

The "Alcohol Language Corpus" (ALC) [183] comprises 162 speakers (84 males, 78 females), whereof a gender balanced subset including 154 speakers was used for the intoxication detection task (IS11SSC). Speakers voluntarily underwent a systematic intoxication test. Prior to the test, every subject decided on the blood alcohol concentration (BAC) within the range between 0.28 and 1.75 per mill. The required amount of alcohol according to the individuals' biometric

data (e. g., body mass, percentage body fat, age, gender) was estimated using the Watson- and Widmark formula [244]. The BAC was measured 20 minutes after alcohol intake. Immediately thereafter, a speech test was performed, which lasted no longer than 15 minutes to avoid significant drop in BAC. The corpus further contains control recordings from the subjects in sober condition, produced in the same automotive acoustic environment. For the classification task, the measured BAC values were dichotomised to 'non-alcoholised' (NAL) and 'alcoholised' (AL) at the threshold value of 0.5 per mill, according to the legal drink-drive limit in Germany. The dataset (12 360 instances, totalling 39.1 hours) was down-sampled from 44.1 kHz to 16 kHz.

2.3.5 Sleepy Language Corpus

The "Sleepy Language Corpus" (SLC) [121] consists of 21.3 hours of recordings collected in sleep deprivation studies involving 99 subjects (43 males, 56 females). The speech recordings include sustained vowels, read stories, commands and control in a driver assistance system, simulated pilot-air traffic controller communication statements, and free speech (e. g., describing pictures, ad-lib using prepared lecture slides). A well-established sleepiness questionnaire based on the Karolinska sleepiness scale (KSS) was completed by the subjects (self-assessment) and additionally by two assessors (observer-assessment). The scores range from 1 (extremely alert) to 10 (cannot stay awake) were mapped to the binary states 'not sleepy' (NSL) and 'sleepy' (SL). The threshold value of 7.5, i. e., between level 7 ('sleepy, some effort to stay awake') and level 8 ('very sleepy, great effort to stay awake'), has been found to be a determining indicator for the occurrence of microsleep events, which entail a marked increase of accident risk [106]. The dataset is composed of 9 089 instances, after having removed the instances comprising linguistic cues to the speaker's sleepy state (e. g., "I'm very tired").

2.3.6 Speaker Personality Corpus

The "Speaker Personality Corpus" (SPC) [140] incorporates 640 clips (total duration 1.7 hours) collected from the French speaking Swiss national broadcast. Each clip belongs to a single speaker out of 322 individuals (263 males, 59 females), where some speakers appeared in two or more clips. Eleven assessors annotated each clip with the speaker's personality traits using the 10-item short version of the Big Five Inventory (BFI-10) [161], within the range of $[-4, 4]$. The Big-Five personality dimensions [15] consist of **O**penness to experience, **C**onscientiousness, **E**xtroversion, **A**greeableness, and **N**euroticism (emotional instability). In the Personality Sub-Challenge (IS12STC), the personality scores from each annotator were discretised into two counterparts as above vs below the average of their ratings for the respective trait X $\in$ { O, C, E, A, N }. The gold-standard labels were obtained through majority voting.

2.3.7 Speaker Likability Database

The "Speaker Likability Database" (SLD) [29] is an age and gender balanced subset comprising 800 speakers (397 males, 403 females) from the German Agender database [28] (cf. Section 2.3.2). The dataset is composed of the longest utterance from each speaker, totalling 40 minutes. Likability ratings of each individual speaker's voice were obtained from 32 annotators using a 7-point Likert scale, where each annotator only undertook three of the age/ gender pairs, excluding the child category (cf. Section 2.3.2). Based on the resulting 16 ratings for each instance, the gold-standard labels were derived from the evaluator weighted estimator (EWE) [79]. In the Likability Sub-Challenge (IS2012STC), the continuous valued labels were discretised into the two classes 'likable' (L) and 'non-likable' (NL) with the median of the EWE scores as threshold. Based on the partition from the IS10PC, the training set was retained, while 20 % of the speakers from the validation set were moved to the test set in a stratified way.

2.3.8 NKI CCRT Speech Corpus

For intelligibility recognition in the IS12STC Pathology Sub-Challenge, the "NKI CCRT Speech Corpus" (NCSC) [32] was employed. It contains 2.1 hours of recordings (44.1 kHz, 24-bit) collected in a sound-treated room from 55 patients (45 males, 10 females) who were treated for cancer of the head and neck. The speech intelligibility on a 7-point scale was rated by 13 speech pathologists. As in the Likability Sub-Challenge, the EWE was computed and discretised into binary class labels, i. e., intelligible (I) vs non-intelligible (NI), divided at the median of the distribution. The dataset (2 386 instances in total) was stratified and partitioned in a roughly 40 % / 30 % / 30 % ratio.

2.3.9 SSPNet Vocalisation Corpus

The ComParE 2013 Social Signals Sub-Challenge was carried out on the "SSPNet Vocalisation Corpus" (SVC) [237], which contains 2 763 audio clips of 11 seconds (total duration 8.4 hours). The corpus was extracted from a collection of 60 phone calls between 120 subjects (57 male, 63 female), who were mutually unacquainted and conversed on specific topics. The audio tracks were recorded with separate microphones of the phones. The data was manually segmented so that each segment contains at least one spontaneous vocalisation (laughter or filler) between $t = 1.5$ seconds and $t = 9.5$ seconds. In total, the dataset contains 2 988 filler events and 1 158 laughter events, which had to be detected on frame-level in the localisation task. The dataset was divided into speaker disjoint partitions by using calls 1–35 for training, calls 36–45 for development, and calls 46–60 for testing.

2.3.10 SSPNet Conflict Corpus

The "SSPNet Conflict Corpus" (SC^2) [118] was extracted from the Canal 9 Corpus [233], a collection of 45 Swiss political debates (in French). The dataset (total duration 11.9 hours) was segmented into 1 430 non-overlapping, 30 seconds long clips in which at least two speakers were involved. From the total of 110 speakers, there were 18 females (1 moderator and 17 participants) and 92 males (1 moderator and 91 participants). The clips were annotated with a questionnaire including physical (objective observation) and inferential (subjective interpretation) questions about the occurrence of conflict [235]. The annotation was performed by 551 labellers via the Amazon crowdsourcing system, where each clip was rated by 10 assessors using the interval scale $[-10, +10]$. A classification task was specified based on the average ratings, which were discretised into 'high' (≥ 0) or 'low' (< 0) presence of conflict. The training set comprises 55 % of the data, the validation set 17 % and the test set 28 %, where a strict speaker-disjoint partition was not feasible on this dataset.

2.3.11 Geneva Multimodal Emotion Portrayals (GEMEP)

In the ComParE 2013 Emotion Sub-Challenge, the "Geneva Multimodal Emotion Portrayals" (GEMEP) corpus [13] served for benchmarking. In contrast to FAU AEC that targets spontaneous emotion, it features 18 different emotional expressions performed by 10 professional actors (gender balanced), including admiration, amusement, anxiety, cold anger, contempt, despair, disgust, elation, hot anger, interest, panic fear, pleasure, pride, relief, sadness, shame, surprise, tenderness. To avoid verbal content in the speech material, sustained vowel phonations and 'nonsensical' phrases were pronounced by the actors with varying degrees of emotional intensity (normal, low intensity, masked). The class labels correspond to the intended emotion portrayals, where 6 under-represented emotion classes were excluded from the 12-class problem. Additionally, the multi-class labels were mapped to positive/negative arousal/valence. The partitioning of the 1 260 utterances complied with text independence, but was not speaker-disjoint.

2.3.12 Child Pathological Speech Database

The "Child Pathological Speech Database" (CPSD) [170] was created at two university hospitals of child and adolescent psychiatry in France. It contains 2 542 speech recordings (total duration 1 hour). From the total 99 subjects (81 males, 18 females) between 6 to 18 years of age, 35 were diagnosed with autism spectrum condition, which is classified into pervasive development disorder (PDD), specific language impairment such as dysphasia (DYS) and PDD non-otherwise specified (NOS). In addition, a control group of 64 typically developing children without any speech, hearing or general learning impairment were recorded. In the Autism Sub-Challenge,

two evaluation tasks were specified: a binary 'Typicality' task (typically vs atypically developing children), and a four-way 'Diagnosis' task (classifying into the above named categories), which was considered as the official baseline.

2.3.13 Cognitive Load with Speech and EGG

The "Cognitive Load with Speech and EGG" (CLSE) database [251] was created for studying how the mental demand experienced while performing a cognitive task affects the voice source. Central to the concept of cognitive load is the notion of 'working memory', which refers to the brain's limited capacity for storing and processing temporary information [11]. The database design was based on three types of tasks: two of these were variations of the Stroop test (saying the ink colour of words denoting names of colours) and the third one was a reading span task (verifying the logic of sentences while memorising a sequence of letters). To induce ternary load levels, the task difficulty was increased by means of incongruence between the font colours and colour names, time pressure or dual-task (as regards the Stroop test), and the number of letters to be recalled (in the reading span task). In total, the CLSE comprises 2 418 utterances, totalling 5.6 hours of recording time, from 26 subjects (10 males, 6 females).

2.3.14 Munich Bio-voice Corpus

The "Munich Bio-voice Corpus" (MBC) [198] contains 0.4 hours of read speech in English/ German from 19 subjects (15 males, 4 females). Beside audio recordings (96 kHz, 24-bit), physiological signals such as heart rate and skin conductivity were measured to monitor the physical condition of the subjects. Based on the sensory data, two physical load states ('low' vs 'high') were induced, i. e., before and after exercises such as fast stair-climbing and running. The dataset (1 088 instances in total) was stratified according to gender, age and body mass index.

2.3.15 Nativeness Corpora

The Degree of Nativeness Sub-Challenge was carried out in realistic cross-corpus settings with four databases containing non-native speech in English. The training data were taken from the "Automatic Web-based Learner-feedback" (AUWL) corpus [100] and the "Interactive Spoken Language Education" (ISLE) corpus [138]. In AUWL, English learners practised pre-scripted dialogues using a web-based dialogue training tool. It comprises 3 732 speech recordings from 31 speakers (18 males, 13 females), totalling 5.5 hours. Five phoneticians rated the nativeness of each utterance in terms of sentence melody and rhythm on a 5-point Likert scale, ranging from 1 (normal) to 5 (very unusual). The prosody scores were computed as the average of individual

ratings (mean 1.7; std. 0.5; range 1.0–3.8). From the ISLE corpus, 0.3 hours of read speech from 36 non-native English speakers (25 males, 11 females) was extracted. Two English teachers annotated the 158 speech files in terms of language proficiency using a 5-point scale ranging from 1 (beginner) to 5 (fluent) (mean 2.1; std. 0.5; range 1.3–3.4). The "Computer-assisted Pronunciation and Dialogue Training" (C-AuDiT) corpus [100] was used as the test set in the 'Brave New Approach' task and as the validation set in the 'Nativeness' task. It comprises 2.7 hours of read speech from 58 speakers (27 males, 31 females). The 999 utterances were annotated by 21 native English speakers with regard to pronunciation quality on a 3-point scale, ranging from 0 for 'good' to 2 for 'bad' (mean 0.5; std. 0.3; range 0.0–1.6). Although this scale differed from the one used for the training set, it was still valid for evaluation (on a common set of 290 sentences, the association between the two scales corresponds to a Spearman's rank correlation coefficient $\rho = 0.73$). The "Nativeness Corpus" (NC) [36] served as the test set for the Challenge baseline. It contains read speech from 54 speakers (26 males, 28 females) with varying degree of English proficiency. This dataset comprises 594 speech files, totalling 1.4 hours of speech. The prosody scores were obtained in the same way as AUWL from 27 native English speakers (mean 2.9; std. 0.7; range 1.1 to 5.0). More details of the databases will be provided in Chapter 5 and Chapter 8.

2.3.16 UdeA Parkinson's Disease Corpus

The "UdeA (Universidad de Antioquia) Parkinson's Disease Corpus" (UPDC) [149] contains 1.9 hours of audio recordings from 50 subjects (gender-balanced) diagnosed with Parkinson's condition. All subjects were Colombian Spanish native speakers. A total of 2 100 speech files (44.1 kHz, 16-bit) were collected in a sound proof booth. For evaluating phonation, articulation and prosody, the database design was targeted at repetition of vowels/ syllables/ words/ sentences, read and spontaneous speech. The labels correspond to the UPDRS scale [215], with a mean of 38.5 and a standard deviation of 19.1 (range 5 to 92). The dataset was partitioned by including 35 of the patients in the training set used for cross-validation, and assigning the other 15 subjects to the test set.

2.3.17 iHEARu-EAT Database

The audio-visual "iHEARu-EAT" database [90] features read and spontaneous speech which was produced at the same time while the subjects were eating different foods. The ground truth labels correspond to the eating condition, i. e., apple, nectarine, banana, crisps, biscuit, gummy bear, or no food. The database comprises 1 414 instances, totalling 2.9 hours of speech data, collected from 30 speakers (gender-balances). In the Eating Condition Sub-Challenge, the

dataset was split into training set (20 speakers) used for leave-one-speaker-out cross-validation (LOSO-CV) and test set (10 speakers), which were stratified by age and gender.

2.3.18 Deceptive Speech Database

The "Deceptive Speech Database" (DSD) was introduced in the ComParE 2016 [204]. In a real scenario, 72 subjects (41 males, 31 females) were instructed to play an 'imposter' by using a falsified identity to 'steal' a key from an office, or an innocent person who took a leaflet from the same office with their true identity. The ensuing interview was conducted by an embodied conversational agent (ECA). The interview process consisted of an introduction part, in which all participants truthfully answered 10 general questions, and an 'interrogation' part, in which the subjects were questioned about the theft by means of e. g., direct accusations and recognition test of the stolen items. In this experimental phase, the 'imposters' were asked to deliberately lie, while the 'innocent' subjects naturally told the truth. Given this layout, the ground truth, i. e., deception (D) vs non-deception (ND), was inferred from the subjects' role and the interview scope. Consequently, there is a roughly 1:3 distribution of the D/ ND classes. The dataset contains 1 555 utterances, totalling 2.8 hours.

2.3.19 Sincerity Speech Corpus

The "Sincerity Speech Corpus" (SSC) was designed to study perceived sincerity in the context of apology [204]. The dataset contains 911 utterances from 32 speakers (15 males, 17 females), with a total duration of 2.3 hours. Each speaker read six sentences of apologies in different prosodic styles, e. g., monotonic, non-monotonic, slow and fast. Each instance was rated in terms of perceived sincerity by multiple (minimum 13, maximum 19) annotators, using a 5-point Likert scale, ranging from 0 (not sincere at all) to 4 (extremely sincere). To eliminate individual biases, the ratings of each annotator were standardised to zero mean and unit standard deviation. The gold-standard labels correspond to the average of standardised ratings of individual annotators.

2.3.20 ETS Corpus of Non-Native Spoken English

The "Educational Testing Service" (ETS) Corpus of Non-Native Spoken English [204] was designed specifically to support the task of native language (L1) identification. It consists of prompted speech sampled evenly from speakers of 11 different L1 backgrounds (cf. Chapter 8), who participated in a global assessment of English language proficiency. A total of 5 132 spoken responses, one per speaker and each 45 seconds long, were collected through ETS' operational test delivery system for the Test of English as a Foreign Language (TOEFL).

2.3.21 Homebank Child/Adult Addressee Corpus

The "Homebank Child/ Adult Addressee Corpus" (HB-CHAAC) was taken from the repository for long-form real-world audio recordings of children [241], which has been created from a longitudinal study of infant learning and development, as well as parent-child interaction. It consists of conversations between toddlers and their caregivers during their daily experiences at home. Recordings were collected using the LENA recording device and software [78] that provides automated identification of 'conversational blocks' (i. e., speech periods delimited by 5 s of non-speech on either end) as well as speaker diarisation. The audio clips containing adult speech were annotated by three trained assessors who determined whether the adult was speaking to a child (child-directed speech, CDS) or another adult (adult-directed speech, ADS), based on acoustic-phonetic information and context. The addressee class labels were obtained by majority voting of the judges; clips deemed to be non-speech, not produced by an adult, or ambiguous between CDS/ ADS were removed from the dataset. In total, the dataset contains 4.8 hours of speech data in 10 886 segments from 61 speakers (32 males, 29 females).

2.3.22 Upper Respiratory Tract Infection Corpus

The "Upper Respiratory Tract Infection Corpus" (URTIC) contains 44.4 hours of recordings from 630 subjects (382 males, 248 females). The recording tasks include reading stories, telling about specific topics, describing pictures, giving a talk etc. The speech files were segmented into 28 652 chunks with a duration between 3 s and 10 s. To check the health condition, each subject underwent a self-assessment using the Wisconsin upper respiratory symptom survey [14], based on a illness severity scale from 0 (not sick) to 7 (severely sick). The ratings were mapped to 'cold' (C) vs 'non-cold' (NC) using the threshold value 6. Speaker-independent partitions were obtained by assigning the same number of speakers to the train/ validation/ test set, stratifying by gender, age and experimenter ID.

2.4 Benchmarks

In computing the baseline results, comparability, transparency and reproducibility, as well as research validation through peer-review have been retained throughout the ComParE series. Comparability was achieved through consistent acoustic feature sets, test bench and evaluation measures. Test data – without ground-truth/ gold-standard labels – were given to the Challenge participants, who had a limited number of attempts to submit their model predictions for the test instances. To underpin transparency, the openSMILE feature extraction toolkit [61, 62] and the open-source implementations of machine learning algorithms provided by the Weka 3 data

mining software [248] have been consistently used over the years. Moreover, a 'recipe' has been released, which embeds the entire workflow from pre-processing (e. g., feature standardisation, up-sampling) over model training to evaluation. In this way, the baseline results on the validation set can be reproduced in a straightforward way.

Depending on the type of recognition task, different evaluation measures have been used in the Challenges. For classification, unweighted average recall (UAR) has been considered. Given a sequence of discrete labels $y_1, \ldots, y_N$ and a sequence of predicted labels $\hat{y}_1, \ldots, \hat{y}_N$ for the N test instances, UAR is defined as:

$$\mathrm{UAR} = \frac{1}{P} \sum_{c=1}^{P} \mathrm{Recall}(c), \tag{2.1}$$

$$\mathrm{Recall}(c) = \frac{\sum_{i:y_i=c} \delta(\hat{y}_i, y_i)}{N_c}, \tag{2.2}$$

where δ is the Kronecker delta, P is the number of classes and $N_c = |\{i : y_i = c\}|$ is the number of test instances with class label c. If the distribution of class labels is non-uniform, UAR can be more meaningful than the straightforward measure of accuracy (ACC):

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N} \delta(\hat{y}_i, y_i)}{N}. \tag{2.3}$$

The rationale is that picking the class with the highest prior N_c/N can result in high accuracy, whereas it leads to chance level UAR.

For binary classification, i. e., $c \in \{0, 1\}$, an alternative evaluation method is to determine the receiver operating characteristic (ROC). A prerequisite for this is that the classifier output $\hat{y}_i$ is the probability of the positive class per instance, as is the case e. g., for logistic regression. By assigning all instances i with $\hat{y}_i \geq \theta$ to the positive class with a threshold value $0 \leq \theta \leq 1$, the true positive rate (TPR) and false positive rate (FPR) can be computed as a function of θ. TPR is equivalent to the recall of the positive class (Eq. 2.2), while FPR can be defined as $1 - \mathrm{Recall}(0)$. The ROC curve is the function resulting from interpolating the (TPR, FPR) points in the area $[0, 1] \times [0, 1]$, which are obtained by increasing θ gradually from 0 to 1. Using a numeric method such as the rectangle rule, one can integrate the ROC curve over the interval $[0, 1]$ to compute the *area under the curve* (AUC) measure, which ranges from 0 to 1. The advantage of AUC is that it is independent of the threshold parameter θ, in contrast to ACC and UAR. For this reason, it is suitable for tasks such as social signal detection, where the positive instances are rare. To generalise the AUC measure to the multi-class scenario, one can consider a 'one-versus-all' approach, picking a class $c \in \{1, \ldots, P\}$ at a time as the 'positive' class and the union of all other classes as the 'negative' class. The average of the class-wise AUC measures can be called UAAUC (unweighted average area under the curve) [197].

For regression tasks, the standard measures are correlation coefficients (CC) and mean errors. A commonly used CC is the Pearson's r, which is defined as:

$$r = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})^2 \cdot \sum_{i=1}^{N}(y_i - \overline{y})^2}}, \tag{2.4}$$

with $\overline{\hat{y}}$ denoting the average prediction, and $\overline{y}$ denoting the average test set label. The above equation can be rewritten as:

$$r = \frac{\mathrm{Cov}(\hat{Y}, Y)}{\sigma_{\hat{Y}}\sigma_Y}, \tag{2.5}$$

where Cov and σ signify the sample covariance and sample standard deviation, respectively. Thus, it holds that $r \in [-1, +1]$.

Spearman's ρ is computed by first ranking the predictions and test labels according to the $\geq$ relation, and then calculating Pearson's CC of the ranks. In case of ties (i. e., one or more elements having equal values), the average rank is assigned (fractional ranking). The main advantage of ρ over r is its robustness to the data distribution, especially the presence of outliers.

The evaluation of a regressor can be complemented by measuring mean errors, of which the generic form is given by:

$$\sqrt[\xi]{\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|^{\xi}}. \tag{2.6}$$

The standard choices for ξ are 1 and 2. $\xi = 1$ yields the mean linear error (MLE) which is also known as the mean absolute error (MAE). $\xi = 2$ results in the root mean squared error (RMSE). Compared to CCs, mean errors are scale variant, i. e., their value depends on the label range. For example, when choosing a five-point scale ranging from -2 to 2 vs a scale normalised to the interval $[-1, 1]$ for determining arousal, the MAE and RMSE are expected to be doubled in the former case. This makes mean errors little use for cross-corpus experiments with different label scales, as given e. g., in the 'Nativeness' task. However, especially the MAE is advantageous in terms of intuitive interpretation, since it measures the average deviation from the label on the chosen scale, such as years for age estimation.

An overview of the benchmarking results is given in Table 2.2.

Table 2.1: Overview of databases used in ComParE 2009 – 2017.

Corpus	Challenge Task	Content	# Inst	hh:mm	#Subj	Label	Cond	kHz/bit
FAU AEC [216]	IS09EC Emotion, Negativity	German free	18 216	09:12	21 m 30 f	GS (5) observed	studio	48 16
aGender [28]	IS10PC Age, Gender	German mixed	65 364	47:00	404 m 410 f 131 x	GT demographic	telephone	8 8
AVIC [187]	IS10PC Affect	English free	3 880	2:17	11 m 10 f	GS (4) observed	studio	44.1 16
ALC [183]	IS11SSC Intoxication	German mixed	12 360	39:05	84 m 78 f	GT measured	automotive	44.1 16
SLC [121]	IS11SSC Sleepiness	German mixed	9 089	21:16	43 m 56 f	GS (2) assessed	studio	44.1 16
SPC [140]	IS12STC Personality	French free	640	01:43	263 m 59 f	GS (11) assessed	broadcast	8 16
SLD [29]	IS12STC Likability	German scripted	800	00:40	397 m 403 f	GS (32) perceived	telephone	8 8
NCSC [32]	IS12STC Intelligibility	Dutch read	2 386	02:05	45 m 10 f	GS (13) perceived	studio	44.1 24
SVC [237]	ComParE 2013 Social Signals	English free	2 763	08:25	57 m 63 f	GS (2) observed	telephone	16 16
SC² [118]	ComParE 2013 Conflict	French free	1 430	11:55	92 m 18 f	GS (551) observed	television	16 16
GEMEP [13]	ComParE 2013 Emotion	nonsensical prompt	1 260	00:52	5 m 5 f	GT enacted	studio	16 16
CPSD [170]	ComParE 2013 Autism	French free	2 542	01:02	81 m 18 f	GS diagnosed	mixed	16 16
CSLE [251]	ComParE 2014 Cognitive Load	English mixed	2 418	05:34	20 m 6 f	GS (1) self-assessed	studio	16 16
MBC [198]	ComParE 2014 Physical Load	German read	1 088	00:22	15 m 4 f	GT measured	studio	96 24
AUWL [100],**ISLE** [138], **C-AuDiT** [100], **NC** [36]	ComParE 2015 Degree of Nativeness	English read	5 483	09:51	96 m 83 f	GS (5–27) perceived	mixed	16 16
UPDC [149]	ComParE 2015 Parkinson	Spanish mixed	2 562	02:37	25 m 25 f	GS diagnosed	studio	44.1 16
iHEARu-EAT [90]	ComParE 2015 Eating Condition	German mixed	1 414	02:53	15 m 15 f	GT exp. cond.	studio	16 16
DSD [204]	ComParE 2016 Deception	English free	1 555	02:47	41 m 31 f	GT exp. cond.	studio	16 16
SSC [204]	ComParE 2016 Sincerity	English read	911	02:20	15 m 17 f	GT (13–15) perceived	normal	16 16
ETS [204]	ComParE 2016 Native Language	English prompt	5 132	66:21	2 570 m 1 938 f 624 NA	GT demographic	normal	16 16
HB-CHAAC [241]	ComParE 2017 Addressee	English free	10 886	04:48	32 m 29 f	GS (3) observed	home	16 16
URTIC [205]	ComParE 2017 Cold	German mixed	28 652	44:24	382 m 248 f	GS (1) self-assessed	studio	44.1 16

Abbreviations: Inst: instances; Subj: subjects (m: male, f: female, x: child); Cond: recording condition; GS (R): gold standard (R = number of annotators); GT: ground-truth; exp. cond.: experimental condition

Table 2.2: Overview of the INTERSPEECH ComParE 2009-2017 Challenges. Evaluation measures: unweighted average recall (UAR [%]); Correlation coefficient (*CC); unweighted average area under curve ($^+$UAAUC [%]) Base: baseline. Best: best participant results.

Challenge	**Tasks**	**Description**	**Labels**	**Base**	**Best**
ComParE 2017	Addressee	Adult talks to adult or child	Adult-Directed (AD), Child-Directed (CD)	70.2	–
	Cold	Upper respiration tract infection	Cold (C), Non-Cold (NC)	71.0	72.0 [73]
ComParE 2016	Deception	Detecting deceptive speech	Deceptive (D), Non-Deceptive (ND)	68.3	72.1 [142]
	Sincerity	Degree of sincerity when apologising	[-2.5, 1.7]	.602*	.640* [114]
	Native Language	Identifying L1 of English L2 speakers	Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR)	47.5	81.3 [1]
ComParE 2015	Nativeness	Scoring prosody of English L2 speakers	[1.0, 5.0]	.415*	.735* [20]
	Parkinson's Condition	Unified Parkinsons disease rating	[5, 92]	.491*	.649* [81]
	Eating Condition	Distinguishing food types along speech	No Food, Apple, Banana, Biscuit, Crisps, Nectarine, Gummi Bear	65.9	83.1 [116]
ComParE 2014	Cognitive Load	Working memory	Low (L1), Medium (L2), High (L3)	61.6	68.9 [207]
	Physical Load	Speaking before and after exercise	Low (L), High (H)	71.9	75.4 [115]
ComParE 2013	Social Signals	Localisation of vocalisation segments	Garbage, Laughter, Filler	83.3$^+$	91.5$^+$ [83]
	Conflict	Conflict in dyadic group discussions	Low $\equiv [10, 0[$, High $\equiv [0, +10]$	80.8	83.9 [164]
	Emotion	Enacted emotion classification	Amusement, Anxiety, Cold Anger, Despair, Elation, Hot Anger, Interest, Panic Fear, Pleasure, Pride, Relief, Sadness	40.9	42.3 [72]
	Autism	Diagnosis	Typically developing (TYP), Pervasive Developmental Disorders (PDD), Non-Otherwise Specified (NOS), Dysphasia (DYS)	67.1	69.4 [9]
IS12STC	Personality	Openness	Open (O), Non-Open (NO)	59.0	62.5
		Conscientiousness	Conscientious (C), Non-Conscientious (NC)	79.1	80.1
		Extraversion	Extroverted (E), Non-Extroverted (NE)	75.3	79.2
		Agreeableness	Agreeable (A), Non-Agreeable (NA)	64.2	66.8
		Neuroticism	Neurotic (N), Non-Neurotic (NN)	64.0	69.2
		Mean	OCEAN	68.3	71.6 [107]
	Likeability	Sentiment towards voice agreeableness	Likable (L), Non-Likable (NL)	59.0	65.8 [141]
	Intelligibility	Intelligibility of pathological speakers	Intelligible (I), Non-Intelligible (NI)	68.9	76.8 [117]
IS11SSC	Intoxication	Blood Alcohol Concentration (BAC)	Alcoholised (AL) $\equiv$ BAC ‰ $\in [0, 0.5]$, Non-Alcoholised (NAL) $\equiv$ BAC ‰ $\in]0.5, 1.75]$	65.9	70.5 [22]
	Sleepiness	Karolinska Sleepiness Scale (KSS)	Sleepy (S) $\equiv$ KSS $\in [1, 7.5]$, Non-Sleepy (NS) $\equiv$ KSS $\in]7.5, 10]$	70.3	71.7 [103]
IS10PC	Age	Age group classification	Child (7-14), Youth (15-24), Adult (25-54), Senior (55-80)	48.9	52.4 [120]
	Gender	Gender classification	Female, Male, Children	81.2	84.3 [137]
	Affect	Continuous-valued level of interest	[-1, 1]	.421*	.428* [110]
IS09EC	Emotion	Spontaneous emotion classification	Anger, Emphatic, Neutral, Positive, and Rest	38.2	41.6 [53]
	Negativity	Negative valence	Idle (IDL), Negative (NEG)	67.7	70.3 [127]

Chapter 3

Machine Analysis of Speaker Characteristics: Background

In A.I., the holy grail was how do you generate internal representations.

GEOFFREY HINTON

This Chapter expounds the machine learning background for paralinguistic analysis. To start with, the steps in the processing chain are described, from getting an input audio to segmentation, to acoustic feature extraction, and to model training. The loop is closed by feeding back the machine- or human-labelled instances into the training database. Then, the methods of feature extraction and selection are detailed and the feature sets used in Computational Paralinguistics are presented. With respect to the objectives of the thesis, the state-of-the-art machine learning techniques for training recognition models, data aggregation and multi-task learning are surveyed.

3.1 Processing Chain

Figure 3.1 depicts the processing chain of the paralinguistic analysis systems, which are composed of the following building blocks. The audio databases provide digitised waveforms of speech, which can be either labelled data for model training or unlabelled data collected from various sources such as the public domain or users of the system. The raw data are preprocessed by means of e. g., energy normalisation, resampling and/ or de-noising in order to increase the robustness to various acoustic and recording conditions.

Next, the input audio is segmented according to a meaningful time unit of analysis, e. g., phonemes, syllables, words, phrases, complete sentences or even entire speaker turns in a dialogue,

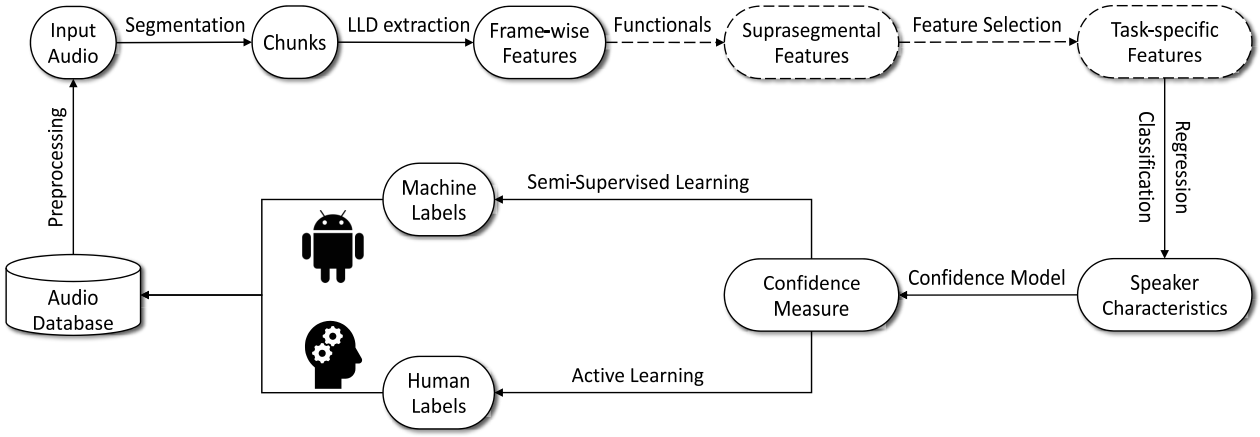


Figure 3.1: Processing chain of paralinguistic analysis systems. Dashed lines indicate optional modules.

which are subsumed under the term 'chunks'. To this end, voice activity detection (VAD) or structural analysis of syntactic or semantic entities is conducted. The appropriate method of segmentation depends on the temporality of the paralinguistic pattern regarded (cf. Chapter 2).

Subsequently, frame-level features are extracted from overlapping fixed-size windows of audio samples, typically at 100 frames per second with a window size of 10–30 ms. Denoting the input audio signal by $x(\tau)$, each frame corresponds to a vector $\mathbf{v}_t = (x(t\Delta w), \ldots, x(t\Delta w + W - 1))$, where τ is the sample index, t is the frame index, Δw is the window shift in samples, and W is the number of samples in each window. The features computed from each frame are also referred to as low-level descriptors (LLDs). The stacking of LLDs yields a F-dimensional vector for each frame, where F is the number of frame-level features (e. g., band energies, fundamental frequency, formants). The column-wise concatenation of these LLD vectors results in a matrix $\mathbf{X} = (x_{f,t}) \in \mathbb{R}^{F \times T}$, where T is the number of frames in a segment. Consequently, the rows of the matrix $\mathbf{X}$ correspond to finite sequences $(x_{f,t})_t = (x_{f,1}, \ldots, x_{f,T})$, which describe the temporal evolution – also called 'contour' – of the LLDs. For noise reduction, the contours can be smoothed by applying a low-pass filter such as a moving average filter.

For short-term paralinguistic phenomena such as social signals or affect, the LLDs or small sliding windows of LLDs can be directly used as acoustic features. However, for most paralinguistic analysis purposes, each LLD contour is represented by a fixed number G of statistic functionals $g_j, j = 1, \ldots, G$ such as various types of means (arithmetic, geometric, etc.), moments (standard deviation, skewness, kurtosis, etc.), percentiles (median, quartiles, etc.). As a result, each speech segment can be represented as a $(F \cdot G)$-dimensional vector $\mathbf{x}$:

$$\begin{aligned} \mathbf{x} = (&g_1(x_{1,1}, \ldots, x_{1,T}), \ldots, g_G(x_{1,1}, \ldots, x_{1,T}), \ldots, \\ &g_1(x_{F,1}, \ldots, x_{F,T}), \ldots, g_G(x_{F,1}, \ldots, x_{F,T}))^\mathsf{T}. \end{aligned} \tag{3.1}$$

The components of $\mathbf{x}$ are referred to as 'supra-segmental' features[1] due to the fact that they model more than one speech sound. The next section elaborates on the standard sets of LLDs and functionals used in Computational Paralinguistics. In addition, the first- or second-order differences $(\Delta x_{f,t})_t$ of the LLDs can be computed to obtain ΔLLD contours. Accordingly, functional extraction as described above is applied to the ΔLLD contours $(\Delta x_{f,t})_t$, yielding a vector $\mathbf{x}_\Delta$ that is concatenated with $\mathbf{x}$ for further processing.

The combination of LLDs/ ΔLLDs with functionals can easily lead to a vast number of features. Therefore, feature selection is usually considered to mitigate the curse of dimensionality and thus improve speed and accuracy of paralinguistic analysis. On the one hand, expert-designed rules can be followed to eliminate meaningless combinations of LLDs and functionals such as the arithmetic mean of Δs which is expected to be zero. On the other hand, statistical criteria such as information gain with respect to nominal labels, correlation with continuous-valued labels, or the recognition performance in a cross-validation setup can be used. An often used method in this thesis is the correlation-based feature selection (CFS), which searches the feature space for relevant features that are highly correlated with the class label, yet uncorrelated with each other [85]. Typically, CFS discards well over half of the features, while achieving better accuracies than using the complete feature set in most cases [85]. It is therefore especially useful for filtering out redundant features resulting from the brute-force feature generation approach. In CFS, the merit Q of a feature subset with k features is given by:

$$Q = \frac{k\ \mathrm{CC}_{\mathrm{cf}}}{\sqrt{k + k(k-1)\mathrm{CC}_{\mathrm{ff}}}}, \tag{3.2}$$

where $\mathrm{CC}_{\mathrm{cf}}$ denotes the mean correlation coefficient (CC) of features with the class label, and $\mathrm{CC}_{\mathrm{ff}}$ is the average CC of features with each other. It is easy to see that a candidate subset that maximises Q will achieve an optimal trade-off between high predictive power regarding the class label (numerator) and low redundancy among the features (denominator).

Besides the above described approach of functional generation followed by feature selection, there are various alternatives to turn the matrix $\mathbf{X}$ with a variable number of columns T into a fixed-length representation, including bags-of-audio words [184], recurrent neural networks [205] and attention [139].

After obtaining a set of task-specific features, a model for classification or regression is trained for pattern recognition (cf. Section 3.3). Besides, model confidence is estimated to capture the reliability of the predicted labels. This is especially useful in semi-supervised learning (cf. Section 3.4.1), where machine-labelled instances are fed back into the training set, and active learning (cf. Section 3.4.2), where the model calls for human intervention in case of uncertainty.

[1] For disambiguation, 'suprasegmentals' in linguistic terms refers to "a vocal effect which extends over more than one sound segment in an utterance" [38]. In phonetics, sound segments denote the smallest perceptible discrete units, i. e., phones, and thus need to be distinguished from audio segments.

3.2 Feature Extraction

For paralinguistic analysis, different feature sets have been devised, which are usually composed of the following feature groups. For reproducibility, in this thesis, feature extraction is performed with the openSMILE feature extraction toolkit [62].

3.2.1 Feature Groups

Spectral Features (*SPEC*)

The extraction of spectral features is a fundamental part of audio signal processing, and is a prerequisite for many of the features described below. The spectral feature extraction process is based on frequency transformations such as the discrete Fourier transform (DFT). The DFT $\mathbf{z} = (z_0, \ldots, z_{W-1}) \in \mathbb{C}^W$ is computed from a sample window $\mathbf{v} = (v_0, \ldots, v_{W-1})$ as the cross correlation of $\mathbf{v}$ with sinusoid functions with varying frequencies:

$$z_k = \sum_{n=0}^{W-1} v_n e^{-\frac{j \cdot 2\pi k n}{W}} = \sum_{n=0}^{W-1} v_n \left[\cos\left(\frac{2\pi k}{W} n\right) - j \sin\left(\frac{2\pi k}{W} n\right) \right], \tag{3.3}$$

where j is the imaginary unit. To eliminate aliasing in the high frequency components, $\mathbf{v}$ is multiplied with a smooth windowing function such as the Hamming or Hann function prior to the transformation (Eq. 3.3). As commonly practised in audio processing, the phase φ of the complex spectrum $\mathbf{z}$ is discarded and only the magnitude spectrum $|\mathbf{z}|$ or the power spectrum $|\mathbf{z}|^2$ are retained. Since the audio samples are real-valued, the DFT obeys the symmetry $z_{W-k} = z_k^*$, where * signifies complex conjugation. Thus, the magnitude/ power spectrum is specified by $\lfloor W/2 \rfloor + 1$ elements.

For the reason that the human hearing works logarithmically in the perception of frequency and magnitude, two transformations are commonly applied. First, the linear frequency scale (in Hz) is warped to a logarithmic scale such as the Mel scale [218]. The main idea is to group DFT frequencies by defining a set of D bins, of which the centre points are equidistantly spaced on the chosen logarithmic scale. Then the components of the DFT magnitude/ power spectrum are processed by a triangular filter bank to obtain the magnitude/ power in each bin, resulting in a D-dimensional vector $\mathbf{m}$. Besides, the logarithmic warping of the frequency axis often serves for feature reduction, since the number of bins is smaller than the number of DFT frequencies [60]. Second, the magnitude or power spectrum is often transformed by a mapping $\mathbf{s} \mapsto \log \mathbf{s}$ (applied element-wise). Alternatively, the range can be compressed through a mapping $\mathbf{s} \mapsto \mathbf{s}^\gamma$ with $\gamma < 1$. For instance, the calculation of the 'auditory spectrum' used in the ComParE feature set applies $\gamma = 0.33$ to the power spectrum [60]. Other spectral LLDs computed from the DFT

spectrum are specified in the MPEG-7 multimedia content description standard [134], including band energies and various statistics [60].

Mel Frequency Cepstral Coefficients (*MFCC*)

MFCCs [42] $\tilde{\mathbf{m}} = (\tilde{m}_0, \ldots, \tilde{m}_{M-1})$ are obtained from the Mel-scale spectrum $\mathbf{m}$ in two steps: (i) taking the logarithm (cf. above), and (ii) a second frequency transform, usually the type II discrete cosine transform (DCT):

$$\tilde{m}_k = \sum_{f=0}^{D-1} m_f \cos\left(\frac{\pi k}{D}(f+0.5)\right), k = 0, \ldots, M-1. \tag{3.4}$$

Consequently, the Mel cepstrum can be understood as the spectrum of the Mel spectrum. The number of MFCCs M can be chosen as $M < D$ to obtain further feature reduction, since most of the signal information is contained in the lower-order coefficients. MFCC features are among the most popular features for automatic speech recognition, music information retrieval, and a wide variety of paralinguistic tasks. For example, the first MFCC ($k = 1$) models the presence of high vs low frequency components in the Mel spectrum, and is highly efficient for predicting arousal in speech, music and sound [246].

Prosodic Features (*PROS*)

Given the importance of sentence melody for the assessment of speaker characteristics, the above features are usually augmented with prosodic features. These are based on pitch (fundamental frequency, F_0), loudness, and energy. In openSMILE, frame-wise pitch estimates are first extracted by the subharmonic summation (SHS) method, which is based on summing up frequency bands shifted by one or more octaves. SHS serves to amplify peaks in the spectrum, which are then selected as F_0 candidates. Then, the pitch contour is smoothed to prevent, e. g., sudden jumps in the fundamental frequency [60]. In addition, an estimate of the perceived loudness is provided by summing the auditory spectrum. Finally, zero-crossing rate (ZCR) and root mean square (RMS) energy are computed from the time-domain windows.

Voice Quality (*VQ*)

Voice quality LLDs describe micro-prosodic variations within short-time frames [60]. The probability of voicing can be computed from the strength of the peak of the pitch candidate in the subharmonic spectrum (cf. above). The harmonics-to-noise ratio (HNR) is generally defined as the (logarithmic) ratio of energy in harmonic components to energy in noise-like components.

Table 3.1: Feature sets used in INTERSPEECH Challenges 2009–2012 and ComParE 2013–2017.

# Features	IS09EC	IS10PC	IS11SSC	IS12STC	ComParE
LLDs	16	38	59	64	65
Functionals	12	21	39	61	139
Total	384	1 582	4 368	6 125	6 373

Table 3.2: The IS09EC feature set.

LLD (Δ)	**Functionals**
ZCR	Mean
RMS Energy	Standard Deviation Energy
F0	Kurtosis, Skewness
HNR	Extremes: Value, Position, Range
MFCC 1-12	Linear Regression: Offset, Slope, MSE

The work [60]) elaborates on the various algorithms for estimating these energies. Jitter and shimmer capture short-term variations of the fundamental frequency period and amplitude for harmonic sounds.

3.2.2 Standard Paralinguistic Feature Sets

In each INTERSPEECH Challenge, an acoustic feature set was provided for the participants to compute the baseline. Over the years, the feature sets used for paralinguistic tasks have been continuously improved and increased (cf. Table 3.1). The IS09EC feature set is composed of 384 features, which are obtained by applying 12 functionals to 16 LLDs and their Δs (cf. Table 3.2). The IS10PC, IS11SSC, and IS12STC feature sets are not detailed here since they are not used in this thesis.

In the ComParE series, the large ComParE set has been consistently used [62], serving as the standard feature set for paralinguistic tasks. It is generated by applying different sets of functionals to two groups of LLDs (cf. Table 3.3a). Group A of LLDs consists of four energy related LLDs and 55 spectral LLDs; group B consists of six voicing related LLDs. A set of 54 functionals is applied to the LLDs of group A, and 46 functionals are applied to the ΔLLDs of this group, resulting in $(54 + 46) \cdot (55 + 4) = 5\,900$ acoustic features. A smaller set of 39 functionals is applied to the LLDs of group B and their ΔLLDs, yielding another $39 \cdot 6 \cdot 2 = 468$ acoustic features. In addition, five temporal statistic descriptors are computed for voiced segments: the mean length, the standard deviation of the segment length, the minimum and the maximum length of the voiced segments, as well as the ratio of the non-zero F_0 values. In total, the ComParE feature set consists of $5\,900 + 468 + 5 = 6\,373$ features. The sets of LLDs and applied functionals are specified in Table 3.3a and Table 3.3b, respectively (cf. [206]).

Table 3.3: ComParE acoustic feature set: LLDs and functionals applied to LLDs [206].

(a) LLDs

4 Energy Related LLD	**Group**
Sum of Auditory Spectrum (Loudness)	Prosodic
Sum of RASTA-Style Filtered Auditory Spectrum	Prosodic
RMS Energy, Zero-Crossing Rate	Prosodic
55 Spectral LLD	**Group**
RASTA-Style Auditory Spectrum, Bands 1–26 (0–8 kHz)	Spectral
MFCC 1–14	Cepstral
Spectral Energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90	Spectral
Spectral Flux, Centroid, Entropy, Slope, Harmonicity	Spectral
Spectral Psychoacoustic Sharpness	Spectral
Spectral Variance, Skewness, Kurtosis	Spectral
6 Voicing Related LLD	**Group**
F_0 (SHS & Viterbi Smoothing)	Prosodic
Probability of Voicing	Voice Quality
Log. HNR, Jitter (Local, Delta), Shimmer (Local)	Voice Quality

(b) Functionals

Mean Values
Arithmetic Mean$^{A^{\triangleq},B}$, Arithmetic Mean of Positive Values$^{A^{\Delta},B}$, Root-Quadratic Mean, Flatness

Moments: Standard Deviation, Skewness, Kurtosis

Temporal Centroid$^{A^{\triangleq},B}$

Percentiles
Quartiles 1–3, Inter-Quartile Ranges 1–2, 2–3, 1–3, 1 %-tile, 99 %-tile, Range 1–99 %

Extrema
Relative Position of Maximum and Minimum, Full Range (Maximum − Minimum)

Peaks and Valleys^{A}
Mean of Peak Amplitudes,
Difference of Mean of Peak Amplitudes to Arithmetic Mean,
Mean of Peak Amplitudes Relative to Arithmetic Mean,
Peak to Peak Distances: Mean and Standard Deviation,
Peak Range Relative to Arithmetic Mean
Range of Peak Amplitude Values,
Range of Valley Amplitude Values Relative to Arithmetic Mean,
Valley-Peak (Rising) Slopes: Mean and Standard Deviation,
Peak-Valley (Falling) Slopes: Mean and Standard Deviation

Up-Level Times: 25 %, 50 %, 75 %, 90 %

Rise and Curvature Time
Relative Time in which Signal is Rising,
Relative Time in which Signal has Left Curvative

Segment Lengths^{A}
Mean, Standard Deviation, Minimum, Maximum

Regression$^{A^{\triangleq},B}$
Linear Regression: Slope, Offset, Quadratic Error,
Quadratic Regression: Coefficients a and b, Offset c, Quadratic Error

Linear Prediction
LP Analysis Gain (Amplitude Error), LP Coefficients 1–5

^{A} Functionals applied only to energy related and spectral LLDs (group A)
^{B} Functionals applied only to voicing related LLDs (group B)
$^{\Delta}$ Functionals applied only to ΔLLDs
$^{\triangleq}$ Functionals **not** applied to ΔLLDs

3.3 Machine Learning Models

This section expounds the machine learning models which are typically used for classification and regression tasks in CP. Support Vector Machines (SVMs) [35] and Support Vector Regression Machines (SVRs) [51] serve as the standard methods for baseline computation in the ComParE due to their robustness to overfitting in large feature spaces. A Deep Neural Network (DNN) is a layered composition of multiple non-linear operations and hence can model complex input-output relationships.

3.3.1 Support Vector Machines

SVMs build on the core idea of binary classifiers with linear decision boundaries [35]. Similar to other linear classifiers, SVMs are defined by a hyperplane $\mathbf{w}^T\mathbf{x} - b = 0$ with the normal vector $\mathbf{w}$ and the scalar bias b. However, the key differentiator of SVM is the definition of the *margin*, which is bordered by two parallel hyperplanes, as shown in Figure 3.2:

$$\begin{aligned} \mathbf{w}^T\mathbf{x} - b &= +1, \\ \mathbf{w}^T\mathbf{x} - b &= -1. \end{aligned} \tag{3.5}$$

Under the assumption that the training data $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}$ are linearly separable, all training data with label $y_i = +1$ are required to lie on one side of the margin, while the data with label $y_i = -1$ must lie on the other side. These requirements can be written as:

$$y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1, \quad i = 1, \ldots, L, \tag{3.6}$$

where L is the number of instances in the labelled training set. Out of the possible hyperplanes that fulfil these conditions, the SVM hyperplane is the one which maximises the width of the margin and thereby achieves the largest separation of the two classes.

From the Hesse normal form of the hyperplane, $\frac{1}{||\mathbf{w}||}(\mathbf{w}^T\mathbf{x} - b) = 0$, it is easy to see that the distance from a point on the margin to the hyperplane is $1/||\mathbf{w}||$ and thus the total width of the margin is $2/||\mathbf{w}||$. Therefore, the problem of finding the separating hyperplane with maximum margin is equivalent to minimising $||\mathbf{w}||$ under the boundary conditions (3.6). Minimising $||\mathbf{w}||$, in turn, is equivalent to minimising $\frac{1}{2}\mathbf{w}^T\mathbf{w}$. This results in an optimisation problem with boundary conditions, which can be transformed into the following convex optimisation problem using Lagrange multipliers α_i:

$$\min_{\mathbf{w},b,\alpha_i \geq 0} \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{L} \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i - b) - 1]. \tag{3.7}$$

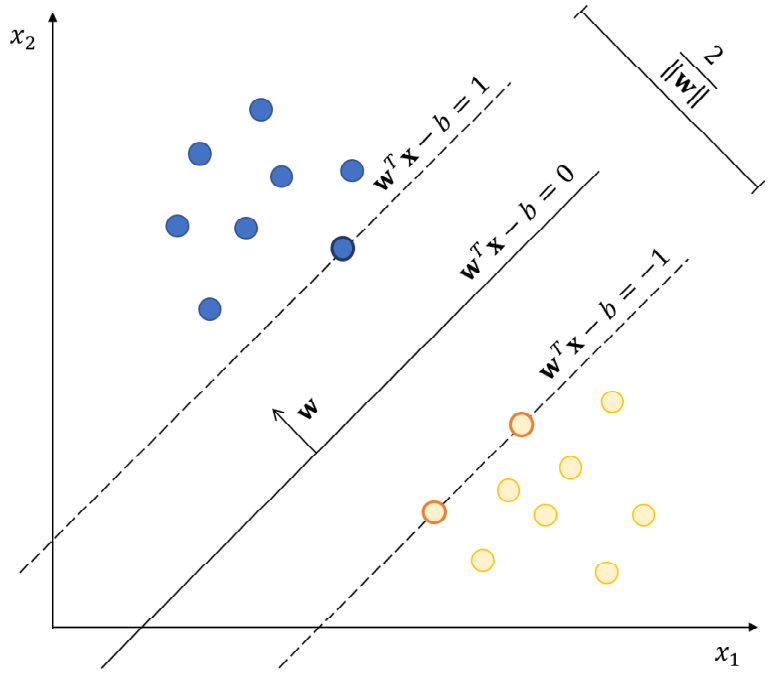


Figure 3.2: Illustrative depiction of a binary SVM classifier trained using the maximum-margin criterion in a two-dimensional feature space (x_1, x_2). Dots indicate training data points; the class membership is indicated by the colour of the dots. Dots with bold lines represent support vectors lying on the boundaries of the margin, which are signified by dashed lines.

Moreover, according to the theory of convex optimisation [23], the solution of this so-called *primal* problem is equivalent to the solution of the corresponding so-called *dual* problem:

$$\max_{\alpha_i \geq 0} \min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{L} \alpha_i [y_i(\mathbf{w}^T\mathbf{x}_i - b) - 1]. \tag{3.8}$$

For fixed α_i, the inner minimisation problem in (3.8) can be solved analytically by setting the gradients w.r.t. $\mathbf{w}$ and b to zero, yielding:

$$\mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{x}_i, \tag{3.9}$$

$$\sum_{i=1}^{L} \alpha_i y_i = 0. \tag{3.10}$$

With this and some algebraic transformations, the dual problem can be written as:

$$\max_{\alpha_i \geq 0, \sum_i \alpha_i y_i = 0} \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j. \tag{3.11}$$

Solving the dual problem (i. e., determining the coefficients α_i) instead of the primal problem (determining the weights directly) is very efficient in case that there are less instances than features (which is often the case in CP). Note that the complexity of computing (3.11) is quadratic in the number of training points, as it contains L^2 terms of the form $\mathbf{x}_i^T\mathbf{x}_j$. In practice, this complexity can be reduced by 'divide-and-conquer' algorithms, which break the dual problem (3.11) into smaller optimisation problems, solve these analytically and finally combine the results. The most well known representative of this class of algorithms is sequential minimal optimisation (SMO), which is introduced in detail by [156].

The dual problem formulation has two interesting properties. First, the 'complementary slackness' condition from convex optimisation theory [23] states that, if α_i is the solution of the above maximisation problem, $\alpha_i > 0$ implies that $y_i(\mathbf{w}^T\mathbf{x}_i - b) - 1 = 0$, i. e., $\mathbf{x}_i$ lies on the margin. Given such $\mathbf{x}_i$ on the margin, the bias of the SVM is easily computed as:

$$b = \mathbf{w}^T\mathbf{x}_i - y_i. \tag{3.12}$$

Second, equation (3.9) states that the weights $\mathbf{w}$ can be expressed as a linear combination of the vectors $\mathbf{x}_i$ with $\alpha_i > 0$. From these observations, it follows that the SVM is entirely determined by the points lying on the margin, which are hence called *support vectors*.

It remains to describe the methodology to apply SVM to non-linearly-separable training data. For this, two methods are usually employed. The first is the introduction of additional 'slack' variables ζ_i into the optimisation problem, which relax the assumption that the training data need to lie on the 'correct' sides of the margin:

$$y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, \quad i = 1, \ldots, L. \tag{3.13}$$

Of course, there should be as little deviation as possible from the ideal case. This results in the following term to minimise:

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{L}\zeta_i, \tag{3.14}$$

subject to the boundary conditions (3.13) and $\zeta_i \geq 0$. C denotes the complexity parameter. Using Lagrange multipliers, this problem can be converted into a primal problem in analogy to above. Moreover, the dual problem for the non-linearly-separable case is equivalent to (3.11), with the additional constraint:

$$\alpha_i \leq C, \quad i = 1, \ldots, L. \tag{3.15}$$

The second concept to deal with non-linearly-separable data is the *kernel trick*. Here, one defines a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ that is used instead of the simple scalar product $\mathbf{x}_i^T\mathbf{x}_j$ in training and testing. For instance, the polynomial kernel of degree p is given by $k_p(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j)^p$.

It can be shown that this kernel and kernels with similar properties correspond to a non-linear transformation ψ of the vectors $\mathbf{x}_i, \mathbf{x}_j$ followed by a simple scalar product: $k(\mathbf{x}_i, \mathbf{x}_j) = \psi(\mathbf{x}_i)^T \psi(\mathbf{x}_j)$ [211]. Thus, effectively the training and testing is carried out in a non-linear feature space that is never computed explicitly. As pointed out by Smola and Schölkopf [211], this non-linear feature space can be of extremely high or even infinite dimension while k is still efficient to compute.

At test time, the classification by SVM can be done as with any other linear classifier, using

$$\hat{y} = \text{sgn}(\mathbf{w}^T \mathbf{x} - b) \tag{3.16}$$

to assign class +1 or -1 to a test instance $\mathbf{x}$, where sgn is the signum function. In case a non-linear kernel k is used, the above is rewritten as:

$$\hat{y} = \text{sgn}\left(\left(\sum_{i=1}^{L} \alpha_i k(\mathbf{x}_i, \mathbf{x})\right) - b\right), \tag{3.17}$$

exploiting the relation (3.9). To handle multi-class problems, a set of binary SVMs can be trained for each pair of classes and the final class label is determined through majority vote.

Support Vector Regression

To transfer the concept of maximum margin hyperplanes to the regression case, the regressor $\hat{y}$ is defined as a hyperplane $\hat{y} = \mathbf{w}^T \mathbf{x} + b$ with a 'tolerance' region as shown in Figure 3.3, requiring all training data points to lie within this region. Again, the margin is delimited by two hyperplanes, and the problem of training the regression function can be cast as a boundary value problem: minimise $\frac{1}{2}\mathbf{w}^T \mathbf{w}$ subject to:

$$\begin{aligned} y_i - (\mathbf{w}^T \mathbf{x}_i + b) &\leq \varepsilon, \\ \mathbf{w}^T \mathbf{x}_i + b - y_i &\leq \varepsilon, \quad i = 1, \ldots, L \end{aligned} \tag{3.18}$$

where y_i are the (now continuous-valued) labels of the training vectors $\mathbf{x}_i$ and $\varepsilon > 0$. Similar to the classification case, to cope with the situation that the above problem is infeasible, variables ζ_i, ζ_i^* can be introduced to allow larger deviations than ε:

$$\begin{aligned} y_i - (\mathbf{w}^T \mathbf{x}_i + b) &\leq \varepsilon + \zeta_i, \\ \mathbf{w}^T \mathbf{x}_i + b - y_i &\leq \varepsilon + \zeta_i^*, \\ \zeta_i, \zeta_i^* &\geq 0, \quad i = 1, \ldots, L. \end{aligned} \tag{3.19}$$

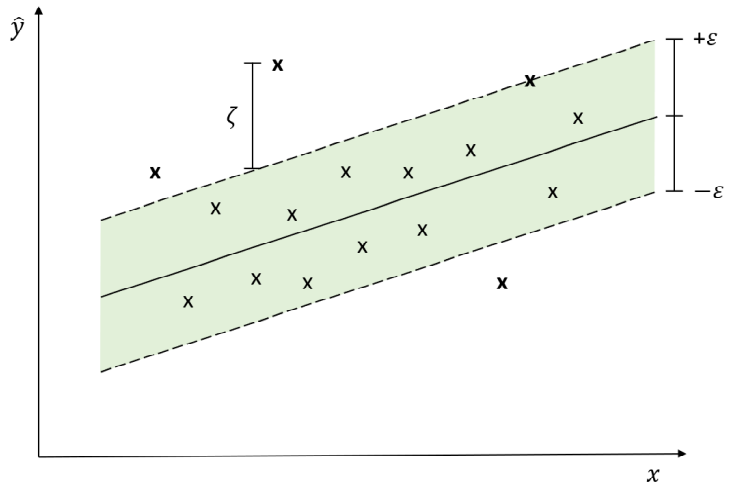


Figure 3.3: Illustrative depiction of a SVR trained using the maximum-margin criterion in a two-dimensional space (x, y). Crosses indicate training data points. Crosses with bold lines represent support vectors lying on or outside of the margin, which is signified by dashed lines.

Using Lagrange multipliers $\alpha_i, \alpha_i^*, \beta_i, \beta_i^*$, this yields the dual SVR optimisation problem:

$$\begin{aligned}\max_{\alpha_i,\alpha_i^*,\beta_i,\beta_i^*} \min_{\mathbf{w},b,\zeta_i,\zeta_i^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{L}(\zeta_i+\zeta_i^*) - \sum_{i=1}^{L}(\beta_i\zeta_i+\beta_i^*\zeta_i^*) \\ - \sum_{i=1}^{L}\alpha_i(\varepsilon+\zeta_i-y_i+\mathbf{w}^T\mathbf{x}_i+b) \\ - \sum_{i=1}^{L}\alpha_i^*(\varepsilon+\zeta_i^*+y_i-\mathbf{w}^T\mathbf{x}_i-b). \end{aligned} \tag{3.20}$$

In analogy to the classification case, the inner minimum can be determined analytically by setting the gradients to zero, yielding:

$$\mathbf{w} = \sum_{i=1}^{L}(\alpha_i-\alpha_i^*)\mathbf{x}_i, \tag{3.21}$$

$$\sum_{i=1}^{L}(\alpha_i-\alpha_i^*) = 0, \tag{3.22}$$

$$C-\beta_i-\alpha_i = 0, \tag{3.23}$$

$$C-\beta_i^*-\alpha_i^* = 0. \tag{3.24}$$

With this, the dual SVR optimisation problem can be written as:

$$\max_{\sum_{i=1}^{L}(\alpha_i-\alpha_i^*)=0;0\le\alpha_i,\alpha_i^*\le C} \sum_{i=1}^{L} y_i(\alpha_i-\alpha_i^*) - \varepsilon\sum_{i=1}^{L}(\alpha_i+\alpha_i^*) - \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\mathbf{x}_i^T\mathbf{x}_j. \tag{3.25}$$

Since $\alpha_i > 0$ implies $y_i - (\mathbf{w}^T\mathbf{x}_i - b) = \varepsilon + \zeta_i$, and $\alpha_i^* > 0$ implies $\mathbf{w}^T\mathbf{x}_i + b - y_i = \varepsilon + \zeta_i^*$, the points $\mathbf{x}_i$ with $\alpha_i > 0$ or $\alpha_i^* > 0$ are the ones lying on the boundaries of the ε-margin or outside; in analogy to the classification case, these can be interpreted as support vectors, since they define the regression coefficients $\mathbf{w}$ and b.

3.3.2 Deep Neural Networks

Deep learning enables the computer to learn from a hierarchy of concepts, with each concept built on simpler concepts [71]. These concepts are generally described with deep computational graphs with many layers. The simplest form of a deep learning model is the DNN, or multilayer perceptron (MLP).

The performance of machine learning algorithms strongly depends on the feature representation of the data they are given, where learnt representations often lead to better performance than those hand-designed features. The quintessential example for representation learning is the autoencoder, which learns to compress the input data into a compact representation (encoding), and then converts that representation back into a form that closely matches the original data (decoding). The key differentiator distinguishing deep learning from other machine learning areas is the joint learning of feature representation and classification/ regression in a single model.

In the last decade, graph-based approaches to deep learning have become the de facto standard, amplified by the success of open-source libraries such as TensorFlow [2] implementing this paradigm. In a graph-based approach, model parameters, input features, internal representations, predictions and the loss function are treated as data. The directed, acyclic graph $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ represents the data flow between the vertices $\mathcal{V}$ along the edges $\mathcal{E}$. The data are usually stored in matrix or generally tensor form, except for the loss J which is a scalar.

The *forward pass* computes the predictions of the deep learning model and the corresponding loss J. Here, we assume that the input of the DNN is a minibatch of training vectors x_i given in the rows of a matrix $\mathbf{X} = [\mathbf{x}_1; \cdots; \mathbf{x}_B]$, where B is the minibatch size. The forward pass is executed by traversing the graph $\mathcal{D}$ along its edges $\mathcal{E}$ starting from the input nodes and computing:

$$\mathbf{Y}_k = \mathsf{op}_k(\mathbf{Y}_{j_1}, \mathbf{Y}_{j_2}, \dots), \tag{3.26}$$

where $\mathbf{Y}_k$ denotes the data associated with vertex k, op_k is the corresponding operation, and $(j_1, k), (j_2, k), \cdots \in \mathcal{E}$.

To construct the computational graph for a 'classical' DNN with fully connected layers (i. e., where each neuron is connected to all neurons from the previous layer), the mapping from the input features $\mathbf{X}$ to the output $\hat{\mathbf{Y}}$ given by the DNN is described as a composition of affine

transformations and non-linear activation functions. The activations $\mathbf{H}^{(l)}$ of the DNN with H hidden layers are computed by applying:

$$\mathbf{A}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}, \tag{3.27}$$

$$\mathbf{H}^{(l)} = u^{(l)}(\mathbf{A}^{(l)}), \tag{3.28}$$

for $l = 1, \ldots, H+1$, with per-layer weight matrices $\mathbf{W}^{(l)}$, linear activations $\mathbf{A}^{(l)}$ and activation functions $u^{(l)}$. $\mathbf{H}^{(0)}$ is defined as the input features $\mathbf{X}$ (containing feature vectors in its rows), while the output $\hat{\mathbf{Y}}$ equals the last layer activation (containing prediction vectors in its rows).

For the hidden layers of a DNN, arguably the most popular activation function is the rectifier:

$$u(a) = \max(0, a), \tag{3.29}$$

which is applied elementwise. The choice of the rectified linear activation function is biologically inspired [84] and, on a practical level, alleviates the vanishing gradient problem in training DNNs, thus allowing for dispensing with pre-training [45].

The activation function used for the *output* layer as well as the appropriate loss J depend on the task: For classification, the softmax function with cross-entropy loss is favoured, while the linear activation function with mean squared error loss is most suitable for regression [71]. Besides, for classification with the softmax function, the output layer needs to have a dimensionality of P for a classification task with P classes. For regression, the dimensionality O of the output layer corresponds to the number of regression variables. The softmax function is given by:

$$\hat{y}_{i,j} = \frac{\exp(a_{i,j})}{\sum_{j=1}^{P} \exp(a_{i,j})}. \tag{3.30}$$

Since the activations $\hat{y}_{i,j}$ lie in the interval $[0, 1]$ and sum to one for each i, they can be interpreted as posterior probabilities of the classes $1, \ldots, P$ for each instance i.

The cross-entropy function for discrete class labels $y_i \in \{1, \ldots, P\}$ is defined as:

$$J_{\text{CE}} = -\frac{1}{B} \sum_{i=1}^{B} \log \hat{y}_{i,y_i}. \tag{3.31}$$

Note that the combination of softmax with cross-entropy has a simple gradient, thus this is often subsumed in a single operation. The mean squared error (MSE) loss is the mean squared deviation of the outputs from the regression targets $\mathbf{y}_i$:

$$J_{\text{MSE}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{O} (y_{i,j} - \hat{y}_{i,j})^2. \tag{3.32}$$

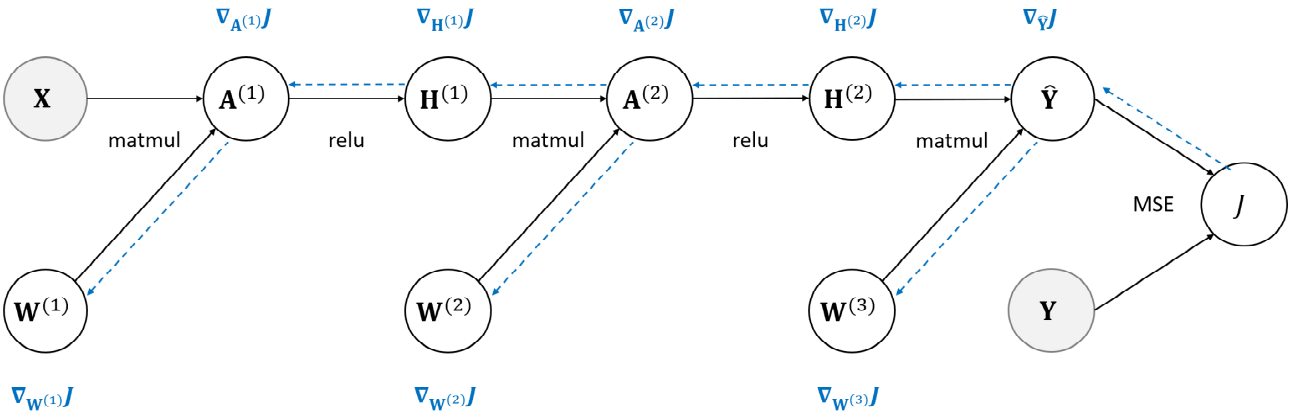


Figure 3.4: Illustrative computational graph of a DNN with two hidden layers. matmul denotes matrix multiplication while relu stands for the rectified linear activation function. Solid lines indicate the forward pass and dashed lines the backward pass. Gray nodes indicate constants.

Figure 3.4 shows a computational graph for an exemplary DNN with $H = 2$ hidden layers. Biases are omitted here for the sake of clarity.

DNN Training

The parameter vector $\mathbf{w}$ of a deep learning model, which contains all trainable parameters (in the above example, the weights), is commonly determined by minimising the loss J by stochastic gradient descent (SGD). The gradient, in turn, is obtained by means of error backpropagation, which is implemented as a backward pass through the computational graph $\mathcal{D}$.

Since the forward pass of the deep learning model is a composition of differentiable operations, the backward pass is based on applying the chain rule of calculus. Effectively, the computational graph is traversed in reverse order, annotating each vertex j with the gradient w.r.t. $\mathbf{Y}_j$:

$$\nabla_{\mathbf{Y}_j} J = \sum_{(j,k)\in\mathcal{E}} \left(\frac{\partial \mathbf{Y}_k}{\partial \mathbf{Y}_j}\right)^T \nabla_{\mathbf{Y}_k} J, \tag{3.33}$$

where $\frac{\partial \mathbf{Y}_k}{\partial \mathbf{Y}_j}$ is a Jacobian matrix (for this, let us assume that tensor elements are stacked into a vector). In Figure 3.4, the gradients occurring in the backward pass on the example DNN are highlighted in blue colour. Once all the vertices containing trainable parameters (in the above case, the weight matrices $\mathbf{W}^{(k)}$) have been visited by the graph traversal, the entire gradient $\nabla_{\mathbf{w}} J$ is known to the optimiser.

The SGD optimiser performs the following parameter updates:

$$\mathbf{w}^{(q+1)} = \mathbf{w}^{(q)} - \eta \nabla_{\mathbf{w}} J\left(\mathbf{w}^{(q)}\right), \tag{3.34}$$

with training step q and learning rate $\eta > 0$. To minimise oscillations of the loss between minibatches, a momentum term [159] is often added to the parameter update:

$$\Delta \mathbf{w}^{(q+1)} = \mu \Delta \mathbf{w}^{(q)} - \eta \nabla_{\mathbf{w}} J\left(\mathbf{w}^{(q)}\right), \tag{3.35}$$

$$\mathbf{w}^{(q+1)} = \mathbf{w}^{(q)} + \Delta \mathbf{w}^{(q+1)}, \tag{3.36}$$

with $\Delta \mathbf{w}^{(0)} = \mathbf{0}$ and $0 < \mu \leq 1$ being the momentum coefficient. In the Nesterov accelerated gradient method [146], (3.35) is modified as:

$$\Delta \mathbf{w}^{(q+1)} = \mu \Delta \mathbf{w}^{(q)} - \eta \nabla_{\mathbf{w}} J\left(\mathbf{w}^{(q)} + \mu \Delta \mathbf{w}^{(q)}\right), \tag{3.37}$$

i. e., the gradient computation already anticipates the next parameter update.

With multiple non-linear hidden layers, DNNs are very powerful models in learning complex relationships between their inputs and outputs. In the setting of limited training data, however, they are likely to over-adapt to spurious patterns which exist in the training data but not in the test data (overfitting). To mitigate the problem of overfitting and improve generalisation, three common heuristic methods are used in this thesis.

Dropout randomly drops units (along with their incoming and outgoing connections) from the neural network to prevent co-adaptations between units [214]. During training, dropout samples from an exponential number of different thinned networks, effectively performing model combination and thus improving robustness. *Data shuffling* randomly generates mini-batches to prevent the model from learning the order of the training instances. Shuffling data at each epoch has immediate impact on SGD convergence since it helps reduce susceptibility to local minima in the error function. *Early stopping* uses a 'held-out' set to evaluate the convergence criterion for the loss function independently from the training set. The training process is stopped once the model incurs larger generalisation error on the validation set.

Optimising DNN training is still an active area of research, and more regularisation 'recipes' to improve generalisation are presented in [143]. Despite all these 'tricks of the trade', DNNs are still very thirsty for data due to the large number of parameters which need to be learnt. For example, the connections between two hidden layers with 1 000 nodes each amount to 1 M weights. In contrast, the number of parameters to be learnt in SVM training is bounded by the number of training vectors. Given the scarcity of training data in CP, DNNs have not yet supplanted SVMs as in other areas of machine learning. In theory, it is possible to model the entire processing chain from raw audio to predictions (cf. Section 3.1) as a single deep learning model, also known as end-to-end learning. However, due to the aforementioned reasons, this approach often leads to inferior results in CP [205]. Therefore, in this thesis, DNNs are used for prediction based on predefined feature sets.

3.4 Machine Learning Tasks

This section deals with the machine learning tasks which form the basis for the methods proposed in this thesis. Semi-supervised learning (SSL) and active learning (AL) both aim to exploit the vast amount of unlabelled data that can be easily acquired but are difficult, time-consuming, or costly to label. While SSL sidesteps human annotation completely by means of machine labelling, AL enables selective sampling of only the interesting data points, thereby reducing labelling costs. Multi-task learning (MTL) and transfer learning (TL) are machine learning methods which aim to improve generalisation by learning multiple tasks jointly and as such they are of central importance for holistic analysis.

3.4.1 Semi-Supervised Learning

SSL exploits a large amount of unlabelled data to improve a model trained on a limited amount of labelled data. It is thus halfway between supervised learning (SL – with completely labelled training data) and unsupervised learning (without any labelled training data).

Self-training is a practical wrapper method for SSL, which can be applied to existing (complex) learning algorithms. The learning circle starts with training a model on a small labelled dataset. Then, it is used to make predictions for the unlabelled data. Commonly, the instances the model is most confident about, together with their predicted labels, are added to the labelled data. Using the updated training set, the model is retrained and the whole process is repeated until there are no unlabelled data left in the experimental setting or a specific stopping criterion is fulfilled, e. g., when the confidence value falls below a threshold. A pseudo-code description of the self-training algorithm is shown in Figure 3.5.

Conceivably, a drawback of self-training is its proneness to error reinforcement. This happens when wrong predictions are made with high confidence and thus erroneous labels are used for retraining in the following iterations. Therefore, the performance of self-training highly depends on the quality of the confidence measure used and the initial training data. To mitigate the problem of error accumulation, co-training [21] and multi-view learning [43] can be applied. Based on ensemble learning, they allow different models to generate labelled examples for one another or to reach an agreement.

Other SSL algorithms include generative models such as expectation-maximisation (EM) with generative mixture models; low-density separation methods such as transductive SVMs; and graph-based algorithms [31, 265].

Algorithm: *Self-Training*
Input: Original training set $\mathcal{L} = \{(\mathbf{x}_i, y_i) \mid y_i \neq \perp\}$
Unlabelled set $\mathcal{U} = \{(\mathbf{x}_i, y_i) \mid y_i = \perp\}$
Output: Labelled data $\{(\mathbf{x}_j, y_j) \mid \mathbf{x}_j \in \mathcal{U}\}$
Final model h_{end}
Do:
$\mathcal{U} := \{\mathbf{x}_j \mid y_j = \perp\}$
$\mathcal{L} := \{\mathbf{x}_i \mid y_i \neq \perp\}$
$h := \text{Train}(\mathcal{L})$
$\hat{\mathcal{Y}}, Z := \text{Predict}(h, \mathcal{U})$ // Prediction + confidence measure
$\mathcal{J} := \text{Select}(\mathcal{U}, Z)$ // Instance selection
For $j \in \mathcal{J}$:
$y_j := \hat{y}_j$
While $\mathcal{U} \neq \emptyset$
$h_{\text{end}} := h$

Figure 3.5: Pseudo-code description of the self-training algorithm.

3.4.2 Active Learning

Like SSL, AL aims to mitigate the bottleneck problem of data annotation. The key idea behind AL is that a learning system can achieve higher cost efficiency, i. e., equal accuracy with fewer human annotations, by selecting the data from which it learns [208]. AL algorithms are generally evaluated by means of learning curves, which plot the performance measure (e. g., accuracy) as a function of the number of new queries.

The three main AL scenarios are (i) membership query synthesis, which generates queries de novo; (ii) stream-based selective sampling, which sequentially samples one instance at a time and let the model decide whether to label or discard it; and (iii) pool-based sampling, which assumes a small labelled training set $\mathcal{L}$ and a large pool of unlabelled data $\mathcal{U}$ [208]. The pool-based setting [130] is considered as the realistic scenario in CP since speech data are abundant but labelled training data are scarce.

Figure 3.6 depicts the pool-based AL cycle. The active learner begins with training a model using the labelled training set $\mathcal{L}$. This model is used to make predictions for the unlabelled instances in $\mathcal{U}$. Then, based on the query strategy, the model selects the most informative instances to be labelled by human annotators ('oracles'). The new labelled instances are added to the training set $\mathcal{L}$ and the learning process is iterated from there. This results in a similar iterative algorithm as the self-training algorithm (cf. Figure 3.5), with the difference that the labels $y_j, j \in \mathcal{J}$ of the selected instances are obtained from human oracles. Consequently, in contrast to SSL, AL does not suffer from error reinforcement due to human in the loop.

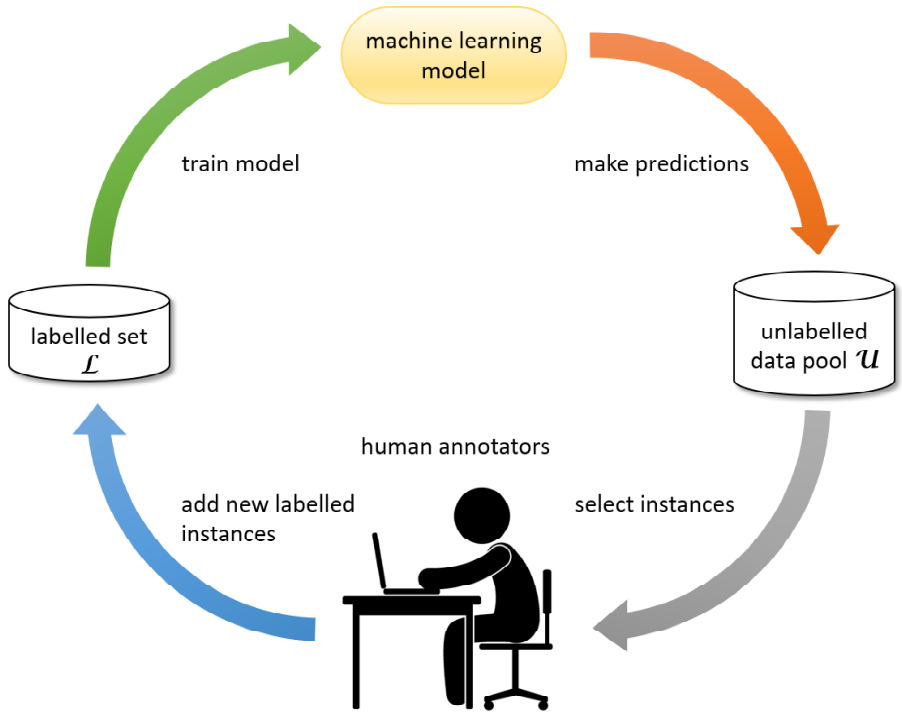


Figure 3.6: Active learning circle in a pool-based setting (cf. [208]).

There are various query strategies by which the informativeness of unlabelled instances can be assessed. A commonly used technique is uncertainty sampling [130], where the instances predicted with low confidence $\mathbf{x}_{LC}$ are selected. For binary classification, when using a probabilistic learning model, the instances are selected which are predicted with the posterior probabilities closest to 0.5 for either of the classes. For multi-class problems, the model confidence can be derived from the posterior probability of the predicted class $\hat{y} = \arg\max_y P(y|\mathbf{x})$ of an instance $\mathbf{x}$:

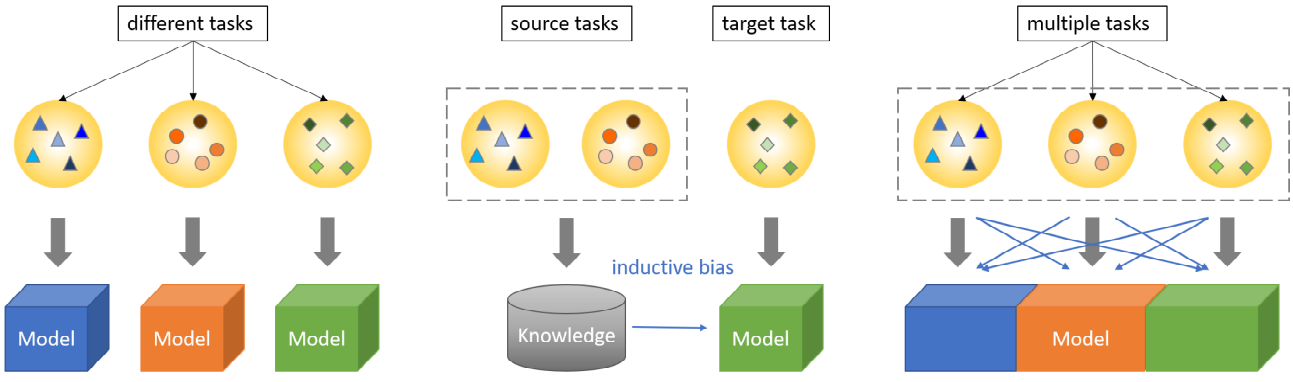
$$\mathbf{x}_{LC} = \arg\min_{\mathbf{x}} P(\hat{y}|\mathbf{x}). \tag{3.38}$$

Since this query strategy only considers the class with the highest probability, which comes with information loss about the probability distribution of the remaining labels, a variant of uncertainty sampling, also known as margin sampling [178], has been proposed to include the two most probable class labels $\hat{y}_1$ and $\hat{y}_2$:

$$\mathbf{x}_{LC} = \arg\min_{\mathbf{x}} |P(\hat{y}_1|\mathbf{x}) - P(\hat{y}_2|\mathbf{x})|. \tag{3.39}$$

For multi-class problems with a large number of classes, where the probability distribution is still not properly captured by margin sampling, the Shannon entropy measure can be used:

$$\mathbf{x}_{LC} = \arg\max_{\mathbf{x}} H(\hat{y}|\mathbf{x}) = \arg\max_{\mathbf{x}} -\sum_j P(\hat{y}_j|\mathbf{x}) \log P(\hat{y}_j|\mathbf{x}). \tag{3.40}$$



(a) STL (b) TL (c) MTL

Figure 3.7: Different learning processes of single-task learning (STL), transfer learning (TL) and multi-task learning (MTL).

Other query strategies include query-by-committee, which selects the instances with maximum disagreement among different models trained on the labelled set; and expected model change, which selects the instances with the highest estimated impact on the current model.

Since SSL and AL both aim to make the most of unlabelled data while reducing annotation effort, it is natural to combine the two learning tasks. In the works [144, 230, 262, 263], hybrid approaches have been applied to various real-world problems such as web page classification, spoken language understanding, content-based image retrieval and speech emotion recognition. Conceptually, SSL and AL address the same problem from opposite directions: SSL exploits what the learner thinks it knows about the unlabelled data, whereas AL focusses on the unknown aspects with the help of human oracles [208]. Accordingly, self-training and multi-view learning can be regarded as complementary to uncertainty sampling and query-by-committee, respectively.

3.4.3 Multi-Task and Transfer Learning

The standard methodology in machine learning is to perform single-task learning (STL), i. e., learning just one task at a time. Large problems are typically broken into small sub-problems that are tackled separately and then recombined. The assumption is that for each target task, enough training data is available that matches the distribution of the test data.

Inspired by human learning, TL deals with the problem of transferring knowledge from one or more source tasks to a target task which can have a different feature space or distribution than the source tasks. TL is especially useful when the target task has only few training data. While TL focusses on the target task, MTL aims to improve the performance on multiple tasks simultaneously. Generally speaking, MTL algorithms are based on the central idea of “sharing

what is learned by different tasks while tasks are trained in parallel" [30]. In this sense, the MTL paradigm is perfectly aligned with the concept of holistic analysis. Specifically, Caruana [30] succinctly described MTL as "an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of *related* tasks as an inductive bias". As opposed to general TL [150], which can be understood as *serial* inductive transfer, MTL emphasises on the *parallel* knowledge transfer between multiple tasks. This has two main advantages: First, the order in which the tasks are learnt is of little account for the learning performance; second, tasks can mutually benefit each other, whereas serial transfer prevents the backward knowledge transfer from newly learnt tasks to previously learnt tasks. Figure 3.7 depicts the general concepts of STL, TL, and MTL.

One of the most important instantiations of MTL is the shared hidden layer approach, which can be applied to DNNs trained with backpropagation (cf. Section 3.3.2). In this so-called hard parameter sharing approach, output layers are defined for all the tasks, reading from the same hidden layer. Caruana [30] elucidated the underlying mechanisms which explain how MTL in neural network models works. *Statistical data amplification* is an effective increase in sample size by uniting different datasets for training a joint model. *Attribute selection* coerces the learner to attend to the common hidden layer features relevant to all tasks. *Eavesdropping* can be useful if related tasks give hints about important features which would be otherwise ignored by a task. *Representation bias* causes a learner to prefer some hypotheses over other hypotheses, e. g., a feature representation that is also suitable to predict other targets of interest. All these mechanisms are derived from the summing of error gradients from the task-specific output layers when calculating the gradient of the shared hidden layers, which is then propagated further down the network [30] (cf. Eq. 3.33). Effectively, the gradient of each task-specific loss is augmented by a regularisation term, which ensures that no steps can be taken in a direction that would hurt the performance on the other tasks, and conversely, steps are encouraged in directions that improve the performance on more than one task.

Part II

Efficient Methods for Data Annotation

Chapter 4

Dynamic Active Learning

Learning is more effective when it is an active rather than a passive process.

KURT LEWIN

The task of obtaining meaningful annotations of speaker characteristics is a quite tedious work, connected with high costs and time consumption. Due to the subjectivity of many paralinguistic speech phenomena, standard annotation procedures acquire the same number of human labels for all training points so as to derive a gold standard by means of majority voting or averaging. In this Chapter, a dynamic active learning (DAL) algorithm is proposed, which decides for each instance whether it is subject to manual labelling, and if this is the case, how many human annotators enlisted in a queuing system are needed to establish an interrater agreement. To this end, an early stopping criterion is enabled, which stops the sequential query process when a predefined agreement level has been reached. Moreover, a DAL variant is introduced, taking into account the interrater reliability in the order the annotators are queried. Applying the DAL method to speech emotion recognition, the evaluation will show that the shift from absolute majority to agreement- and reliability-based early stopping leads to considerable reduction of human labelling work at the same accuracy.

4.1 The Sub-optimality of Majority Voting in Data Annotation

With the massive growth of data created every instant, manual annotation has become the major bottleneck of data processing for pattern recognition due to the prohibitive costs and enormous time consumption involved. The main challenge inherent to annotation tasks is label

subjectivity, which naturally arises with perceptual assessment of ambiguous patterns such as emotion [174]. For this reason, in contrast to the *ground truth* (e. g., age, gender) that can be objectively determined, subjective annotations lead to what is known as the *gold standard*, i. e., the pseudo-truth obtained from a collection of possibly noisy annotations. To this end, majority voting (for nominal class labels) or measures of central tendency such as median or mean (for ordinal/ interval label scales) are usually applied to approach the underlying collective opinion, mitigating subjective bias. However, these interrater agreement procedures make a set of assumptions which are not suitable for real-world settings, namely that the majority of a large number of annotators is requisite for identifying the 'hidden true' label, and that the annotators are independent and equally good.

As aforementioned, in practice, a consensus can hardly be attained for subjective tasks due to the plurality and dynamics of individuals' opinions, yet, the ambiguity about each specific instance varies and thus the difficulty of labelling it. Hence, it is likely more efficient, in terms of labelling costs, to apply an adaptive query strategy. In establishing the gold standard, the focus should be shifted from the majority of annotators to a certain extent of agreement among them so as to dynamically allocate human resources based on the difficulty of annotating an instance.

To overcome the second shortcoming, the fastest way to reach an agreement is to query the annotators in the order of their reliability rank. The reliability aspect is especially important due to the fact that, besides experts, annotators can be inexperienced, indifferent or malicious. This phenomenon has been amplified by the emergence of crowdsourcing and concomitantly the recent trend from small-scale, expensive expert annotation towards large-scale, low-cost labelling by clickworkers, who usually have no formal training at the task at hand [102]. This so-called Vox Populi approach [132] allows for outsourcing small tasks to web-based services like Amazon's Mechanical Turk, employing masses of ordinary people at the risk of noisy labelling.

In this Chapter, a DAL method is proposed, which is aimed at improving the efficiency of data annotation by striving to reach an agreement with as little human intervention as possible. To this end, the standard AL technique is augmented with the measures of interrater agreement and reliability. Throughout this thesis, the words "annotator", "rater", "oracle" and "labeller" are used interchangeably.

Related Work The construct of gold standard is often modelled by the concepts of *interrater agreement* (IRA), i. e., the absolute rater consensus based on estimation of within-group rating dispersion, and the *interrater reliability* (IRR), which refers to the relative rater consistency indexed via correlation coefficients [126].

Interrater agreement was first investigated in a study [98], which proposed a model that predicted the performance of a regression system as a function of the number of annotators and different

model parameters such as the difficulty of annotating a task and other system dependencies. For the specific task of non-native English prosody regression, they advocated a rule of thumb, namely that the performance gain is the highest from one to five labellers, and still clearly visible from six to some ten. However, this general rule does not necessarily apply to other subjective tasks, nor does it differentiate the label ambiguity of individual instances. Nevertheless, their study provides evidence that the number of annotators is a crucial variable for achieving the best tradeoff between system performance and cost efficiency. Sheng et al. [210] proposed repeated labelling for some of the training instances when the labels provided by crowdworkers are noisy. However, they made the unrealistic assumptions that all labellers have the same quality and the labelling difficulty is the same for all instances.

To diminish the individual bias inherent to subjective annotations, some measures such as the EWE [79] use evaluator-dependent weights derived from the correlation between the individual evaluators' outputs and the sequence of maximum likelihood estimates. Yet the EWE is computed a posteriori to obtain a weighted gold standard and thus does not improve the efficiency of the annotation process itself. Raykar et al. [166, 167] proposed a maximum likelihood estimator that jointly learns the classier, the experts' performance, and the actual true label, in the scenario that multiple annotators provide labels but no absolute gold standard. By contrast with the EWE, the probabilistic algorithm iteratively establishes a particular gold standard, measures the performance of the annotators given that gold standard, and then refines the gold standard based on the performance measures. A particular advantage of their approach is that it can handle binary/ categorical/ ordinal classification and regression problems. Donmez et al. [48] proposed an algorithm for estimating the accuracy of multiple noisy labellers and filtering out the best one(s) for active learning, showing that the annotator-aware approach outperforms baselines such as asking all available experts and random expert selection. Yan et al. [250] suggested to partition the feature space into regions, of which each can be reliably assessed by a single annotator in an active learning setting. For each new training sample to be labelled, their probabilistic model chooses the appropriate labeller who will provide the highest confidence. However, their approach is limited to settings where there exists at least one expert annotator who provides more reliable labels than the consensus of multiple annotators. The novel DAL algorithm differs from this previous body of work in the following aspects:

- A sequential query strategy is introduced, which successively acquires new labels for an instance identified as informative by the active learner.
- The dynamic active learner adapts the number of annotators to the required level of interrater agreement, focusing human resources on those ambiguous instances that are difficult to label.
- The queuing system gives precedence to the most reliable annotators, where the prior knowledge of their reliability is gained from an initial training set labelled by all oracles.

4.2 Dynamic Query Strategies

The DAL method is aimed at reducing the amount of human labelling effort to a minimum in reaching three goals: (i) To selectively label new training instances and add them to the training set for continuous model improvement (standard AL); (ii) To enable early stopping when enough annotations have been obtained to reach a defined interrater agreement level for each instance; (iii) To query the annotators in descending order of their prior reliability so as to mitigate the effect of labelling noise. The next three subsections present the dynamic query strategies to achieve these goals. A pseudo-code description of the DAL algorithm is provided in Figure 4.1. For simplicity, the DAL method is considered for binary classification in this work. The generalisation to regression problems will be presented in Chapter 5.

4.2.1 Uncertainty Sampling: Which Instances to Select?

Similar to standard AL (SAL), the dynamic active learner selects the most useful data samples from which it learns, where the 'informativeness' is determined by means of model confidence. To obtain confidence measures for a binary classifier with a decision function $f(\mathbf{x})$, we compute posterior probabilities by means of logistic regression:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(Af(\mathbf{x}) + B))}, \tag{4.1}$$

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}). \tag{4.2}$$

In the case of SVM, the value of $f(\mathbf{x})$ corresponds to the distance from an instance $\mathbf{x}$ to the maximum-margin hyperplane, which represents the largest separation between the two classes. The confidence value assigned to the predicted label of $\mathbf{x}$ is derived by forming the difference between the probability values of the classes '0' and '1':

$$Z(\mathbf{x}) = |P(y = 1|\mathbf{x}) - P(y = 0|\mathbf{x})|. \tag{4.3}$$

This confidence measure takes its maximum value of 1 if $P(y = 1|\mathbf{x}) = 1$ or $P(y = 0|\mathbf{x}) = 1$, and its minimum value of 0 if $P(y = 1|\mathbf{x}) = P(y = 0|\mathbf{x}) = 0.5$. According to previous findings [253, 262], the instances predicted with medium certainty (MC) seem to be most suitable for human inspection, whereas the least certainty (LC) strategy tends to select noisy data points. In addition, a threshold $U_{\min}$ on the number of instances remaining in $\mathcal{U}$ can be defined to avoid selecting noisy data. With Z_m as the median confidence value, the query function is defined as:

$$\mathbf{x}_{MC} = \arg\min_{\mathbf{x}} |Z(\mathbf{x}) - Z_m|. \tag{4.4}$$

4.2.2 Early Stopping Based on Interrater Agreement: How Many Annotators to Ask?

The crux of the DAL algorithm is the definition of an early stopping criterion, which ends a sequential annotation process once the required level of agreement has been reached for an instance. In the binary case, the agreement level a can be defined as the minimum number of votes which either of the two categories receives first. Given the number R of available annotators from a pool $\mathcal{R}$, meaningful values for a range from $2, \ldots, \lfloor \frac{R+1}{2} \rfloor$, where the upper limit of the interval is rounded down if R is an even number. From this, it can be deduced that using the highest value of a automatically results in the same gold-standard labels obtained from majority voting of R annotators. For example, assuming $R = 5$ annotators are available and $a = 3$, if three of them have already agreed on a class label, the gold standard will not further change, thus the remaining annotations are dispensable. Consequently, $R' \in \{a, ..., 2a - 1\}, R' \in \mathbb{N}$ annotators are potentially required depending on the label ambiguity, where in practice, the first a annotations can be performed all at once. The advantage of this dynamic query strategy is that the number of annotators is fine-tuned on a per instance level and thus not bound to a specific task. Conceivably, this approach mainly targets at saving labelling costs for the 'mildly' ambiguous instances for which a consensus can be reached fast.

4.2.3 Query Order Based on Rater Reliability: Whom to Ask First?

In order to gain prior knowledge about the rater reliability, the initial training set $\mathcal{L}$ containing all labeller tracks is used. For each annotator $o \in \mathcal{R}$, a measure of similarity between the labels $y_{1,o}, \ldots, y_{|\mathcal{L}|,o}$ and the gold standard $y_1, \ldots, y_{|\mathcal{L}|}$ is computed. In the case of binary classification, the gold standard y_i for instance $i = 1, \ldots, |\mathcal{L}|$ is the mode of the assigned labels $y_{i,1}, \ldots, y_{i,R}$. To access the reliability r_o of each annotator o, the most popular measure is the Pearson's r [126]. Accordingly, the set of annotators $\mathcal{R}$ is sorted by r_o in descending order. Henceforth, the reliability-based variant of the DAL algorithm is referred to as rDAL, whereas the basic variant with random query order is termed DAL.

4.3 Experiments

The DAL method was applied to spontaneous emotion recognition, which presents a typical subjective task. For experimental purposes, the FAU AEC (cf. Section 2.3.1) was used. Speaker independence was ensured by using the speech material of one school for training and the data collected at the other school for validation. Note that the training data contains both the labelled training set $\mathcal{L}$ and the unlabelled data pool $\mathcal{U}$ (cf. Table 4.1).

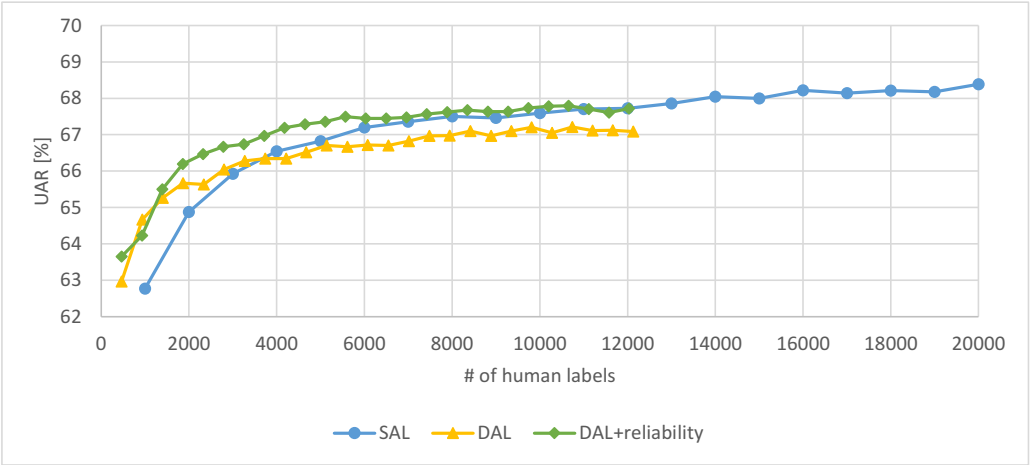
Algorithm: *Dynamic Active Learning*
Input: Original training set $\mathcal{L} = \{(\mathbf{x}_i, y_i) \mid y_i \neq \perp\}$
Unlabelled set $\mathcal{U} = \{(\mathbf{x}_i, y_i) \mid y_i = \perp\}$
Pool of raters $\mathcal{R}$
Agreement level $a \in \{1, \ldots, \lfloor \frac{R+1}{2} \rfloor\}$ for early stopping
Min. number of unlabelled instances U_{min}
Output: Labelled data $\{(\mathbf{x}_j, y_j) \mid \mathbf{x}_j \in \mathcal{U}\}$
Final model h_{end}
Do:
$\mathcal{U} := \{\mathbf{x}_j \mid y_j = \perp\}$
$\mathcal{L} := \{\mathbf{x}_i \mid y_i \neq \perp\}$
$h := \text{Train}(\mathcal{L})$
$Z := \text{Predict}(h, \mathcal{U})$ // Confidence measure
$\mathcal{J} := \text{Select}(\mathcal{U}, Z)$ // Instance selection
For $j \in \mathcal{J}$: // Human labelling
Do:
$o = \text{SelectRater}(\mathcal{R})$ // random or by reliability
$y_{j,o} = \text{Query}(o, x_j)$ // Ask oracle o about instance $\mathbf{x}_j$
$\mathcal{Y}_j := \mathcal{Y}_j \cup \{y_{j,o}\}$ // Set of human labels
While $(\neg\text{Agreement}(a, \mathcal{Y}_j))$
$y_j := \text{Mode}(\mathcal{Y}_j)$
While $|\mathcal{U}| > U_{\text{min}}$
$h_{\text{end}} := h$

Figure 4.1: Pseudo-code description of the dynamic active learning (DAL) algorithm.

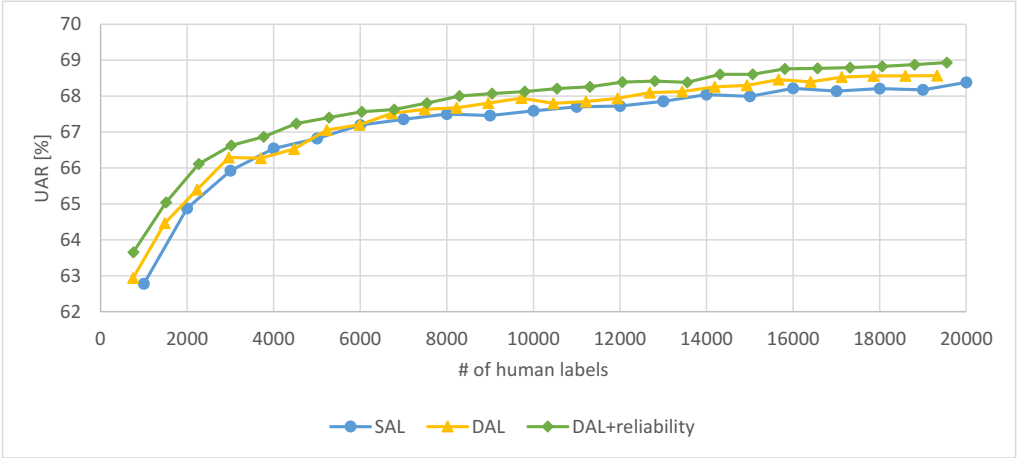
Table 4.1: Distribution of speakers and instances per partition of the FAU AEC. M: male; F: female; NEG: negative emotions; IDL: neutral and positive emotions.

	# speakers		# instances		
FAU AEC	M	F	NEG	IDL	Σ
Training	13	13	3 358	6 601	9 959
Validation	8	17	2 465	5 792	8 257
Σ	21	30	5 823	12 393	18 216

In the experiments, we used the same acoustic features as well as the same SVM implementation in Weka 3 [86] as in the IS09EC (cf. Section 3.2, Table 3.2). The SVM classifier with linear kernel was trained using the SMO algorithm and a constant complexity value $C = 0.05$. The initial training set $\mathcal{L}$ consists of 200 randomly selected training instances, while the rest of the training data were gathered to simulate the pool of unlabelled data. In each training iteration, 200 instances predicted with medium model confidence were forwarded to human inspection. Instead of setting a U_{min} value, the model performance can also be monitored on the validation set in order to stop the learning process once the learning curve ceases to increase. The training process was evaluated in terms of UAR and repeated in 20 independent runs.



(a) $a = 2$



(b) $a = 3$

Figure 4.2: Comparison between standard active learning (SAL), dynamic active learning (DAL) and reliability-based DAL (rDAL): The performance measures show the UAR values averaged across 20 runs of the algorithm over the number of human labels for the agreement levels $a = 2, 3$.

4.4 Results and Discussion

Figure 4.2 depicts the learning curves of the algorithms DAL and rDAL, in comparison with the SAL baseline. As a common characteristic, the model accuracy increases with the amount of newly labelled instances added to the training set, where the learning speed is the fastest at the beginning and levels out as the learning process continues. For the agreement levels $a = 2, 3$, the rDAL algorithm consistently outperforms the SAL baseline. From the vertical distances between the learning curves, i. e., when comparing the different UAR values at the same number of human annotations, it can be seen that the rDAL method makes better use of human workforce than SAL, particularly for $a = 3$. For $a = 3$, with 16 k human labels, the improvements from SAL to DAL, as well as from DAL to rDAL, are significant at the 5% level according to a one-tailed t-test (mean UAR across 20 independent runs). As expected, the

annotator-aware approach rDAL is also more efficient in terms of annotation costs compared to the basic DAL method with random assignment of annotators. The advantage of rDAL becomes evident from the finding that the basic DAL algorithm does not outperform SAL at $a = 2$, which can be explained by insufficient reliability of randomly assigned annotators. For the agreement level $a = 3$, the accuracy improvement by rDAL compared to basic DAL is minor since in this case it is guaranteed that the same gold-standard labels are acquired.

To substantiate the findings, the annotation cost to achieve a given UAR = 68.0 % was compared among the different methods ($a = 3$). With DAL, a cost reduction of 31 % can be achieved compared to SAL. In accordance with the previous findings, further improvement up to 41 % cost reduction can be achieved with rDAL. Overall, the results clearly demonstrate that SAL falls short of making best use of human workforce.

In summary, a novel DAL method has been proposed to reduce human effort in annotating subjective tasks. The DAL algorithms incorporate dynamic query strategies in an active learning setting, thus shifting the viewpoint from majority voting among a fixed number of annotators to sequential annotation procedures. In the next Chapter, the DAL method will be extended to regression problems, using more generic measures of model confidence and interrater agreement. In future work, the rDAL algorithm will be enhanced by updating the rater reliability after each iteration. To avoid repeatedly querying the same annotators while others never get their turn, it is considered to combine DAL and rDAL in a meaningful way.

Chapter 5

A Generic Human-Machine Annotation Framework

We are drowning in information but starved for knowledge.

John Naisbitt

In the previous Chapter, DAL is proposed as a technique to reduce the cost of labelling data for subjective tasks by striving to reach an interrater agreement with as few human annotators as possible in a query queue. In this Chapter, DAL is defined as an instance of a generic learning algorithm, which is aimed at leveraging the abundant resources of unlabelled data. Based on the cooperative learning approach, a novel annotation framework is developed, which efficiently distributes the labelling work among human and machine to achieve the optimal tradeoff between label quality and cost reduction. To this end, the dropout method is applied in DNN training to capture model uncertainty for human-machine arbitration. In this way, the most informative instances by uncertainty sampling are passed to human assessment, whereas those instances predicted with high model confidence are automatically labelled by machines. In contrast to the existing algorithms that are only applicable to binary classification tasks, the proposed deep learning based confidence measure is suitable for both classification and regression tasks. In order to adapt the early-stopping criterion of DAL to regression tasks, a variance-based interrater agreement threshold is defined to determine when enough ratings have been obtained to establish the gold standard for each instance. The effectiveness of the proposed method is demonstrated on the 'Degree of Nativeness' task of the ComParE 2015. In the result, the novel dynamic cooperative learning algorithm yields .424 Spearman's ρ compared to .413 with passive learning, while reducing the number of human annotations by 74 %.

5.1 Learning From Limited Labelled Data (But Unlimited Unlabelled Data)

Within the field of machine learning, it is widely acknowledged that "there is no data like more data" (Bob Mercer, 1985). Modern deep learning techniques have achieved new state-of-the-art performances on a vast array of tasks. However, these gains are often difficult to translate into real-world settings as they require large hand-labelled training sets, which are expensive and laborious to collect. In the "Big Data" era with the impending rise of the "Internet of Things", there is an ongoing data explosion. To take advantage of it, recent research has been focused on leveraging massive amounts of unlabelled data available in the public domain, e. g., social media. Techniques for model training using unlabelled data include SSL [265], AL [52, 208] and some combinations thereof (cf. Section 3.4). The overarching goal of such approaches is to iteratively train a model by progressively adding new (machine or human) labelled instances to the training set. Without any need for human intervention, the SSL techniques (e. g., self-training [172]) exploit unlabelled data by using machine predictions. Complementary to SSL, the AL query strategies (e. g., uncertainty sampling) select the most informative data instances for human inspection [264]. Cooperative learning (CL) is a recently proposed approach to combine SSL and AL in a joint human-machine effort, leading to a considerable reduction of human annotations without deteriorating classifier accuracy. For the purpose of arbitration, i. e., to decide which instances can be reliably labelled by machine and which require human inspection, an accurate confidence measure is key to the optimal tradeoff between label quality and cost reduction. For classification tasks, the model confidence can be derived from posterior probabilities of SVMs or other classifiers. It should be noted though that model confidence does not necessarily reflect model uncertainty (e. g., if an out-of-distribution example is extrapolated far from the decision hyperplane, the logistic regression model would result in an unjustified high confidence). Next, for the instances selected for human labelling, the number of annotators can be further reduced by means of an early-stopping criterion as defined in DAL. In case of majority voting, the interrater agreement threshold is simply defined as the minimum number of annotators voting for the same class label. However, the existing CL and DAL algorithms based on SVMs are primarily applicable to binary classification and cannot be generalised to multi-class or regression tasks in a straight forward way.

In order to obtain a generic measure, dropout is applied in DNN training to capture model uncertainty [68]. The distinct advantage of using dropout DNNs is that it enables the estimation of model uncertainty in both classification and regression systems. Moreover, a variance-based agreement threshold is used for early-stopping on the DAL part. The instantiations of the generic learning framework are evaluated on the ComParE 2015 'Degree of Nativeness' task [202], in which the naturalness of non-native English prosody needs to be estimated on a continuous scale (cf. Section 2.3.15).

Overall, the main contributions of this work lie in four aspects:

- The DAL algorithm is extended to the general case of multi-class and regression problems by using dropout training in DNNs.
- A novel dynamic cooperative learning (DCL) algorithm is introduced, combining DAL and SSL.
- A generic human-machine annotation framework is presented, encompassing (dynamic) active learning, passive learning, semi-supervised learning, and (dynamic) cooperative learning in a generic learning paradigm. The framework is applicable to both categorical and continuous-valued labelling schemes.
- The framework is applied to prosody regression, resulting in a substantial reduction of human labelling effort at the same accuracy.

Related Work The generic learning framework employs a measure of model confidence and an interrater agreement threshold for a two-tier decision making process: First, which instance is subject to human inspection; and second, how many annotators are required to ensure label quality for each instance. The DAL algorithm introduced in the previous Chapter takes the first step in this direction by combining AL with dynamic query strategies based on interrater agreement and reliability. Cooperative learning for speech emotion recognition was first proposed by Zhang et al. [262], combining AL with SSL to distribute the labelling work among human and machine oracles. However, their approach is subject to the restriction that its confidence measure is only suitable for binary classification tasks.

Capturing model uncertainty to estimate the generalisation error is of paramount importance in probabilistic machine learning [69, 124]. In classification models such as DNNs, hidden Markov models (HMMs) and SVMs, pseudo-posteriors obtained from softmax or maximum entropy functions ([157, 247]) are often erroneously interpreted as model uncertainty [68]. Regression models output a single vector, which regresses to the mean of the training data, but does not directly reflect model uncertainty [67]. To represent uncertainty in deep learning, Gal and Ghahramani [68] advocated the usage of dropout training in DNNs as a Bayesian approximation to a deep Gaussian process [40, 165] (marginalised over its covariance function parameters). As shown in their work, dropout-based uncertainty achieves considerable improvement in predictive log-likelihood and root mean square error (RMSE) over other methods specially developed for assessing uncertainty, such as probabilistic back-propagation [92] and variational inference [76]. An approach similar to dropout makes use of neural network ensembles in an AL scheme for regression [122]. Based on query by committee [26], the variation of the output of ensemble members averaged over unlabelled data is used to select new training data to be labelled. An evident drawback of this ensemble method is, however, the high computational cost.

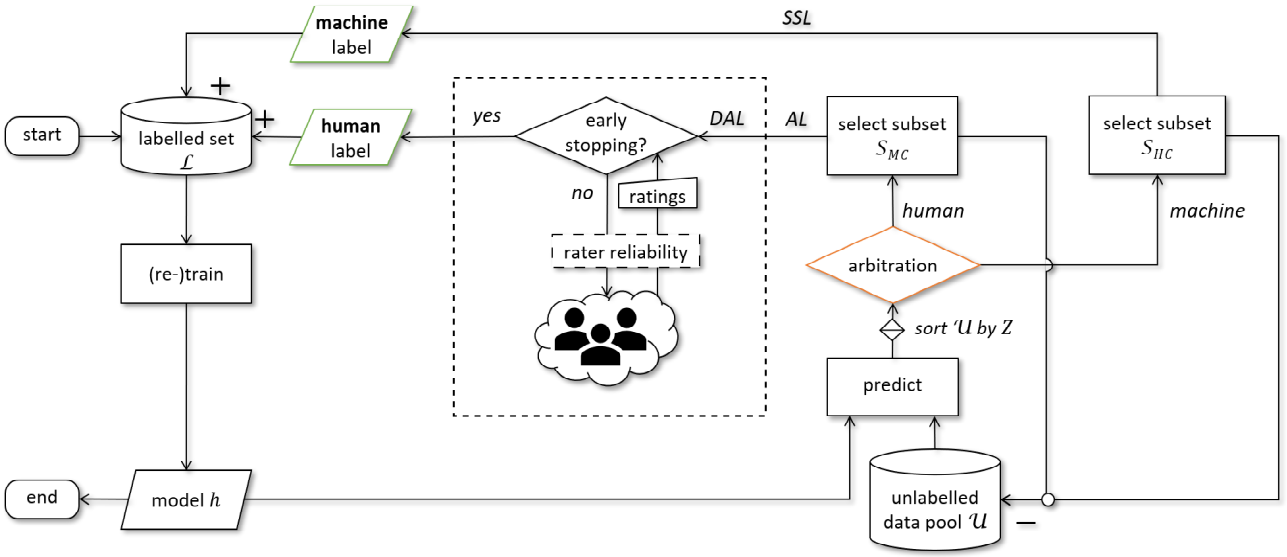


Figure 5.1: Flowchart of the human-machine annotation framework with support for various learning paradigms including semi-supervised learning (SSL), active learning (AL), dynamic active learning (DAL), cooperative learning (CL), and dynamic cooperative learning (DCL); the instances predicted with high certainty (HC) are directly tagged with machine labels, whereas the instances with medium certainty (MC) are subject to human inspection.

5.2 Dynamic Cooperative Learning Framework

In this section, the generic human-machine annotation framework is presented, which can be configured to task-specific requirements within the following degrees of freedom: (i) learning scheme; (ii) confidence measure for efficiently distributing the work amongst human and machine; (iii) optionally enabling early stopping for dynamic annotation procedure based on interrater agreement; and (iv) optionally querying the raters in the order of their reliability ranks.

5.2.1 Learning Paradigm

The functional diagram of the proposed framework is depicted in Figure 5.1. The learning cycle starts with training a model h on a small labelled set $\mathcal{L}$. Then, h is applied to a large unlabelled dataset $\mathcal{U}$, of which the instances are ranked into quantiles according to the model confidence Z. The crux of the cooperative learning framework is the human-machine arbitration, combining self-training and uncertainty sampling (cf. Section 3.4). At this stage, the system either adopts the machine-generated labels (predictions of h) for the instances ranked with high confidence (S_{HC}), or queries human oracles if the model is uncertain. In the latter case, usually, instances with medium certainty are selected for human annotation (S_{MC}) (cf. Section 4.2.1). In each iteration, the selected instances, with labels assigned by either machine or human, are added to the training set $\mathcal{L}$ and removed from $\mathcal{U}$. The model h is then retrained on the increased training

Table 5.1: Instantiations of the generic learning framework. $\mathcal{J}_M$ and $\mathcal{J}_H$ are the sets of instances labelled by machine or human; $Q_{\min}$ denotes the minimum number of queries per instance.

Algorithm	$\mathcal{J}_M$	$\mathcal{J}_H$	$Q_{\min}$
PL	$\emptyset$	Random Sampling	$= R$
SSL	High conf.	$\emptyset$	N/A
AL	$\emptyset$	Medium/ Low Certainty	$= R$
DAL	$\emptyset$	Medium/ Low Certainty	$< R$
CL (AL + SSL)	High Conf.	Medium/ Low Certainty	$= R$
DCL (DAL + SSL)	High Conf.	Medium/ Low Certainty	$< R$

set and evaluated on a speaker-disjunct test set $\mathcal{T}$. For the sake of clarity, $\mathcal{T}$ is not illustrated in the flowchart. The learning process finishes by producing a final model h_{end} when all instances in $\mathcal{U}$ are annotated if not otherwise specified. However, since the model accuracy usually starts to deteriorate at some point due to the noise samples left in the end, the labelling process can be terminated after having reached the best performance on a held-out set, or if the number of remaining instances in the unlabelled pool falls below a threshold $U_{\min}$.

Table 5.1 presents the instantiations of the generic learning framework, which are configured by (i) the work distribution between human and machine, and (ii) the minimum required number of queries $Q_{\min}$ per instance. In AL, the most informative examples are subject to human inspection, while on the other hand, SSL selects the data points with high model confidence for automatic machine labelling. The DAL method varies the number of manual annotations for each instance depending on a predefined agreement level (cf. Chapter 4). For the utmost reduction of human labelling effort, we introduce the DCL algorithm (Figure 5.2), which augments the previous CL method with the DAL strategies. As an exemption, the PL baseline does not rely on any confidence measure by randomly passing unlabelled instances to human annotators.

5.2.2 Dropout as a Generic Uncertainty Measure

To obtain an uncertainty measure from DNNs, we advocate the usage of a Bayesian approximation that was first proposed by Gal and Ghahramani [68], based on a probabilistic interpretation of the dropout method. As described in Section 3.3.2, dropout is a regularisation technique which randomly drops units (along with their connections) from the neural network during training [214]. At test time, dropout can also be used to estimate a model's prediction uncertainty. Conceivably, if the DNN outputs a substantially different result just by omitting single units, the prediction uncertainty of the model is rather high.

Formally, according to the work [68], after training a DNN with dropout, dropout is applied at test time to compute Monte Carlo estimates of the mean and variance of the distribution of the prediction $\hat{y}$ for an input $\mathbf{x}$:

Algorithm: *Generic Form of the DCL Algorithm*
Input: Original training set $\mathcal{L} = \{(\mathbf{x}_i, y_i) \mid y_i \neq \perp\}$
Unlabelled set $\mathcal{U} = \{(\mathbf{x}_i, y_i) \mid y_i = \perp\}$
Pool of raters $\mathcal{R}$
Min. number of queries $Q_{\min}$
Max. number of queries $Q_{\max} \leq |\mathcal{R}|$
Agreement level a for early stopping
Min. number of unlabelled instances $U_{\min}$
Output: Labelled data $\{(\mathbf{x}_j, y_j) \mid \mathbf{x}_j \in \mathcal{U}\}$
Final model h_{end}
Do:
$\mathcal{U} := \{\mathbf{x}_j \mid y_j = \perp\}$
$\mathcal{L} := \{\mathbf{x}_i \mid y_i \neq \perp\}$
$h := \text{Train}(\mathcal{L})$
$\hat{\mathcal{Y}}, C := \text{Predict}(h, \mathcal{U})$ // Prediction + confidence measure
$\mathcal{J}_M, \mathcal{J}_H := \text{Select}(\mathcal{U}, C)$ // Arbitration + instance selection
For $j \in \mathcal{J}_M$: // Machine labelling
$y_j = \hat{y}_j$
For $j \in \mathcal{J}_H$: // Human labelling
Do:
$o = \text{Randomise}(\mathcal{R})$ // Optional: sort $\mathcal{R}$ by reliability
$y_{j,o} = \text{Query}(o, x_j)$ // Ask oracle o about instance $\mathbf{x}_j$
$\mathcal{Y}_j := \mathcal{Y}_j \cup \{y_{j,o}\}$ // Set of human labels
If $(\text{Agreement}(a, \mathcal{Y}_j) \wedge |\mathcal{Y}_j| \geq Q_{\min})$: **Break**
While $(|\mathcal{Y}_j| < Q_{\max})$
$y_j := \text{Gold standard}(\mathcal{Y}_j)$ // e. g., mean, majority vote
While $|\mathcal{U}| > U_{\min}$
$h_{\text{end}} := h$

Figure 5.2: Pseudo-code description of the generic dynamic cooperative learning (DCL) algorithm, which combines semi-supervised learning (SSL) and dynamic active learning (DAL) by means of human-machine collaboration.

$$p(\hat{y}|\mathbf{x}) = \int p(\hat{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}. \tag{5.1}$$

Then, $p(\mathbf{w})$ is approximated by a distribution $q(\mathbf{w})$ over weight matrices with randomly applied dropout. The plausibility of this approximation is corroborated by the study [68]. In fact, neural network training with dropout and L2 regularisation is equivalent to minimising the Kullback-Leibler (KL) divergence between $q(\mathbf{w})$ and a deep Gaussian process $p(\mathbf{w})$ [68]. Sampling K sets of weight vectors $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)}$ from $q(\mathbf{w})$, the Monte Carlo estimate of the mean $\overline{\hat{y}}$ is given by:

$$\overline{\hat{y}} = \frac{1}{K} \sum_{k=1}^{K} \hat{y}(\mathbf{x}, \mathbf{w}^{(k)}). \tag{5.2}$$

In practice, the sampling from $q(\mathbf{w})$ is implemented as performing K forward passes using random dropout masks. The optimal number of K passes mainly depends on the size of the model, the data provided, and the percentage of dropout. A small value for K speeds up the computation, however, it can affect the accuracy of the results and the quality of the uncertainty measure. In analogy, the variance $\sigma_{\hat{y}}$ is estimated as:

$$\sigma_{\hat{y}} = \epsilon + \frac{1}{K}\sum_{k=1}^{K} \hat{y}^2(\mathbf{x}, \mathbf{w}^{(k)}) - \overline{\hat{y}}^2, \tag{5.3}$$

where ϵ is a constant that depends on the model precision (uncertainty of the output). In the human-machine annotation framework, the automatically predicted label is equivalent to $\overline{\hat{y}}$. In the case of regression, high variances are directly interpretable as high uncertainty (low confidence of the DNN in its prediction), and vice versa. This is convenient since only the rank of the instances w.r.t. the confidence measure is important for human-machine arbitration, whereas the value of ϵ, as well as the actual range of the variances – which depend on the data – are irrelevant in this context. Note that the dropout method can be applied for arbitrary network topologies, and in particular, both for regression and classification (softmax) output layers. For classification, the methods as described in the survey [68] such as predictive entropy [209] can be applied to extract uncertainty information from the softmax output layer.

5.2.3 Generic Early Stopping Criterion

A key feature of the dynamic learning framework proposed in this work is *early stopping*, i. e., terminating the rating procedure once the gold standard has been established with sufficient consensus. The definition of the early stopping criterion depends on the labelling scheme, i. e., nominal level or ordinal/ interval scale. For discrete label categories, the previous Chapter defines a deterministic early stopping condition based on the properties of majority voting: if $\mathcal{R}$ or $Q_{\max}$ are given, the voting result cannot further change once a certain number of raters have agreed on a common class. This also applies for ordinal scales if the gold standard is determined by the median. For regression problems, however, the labels are computed as arithmetic mean or EWE and thus can always change with more ratings. Still, the concept of agreement levels can be generalised by means of interrater agreement measures such as the within-group agreement index r_{wg} [108] as defined by:

$$r_{wg} = 1 - \frac{\sigma_Y^2}{\sigma_E^2}, \tag{5.4}$$

where σ_Y^2 is the sample variance of the ratings and σ_E^2 is the expected variance of the random ratings drawn from a null distribution, which is generally assumed to be a uniform distribution [126]. In the case of early stopping, the rating procedure is stopped once the value of r_{wg} rises above a certain threshold a (cf. Figure 5.2). This bears some similarity to the early

stopping technique used in neural network training, where the training procedure is stopped once the generalisation capability of the network is deemed to stagnate. Just as its deep learning analogy, the early stopping of the rating procedure remains a heuristic – there is no guaranteed convergence of r_{wg}. Hence, the convergence of r_{wg} postulates that the considered task is grounded to a certain presence of common-sense norms. In other words, the agreement-based query procedure is targeted at subjective tasks, yet, it is not suitable for annotation schemes that involve strong ambiguities such as voice likeability where divergent views are by all means realistic [203]. Apart from the regression task studied below, appropriate tasks within the realm of Computational Paralinguistics include emotion recognition [192], and interest detection [196], to name but two further examples of subjective ratings. Finally, it is noteworthy that early stopping introduces inherent sequentialism into the rating process: Whether an annotator is queried or not depends on the previous outcomes. Thus, in practical settings, the rating procedure can be implemented by managing annotation requests in a queueing system.

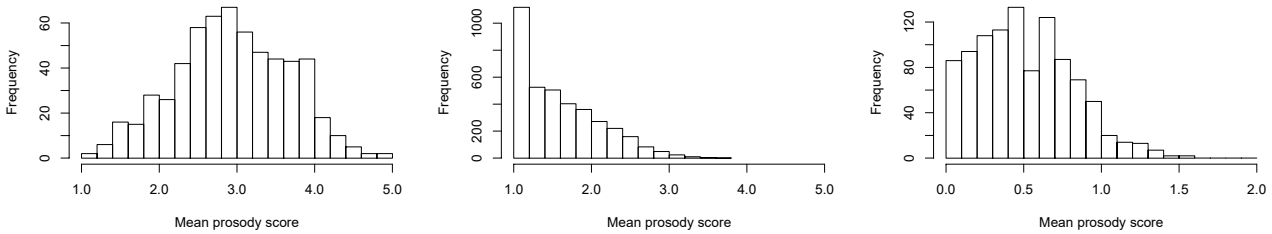
5.3 Experiments

This section reports on the experiments, in which the various instantiations of the proposed framework are applied to a typical regression problem, i. e., non-native prosody regression on speakers learning English as a second language (L2). Located on the word level and above, prosodic speech phenomena encompass word accent position, syntactic-prosodic boundaries, sentence melody and rhythm [101].

5.3.1 Datasets

As in the ComParE 2015 Degree of Nativeness Sub-Challenge [202], cross-corpus experiments were conducted on the NC, AUWL, and C-AuDiT databases, which differed in recording conditions, annotation procedures and speech material. As the initial training set $\mathcal{L}$, the Nativeness Corpus (NC) [36] is used in the first training iteration. To simulate the large unlabelled data pool $\mathcal{U}$, the AUWL corpus [100] is employed. The computer-assisted pronunciation and dialogue training (C-AuDiT) corpus [100] serves as the test set.

Figure 5.3 depicts the distribution of the prosody scores in the data partitions (training set, unlabelled set, and test set). It can be observed that the ratings in the NC database are approximately normally distributed, while the distributions of ratings in both the AUWL and the C-AuDiT databases are positively skewed, i. e., imbalanced towards ‘good’ scores. This shows a typical training/ test data mismatch encountered in cross-corpus settings. In the evaluation, the number of queries per instance only varies on the AUWL database, which represents the



(a) NC (b) AUWL (c) C-AuDiT

Figure 5.3: Distribution of prosody scores in NC (training set $\mathcal{L}$), AUWL (unlabelled set $\mathcal{U}$), and C-AuDiT (test set $\mathcal{T}$) datasets.

unlabelled set $\mathcal{U}$. In particular, for the evaluation on the test set, the gold-standard label, serving as a reference for the predictions, is fixed, i. e., it does not depend on the number of raters requested during iterative training, thus allowing for a fair comparison of the methods employing static or dynamic numbers of queries.

5.3.2 Acoustic Features for Prosody Prediction

The ComParE set of supra-segmental acoustic features [246] is a well-evolved set for automatic recognition of paralinguistic speech phenomena (cf. Section 3.2). It contains 6 373 static features obtained from the computation of various functionals over LLD contours. Of particular importance for assessing the sentence melody and rhythm are the prosodic features based on loudness, energy, pitch, and pauses (cf. Section 3.2.1). A detailed analysis of the relevance of various feature groups for prosody prediction will be carried out in Chapter 8. Further, to reduce the feature space dimension and thus to mitigate the problem of overfitting, CFS (cf. Section 3.1) was applied. According to the BestFirst algorithm [248], feature subsets are constructed in a forward search starting from an empty set by greedy hillclimbing augmented with backtracking. The merit Q is calculated according to Eq. 3.2. The backtracking level (number of iterations without improvement of Q) is set to five. Feature selection is performed on the labelled dataset $\mathcal{L}$, resulting in 1 084 acoustic features. While using the reduced feature set did not improve performance in the experiments compared to the full feature set, it decreased the number of parameters in the rectifier DNN and thereby accelerated the training and evaluation process.

5.3.3 Experimental Setup

The deep rectifier network consists of four hidden layers with 1 000, 750, 500, and 150 units as well as an output layer with a single neuron. The loss function is given by the mean squared error. Optimisation is done via SGD with a constant learning rate of 10^{-4} and a batch size of 64 samples. The initial network is trained for 200 epochs on the labelled set $\mathcal{L}$. The features of the

training set and the unlabelled set are jointly standardised to zero mean and unit variance; the test set is standardised using the same scales and offsets. Re-training after each labelling step is done for 30 epochs. To enable the comparison of the different instantiations of the framework, we save and reuse the initial model (i. e., its architecture and weight parameters). In this way, it is ensured that the learning curves all start at the same point and that the experiments are reproducible. The dropout rate is set to 20 % both in the training and prediction stage. $K = 500$ forward passes with random dropout are used to obtain prosody estimates and confidences on the unlabelled set $\mathcal{U}$. The network topology as well as the training hyper-parameters were determined by a cross-validation experiment on the labelled set $\mathcal{L}$.

In each labelling iteration, 200 instances are labelled by using the proposed framework; in the case of cooperative learning, 100 instances with high confidence are machine-labelled and 100 instances with medium confidence are human-labelled. For early stopping, the agreement threshold a is set to 0.75, representing high agreement as the r_{wg} coefficient is bounded by 1, and $Q_{\min}$ to 2 (the smallest number for which r_{wg} is meaningful). These numbers were validated in a preliminary experiment on the AUWL database. The variance of the null distribution σ_E^2 in Eq. 5.4, which is a discrete uniform distribution on $\{1, 2, 3, 4, 5\}$, is $((5 - 1 + 1)^2 - 1)/12 = 2$. Furthermore, $U_{\min}$ is set so as to label a maximum of 2 000 instances from the AUWL database, i. e., $U_{\min} = 3\,732 - 2\,000 = 1\,732$.

5.4 Results and Discussion

As described in Section 5.2.1, the key component in the generic framework for achieving the optimal accuracy-efficiency tradeoff is the human-machine arbitration based on model confidence. For instance, in SSL, if the model is confident in the case of erroneous predictions, it will train itself with incorrect label predictions, thus gradually deteriorate in performance due to error accumulation. Therefore, it is of crucial importance that the value of uncertainty reflects the magnitude of prediction errors. To validate the uncertainty measure, we train a DNN of the above-mentioned topology for 200 epochs on the NC database and test its predictions and uncertainties against the gold standard (obtained from all raters) on the AUWL database.

The corresponding scatter diagram in Figure 5.4 depicts the positive correlation between the absolute prediction error and the estimated model uncertainty. The ρ of the two variables is 0.658, which is sufficient for the purpose of uncertainty measure. The rationale is that in each iteration a subset of instances is selected based on their uncertainty quantiles (high or medium or least certainty) and hence the exact ranking order is less relevant for the purpose of sampling. On the other hand, it is not reasonable to expect nearly perfect correlation ($\rho = 100\,\%$) since this would imply that the model would be able to correct its own predictions.

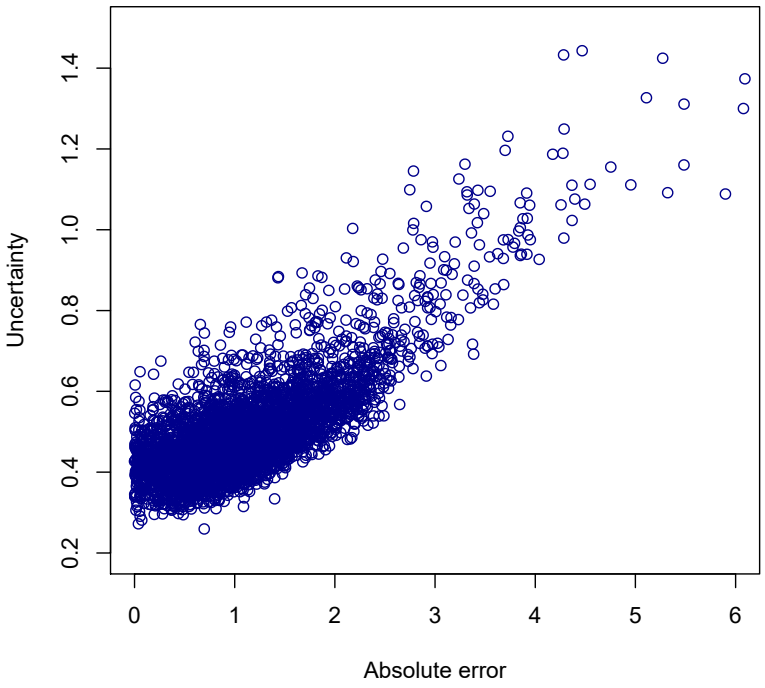


Figure 5.4: Scatter diagram representing the correlation between the absolute prediction error and the estimated model uncertainty on the unlabelled set $\mathcal{U}$ (Spearman's $\rho = 0.658$).

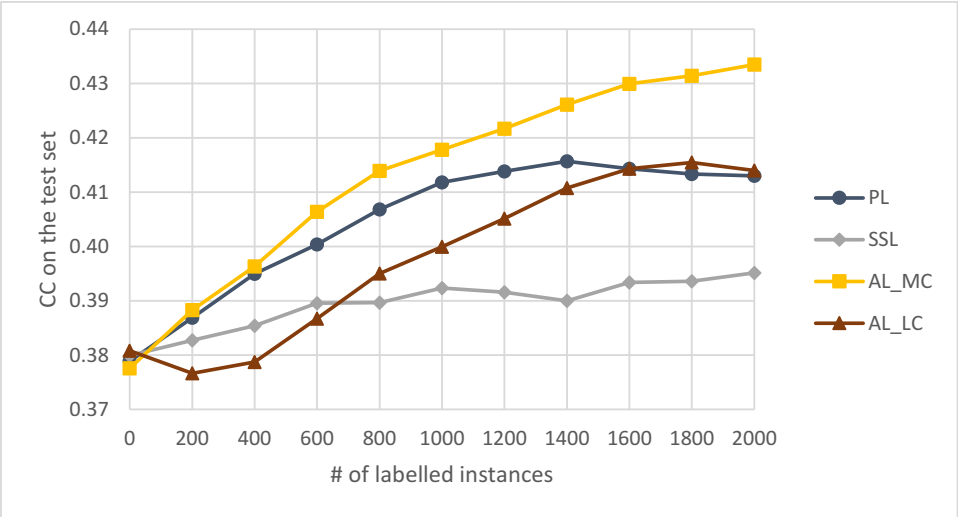


Figure 5.5: Active learning (AL) with the medium certainty (MC) or least certainty (LC) query strategy vs semi-supervised learning (SSL) vs passive learning (PL): Performance in terms of the Spearman's ρ values on the C-AuDiT test set vs the number of labelled instances from the AUWL database.

In Figure 5.5, the learning curves of the standard techniques (without the DAL part which will be investigated later) are compared by plotting the Spearman's ρ as a function of the total number of labelled instances. As can be seen, the sequential addition of labelled instances into the training set $\mathcal{L}$ (200 per iteration) leads to continuous improvement in the recognition accuracy. Overall, the superior performance is achieved by AL with the medium certainty

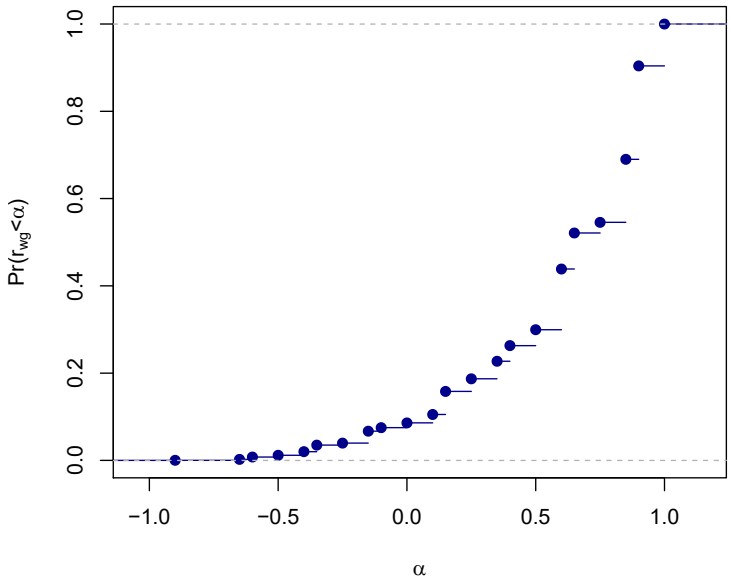


Figure 5.6: Cumulative distribution function of r_{wg} measured on the ratings of the instances in the AUWL database.

strategy, markedly outperforming PL, SSL, and AL with the least certainty strategy in terms of accuracy and smoothness of curves. This initial performance drop of AL with LC can be attributed to the fact that this query strategy is prone to selecting noisy instances, as has been found in the previous works [253, 262] (cf. Section 3.4.2 and Section 4.2.1). The deficient performance (slow rise and unstable curve progression) of SSL can be explained by the fact that prosody recognition represents a typical subjective task; hence, involving human in the loop conceivably enhances the system performance.

Next, we evaluate the early stopping criterion based on the agreement level a, which significantly affects the tradeoff between label reliability and annotation costs. First, we investigate the functional relationship between a given agreement level a and the percentage of rating sequences that would trigger early stopping at this threshold. The probability that the r_{wg} of possible rating permutations takes a value above the threshold a can be approximated by:

$$Pr(r_{wg} > a) = 1 - Pr(r_{wg} \leq a) = 1 - F_{r_{wg}}(a), \tag{5.5}$$

where $F_{r_{wg}}$ denotes the cumulative distribution function (CDF) of the r_{wg} measure Eq. 5.4. The monotonically increasing shape of the discrete CDF estimated on the instances of the AUWL database is shown in Figure 5.6. According to Eq. 5.5, we can estimate the probability of early stopping at 45 % for the threshold $a = 0.75$. It is noted that this is an approximation, as not all rating combinations with a certain r_{wg} value could be encountered in an actual run of the DAL/ DCL algorithm. For instance, the sequence $(3, 3, 4, 5, 3)$ would not occur at $Q_{\min} = 2$ since early stopping would be initiated already after the first two queries.

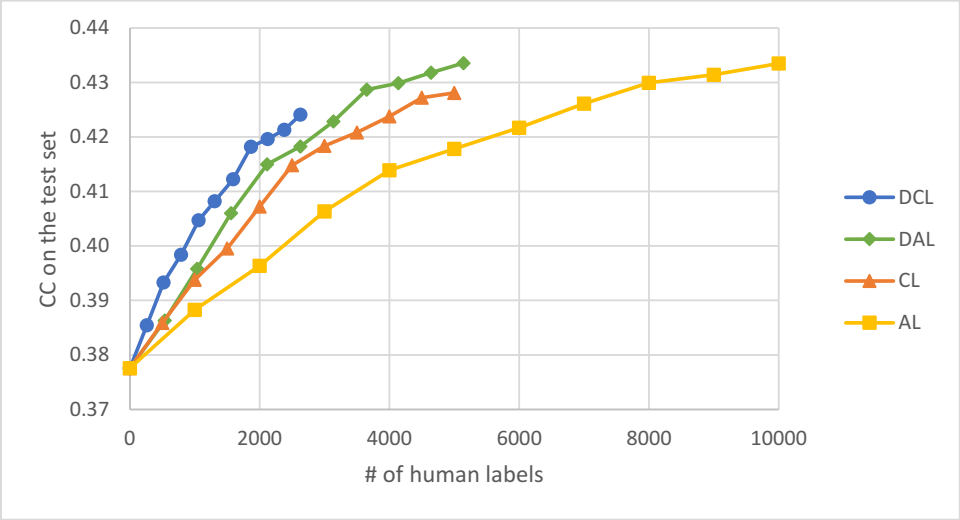


Figure 5.7: Comparison between dynamic cooperative learning (DCL)/ dynamic active learning (DAL), which adapts the number of queries to an agreement level a, and cooperative learning (CL)/ active learning (AL), which sets a fixed number of annotations per instance: Performance in terms of the Spearman's ρ values on the C-AuDiT test set vs the number of human labels.

Taking the AL (MC) method as a baseline, the more elaborate techniques to reduce human labelling work are now evaluated using the MC strategy. As can be observed in Figure 5.7, CL, joining human and machine effort, converges faster to a performance competitive with AL. This coincides with the findings in the work [262]. Augmenting AL and CL with the dynamic query strategy based on early stopping (i. e., using DAL and DCL) markedly reduces the amount of human annotation effort while reaching the same performance. In particular, it can been seen that DCL yields the highest cost efficiency since its learning curve dominates all others.

Table 5.2 allows to interpret the tradeoff between the regression correlation (ρ) and the number of queries to human annotators, i. e., annotation cost, in more detail. Comparing the numbers achieved by (D)AL and (D)CL, it can be seen that the early stopping technique to dynamically adjust the number of raters on a per instance level provides an annotation cost reduction (CR) by roughly 1/2 at similar performance. In particular, the method is found to be superior to the approach of CL [262], which reduces the human effort by combining AL with self-training. In the result, the proposed DCL algorithm is capable of a CR of almost 3/4, demonstrating a multiplicative efficiency gain by combining CL with the dynamic query strategy. Finally, it is noteworthy that the empirically found cost reduction by 1/2 is reasonably close to the 45 % probability of early stopping, which was approximated a priori based on the consideration of the r_{wg} cumulative distribution (Figure 5.6).

In summary, a novel algorithmic framework for human-machine annotation of large databases in paralinguistic dimensions has been presented. As a highlight, the generic annotation framework based on the proposed dynamic cooperative learning technique is able to automatically distribute the labelling work between humans and machines, combining uncertainty sampling (AL) and

Table 5.2: Efficiency of the iterative learning methods with static and dynamic number of annotators, in terms of cost reduction (CR) vs Spearman's ρ on the C-AuDiT test set after 10 iterations, i. e., 2 000 labelled training instances from the AUWL database.

	ρ	# Queries	CR [%]
Static # of annotators per instance			
PL	.413	10 000	0
AL	.433	10 000	0
CL	.428	5 000	50
Dynamic # of annotators per instance			
DAL	**.434**	5 143	49
DCL	.424	**2 630**	**74**

self-training (SSL). For arbitration, the model confidence is assessed by applying dropout to DNNs, which is suitable for nominal level and ordinal/ interval scale annotation tasks. In realistic cross-corpus experiments, the empirical results on the ComParE task of scoring the language proficiency of non-native English speakers substantiate that DCL considerably reduces human labelling effort for 'mildly' ambiguous tasks, for which a consensus among raters conforming to social norms can be presumed. Moreover, the performance of the DNN regressors on the C-AuDiT test set has proven competitive with the Challenge baseline ($\rho = .415$). Finally, it remains to be mentioned that the cooperative learning algorithms have already found use in practical applications as demonstrated in the work [240].

Part III

Holistic Speaker Analysis

Chapter 6

A Guide to the Paralinguistic Universe

The Total Perspective Vortex derives its picture of the whole Universe [...] from, say, one small piece of fairy cake.

The Hitchhiker's Guide to the Galaxy
Douglas Adams

In Chapter 2, the topology of paralinguistic patterns is visualised in a schematic representation, depicting the ComParE tasks in a two-dimensional space spanned by the axes 'time variance' and 'label ambiguity'. In this Chapter, we revisit the paralinguistic space using a data-driven approach, with the aim to discern a big picture, of which the dimensional scales are still to be construed. To this end, a formal notion of task relatedness is devised by computing the intertask correlations from the feature-label correlations. By means of non-metric dimensional scaling, the ComParE tasks are plotted according to different measures of task similarity. Based on their distances in the paralinguistic space, related tasks are discovered. This is conducive to positive knowledge transfer, which is key to high generalisation performance when applying MTL methods. Thus, the paralinguistic atlas presented in this work serves as a guidance for holistic speaker analysis, thereby laying the foundation for the following Chapters.

6.1 The Whole and Its Parts

The thesis posits that paralinguistic phenomena should be viewed in a holistic context, not just as a collection of isolated entities. In search of the big picture, like playing a puzzle, it makes sense to tessellate the coherent pieces of recognisable parts first before assembling larger patterns. Accordingly, the pivotal question to address regarding holistic analysis is which tasks are related and may mutually benefit from joint learning. What might sound like 'cherry picking' in a paradoxical sense (as holism espouses an all-encompassing view), is actually a fundamental

aspect of human cognition. For instance, intuitively, emotion recognition is more related to personality assessment or conflict detection than accent identification. To some extent, all paralinguistic phenomena can be seen as directly or indirectly related, yet, the strength of the interconnections varies.

A substantial body of research supports the premise that task relatedness is the crux of inductive transfer learning [54, 150, 173], and multi-task learning [16, 18, 125]. In particular, it has been evidenced that learning dissimilar tasks together can be even counter-productive due to *negative transfer*, which happens when the performance on these tasks deteriorates. Thus, to reap the benefits of MTL, it is crucial to know a priori which tasks are suitable to be learnt in parallel.

As Caruana [30] stated, one of the most important open problems in MTL is to find an adequate notion of task relatedness. In the schematic diagram (cf. Figure 2.2, Chapter 2), a taxonomic space in the dimensions 'time variance' and 'label subjectivity' has been visualised to surmise the topology of paralinguistic phenomena. Hypothetical task relations are indicated as connecting lines, where the intertask distances are not quantitatively ascertained, but result from the rough positioning of the tasks in dependence on personal assessment.

To substantiate the assumption that a paralinguistic space emerges from the phenomena studied so far, a data-driven approach based on non-metric dimensional scaling is proposed. In contrast to the heuristic method, the dimensional scales of the new maps are not predetermined, allowing the flexibility to explore the topological space using different distance measures. To this end, a $S \times S$ similarity matrix is computed from the feature-label correlations of individual tasks, representing their underlying interrelations. The formed clusters of tasks reveal which of these are closely related and would potentially benefit from each other. By this, it is shown that the parts function towards the whole, which, in turn, gives a conformation or structure to its parts; this precisely corresponds to the quintessence of holism that says "the whole is in the parts and the parts are in the whole" [212].

Related Work Within the active research area of "learning to learn" [224], task relatedness has been extensively studied for MTL to provide a guidance on how to achieve better generalisation performance in avoiding negative knowledge transfer. Evgeniou and Pontil [59] suggested a kernel method based on the minimisation of regularisation functionals, which uses a task-coupling parameter to model the task relations. Kumar and Daume [125] proposed an approach to learning task grouping and overlap structure [8], with the aim to selectively share the information among related tasks. Following the assumption that task parameters lie in a low dimensional subspace, their model allows the tasks in different groups to overlap with each other in one or more bases. Another Bayesian task clustering technique through the design of prior distributions was implemented by Bakker and Heskes [12], who advocated more differentiation in the similarities between tasks. Lee et al. [128] devised an algorithm for simultaneously learning the informative

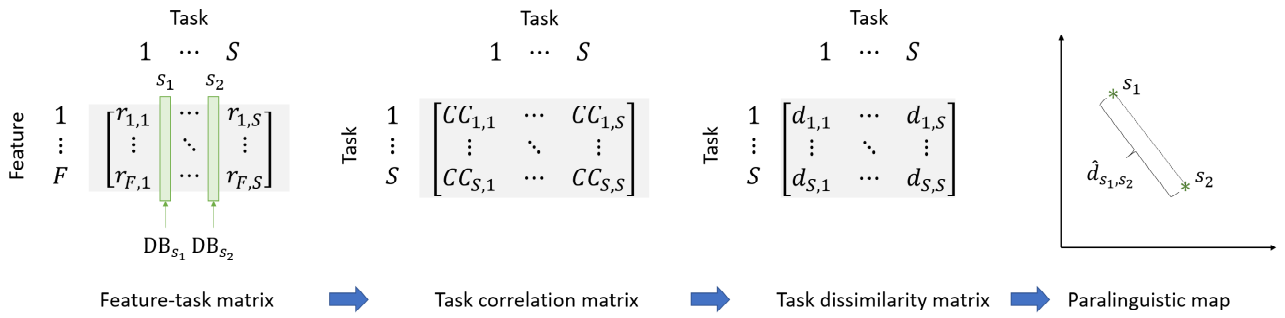


Figure 6.1: Schematic of the proposed method for visualising task relatedness.

prior on feature relevance from an ensemble of related prediction tasks that share a similar relevance structure. Ben-David and Schuller [18] introduced a notion of task relatedness based on a data generation model for learning situations in which data for the same classification task is collected using different recording devices. However, to the author's best knowledge, there is no systematic study on task relatedness in CP, and neither has it been attempted to visualise the topology of paralinguistic phenomena in a data-driven way.

6.2 Measures of Task Relatedness

The holistic concept builds on the assumption that all paralinguistic tasks are related, but some are more strongly interlinked than others.

The proposed measures of task relatedness are established in the acoustic feature space $\mathcal{X}$. Let us denote the total number of paralinguistic tasks by S. Then, for each task $s \in \{1, \ldots, S\}$ and acoustic feature $f \in \{1, \ldots, F\}$, the correlations $r_{f,s}$ between features and labels are computed on the task-specific databases. It is assumed that each task is either a binary classification task or a regression task, ensuring that the feature-label correlation is well-defined.

In case of an F-dimensional feature space, this yields an $F \times S$ feature-task matrix (cf. Figure 6.1), of which the s-th column is an F-dimensional vector $\mathbf{r}_s$ describing task s in terms of feature correlations. If the feature-task matrix consists of absolute correlation coefficients $|r_{f,s}|$, the matrix describes each task in terms of feature *relevance*, where high relevance covers both strong positive and strong negative correlation. Based on the feature-task matrix, we can establish a task correlation matrix $\mathbf{C} = (CC_{s_1,s_2})$, where CC_{s_1,s_2} is the correlation coefficient of $\mathbf{r}_{s_1}$ and $\mathbf{r}_{s_2}$. Thus, a value of CC_{s_1,s_2} near 1 indicates that task s_1 and s_2 have very similar acoustic correlates (or acoustic feature relevance). Conversely, if CC_{s_1,s_2} is near -1, the tasks can be considered 'antipodes'.

Note that task antipodes in the case of feature *correlation* would be strongly related if the labels of one task were swapped (in case of binary classification), or the sign of the labels was reversed

(in case of regression). Thus, one can still efficiently learn such tasks jointly after transforming the labels. However, for task antipodes in the case of feature *relevance* it is expected that they *cannot* be learnt effectively in a common feature space.

Finally, based on the task correlation matrix, two different notions of task dissimilarity can be established. One is to compute the task dissimilarity matrix based on *signed* correlations as $\mathbf{D} = \mathbf{1} - \mathbf{C}$, where $\mathbf{1}$ is a matrix of ones. In this case, the task dissimilarity d_{s_1,s_2} is in the interval $[0, 2]$, and the highest dissimilarity '2' is obtained for the antipodes, while medium dissimilarity '1' is obtained for unrelated tasks. Another possibility is to take the *unsigned* task correlation, where the absolute value is employed: $\mathbf{D} = \mathbf{1} - |\mathbf{C}|$. This yields dissimilarities in the interval $[0, 1]$, of which the maximum of 1 indicates no relatedness. For the reasons outlined above, the latter measure is only employed for the case of feature correlation, as task antipodes for feature relevance cannot be considered in practical holistic analysis.

Figure 6.1 shows the process of computing the task correlation and the task dissimilarity.

6.3 Non-Metric Dimensional Scaling

After obtaining the task dissimilarities, non-metric dimensional scaling (NMDS) [123, 177] is used for visualising the tasks as points in a dimensional space, where distances approximate the dissimilarities.

NMDS is based on classical multidimensional scaling (MDS). In MDS, given a distance matrix, a corresponding set of points in an Euclidean space is obtained analytically by centring the distance matrix and subsequent eigenvalue decomposition. In this Chapter, the dimensionality of the Euclidean space is set to 2 for the purpose of visualisation.

In NMDS, one operates on a matrix $\mathbf{D}$ of general dissimilarities instead of distances. Thus, in addition to the optimal configuration of points one also looks for a (potentially non-linear) transformation of the dissimilarities into distances. To this end, a loss function called *stress* is minimised w.r.t. the distances $\hat{d}_{s_1,s_2} = |\mathbf{y}_{s_1} - \mathbf{y}_{s_2}|^2$ (cf. Figure 6.1) to obtain the configuration of points $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_S)$. The variant of NMDS used in this Chapter is Sammon's non-linear mapping [177]. In accordance with this work, the stress is defined as

$$\text{Stress}(\mathbf{Y}) = \frac{1}{\sum_{s_1<s_2} d_{s_1,s_2}} \sum_{s_1<s_2} \frac{(d_{s_1,s_2} - \hat{d}_{s_1,s_2})^2}{d_{s_1,s_2}}. \tag{6.1}$$

This loss function is minimised by steepest descent [177]. In order to improve convergence speed, the initial configuration of $\mathbf{Y}$ is set to the classical MDS solution where the dissimilarities are interpreted as distances.

6.4 The Paralinguistic Atlas

The paralinguistic atlas is obtained from 21 ComParE tasks ($S = 21$) (cf. Section 2.4). For binary tasks, the nominal labels are converted to 0 and 1, indicating 'absence' or 'presence' of the states and traits depicted in Figure 6.2. The Addressee task is the only exception, where adult-directed speech is signified by 1. In regard to the regression tasks, the numeric labels are used directly, quantifying the expression of the attribute.

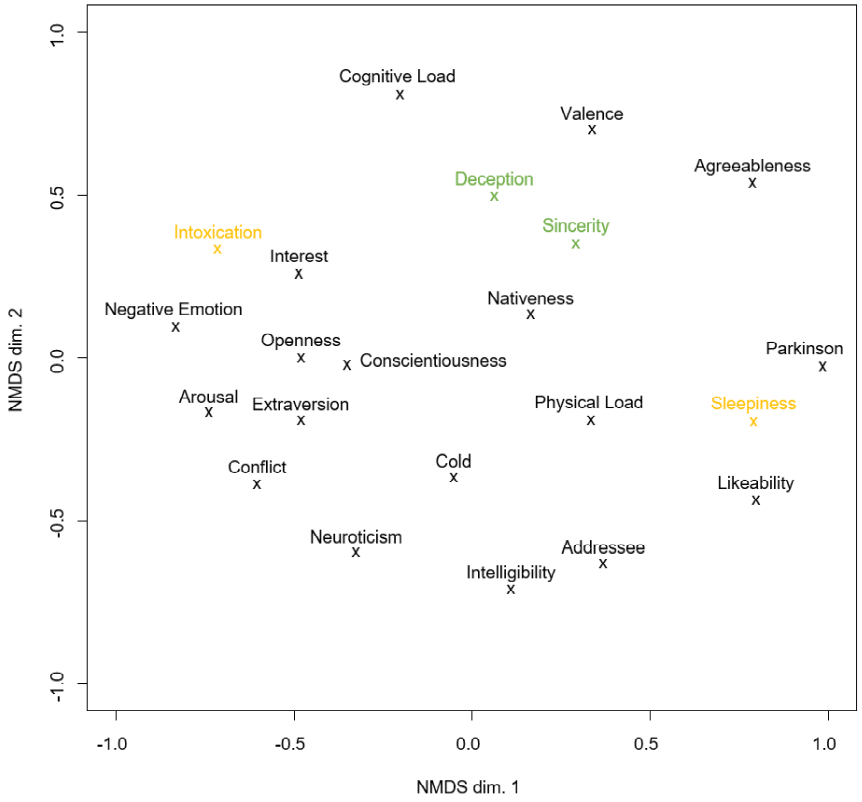
The ComParE acoustic feature set is used ($F = 6\,373$). The feature-label correlations $r_{f,s}$ are computed using Pearson's formula (Eq. 2.4). Both feature correlation and feature relevance are explored to obtain feature-task matrices. Then, the signed task correlation matrices (based on Spearman's ρ) are used to compute the task dissimilarity matrices. For the case of feature correlation, the unsigned task correlation matrix is investigated as well. All in all, this results in three different dissimilarity matrices, each of which is then processed by NMDS to obtain the configurations shown in Figure 6.2. The configurations are sorted by increasing stress of the NMDS solution. In all three cases, the stress is smaller than 0.1, indicating good fit [123].

The configurations show a plethora of interesting relationships worth of further discussion. For instance, in terms of feature correlation and signed task correlation (Figure 6.2a), the personality trait of Agreeableness is an antipode of Neuroticism, which is in accordance with the intuitive relationship of these two traits. However, as expected from the discussion of antipodes above, they lie closely together when moving to the unsigned task correlation or feature relevance points of view (Figures 6.2b, 6.2c). In contrast, the high distance of Agreeableness and Likability is counter-intuitive.

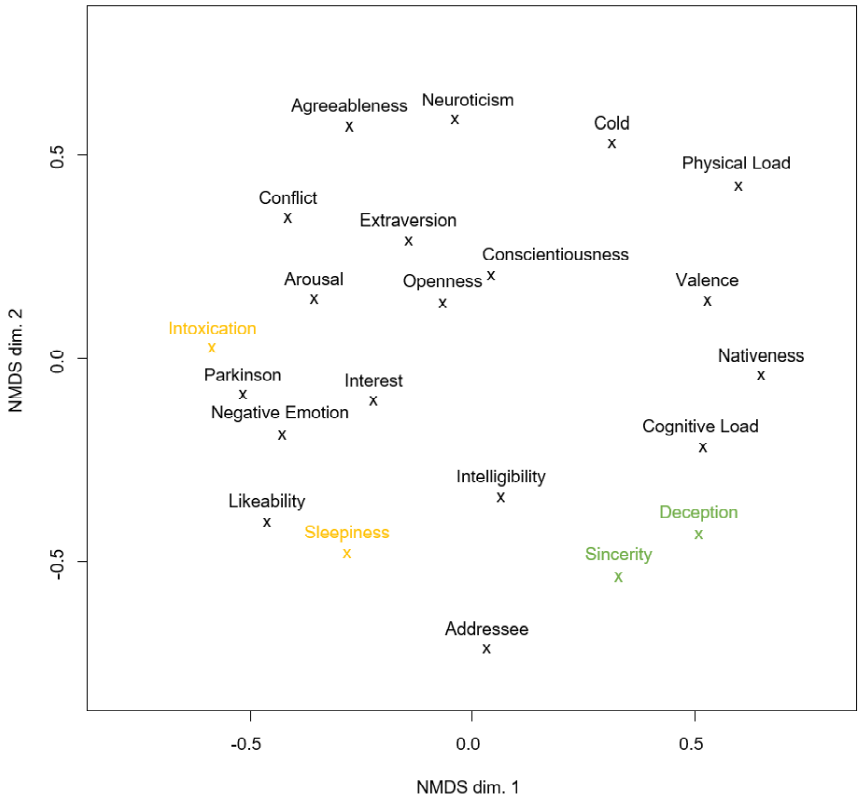
Moreover, the intuitively related tasks of Deception and Sincerity can be identified as candidates for joint learning, as they can be found near each other in all three configurations. The latter is interesting insofar as one would intuitively conjecture that they are negatively correlated. Accordingly, Chapter 8 will dig deeper into the relationship of these speaker states, providing further evidence that this intuitive hypothesis is not true in its generality.

Next, regarding the traits of Intoxication and Sleepiness, it is observed that they are closely related in the configuration computed from feature relevance, yet rather far from each other in terms of feature correlation. This clearly motivates for joint learning of these speaker states in combination with feature relevance analysis, as will be carried out in Chapter 7.

Finally, a 'cluster' of states and traits that can be roughly interpreted as 'active' (Interest, Openness, Extraversion, Arousal, Conflict) can be identified, particularly in the configuration shown in Figure 6.2a. In the same area, we can also find Negative Emotion, which is intuitively related to Arousal given the data collection scenario of the FAU Aibo corpus (cf. Section 2.3.1). In Chapter 10, the joint learning of affect-related states will be presented, which also includes

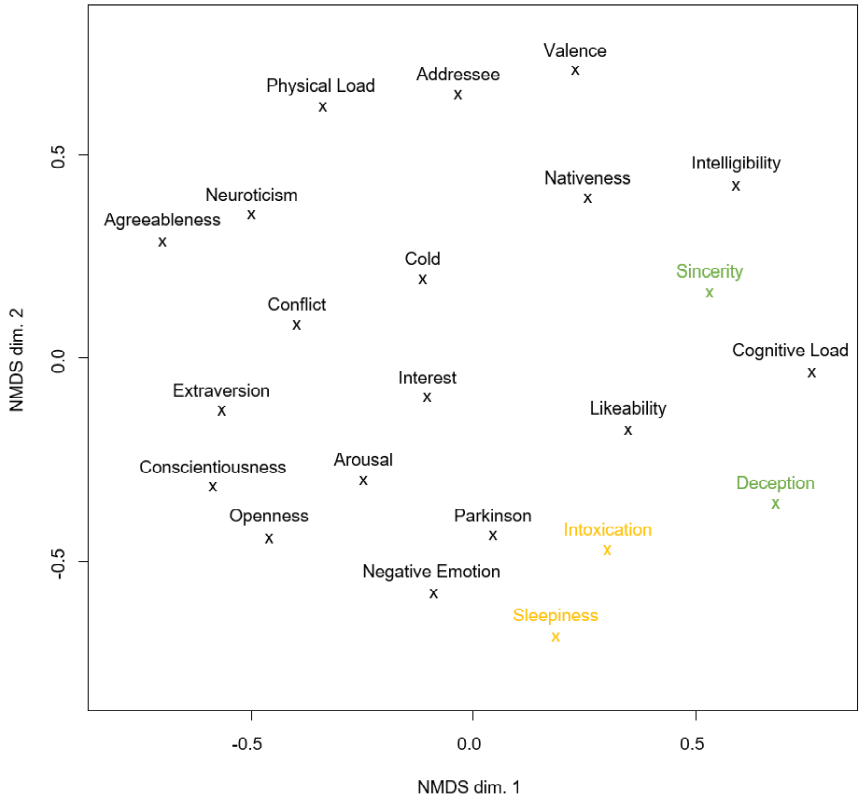


(a) Feature correlation, signed task correlation (stress = 0.05884)



(b) Feature correlation, unsigned task correlation (stress = 0.08719)

Figure 6.2: The paralinguistic atlas based on NMDS applied to various task dissimilarity measures.



(c) Feature relevance, signed task correlation (stress = 0.09551)

Figure 6.2: The paralinguistic atlas based on NMDS applied to various task dissimilarity measures.

multi-class emotion classification which cannot be assessed with the present visualisation methodology.

Summing up, the paralinguistic atlas provides some insights into the topology of speaker attributes, serving as a guidance for applying the adequate MTL techniques to the related tasks. Specifically, the presented maps serve to scrutinise the type of task relations: Figure 6.2a can be consulted for direct data aggregation based on label mapping; Figure 6.2b can be used to identify related tasks for shared hidden layer representation using neural networks (cf. Chapter 10); Figure 6.2c is useful for selecting tasks for joint learning in combination with feature relevance analysis (cf. Chapter 7). These maps can be seen as complementary to the knowledge-based chart with predefined axes (Figure 2.2) insofar as the dimensional scales 'time variance' and 'label subjectivity' cannot be obtained by the data-driven visualisation, where each instance is associated with a time-invariant gold-standard label. In future work, the proposed method will be extended to multi-class problems based on analysis of variance (ANOVA).

Chapter 7

Drowsiness Detection

Much learning does not teach understanding.

HERACLITUS

Motivated by the findings of Chapter 6, this Chapter takes a closer look at the relatedness between alcohol intoxication and sleepiness, which were separately investigated in the INTERSPEECH 2011 Speaker State Challenge. To this end, cross-corpus classification of alcohol intoxication and sleepiness is performed, thereby leveraging the acoustic similarities of these speech phenomena for joint learning. Through feature relevance analysis, task relatedness is exploited by determining the most relevant features on both tasks. In the result, cross-domain classification combined with feature selection yields up to 60.3 % UAR, which is significantly above-chance (50 %) and highly notable given the mismatch in the training and test data. In particular, it is demonstrated that an effective, general drowsiness classifier can be obtained by aggregating the training data from both domains.

7.1 Alcohol Intoxication and Sleep Deprivation

Drowsiness can be caused by inebriation, sleep deprivation and other health issues. The severity of the drowsiness-concomitant cognitive and physiological impairment can be observed in many situations. Drowsy driving due to alcohol intoxication or sleep deprivation frequently results in traffic accidents and is recorded as a high risk factor. Given the nature of the medium-term speaker states which usually continue over a certain period, people are often unaware of the onset of drowsiness, e. g., after alcohol intake. For this reason, automatic monitoring and warning systems which detect diminishing vigilance and performance impairment are of vital importance. Other possible applications are found in the digital health care domain, forensics, and work contexts, especially in safety- and security- critical areas.

Drowsiness exerts a distinct pattern of effects on speech production due to reduced control and coordination of articulation, phonation and respiration, as well as response latency [34]. Perceptual studies have shown that subjects listening to speech samples can reliably discriminate between sober and intoxicated state by means of various linguistic and phonetic characteristics [113, 155, 180]. To measure the effects of alcohol on human speech, continuing research effort has been undertaken for analysing acoustic features [24, 96, 119], phonetic features [17, 213, 229], and emotional changes [39]. From these studies, it has been found that the rate of speech and the overall amplitude decrease after alcohol consumption, whereas the pitch variability, the mean fundamental frequency (F0), and the sentence duration increase [34].

Likewise, sleepy speech manifests changes in prosody (e. g., monotonic intonation, shifted speech rate, reduced syllable duration due to retarded cognitive speech planning), articulation (e. g., slurred pronunciation, speech errors, disfluency), and voice quality (e. g., nasal or breathy speech), as well as linguistic irregularities in syntactic and semantic structures [200]. In accordance with the findings of alcohol intoxication, the sleepiness-induced changes of speech parameters include a general decrease in speech rate [238] and a higher average absolute deviation of intensity [121]. On the other hand, quite opposite effects, for instance a decreased mean fundamental frequency (F0), have been observed [112, 148].

Related Work One of the first studies on automatic alcohol detection was conducted by Levit et al. [129], who reported an accuracy of 69 % in a binary classification task (alcoholised vs non-alcoholised at the alcohol concentration (BAC) threshold of 0.8 ‰). In-car alcohol detection was addressed by Schiel and Heinrich [181] and Schiel et al. [182], using rhythm and formant features. Sleepiness detection from speech has been investigated by Krajewski et al. [121], who obtained an accuracy of 70 %. In clinical studies, the two speaker states have often been investigated in one context regarding their physiological (synergistic and antagonistic) effects. However, the acoustic correlates between alcohol intoxication and sleepiness have not been considered so far. Thus, this study is focussed on the relation between the two prediction tasks, as indicated in the paralinguistic atlas (cf. Figure 6.2c). Given the fact that the actual cause of drowsiness is often unknown in practical situations, it seems meaningful to aggregate the data from both domains to train a general drowsiness classifier. This approach may be particularly useful given the scarcity of genuine intoxicated speech data, since usually plenty of speech is available from speakers being sober, but rarely inebriated.

7.2 Feature Relevance Analysis

Feature selection has been found to be crucial for the performance of cross-domain experiments. A widely used method to reduce the feature space dimension is CFS (cf. Section 3.2), which

serves as a baseline method in this Chapter. To apply CFS for identifying suitable cross-domain features, the merit is measured on the union of the training sets of both domains.

In addition, the cross-domain correlation coefficient (CDCC) measure [246] is applied, which is particularly tailored to cross-domain experiments. The purpose of the CDCC measure is to weigh high correlation in single domains against correlation deviations across different domains. Denoting by i and j the domains (here: intoxication and sleepiness), CDCC is defined as:

$$\mathrm{CDCC}_{f,i,j} = \frac{\left|r_f^{(i)} + r_f^{(j)}\right| - \left|r_f^{(i)} - r_f^{(j)}\right|}{2}, \tag{7.1}$$

where $r_f^{(i)}$ is the correlation of feature f with the domain i. It is obvious that the CDCC measure is symmetric in the sense that $\mathrm{CDCC}_{f,i,j} = \mathrm{CDCC}_{f,j,i}$, and that it ranges from -1 to 1. If a feature f exhibits either strong positive or strong negative correlation with both domains, the CDCC will be near one, whereas it will be near -1 if a feature is strongly positively correlated with one domain yet strongly negatively correlated with the other. A CDCC near zero indicates that the feature is not significantly correlated with both domains (although it might still be correlated with either one). Thus, similar performance across domains can be expected when selecting the features with high CDCC.

7.3 Experiments

For evaluation purposes, the databases ALC and SLC are used, containing in total 60 hours of genuine intoxicated and sleepy speech. The data collection is described in Section 2.3. The distribution of the instances per class is shown in Table 7.1. Due to the skewed class distribution in both the ALC and SLC, the instances of the minority class (NAL and NSL) in the training data are duplicated once, so as to reach an approximately balanced class distribution while ensuring reproducibility.

As in the Challenge, the SVM implementation of the Weka 3 data mining toolkit [86] is used. To mitigate the problem of domain mismatch, feature standardisation is performed by following a straightforward scheme that does not require batch processing at test time. The training set of each corpus (ALC, SLC) is standardised to zero mean and unit variance. When evaluating on a development set, it is standardised using the scales and offsets computed on the training set of the same database. When evaluating on the test set, the scales and offsets are derived from the union of training and development set of the same database. No further standardisation is applied during training and testing.

The first experiment analyses the predictive power of the acoustic features for within-domain and cross-domain classification of alcohol intoxication and sleepiness. To this end, the CDCC

Table 7.1: Partitioning into train, development, and test set according to the setup of the IS11SSC. 'NAL/ AL' and 'NSL/ SL' denote recordings of non-alcoholised/ alcoholised and non-sleepy/ sleepy speakers.

(a) ALC

#	**NAL**	**AL**	Σ
Train	3 750	1 650	5 400
Develop	2 790	1 170	3 960
Test	1 620	1 380	3 000
Σ	8 160	4 200	12 360

(b) SLC

#	**NSL**	**SL**	Σ
Train	2 125	1 241	3 366
Develop	1 836	1 079	2 915
Test	1 957	851	2 808
Σ	5 918	3 171	9 089

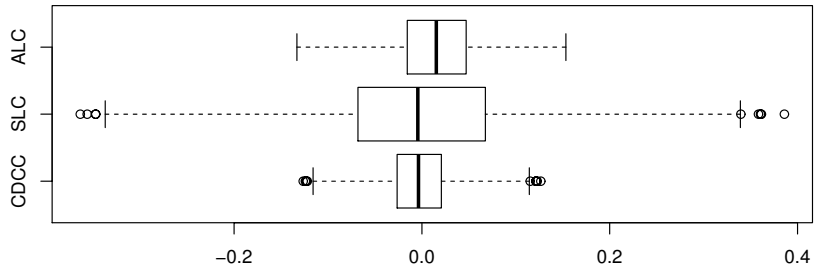
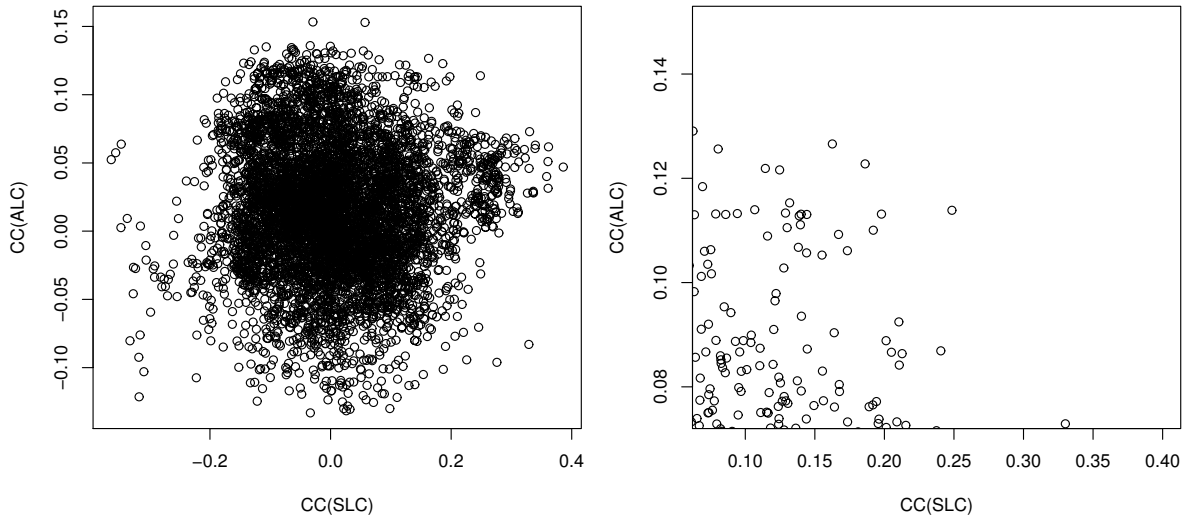


Figure 7.1: Box-and-whisker plots of the distribution of CCs of the 6 373 ComParE acoustic features on the ALC and SLC data sets, as well as the corresponding CDCCs.

and the point-biserial CC of each acoustic feature with the target labels are computed. Here, the CC measure is equivalent to the Pearson's r with the nominal target label converted to a numeric label with the values 0 (non-alcoholised/ non-sleepy) and 1 (alcoholised/ sleepy). The distribution of the three CCs across the acoustic features is shown in Figure 7.1. It can be seen that the correlations on the ALC are much lower than on the SLC; and the maximum CDCC is comparable to the absolute maximum CC on the ALC. Nevertheless, previous studies using the ALC corpus [200] have shown that reasonably robust classification is possible in spite of the generally low feature-label correlations.

Next, to shed light on the cross-domain consistency of the predictive power of the acoustic features, the plot of the CCs on the ALC and SLC training sets is examined (Figure 7.2). It can be seen that measured across all features, there is no strong coherence of the tasks in the acoustic space; yet it has to be noted that the correlation of the CCs on ALC and SLC ($\rho = -0.0270$) is significant at the 5 % level according to a two-tailed t-test. Moreover, there do exist features that are either positively or negatively correlated with both domains. This observation militates in favour of the CDCC-based feature selection.



(a) All features (6 373) (b) 122 features with $CC > 0.075$ on each task

Figure 7.2: Scatterplots of the point-biserial correlation coefficients (CC) of ComParE acoustic features with the target labels of the ALC and SLC training sets.

7.4 Results and Discussion

The cross-domain classification is performed by training on the ALC training data and testing on the SLC test data and vice versa. The obtained results are compared with the in-domain Challenge baseline, i. e., training and testing on disjoint datasets of the the same corpus (ALC or SLC). The evaluation measure is the same as in the Challenge (UAR). The complexity parameter $C \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ is tuned on the development set. Further, the number of CDCC selected features (50, 100, 200) is evaluated on the ALC development set (Figure 7.3). By comparison, the 100 features with the highest CDCC lead to the best performance, which corroborates the assumption that only a certain number of features are correlated with both tasks and thus provide 'reasonable' results (cf. Figure 7.2).

Breaking down the CDCC selected features into feature groups, a generally increasing importance of cepstral features is noticed when gradually reducing the number of features, while the relative importance of prosodic features remains constant (Figure 7.4). At the same time, the share of spectral features continues decreasing, while the voice quality features are not selected at all by the CDCC method.

Table 7.2 summarises the results of the in-domain baselines and cross-domain classification in UAR. Note that the baseline results are slightly better than those reported in the IS11SSC, which can be attributed to the usage of the ComParE feature set, which is larger and more carefully designed than the IS11SSC feature set. As a rule of thumb derived from a one-tailed z-test (at the 5 % level) [47], UARs above 52.2 % can be considered significantly above chance (50 %), which holds for both the ALC and SLC test sets due to their similar size.

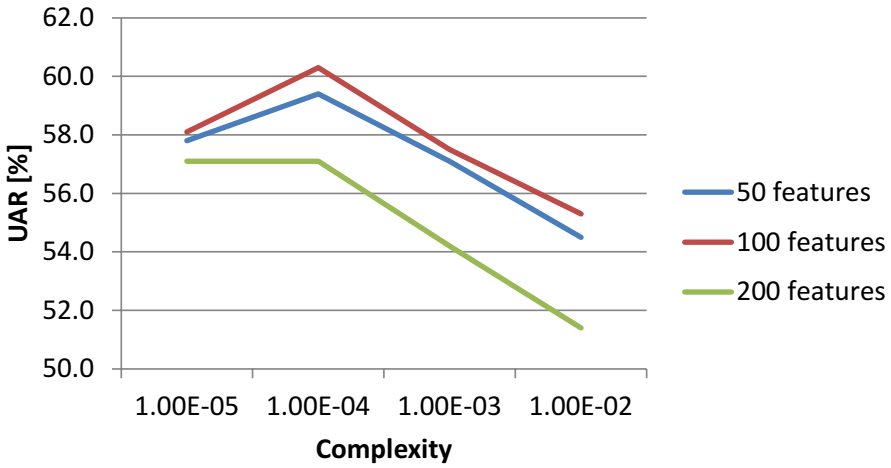


Figure 7.3: Intoxication classification results by percentage of unweighted average recall (UAR) on the ALC development set when training on the SLC and using 50, 100, or 200 features selected by best CDCC.

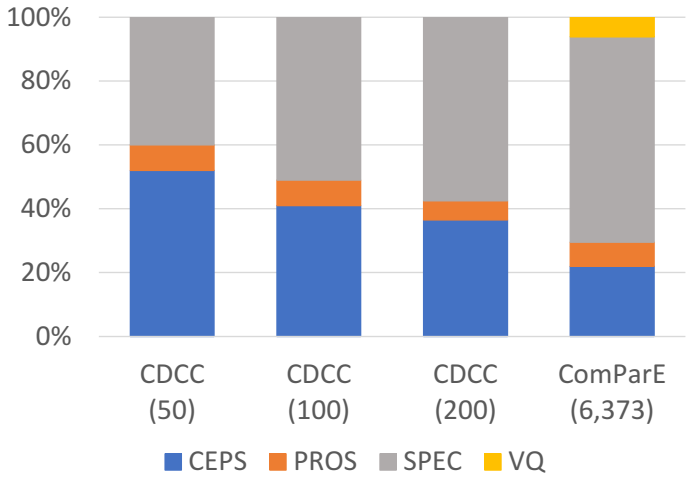


Figure 7.4: Breakdown of feature sets into percentages of CEPStral, PROSodic, SPECtral and Voice Quality features. CDCC(n): Feature selection of n features by highest CDCC.

Summarising the results, it can be noted that on the development set of the ALC, 60.3 % UAR is obtained by training on the SLC training set and using CDCC-based feature selection. This significantly outperforms the result obtained by CFS. Using the entire feature set, only chance level results can be obtained. The picture is similar on the ALC test set, but overall performance is lower – still, the CDCC result is significantly above chance. On the SLC development set, none of the cross-domain classification methods yields a result above chance, yet on the SLC test set CFS gives 59.5 % UAR when using the ALC training data.

Through aggregating the ALC and SLC training data, 63.3 % on the ALC test set and 69.0 % on the SLC test set could be obtained. Although the classification performance on the test sets is worse than when training on in-domain data only, these results argue for a general classification of drowsiness where the specific cause (inebriation or sleepiness) is unknown at test time.

Table 7.2: In- and cross-domain classification results by unweighted average recall (UAR) [%]. Feature selection by correlation-based feature selection (CFS) and cross-domain correlation coefficient (CDCC, top 100 features). C: Complexity of SVM training, tuned on the ALC/ SLC development set.

(a) ALC, training on the ALC (in-domain), SLC (cross-domain), or joint training

Train on	C	Dev	Test
ALC	10^{-4}	66.1	66.5
SLC	10^{-5}	50.1	47.7
SLC (CFS)	10^{-5}	54.7	51.4
SLC (CDCC)	10^{-4}	60.3	55.4
ALC+SLC	10^{-4}	65.5	63.3

(b) SLC, training on the SLC (in-domain), ALC (cross-domain), or joint training

Train on	C	Dev	Test
SLC	10^{-4}	66.5	72.0
ALC	10^{-5}	47.3	50.2
ALC (CFS)	10^{-3}	50.7	59.5
ALC (CDCC)	10^{-2}	51.6	54.7
ALC+SLC	10^{-4}	62.7	69.0

Averaging across the ALC and SLC test sets, data aggregation yields 66.2 % UAR ((63.3+69.0)/2) compared to 58.4 % ((66.5 + 50.2)/2) or 59.9 % ((72.0 + 47.7)/2) (training only on ALC or SLC, test on both), without requiring prior knowledge of the reason for drowsiness.

In conclusion, this study provides evidence for the acoustic correlation of alcohol intoxication and sleepiness. While the overall correlation of acoustic features with the respective target labels severely differs between the two, the usage of cross-domain feature selection by CFS or CDCC enabled classification accuracies significantly above chance on the ALC and SLC development and test sets, employing only out-of-domain training data. These results are all the more notable as the ALC and SLC data comprise genuine intoxicated and sleepy speech recordings, and only a fraction of the baseline features (100 vs 6.373) are used. Finally, automatic detection of general drowsiness can be realised through joint training across the two domains.

Chapter 8

Language Proficiency Assessment

The limits of my language mean the limits of my world.

Ludwig Wittgenstein

This Chapter gives an in-depth analysis of the link between the non-native prosody and the native language (L1) of speakers learning English as a second language (L2). Previously, these two speaker traits were separately addressed in the 'Degree of Nativeness' task and the 'Native Language' task of the Interspeech 2015 and 2016 ComParE. In this Chapter, a multi-task learning scheme based on auxiliary attributes is proposed for jointly learning the tasks of L1 classification and prosody score regression. The effectiveness of this approach is demonstrated in multiple experimental runs using the feature sets of prosodic, cepstral, spectral, and voice quality descriptors described in Section 3.2, as well as automatic feature selection. In the result, it is shown that the prediction of both prosody score and L1 can be improved by combining both tasks. In particular, an 11 % relative gain is achieved in regression performance (Spearman's ρ) on prosody scores, when comparing the best multi- and single-task learning results.

8.1 Non-native Prosody and Native Language

As spoken language applications become more frequent in global business and commerce, it is envisioned that automatic systems performing both language proficiency assessment and native language identification from speech will be increasingly in demand. Non-native speakers of English differ from native English speakers in terms of linguistic (e. g., morphology, syntax, lexicon) and phonetic aspects, including segmental and supra-segmental (prosodic) traits. Located on word level and above, prosodic patterns encompass lexical stress, syntactic timing, intonation and rhythm, hence may be related to the language proficiency in a second language and by that, mutual understanding [101]. Humans are naturally able to adapt to the language

proficiency of their interlocutors, especially in situations where language learners struggle with communication difficulties. Such abilities are required in application areas involving close interactions between human and machines. For example, context-aware spoken dialogue systems may exploit accent-specific acoustic models, adapt the speech rate to the language proficiency of individual speakers, or even switch to a user's native language. Another relevant application field is computer-assisted pronunciation training for providing corrective feedback to language learners [145].

Related Work In the 'Degree of Nativeness' task of the ComParE 2015 [202], the prosodic naturalness of English L2 speakers was to be estimated. Related works targeting automatic recognition of non-native prosody include those by Hönig et al. [99, 101] and Teixeira et al. [223]. In the 'Native Language' task of the following ComParE 2016 [204], the native language (L1) of non-native English L2 speakers from eleven L1 backgrounds was to be automatically identified. In L1 identification from text, the standard approaches make use of features related to spelling errors and the quality of writing, such as character, word and parts of speech (POS) n-grams, function words, and dependency relations [109]. However, the literature on native language identification from speech is still scarce, despite its similarity to the task of spoken language identification (in which a system distinguishes between different languages) [131], and dialect or accent identification (in which a system recognises regional accents of native speakers of the same L1, e. g., American vs British English) [87]. Nevertheless, there exist a number of works on the very similar task of foreign accent identification, using prosodic parameters [88, 152] and phone-based trajectory models [6].

In the general case, the level of accent exhibited in L2 pronunciation depends on a number of other speaker related factors such as the age of L2 learning, the amount of interactive contact with L2, formal instruction, language learning aptitude and motivation [154]. In particular, the L1 background has been found to have a strong influence on L2 foreign accents, occurring due to a negative transfer from the mother tongue to an acquired second language [6]. From a linguistic and phonetic perspective, the global language system can be divided into different language groups. Accordingly, differences in rhythm (isochrony) can be observed between syllable-timed languages such as French, and stress-timed languages such as English [3]. A number of studies [74, 162] indicate that syllables that are weak in stress-timed languages are pronounced stronger in syllable-timed languages, resulting in L1-specific idiosyncrasies and prosodic traits. Moreover, the pitch contours that define intonational functions vary with the L1, e. g., the L1 lexical tone has been found to have an impact on L2 intonation in a tone language such as Chinese [105]. Therefore, it is likely that English L2 speakers from different L1 background show a disposition to produce more or less natural pronunciation [154]. Thus, in this work, the automatic analysis of non-native English speech is carried out by jointly learning both speaker traits to improve the recognition performance on each task.

8.2 Auxiliary Attributes

To exploit the interdependency between non-native prosody and L1 of English L2 speakers, auxiliary features are concatenated in a classifier chain, similar to the work [168]. The idea is to append 'L1' as attribute to the feature vector associated with the target label 'prosody' in the 'Degree of Nativeness' task (DN), and vice versa for the 'Native Language' task (N). Let us first introduce some required notation: $\mathbf{x}_i^{(s)} \in \mathcal{X}$ denotes the i-th feature vector for classification task s, while $y_i^{(s)} \in \mathcal{Y}^{(s)}$ denotes its gold-standard label for task s, where $\mathcal{X}$ is the acoustic feature space and $\mathcal{Y}^{(s)}$ is the label space for task s. $[\cdot;\cdot]$ denotes the concatenation of features. In the case that the ground truth of the auxiliary attribute $y_i^{(s')}$ is given (e. g., collected as meta-data of a database), it can be directly used. Otherwise, if the auxiliary label is missing, a predicted label can be obtained by means of SSL, where a classifier $\hat{y}^{(s')} : \mathcal{X} \rightarrow \mathcal{Y}^{(s')}$ is trained on the original training set of task s'. In this Chapter, $s, s' \in \{\mathrm{DN}, \mathrm{N}\}$. Accordingly, the training data for multi-task learning can be defined as pairs of feature vectors and labels $[\mathbf{x}_i^{(s)}; \hat{y}_i^{(s')}], y_i^{(s)}$, $i = 1, \ldots, L$ and the corresponding test data as $[\mathbf{x}_k^{(s)}; \hat{y}_k^{(s')}], y_k^{(s)}$, $k = 1, \ldots, N$. As the multi-task learning scheme is based on combining acoustic features with auxiliary labels in the feature space, different acoustic feature sets are used to evaluate the complementarity.

8.3 Experiments

To evaluate the multi-task learning method, the following non-native English speech databases from ComParE 2015 (cf. Section 2.3.15)and ComParE 2016 (cf. Section 2.3.20) are employed.

8.3.1 Non-native Speech Databases

The ETS Corpus from the Native Language Sub-Challenge comprises 5 132 non-native English speakers, with eleven different L1 backgrounds (Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR); cf. Table 8.1). Unlike the ComParE 2015 databases, a prosody score is not provided. As in the Degree of Nativeness Sub-Challenge, the AUWL and ISLE databases are used for training while the C-AuDiT database is used for testing. The AUWL corpus [100] comprises 31 speakers, who speak the following native languages: 16 GER, 4 ITA, 3 CHI, 3 JAP, 2 ARA, 1 Portuguese (POR), 1 Hungarian (HUN) (cf. Table 8.1). In the ISLE dataset [138], there are 30 German and 16 Italian native speakers. The C-AuDiT database [97] comprises 58 speakers (26 GER, 10 FRE, 10 ITA, 10 SPA, 2 HIN). Figure 8.1 shows the distribution of prosody scores for each L1 in the AUWL, ISLE, C-AuDiT database family.

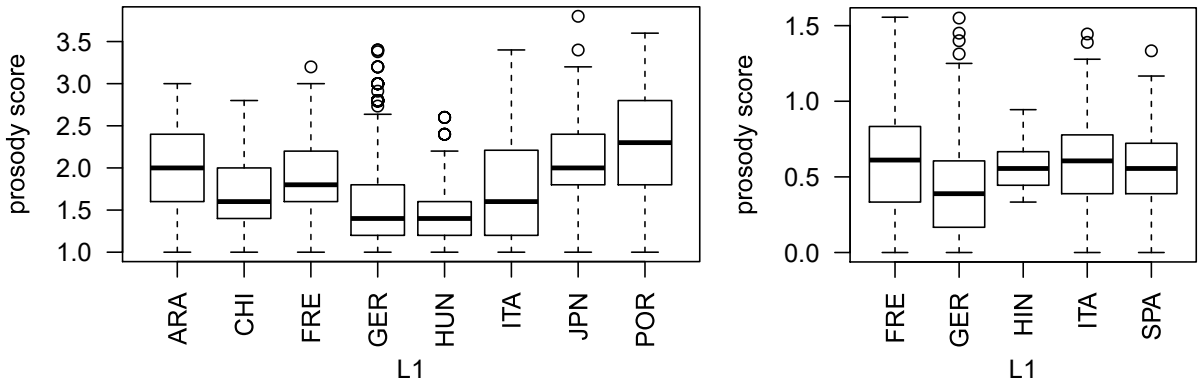
Table 8.1: Databases of non-native Spoken English: Number of instances per class in the train/ devel/ test split used for the Challenges.

#	*ComParE 2015*			*ComParE 2016*			
	Train	Test	Σ	Train	Devel	Test	Σ
ARA	62	-	62	300	86	80	466
CHI	203	-	203	300	84	74	458
FRE	257	166	423	300	80	78	458
GER	2 208	451	2 659	300	85	75	460
HIN	-	37	37	300	83	82	465
HUN	216	-	216	-	-	-	-
ITA	290	176	466	300	94	68	462
JAP	406	-	406	300	85	75	460
KOR	-	-	-	300	90	80	470
POR	248	-	248	-	-	-	-
SPA	-	169	169	300	100	77	477
TEL	-	-	-	300	83	88	471
TUR	-	-	-	300	95	90	485
Σ	3 890	999	4 889	3 300	965	867	5 132

8.3.2 Experimental Setup

As in the ComParE 2015/16, the ComParE set of acoustic features (cf. Section 3.2) is used for prosody regression and L1 identification. Furthermore, relevant subgroups of the ComParE set, i. e., prosodic (PROS), spectral (SPEC), cepstral (MFCC), and voice quality (VQ) features (cf. Section 3.2), are compared with each other. Again, CFS (cf. Section 3.1) is applied for selecting a set of task-specific features.

In the experiments, SVM is used for L1 classification, and SVR for prosody regression; both with linear kernels (cf. Section 3.3.1). For model training, the SMO algorithm is applied. For transparency and reproducibility, open-source implementations of Weka 3 are employed [86]. The epsilon-insensitive loss parameter ϵ is set at 1.0, while the complexity parameter C is optimised through cross-validation on the training set. However, due to the mismatching text materials and recording conditions, the parameters optimised on the validation set may not lead to the best results on the test data. Features are scaled to zero mean and unit standard deviation, using the scales and offsets from the training set. Given the cross-corpus setting of the 'Degree of Nativeness' task, the cross-validation is performed in a 4-fold double nested loop over speakers and texts, in accordance with the ComParE 2015 baseline experiments [202]. In the experiments, the same evaluation measures as in the Challenges are used, i. e., Spearman's ρ for the regression task, and UAR for the classification task. To evaluate the performance in the scenario that the L1 of a test speaker has already been seen in training, additional experiments are conducted on the test set containing only the intersection of the L1 backgrounds found in the AUWL, ISLE, and C-AuDiT databases (FRE, GER, and ITA), for a total of 793 instances.



(a) AUWL-ISLE (b) C-AUDIT

Figure 8.1: Distribution of prosody scores in the AUWL-ISLE and C-AUDIT databases.

Table 8.2: Degree of Nativeness (DN) and Native Language (N); Performance on the test set in terms of ρ and unweighted average recall (UAR), by single-task learning (STL), and by multi-task learning (MTL).

Feature set	#	*DN (ρ)*		*N (UAR[%])*	
		STL	MTL	STL	MTL
ComParE	6 373	.427	.447	47.5	**47.8**
SPEC	2 600	.326	.335	40.8	40.0
MFCC	1 400	.412	.444	39.4	39.8
PROS	483	.366	.474	33.5	34.8
VQ	390	.340	.382	26.8	25.5
CFS	98	.409	**.476**	41.2	42.1

8.4 Results and Discussion

In this section, the findings from the experiments are discussed, where STL using the ComParE set serves as the baseline. From the results in Table 8.2, it can be clearly seen that on the 'Degree of Nativeness' task, MTL consistently outperforms STL for each feature set considered. In fact, the improvement is significant at the 5 % level according to a paired t-test. In particular, the best MTL result yields an 11 % relative gain over the best STL result. The performance gain by MTL over STL is the largest for the PROS feature set (30 % relative). In contrast, only a slight improvement through MTL can be observed for the 'Native Language' task, which is not statistically significant. Further analysis shows that the predicted prosody scores on the ETS corpus all fall in a tight range. This can be explained by a comparable English proficiency of all the TOEFL examinees considered in the study, or a poor performance of the prosody regressor when applied to the ETS corpus.

As regards the performance with different feature sets, the feature selection by CFS has been found particularly effective for prosody regression. The relevance of the auxiliary attributes is brought out by the fact that they are selected by the CFS method out of 6 374 (6 373 + 1)

Table 8.3: Degree of Nativeness (DN) performance by ρ for various feature sets (#: number of features) with STL, MTL, as well as dummy regression corresponding to training mean prediction, on a subset of the C-AUDIT database comprising speakers whose L1 background is also found in the AUWL-ISLE set.

Feature set	#	STL	MTL
SVR on acoustic features			
ComParE	6 373	.435	.459
SPEC	2 600	.338	.348
MFCC	1 400	.416	.458
PROS	483	.402	**.511**
VQ	390	.364	.409
CFS	98	.431	.495
Training mean prediction			
mean pros.	1	0.273	

features, remaining as one of the 136 features for L1 classification and 98 for prosody score regression, respectively. Further, the obtained results corroborate the findings by Coutinho et al. [36] that prosodic features are suitable for assessing the naturalness of English produced by non-native speakers. Interestingly, it has been also found that the various feature groups ranked differently as to their worth for prosody regression vs L1 classification.

In Table 8.3, the results of the 'Degree of Nativeness' task are shown, obtained on the reduced test set of L1 backgrounds that match those of the training set. As expected, the performance is higher on this test set than on the full test set, due to a better match of training and test data. Furthermore, the gain by MTL is similar on both test sets, indicating stable improvement even if a language has not been seen in training. In the result, the usage of prosodic features and auxiliary attributes leads to .511 rank correlation, with a significant gain by MTL over STL according to a z-test ($p < .001$). It remains to investigate if the large gain by including L1 is due to the regressor simply learning the mean score of each L1 group, which can be conjectured due to the unequal distributions per L1 group (cf. Figure 8.1). Table 8.3 also provides the performance of a 'dummy' regressor that, given the L1 of a test speaker, outputs the mean prosody score among the training speakers with the same L1. Note that by design, the performance of the dummy regressor can only be given on the reduced test set. The obtained $\rho = .273$ is significantly lower ($p < .005$ according to a z-test) than any of the performances with acoustic features achieved on the same test set. In addition, Spearman's ρ was evaluated within each L1 group in the test set, using PROS as the best feature group. There, an improved ρ by MTL within each of the French, German, and Italian L1 groups was found, although the difference was smaller in magnitude (up to .044) than the total gain by MTL on the whole test set (.109). Altogether, these findings confirm that the L1 feature helps the regression beyond a simple mean adjustment per L1 group. Thus, it can be concluded that the notable regression performances by acoustic features cannot be fully explained by implicit L1 classification.

Chapter 9

Two Sides of the Same Coin?

The true hypocrite is the one who ceases to perceive his deception, the one who lies with sincerity.

André Gide

This Chapter looks at an exemplary pair of related, but intuitively opposite tasks: Deception (inferable from facts) and perceived sincerity in speech, as addressed in the ComParE 2016. It is shown that the naïve label mapping method based on the assumption that sincerity and deception are just 'two sides of the same coin', i. e., taking deceptive speech as equivalent to non-sincere speech and vice versa, does not yield satisfactory results. However, the interrelation between these characteristics can be exploited for multi-task learning based on joint analysis. To this end, a cross-task labelling technique is proposed to obtain auxiliary attributes as introduced in Chapter 8. By combining data aggregation and multi-task learning via auxiliary features, an effective approach is demonstrated to achieve higher classification accuracy on these tasks. In the result, the proposed method surpasses all participants' approaches on the 'Deception' task.

9.1 Deception and Sincerity

The previous Chapter provides evidence that the joint analysis of non-native prosody and native language of English L2 speakers can be useful for language proficiency assessment. There, the auxiliary attributes are either given as demographic data, or predicted by a classifier trained on a labelled database. Based on the idea of mutual label completion, the existing SSL technique is now extended to an automatic cross-task labelling technique that automatically determines the missing labels in the databases considered for joint learning. To this end, single-task classifiers are trained on the task-specific databases and used to make predictions for the unlabelled data. In this way, a multidimensional label space containing various speaker characteristics can

be generated for holistic paralinguistic analysis using multi-task learning schemes. Moreover, the paralinguistic databases labelled in the same dimensions can be aggregated to gain more training data. A particular advantage of the proposed method is that it is generally applicable to ensembles of related tasks, for which the auxiliary labels are not at hand, as shown e. g., for the classification of personality, emotion and likeability in the previous work [258]. By this means, this approach meets the challenge of multi-label data poverty since there exist numerous speech databases which are designed for a vast array of 'isolated' tasks and thus are annotated in single paralinguistic dimensions (cf. Section 2.3).

The general methodology is benchmarked on an exemplary pair of interrelated yet intuitively opposite tasks: Recognising deception and sincerity in speech as featured in the Interspeech ComParE 2016 [204]. The subtle difference between these tasks is that, beside their antonymous nature, the deceptive speech was collected ad-hoc and is grounded in facts, whereas the sincere speech was perceived and estimated by human judges. As indicated in the paralinguistic atlas (cf. Figure 6.2, Chapter 6), these two tasks are closely related in terms of acoustic features. This finding motivates for the usage of MTL techniques in order to improve the recognition performance on both tasks. To this end, the missing auxiliary labels need to be predicted since the training databases for each task are only annotated in the specific target dimension. Besides improving performance, the study is aimed at shedding light on the question if these speaker traits can be directly mapped to each other (as is often done in emotion recognition, where data associated with different emotion representations are united by label discretisation [190]), or, if there is a more complex interplay, justifying treating them as separate tasks. For this purpose, the effectiveness of heuristic label mapping vs cross-task labelling is investigated.

Related Work Whether for forensics, anti-terrorism or fraud prevention, lie detection has many advantages, if ethically applied. According to Coleman and Kay [33], a lie is defined under the aspects: (a) omission or falsification of content; (b) awareness of such falsehood; and (c) intention to deceive the receiver. Although lie detection is an objective task, there is no clear indicator of a statement's truth or falsehood like Pinocchio's nose, growing longer with each fib that Pinocchio told. Nevertheless, previous research has shown that speech holds important verbal and non-verbal cues to deception [46, 266]. The literature is replete with evidence that deceptive speech exhibits slower tempo, disfluencies, increased intensity range and pitch variance, less pleasant vocal quality, and filled pauses (e. g., um, uh), as well as shorter message length than truthful speech [46, 171]. Of special interest, a higher pitch when lying has been repeatedly reported in earlier works [56, 219]. Based on these findings, the bulk of the studies on deceptive speech detection make use of acoustic [142, 232], vocal tract [225] and lexical features [147], as well as different combinations thereof [75, 95]. Some other approaches [242] capture non-verbal leakage, following the presumption that deception influences the emotional state, the cognitive load and physiological responses the liars may experience [133, 239].

In contrast to deception which is grounded in facts, the perceived sincerity is a highly subjective speech phenomenon due to cultural, interpersonal and other factors. From a philosophical point of view, a speech act is sincere if the speaker is in the state of mind that the speech act functions to express [77]. Another notion of sincerity in speech as spontaneity can be found in [58]. According to Fetzer [63], sincerity is anchored in the speaker's and hearer's attitudes towards the propositions and thus in the illocutionary force. In apology-making context, "I'm so sorry" may sound wholehearted in some situations, ironic in others. In terms of intonation, Alexander [5] found that a high pitch accent and a low boundary tone are associated with sincere apologies while a double pitch accent and a high boundary tone are related to ostensible apologies. Fish et al. [65] found significant acoustic differences in both pitch and intensity; utterances perceived as being sincere are spoken faster and begin with a higher pitch than those that sound insincere, while utterances rated as insincere tend to get louder as the utterance unfolds. Rao [163] endorsed this finding for Mexican Spanish that F0 and intensity increases can be linked to sincerity, whereas sarcasm (one form of insincerity) features a lower speech rate (and hence higher syllable length) and lower mean F0. These findings demonstrate the importance of vocal cues in disambiguating sincere from insincere speech, in an analogous manner to using prosody to infer deception. Yet in the literature, there are hardly any studies on the automatic estimation of perceived sincerity, apart from the participants' contributions to the Sincerity Sub-Challenge.

Notwithstanding the prosodic correlates between deception and sincerity, only very few studies [64, 169] consider these two speech patterns as one task, analysing what prosodic markers and brain processes allow listeners to detect when a spoken compliment is meant to be sincere (true compliment) or not ('white lie'), e. g., "What do you think of my talk?"/ "I find it interesting". Here, the underlying hypothesis is that insincerity is reflected in prosocial lies which are meant to prevent others from being hurt by negative truths. However, to the author's best knowledge, no previous work has been carried out on applying multi-task learning techniques to deception and sincerity classification as stand-alone tasks.

Methodologically, it remains to mention that in a preliminary study [201], 'oracle' experiments were conducted on the Likability database of the ComParE 2012, which contains the labels of the perceived voice likability, as well as the speakers' demographics such as age and gender. Specifically, single-task models were trained for the classification of likability, age and gender, where the ground-truth/ gold-standard labels were used as auxiliary attributes to help improve the accuracy on the target task. As the obtained results show clear signs of over-fitting, poor generation performance can be attributed to very limited amount of training data (800 instances). This motivates for unifying the data resources labelled in single paralinguistic dimensions for joint training by finding ways to handle the missing labels, rather than relying on small multi-label databases. One possibility is cross-corpus data aggregation as advocated by Schuller et al. [194], where multiple emotional speech databases were united for joint training.

Table 9.1: Relationship between different learning settings.

	Training data	Cross-task training data	Test data
Label mapping	–	$\mathbf{x}_j^{(s')}, h^{s'\to s}(y_j^{(s')})$	
Data aggregation:			
– label mapping	$\mathbf{x}_i^{(s)}, y_i^{(s)}$	$\mathbf{x}_j^{(s')}, h^{s'\to s}(y_j^{(s')})$	$\mathbf{x}_k^{(s)}, y_k^{(s)}$
– semi-supervised learning	$\mathbf{x}_i^{(s)}, y_i^{(s)}$	$\mathbf{x}_j^{(s')}, \hat{y}^{(s)}(\mathbf{x}_j^{(s')})$	
– CTL, single-task learning	$\mathbf{x}_i^{(s)}, y_i^{(s)}$	$\mathbf{x}_j^{(s')}, \tilde{y}_j^{(s)}$	
– CTL, multi-task learning	$[\mathbf{x}_i^{(s)}; \tilde{y}_i^{(s')}], y_i^{(s)}$	$[\mathbf{x}_j^{(s')}; y_j^{(s')}], \tilde{y}_j^{(s)}$	$[\mathbf{x}_k^{(s)}; \tilde{y}_k^{(s')}], y_k^{(s)}$

9.2 Learning Schemes for Related Tasks

In this work, different learning schemes are applied to exploit the presumed task relatedness. To start with, a heuristic approach by label mapping is tested. The baseline method is then augmented with data aggregation, while different strategies are used to predict the missing labels in the cross-task training data (cf. Table 9.1). To this end, a cross-task labelling (CTL) technique is introduced, which automatically completes the missing auxiliary labels (cf. Section 8.2) in an ensemble of databases considered for MTL. On top, the obtained auxiliary labels are used to combine multi-task learning and data aggregation.

Let us now introduce some required notation: Assuming a set of S different classification tasks to be performed, $\mathbf{x}_i^{(s)} \in \mathcal{X}$ denotes the i-th training feature vector for classification task s, while $y_i^{(s)} \in \mathcal{Y}^{(s)}$ denotes its ground-truth/ gold-standard label and $\mathcal{Y}^{(s)}$ the set of possible labels for task s. $[\cdot;\cdot]$ denotes the concatenation of features. For the purpose of exemplification and according to the ComParE settings, it is assumed that each task is associated with a task-specific database labelled in the respective target dimension.

Table 9.1 shows a summary of the learning schemes presented in the following sub-sections. It depicts the pairs of feature vectors and labels used from the training data of the target task s, as well as a task s' that is used for cross-task experiments. In this Chapter, $s, s' \in \{\text{Deception}, \text{Sincerity}\}$. In particular, when referring to experiments on the 'Deception' task, the index s represents Deception and the index s' (cross-task) represents Sincerity; in the context of the 'Sincerity' task, the roles of s and s' are swapped.

9.2.1 Label Mapping

The label mapping experiment is based on the assumption that the 'Sincerity' and 'Deception' tasks represent 'two sides of the same coin', i. e., insincere speech is the same as deceptive speech and vice versa. Such heuristic rules can be defined as mapping functions $h^{s'\to s} : \mathcal{Y}^{(s')} \to \mathcal{Y}^{(s)}$ that map a label for task s' to the corresponding label for task s, $s' \neq s$. Using $h^{s'\to s}$, a

Algorithm: *Cross-Task Labelling*
Input: Original data sets $\{(\mathbf{x}_i^{(s)}, y_i^{(s)})\}$, $1 \leq s \leq S$
Output: Cross-labelled data sets $\{(\mathbf{x}_i^{(s)}, \tilde{\mathbf{y}}_i = (\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(S)})^\intercal)\}$, $1 \leq s \leq S$
Initialisation: **For all** i, s':

$$\tilde{y}_i^{(s')} := \begin{cases} y_i^{(s)} & s = s' \\ \perp & \text{otherwise} \end{cases}$$

For $s = 1, \ldots, S$:
  Do:
    $\mathcal{U}^{(s)} := \{\mathbf{x}_i^{(s')} \mid 1 \leq s' \leq S, \tilde{y}_i^{(s)} = \perp\}$
    $\mathcal{L}^{(s)} := \{(\mathbf{x}_i^{(s')}, \tilde{y}_i^{(s)}) \mid 1 \leq s' \leq S, \tilde{y}_i^{(s)} \neq \perp\}$
    $\hat{y}^{(s)} := \text{Train}(\mathcal{L}^{(s)})$
    $\mathcal{J} := \text{Select}(\mathcal{U}^{(s)}, \hat{y}^{(s)})$ // according to high confidence
    For $j \in \mathcal{J}$: $\tilde{y}_j^{(s)} := \hat{y}^{(s)}(\mathbf{x}_j^{(s_j)})$
  While $\mathcal{U}^{(s)} \neq \emptyset$

Figure 9.1: Pseudocode description of the cross-task labelling (CTL) algorithm.

straightforward label mapping scheme can be derived for the task s of interest, by re-labelling the training data from task s'. To obtain discrete labels, the sincerity ratings provided in the Challenge are dichotomised to sincere (S) vs non-sincere (NS) at the threshold of zero. Note that, this is similar in approach to the strategies proposed for cross-corpus emotion recognition in the work [190], where a heuristic mapping between emotional descriptors in various categories and dimensions was defined and applied to a set of corpora prior to joining them to obtain best performance.

9.2.2 Data Aggregation: Baseline Strategies

Label mapping is mainly useful when there are little or no labelled training data for the target task. However, in the scope of this study, it can be assumed that enough labelled training data are available for both domains, and the question is how to combine them for best performance. As a baseline strategy, the data resources are unified by label mapping, and the training sets are used jointly. Furthermore, simple SSL can be applied, exploiting a classifier $\hat{y}^{(s)} : \mathcal{X} \to \mathcal{Y}^{(s)}$ trained on the original training set of task s.

9.2.3 Data Aggregation: Cross-Task Labelling

CTL can be understood as a generalisation of SSL, with the aim to create a S-dimensional multi-label database, where each dimension corresponds to a paralinguistic task. The general idea is to use a set of iteratively trained task-specific classifiers to mutually complete the missing labels in $\mathcal{Y}^{(s)}, 1 \leq s \leq S$ for all the databases considered. The algorithm is depicted in Figure 9.1.

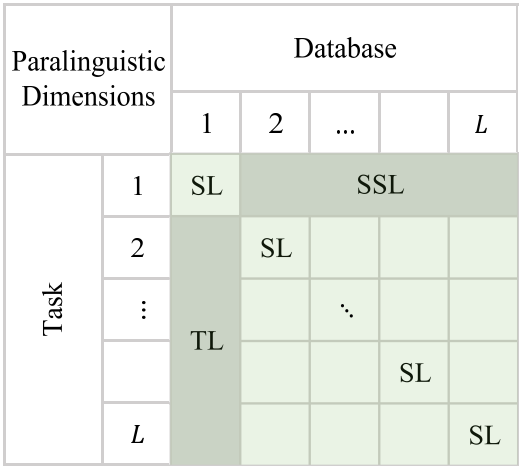


Figure 9.2: Holistic paralinguistic matrix: Cross-task labelling (CTL) applied to a set of S task-specific databases labelled in the target dimension to create a holistic multi-label database (depicted as area in green); SL: supervised learning; SSL: semi-supervised learning; TL: transfer learning.

Here, $\mathcal{L}^{(s)}$ denotes the labelled training set of a specific task s. $\mathcal{U}^{(s)}$ comprises all the data where the labels for the task s are not provided ($\perp$). In a double nested loop, a form of iterative SSL is applied to each task. In the inner loop, refined classifiers $\hat{y}^{(s)}$ are constructed, based on a training algorithm denoted by Train(). The Select() function returns the indices of instances from the set $\mathcal{U}^{(s)}$ where the classifier outputs have a high confidence. As confidence measure, the absolute difference of the posterior probabilities for the two classes of the binary task is used, as implemented in the DAL algorithm (cf. Chapter 4). For simplicity, it is assumed that all instance indices are unique and every instance j belongs to exactly one 'original' task, which is denoted by s_j. For the selected instances, the labels obtained by the classifier $\hat{y}^{(s)}$ are added to the multi-dimensional label vector $\tilde{\mathbf{y}}_j$ for instance j. Starting from the labelled training data of separate tasks, a holistic multi-label database can be constructed, whose feature vectors comprise the union of instances from all the tasks, and whose labels are defined for all instances in all dimensions (Figure 9.2).

Tables 9.2 and 9.3 show the number of instances in the DSD and SSC (cf. Chapter 2, Section 2.3.18 and Section 2.3.19), as well as those that are added by the CTL algorithm to either of the classes. If we compare the numbers in the column 'Added by CTL' of Table 9.2 with the ones we would obtain by direct mapping (from NS to D, as well as from S to ND), it becomes apparent that there is a 'drift' of NS instances (434) away from the D (297) and into the ND (614) class. It is notable that, a similar tide (with swapped roles of sincerity and deception) is also to be observed in the 'Sincerity' task (Table 9.3). From this observation, it is clear that the CTL algorithm performs labelling against the initial hypothesis of a direct mapping between the classes of the 'Sincerity' and 'Deception' tasks. The classification results below elucidates which of these 'competing' labelling schemes is preferable in terms of accuracy.

Table 9.2: Class distribution of the original instances in the "Deception Speech Database" (DSD), and those from the "Sincerity Speech Corpus" (SSC) added by CTL.

#	Train	Test	Σ	Added by CTL	Σ
D	311	121	432	297	729
ND	747	376	1 123	614	1 737
Σ	1 058	497	1 555	911	2 466

Table 9.3: Class distribution of the original instances in the "Sincerity Speech Corpus" (SSC), and those from the "Deception Speech Database" (DSD) added by CTL.

#	Train	Test	Σ	Added by CTL	Σ
S	326	151	477	661	1 138
NS	329	105	434	894	1 328
Σ	655	256	911	1 555	2 466

9.2.4 Multi-Task Learning

While all the above mentioned methods can be used for conventional single-task learning, it seems promising to combine data aggregation and multi-task learning using the auxiliary features obtained by CTL. The idea is to append the labels for a task s' that is expected to be related to task s as features to the vectors $\mathbf{x}^{(s)}$. In the case that the feature vector belongs to the training or test set for task s, we have to resort to the predicted labels $\tilde{y}^{(s')}$. The test set is separately labelled from the training set. For the cross-task training set (from task s'), the existing annotations for s' ($y^{(s')}$) are used.

9.2.5 Knowledge-Based Instance Selection in CTL

The mapping $h^{s' \to s}$, here from deception to sincerity labels and vice versa, can also be interpreted as an expert rule that can be exploited for instance selection in CTL. In doing so, we assume that instances that violate the constraint imposed by $h^{s' \to s}$, i. e., $\tilde{y}^{(s)}(\mathbf{x}_j^{(s')}) \neq h^{s' \to s}(y_j^{(s')})$, should be dropped from the cross-task training set, as they are considered as outliers.

9.3 Experiments

In the experiments, SVMs are consistently used as classifier (cf. Section 3.3.1). In the first set of experiments, the complexity parameter is kept constant at $C = 10^{-4}$ (the optimised value set in the Challenge baseline). Due to the skewed class distribution in the DSD, the training instances of the D class are duplicated so as to reach a more balanced distribution on the training set.

Table 9.4: Unweighted average recalls (UARs) of the D/ ND classes (Deception task) and S/ NS classes (Sincerity task), using the learning schemes depicted in Table 9.1 with SVMs. Note that, the Sincerity baseline is not an official baseline, as it uses discrete instead of continuous labels.

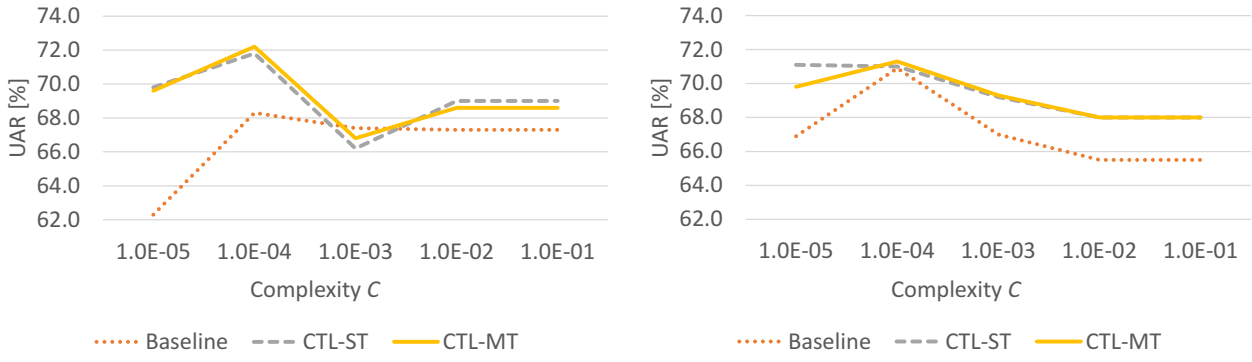
UAR [%]	Deception	Sincerity
Baseline	68.3	70.9
Cross-task methods		
Label mapping	50.7	51.2
Data aggregation:		
– label mapping	66.8	69.4
– semi-supervised learning	69.7	69.1
– CTL, single-task learning	71.8	71.0
– CTL, multi-task learning	**72.2**	**71.3**
– with instance selection	68.9	70.4

Feature standardisation is crucial in cross-task experiments [91, 190]. Here, a straightforward scheme is considered that lends itself to instance-based processing at test time (i. e., no batch processing required). The training set of task s is standardised to zero mean and unit variance. The test set of task s is standardised using the same scales and offsets. Then, the entire data set of task s' is standardised separately to zero mean and unit variance prior to applying the labelling schemes (direct mapping, SSL or CTL). Finally, the standardised training set and the cross-task data set are jointly used to train a classifier. No further standardisation is applied during training and testing.

9.4 Results and Discussion

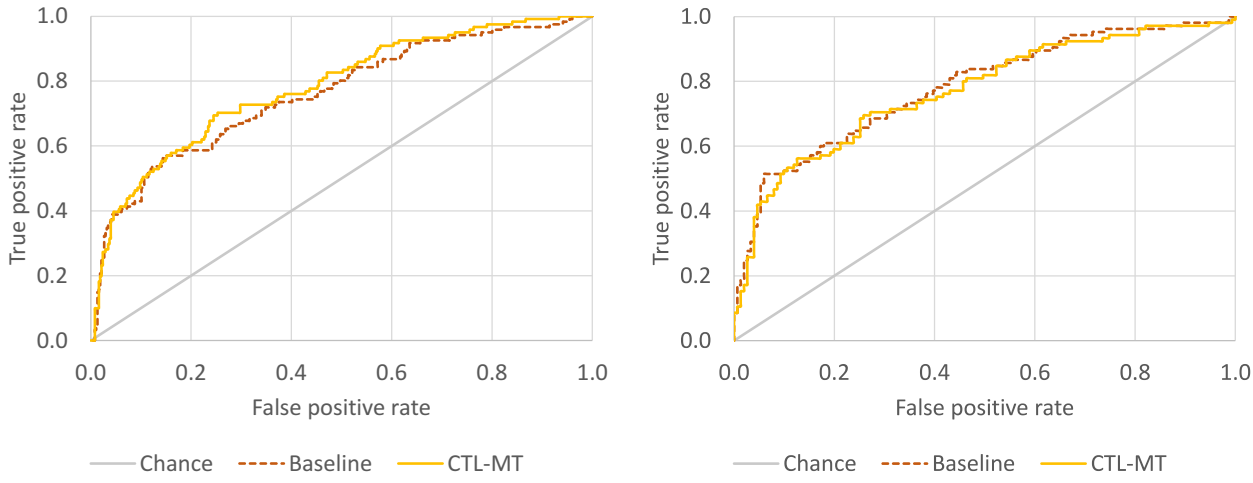
Table 9.4 shows the results in UAR, where the average is taken over the D/ ND and S/ NS classes, respectively. First, it can be seen that label mapping results hover around chance level. It is therefore a fitting result that adding original and re-labelled training sets, as done by the 'data aggregation through label mapping' strategy, merely restores the baseline performance on both tasks.

For the 'Deception' task, simple SSL slightly outperforms the baseline, and CTL yields another gain on top. This can be attributed to the iterative self-training present in the CTL method. The best performance is obtained by combining CTL with multi-task learning, reaching 72.2 % UAR on the official test set of the Deception Sub-Challenge. The improvement from the baseline (68.3 % UAR) to the CTL-MTL method corresponds to a 12 % relative error rate reduction and is significant at $\alpha = 0.1$ according to a one sided z-test. Finally, the results with knowledge-based instance selection fall considerably behind the CTL-MTL UAR. This indicates the importance of those instances that are 'inconsistent' with the intuitive assumption. On the 'Sincerity' task, the baseline is considerably harder to surpass by the data aggregation methods.



(a) Deception (b) Sincerity

Figure 9.3: UAR by SVM complexity C on the Deception and Sincerity task: baseline (ComParE training sets), CTL-STL (cross-labelled training set, single-task learning), CTL-MTL (cross-labelled training set, multi-task learning).



(a) Deception (b) Non-sincerity

Figure 9.4: ROC of deception and non-sincerity detection using the baseline and CTL-MTL methods.

Figure 9.3 depicts the UAR obtained by using different complexity values in SVM training on both tasks. In particular, Figure 9.3b shows that, while the optimal choice of C for the CTL (STL/ MTL) methods leads to slightly better performance than the optimal choice of C for the baseline, for other C values the performance gain is more substantial. Therefore, the marginal improvement on the 'Sincerity' task can be attributed to a sub-optimal tuning of the complexity parameter C (cf. Section 3.3.1). Moreover, the superior performance of CTL (STL/ MTL) at $C = 10^{-5}$ provides evidence for the added value of the cross-labelled instances when the optimisation algorithm is coerced to select fewer support vectors.

Given the fact that there are less deceptive speech utterances than non-deceptive ones, the performance of deception detection is further scrutinised on the basis of ROC curves (Figure 9.4a).

In line with the results above, CTL-MTL yields superior or equivalent true positive rates at a given false positive rate w.r.t. the baseline, resulting in areas under the ROC curve (AUC) of 0.781 (CTL-MTL) vs 0.760 (baseline). For the sake of completeness, the ROC plot for non-sincerity detection is shown in Figure 9.4b.

In conclusion, a study on the automatic classification of deception and sincerity is presented, where the methodology is motivated by the finding from the paralinguistic atlas that these tasks are related. The hypothesis that de facto deception and perceived sincerity are 'two sides of the same coin' is falsified in its generality by the inferior performance of simple relabelling based on this hypothesis. In contrast, the performance analysis of the CTL method demonstrates that an iteratively self-trained classifier is able to select 'appropriate' instances from another task that supplement the available training data. Last but not least, it turns out that the labels obtained by CTL considerably differ from those obtained by the simple label mapping.

Chapter 10

Holistic Emotion Recognition

Colours, like features, follow the changes of the emotions.

Pablo Picasso

Emotion representations are psychological constructs for modelling, recognising, analysing and synthesising human emotions. Due to the subtlety and complexity of emotional experience, the theoretical boundaries between different emotion concepts are often fuzzy, as reflected in the fact that the existing emotion databases are associated with inconsistent annotation schemes. When facing the scarcity of labelled data as an ever present issue for emotion recognition, the straightforward method for joining the existing data resources is to map various emotion labels to one common dimensional space; this, however, comes with considerable information loss. To enable data aggregation whilst avoiding label discretisation, in this Chapter, multi-task shared-hidden-layer deep neural networks (MT-SHL-DNN) are applied, in which the feature transformations are shared across multiple emotion representations, whereas the output layers are separately assigned to each emotion recognition task. In this way, the original meaning and interrelations of manifold emotion descriptors can be exploited for holistic emotion recognition. Through evaluation on nine frequently used emotional speech databases, it is demonstrated that the MT-SHL-DNN method outperforms the single-task DNNs that are trained with one emotion representation at a time.

10.1 Breaking Down the Walls Between Emotion Representations

Emotion has been widely studied in various scientific disciplines like psychology, neuroscience, and cognitive science since long before the advent of Affective Computing in the 1990s [41].

Within the realm of emotion research, the most prevalent emotion concepts stem from the categorical, dimensional and appraisal-based approaches [82].

The categorical emotion theory posits that the affect system is constituted on a limited set of universal basic emotions (e. g., happiness, surprise, fear, sadness, anger, disgust) [55]. However, due to the multi-faceted nature of emotion, many complex and high level[1] affective states such as interest and elation cannot be adequately covered by a few emotion classes. Meanwhile, other evidence construes that affective states are not isolated, discrete entities, but rather present a continuum of intertwining processes, similar to the spectrum of colour [158]. In this light, dimensional models of affect have been developed, mostly drawing on Russell's circumplex model of core affect [175, 176], which represents affective phenomena in a neurophysiological plane spanned by the valence (pleasant – unpleasant) and arousal (activation – deactivation) dimensions. Additionally, a third dimension designating the dominance – submissiveness scale can be integrated, commonly referred to as the PAD (Pleasure, Arousal, Dominance) emotion space [111]. With the emergence of the appraisal theories of emotion [66, 179], focussing on the aspect of emotion elicitation, the fundamental basis for various computational models of emotional processes has been formed. According to the different concepts of emotion, various emotion databases have been designed (Section 10.3.1), using inconsistent annotation schemes and heterogeneous emotion labels [49, 231].

In Machine Learning, the rich sources of information as conveyed by such variety of emotion descriptors are often sacrificed in favour of data aggregation by classifying specific and subtly nuanced emotions to the quadrants of the arousal-by-valence plane. For simplification, some under-represented or ambiguous emotion classes (e. g., pride, surprise) are subsumed under the categories 'other' or 'undefined' [194, 197], and excluded from the recognition task to bypass the problem of label standardisation. The general scarcity of labelled data compounds the problem of label inconsistencies. This particularly affects DNNs without special pre-training, which require large amounts of training data [94].

To break down the wall between different emotion representations across databases, a deep learning based multi-task learning method is proposed to aggregate the valuable emotional data resources for holistic emotion recognition, whilst avoiding information loss owing to label mapping or discretisation.

Related Work The first benchmark for acoustic emotion recognition using DNNs was set by Stuhlsatz et al. [222], showing that the Generalised Discriminant Analysis (GerDA) [221] clearly outperforms the SVM baseline. Like every data-driven tool though, GerDA requires a

[1] According to Buck [25], higher level affects comprise social affects (e.g., pride, guilt, shame, pity, jealousy), cognitive affects (e.g., interest, curiosity, boredom, surprise), and moral affects (admiration and veneration, scorn, and contempt).

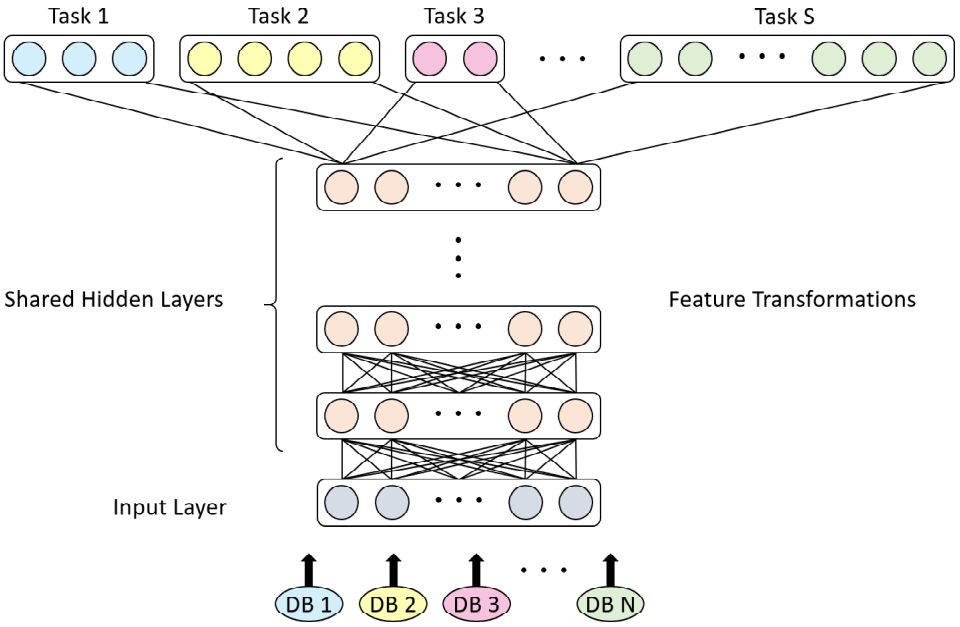


Figure 10.1: Multi-task shared-hidden-layer deep neural network (MT-SHL-DNN) aggregating S task-specific databases (DB), each associated with a different emotion representation.

sufficient amount of data to obtain highly compact and discriminative features, and is thereby inapplicable given a high number of emotion classes with few instances each. Deng et al. [44] introduced SHL autoencoders for learning common feature spaces shared across the training and test set to diminish their discrepancy resulting from different corpora. Unlike the present work, their method does not take advantage of the multi-task learning capabilities of DNNs. For ASR, Huang et al. [104] proposed multilingual DNNs with SHLs to reduce the word error rate over monolingual DNNs trained with language-specific data. Further, it has been shown that the SHLs can be transferred to improve the recognition accuracy of new languages. Moreover, the SHL DNN approach has been effectively applied to language modelling [160]. Recently, recurrent neural network (RNN) based transfer learning for speech emotion recognition from dimensional to categorical emotion attributes has been tested on a single emotion database [70], with moderate success; yet this study did not exploit multi-task learning.

10.2 Multi-Task Shared-Hidden-Layer Nets

Figure 10.1 depicts an instantiation of the multi-task learning framework for holistic emotion recognition. In the MT-SHL-DNN architecture, the input and hidden layers are shared across all emotion representations considered. Extracted from $S = 9$ emotional speech databases, acoustic input features are transformed through the SHLs, while the softmax layers function as log-linear classifiers for different emotion concepts. Accordingly, the number of nodes in each output layer equates to the number of emotion classes for each concept. Unlike the standard DNNs, multiple

output layers are used for emotion classification according to various emotion labelling schemes, such as the valence/ arousal dimensions, or a set of emotion classes.

The motivation for using this network structure is two-fold: First, multi-task learning acts as a regularisation for the network training, since the hidden layer representation is coerced to be predictive for a broad array of emotion representations (cf. Section 3.4.3). Second, it allows an utterance to be interpreted in manifold ways according to various emotion representations. This is unlike the application of shared-hidden-layer DNNs to multi-lingual automatic speech recognition [104] or language modeling [160], where usually utterances are transcribed only in a single language. In the case that such multi-dimensional interpretation of an emotional speech utterance is desired, the proposed multi-task network structure is far more efficient than using a set of single-task networks, since the input-to-hidden and hidden-to-hidden connections have to be computed only once for each input vector, and the number of parameters in each output layer is small.

Mathematically, the output $\hat{\mathbf{y}}$ of the MT-SHL-DNN on an acoustic feature vector $\mathbf{x}$ is composed of sub-vectors for each task $1, \ldots, S$:

$$\hat{\mathbf{y}} = \left[\hat{\mathbf{y}}^{(1)}; \hat{\mathbf{y}}^{(2)}; \ldots; \hat{\mathbf{y}}^{(S)}\right], \tag{10.1}$$

where each $\hat{\mathbf{y}}^{(s)}$, $s = 1, \ldots, S$ corresponds to a transformation of the last hidden layer activation with a task-specific weight matrix $\mathbf{W}^{(s)}$:

$$\hat{\mathbf{y}}^{(s)} = u(\mathbf{W}^{(s)}\mathbf{h}) = u(\mathbf{W}_H^{(s)} u'(\mathbf{W}_{H-1}(\cdots u'(\mathbf{W}_1\mathbf{x})))), \tag{10.2}$$

with an output layer activation function u and a hidden layer activation function u'. Biases are omitted in the above equation for ease of exposition, but are used in the experiments. Optimisation of the parameters $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(S)}\}$ of the MT-SHL-DNN is done via error back-propagation and SGD. Assuming that each acoustic feature vector $\mathbf{x}_i$ in the training set belongs to exactly one task s_i with label y_i, the loss function to be minimised is:

$$J^{\mathrm{MT}} = \sum_i J(\hat{\mathbf{y}}^{(s_i)}, y_i), \tag{10.3}$$

i. e., the outputs for the tasks $s' \neq s_i$ are irrelevant. Hence, the forward and backward propagation mechanism needs to evaluate only one output layer per input vector. In order to increase efficiency, one can restrict each mini-batch to only input vectors belonging to a single task, which we will refer to as *homogeneous* mini-batches. Alternatively, according to the theory laid out in Section 3.4.3, one can conjecture that the performance can be improved by using *stratified* mini-batches, where every mini-batch is representative of the full training set in that it contains input vectors from each task.

Table 10.1: Overview of a selection of frequently used emotion databases (cf. [222]).

Corpus	Content	# / Emotion							# Inst	hh:mm	#Subj	Cond	kHz
ABC	German fixed	aggr 95	chee 105	into 33	nerv 93	neut 79	tire 25	-	430	1:15	4 m 4 f	acted studio	16
AVIC	English free	loi1 553	loi2 2278	loi3 170	-	-	-	-	3002	1:47	11 m 10 f	spont normal	44.1
DES	Danish fixed	angr 127	happ 86	neut 85	sad 84	surp 84	-	-	419	0:28	2 m 2 f	acted normal	20
EMOD	German fixed	angr 127	bore 79	disg 38	fear 55	happ 64	neut 78	sadn 53	494	0:22	5 m 5 f	acted studio	16
eNTER	English fixed	angr 215	disg 215	fear 215	happ 207	sadn 210	surp 215	-	1277	1:00	34 m 8 f	acted normal	16
SAL	English free	q1 459	q2 320	q3 564	q4 349	-	-	-	1692	1:41	2 m 2 f	spont normal	16
Smart	German free	angr 220	help 161	joy 284	neut 2179	pond 643	surp 70	unid 266	3823	7:08	32 m 47 f	spont noisy	16
SUSAS	English fixed	hist 1202	meds 1276	neut 701	scre 414	-	-	-	3593	1:01	4 m 3 f	mixed noisy	8
VAM	German free	q1 21	q2 50	q3 451	q4 424	-	-	-	946	0:47	15 m 32 f	spont norm	16

Abbreviations: Inst: instances; Subj: subjects (m: male, f: female); Cond: condition; spont.: spontaneous speech; aggr: aggressive, angr: anger, bore: boredom, chee: cheerful, disg: disgust, happ: happy, help: helplessness, hist: high stress, into: intoxicated, loi1-3: level of interest 1-3, meds: medium stress, nerv: nervousness, neut: neutral, pond: pondering, q1 - q4: quadrants in the arousal-valence plane, sadn: sadness, surp: surprise, tire: tired, unid: unidentifiable

10.3 Experiments

The experiments are aimed at validating the effectiveness of multi-task vs single-task training, as well as testing different batching schemes and acoustic feature sets (ComParE vs eGeMAPS). For a detailed description of the feature extraction process, the reader is referred to Section 3.2. Considering the highly unbalanced distributions of instances among classes, UAR is suited as evaluation measure. To enable comparability with the previous benchmark [222], the choice of the nine frequently used emotional speech databases is retained, ranging from enacted over induced to spontaneous emotion portrayal. In the following, the emotional speech databases are briefly described.

10.3.1 Emotional Speech Databases

The "Airplane Behaviour Corpus (ABC)" [186] is an audio-visual emotion database created for applications concerning public transport surveillance. The recordings captured mood evoked by prerecorded announcements (e. g., start, turbulences) during a simulated vacation flight. The "Audiovisual Interest Corpus" (AVIC) [187] (cf. Section 2.3.3) features the level of interest of subjects in the scenario in which a salesperson led them through a commercial presentation.

The "Danish Emotional Speech" (DES) database [57] contains recordings of four professional actors, each of them pronounced two words, nine sentences and two passages of fluent text while displaying different emotions. The "Berlin Emotional Speech Database" (EMOD) [27] is targeted at a number of basic emotions expressed by professional actors. The "eNTERFACE" (eNTER) database [135] involves subjects speaking predefined content in English, after having listened to six short stories, each of them eliciting a specific emotion. The "Belfast Sensitive Artificial Listener" (SAL) dataset is a subset of the HUMAINE database [50, 249] comprising audio-visual recordings from natural human-computer conversations. The "SmartKom" (Smart) corpus [217] provides spontaneous speech produced in Wizard-of-Oz dialogues. The "Speech Under Simulated and Actual Stress" (SUSAS) database [89] was designed in four stress-inducing scenarios: Simulated stress, calibrated work load tracking task, acquisition and compensatory tracking task, and amusement park roller-coaster. The "Vera-An-Mittag" (VAM) corpus [80] was extracted from spontaneous and emotionally coloured speech of guests during a German TV talk show.

A comprehensive overview of the databases is given in Table 10.1. From this, the following observations can be made: It can be seen that there are not many overlapping classes, thus any attempt to unify the representations would inevitably result in a loss of information in the label space. Through the MT-SHL-DNN approach, 15 hours of speech data can be aggregated, where each individual database would only yield around one hour of data with the exception of the larger SmartKom database. This clearly illustrates the potential of the proposed multi-task learning approach.

10.3.2 Experimental Setup

In a preprocessing step, the data sets are normalised to zero mean and unit standard deviation. For training multi-task models, two different batching schemes are considered (cf. above). To obtain homogeneous mini-batches, each database was divided into 20 mini-batches; these are shuffled for all tasks in each training epoch. When forming stratified mini-batches, all 15 676 training instances are partitioned into stratified mini-batches containing 100 samples each. The order of the mini-batches was randomised at the beginning of each epoch.

In alignment with the experimental settings of the reference work [222], leave-one-speaker-out (LOSO) validation was conducted for databases with 10 or less subjects, otherwise, namely for AVIC, eNTERFACE, SmartKom, and VAM, leave-one-speaker-group-out (LOSGO) was performed to ensure speaker independence. Here, it is noted that Stuhlsatz et al. [222] selected five speaker groups with equal number of male and female speakers, as well as the same number of speech samples per group for LOSGO evaluation, whereas, in our experiments, the training data are partitioned into five gender-stratified speaker groups.

The network topology consists of five hidden layers ($H = 5$) with rectified linear activation function and output layers with softmax activation function. For the loss function J in the multi-task objective (cf. Eq. 10.3), cross-entropy is used. With the large ComParE feature set, each hidden layer comprises 2 048 neurons, and the training is carried out for 80 epochs. The learning rate is set to 0.01 and a Nesterov momentum term (cf. Eq. 3.37) with coefficient $\mu = 0.9$ is used. The DNNs for the eGeMAPS feature set have 256 neurons in each hidden layer and are trained for 500 epochs. The rationale is that larger networks did not outperform this small topology in preliminary experiments with the reduced feature set. For both feature sets, dropout with probability of 0.5 is applied to the input layer for better generalisation.

In another set of experiments, the pre-trained MT-SHL-DNN is used as a starting point to train improved single-task networks. This is likely to improve performance since the proposed multi-task learning method allows us to leverage multiple training databases and regularise the network training. To verify this, single-task networks are created using the parameters $\mathcal{W}^{(s)} = \{\mathbf{W}_1, \ldots, \mathbf{W}_{H-1}, \mathbf{W}^{(s)}\}$ from the trained shared hidden layers and a single trained output layer for task s. Then, the entire parameter set $\mathcal{W}^{(s)}$ is retrained on the task-specific dataset.

10.4 Results and Discussion

From Table 10.2, two findings can be made. First, training with homogeneous mini-batches seems to be slightly better than stratified mini-batches. This is encouraging since homogeneous mini-batches can be trained more efficiently. Second, regarding both batching schemes, training a task-specific network based on the MT-SHL-DNN initialisation results in better accuracy than the MT-SHL-DNN itself. This is not surprising for the reason that fine-tuning the hidden layer weights for each database leads to a blow-up of the effective number of parameters that is proportional to the number of tasks. Juxtaposing Table 10.2 to Table 10.3, it can be found that the MT-SHL-DNN results are superior to those obtained with the set of single-task DNNs, on average across the nine databases (52.01 vs 51.57 % UAR). Although the improvement is not statistically significant according to a one-sided Wilcoxon signed-rank test, it is remarkable given the compactness of the MT-SHL-DNN compared to the set of nine single-task DNNs. Moreover, Table 10.3 confirms the usefulness of the MT-SHL-DNN as an initialisation for training the set of single-task DNNs, independent of the used feature set. As to the ComParE feature set, the average improvement across databases is significant at the 0.01 level (51.57 to 52.84 % UAR).

Table 10.2: Emotion recognition performance in terms of UAR on the ComParE feature set; Mini-batches are either stratified or homogeneous; SHL: performance of the MT-SHL-DNN; SHL+Retrain: single-task DNNs with MT-SHL-DNN initialisation.

UAR [%]	stratified		homogeneous	
Corpus	SHL	SHL+Retrain	SHL	SHL+Retrain
ABC	**56.11**	54.94	51.26	52.36
AVIC	56.62	**57.23**	56.28	56.08
DES	52.16	54.67	53.96	**55.39**
EMOD	**82.34**	78.63	78.47	81.09
eNTER	66.32	69.11	68.32	**69.19**
SAL	28.78	30.13	**31.05**	30.75
SMART	24.06	24.88	25.70	**25.93**
SUSAS	62.51	63.10	62.41	**64.31**
VAM	38.27	39.39	**40.61**	40.40
Mean	51.91	52.45	52.01	**52.84**

Table 10.3: Single-task (STL) DNNs without pretraining compared to ST DNNs with MT-SHL-DNN initialisation (SHL+Retrain), using the eGeMAPS or ComParE feature set.

UAR [%]	eGeMAPS		ComParE	
Corpus	STL	SHL+Retrain	STL	SHL+Retrain
ABC	46.51	**46.74**	50.80	**52.36**
AVIC	52.39	**54.42**	55.85	**56.08**
DES	53.23	**53.31**	**55.94**	55.39
EMOD	**73.21**	71.30	78.75	**81.09**
eNTER	61.05	**62.94**	68.95	**69.19**
SAL	**32.53**	31.05	28.27	**30.75**
SMART	20.64	**21.95**	24.54	**25.93**
SUSAS	**61.26**	60.58	62.94	**64.31**
VAM	37.78	**38.42**	38.10	**40.40**
Mean	48.73	**48.97**	51.57	**52.84**

In conclusion, the usage of MT-SHL-DNN is advocated for holistic emotion recognition. On nine frequently used emotional speech databases, the MT-SHL-DNN method yields an effective data aggregation scheme with remarkable performance compared to training single-task DNNs on isolated emotion databases. In future work, the goal is to extend the MT-SHL-DNN approach to simultaneous classification and regression for acoustic emotion recognition by using multiple softmax and linear layers.

Chapter 11

Conclusions and Outlook

Continuous improvement is better than delayed perfection.

Mark Twain

11.1 Summary of Thesis Achievements

Paralanguage is a central component of non-verbal communication. Analysing the voice behind the words is thus of paramount importance for intelligent machine listening. This thesis aims at holistic speaker analysis within the realm of Computational Paralinguistics. In this section, the main achievements of this thesis are put in a nutshell.

A prerequisite for recognising various speaker characteristics is the availability of richly labelled databases. Due to the subjective nature of many paralinguistic phenomena, data annotation is commonly accomplished by determining the gold standard based on the personal judgements of multiple annotators. In Chapter 4, the sub-optimality of majority voting has been pointed out and a novel DAL technique has been proposed to adapt the number of annotators to the ambiguity regarding an instance, thereby optimally utilising human resources. Applied to speech emotion recognition, the DAL algorithms achieve up to 41 % reduction of the total number of annotations, compared to standard active learning at the same accuracy. Moreover, to leverage the vast amount of unlabelled data, a human-machine annotation framework has been presented in Chapter 5, reducing the human workload in data annotation to a minimum by means of machine labelling. Using deep learning based confidence measures, the instances are automatically labelled by machine if the model confidence is high, otherwise, they are inspected by human annotators. On the task of prosody estimation, the novel dynamic cooperative learning algorithm yields 74 % cost reduction compared to the baseline algorithm (passive learning), while achieving the same model performance. Thus, the generic cooperative learning framework

provides the basis for efficient, semi-automatic data annotation and thereby paves the way for the holistic analysis of various paralinguistic patterns including speaker characteristics, social signals and other vocal cues to non-verbal behaviours.

Since most of the paralinguistic speech databases only provide the labels for one target attribute, current systems have been mostly designed for single-task learning, i. e., recognising speaker characteristics one by one. In this thesis, it is posited that various speaker traits and states should be analysed within a larger context by modelling the interrelationships between tasks. Thus, to achieve human-like machine perception, the contextual dependencies need to be taken into account as well. To this end, a data-driven approach to modelling task relatedness, using NMDS and different distance metrics, has been described in Chapter 6, projecting the isolated paralinguistic tasks investigated so far onto a topological space. This step represents a main milestone towards holistic Computational Paralinguistics, providing for the first time insights into the paralinguistic universe as a whole.

To exploit the interdependencies between tasks, various approaches to multi-task learning have been applied to related tasks. In particular, feature relevance analysis proves to be an effective way to determine the predictive features across tasks, as shown in Chapter 7 and Chapter 8. Combined with cross-domain data aggregation, an effective training scheme has been obtained to recognise new patterns, e. g., drowsiness caused by alcohol intoxication or sleep deprivation. Moreover, it has been found that by including additional information about the speaker in the form of auxiliary labels, the accuracies on individual tasks can be improved. For example, if the native language of the speaker is known, the accuracy of prosody regression can be increased by 11 % relative to the single-task learning results. This argues for the usage of multi-label databases, which, however, are very scarce. To create more multi-label data for related tasks, the CTL method based on semi-supervised learning has been developed, which automatically completes the missing auxiliary labels. This approach has proven particularly useful for the joint analysis of deception and sincerity (cf. Chapter 9), achieving a 12 % relative error rate reduction on the deception detection task. An alternative to label completion is the multi-task DNN with shared hidden layers, which jointly learns multiple tasks. In contrast to CTL, the MT-SHL-DNN simply feeds all single-label data into the network and computes the feature transformations through the shared hidden layers, before making predictions for the tasks separately. Given the importance of emotion recognition in human-machine interaction, this multi-task learning technique has been successfully applied to holistic emotion recognition (cf. Chapter 10). There, it has been shown that the main advantages of this multi-task network topology are as follows: first, it is far more compact than a set of single-task networks; second, it yields an effective scheme for data aggregation; third, the tasks are learnt together by means of shared hidden layers; fourth, the network can predict multiple targets of interest at the same time; and lastly, the information loss due to label mapping can be avoided. On these grounds, this achievement can be considered as a first milestone towards human-like holistic analysis.

All in all, the works presented in this thesis have demonstrated that the holistic analysis of paralinguistic aspects by means of multi-task learning and large-scale data aggregation yields more general models for practical purposes, better accuracies on individual tasks, and highly compact model parametrisation, thereby translating Aristotle's postulate into effective machine learning methods.

11.2 Towards Holistic Machine Perception

To conclude the thesis, the recent trends in holistic Computational Paralinguistics are outlined. Moreover, some ongoing work is presented.

11.2.1 Human-Machine Cross-Task Data Aggregation

Based on the works carried out in the Part II and III of this thesis, a promising approach is to combine the human-machine annotation framework with the cross-task labelling technique. The aim is to join all the existing and coming paralinguistic corpora into a holistic, continuously expanding database, which is labelled in universal dimensions, thereby achieving large-scale data aggregation. One possibility would be to integrate an additional loop into the cooperative learning algorithm to iterate the labelling process for each task. More efficiently, the new framework can be augmented with the MT-SHL-DNN method applied to holistic emotion recognition. In a first study by the author and her colleagues [255], a multi-task shared-hidden-layer DNN is trained on all the ComParE classification tasks, thereby completing the target labels in the holistic database composed of the single-task databases (cf. Figure 9.2).

11.2.2 Holistic Speaker Recognition and Diarisation

Humans are naturally able to recognise speakers according to various speaker characteristics, which can be seen as voice markers of speaker IDs. Due to the robustness of voice biometric authentication, it is found in many security-critical deployments such as telephone banking. In Figure 2.2, speaker ID is depicted as the inherent attribute that underlies, emerges from and is consolidated by all other speaker characteristics. This motivates for the usage of paralinguistic features as extra information to accomplish speaker identification.

Speaker diarisation is the process of determining "who speaks when" in an audio stream [7]. Based on the assumption that humans are naturally able to tell speakers apart based on their individual traits, a new paralinguistic approach to speaker diarisation has been proposed by the author and her colleagues [260], which is described briefly here as an inspiration for future work.

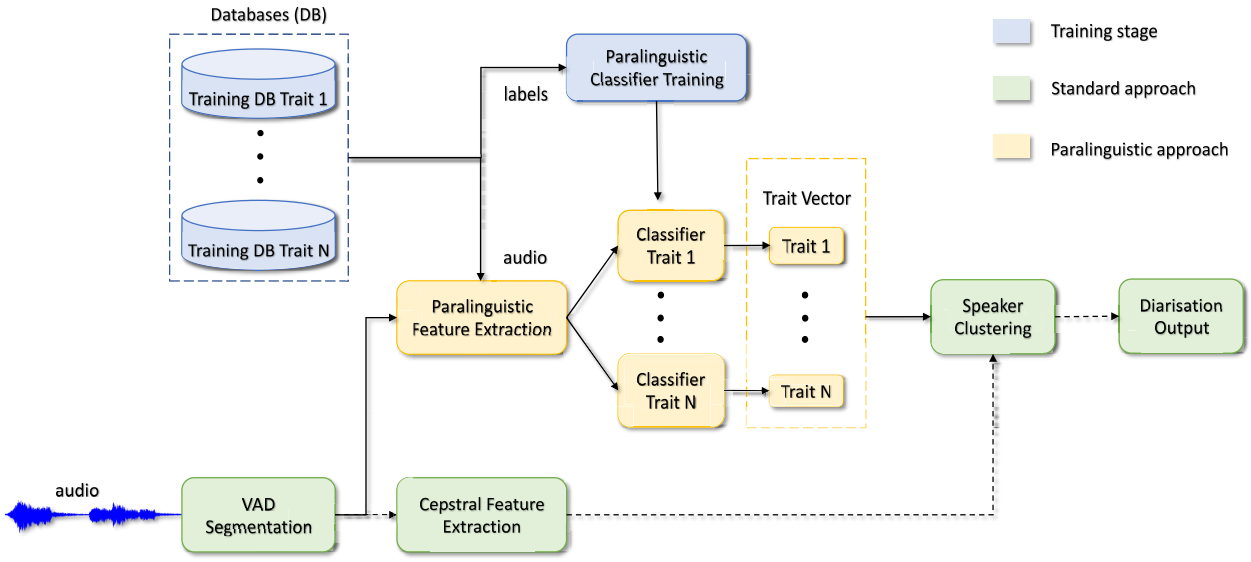


Figure 11.1: Processing chain of paralinguistic (solid line) and standard (dashed line) speaker diarisation [260].

Figure 11.1 sketches the process flow of a standard diarisation system, as well as the paralinguistic analyser connected in parallel. At the front-end, speaker turns are detected and segmented. Subsequently, the processing chain is divided into two pipelines that extract different features for speaker clustering. The paralinguistic module serves to train discriminative classifiers for recognising stable speaker traits. These paralinguistic features can be used on their own for speaker clustering, or they can complement the acoustic features extracted in the standard way. While the segmentation step serves to separate adjacent windows belonging to different speakers, hierarchical clustering operates globally on the audio stream in order to identify and group together the segments for each speaker [7]. On the output side, the clustered segments are tagged with relative speaker labels (e. g., speaker 1) or true identities (e. g., speaker name). The preliminary results [260] show that the paralinguistic method is competitive to the usage of MFCC features, suggesting that this is a promising research endeavour worth pursuing.

11.2.3 Multi-Modal Holistic Machine Perception

Resuming Albert Mehrabian's formula "Total feeling = 7% verbal feeling + 38% vocal feeling + 55% facial feeling", the long-term goal towards human-like computing would be to unite the core areas of machine perception for holistic analysis. To reach this goal, an end-to-end multi-modal multi-task framework is envisioned, which promises to achieve holistic machine perception by combining auditory, visual and physiological analysis on the input side, and recognising various non-verbal behavioural patterns on the output side. Since emotion is fundamental to human experience, most multi-modal databases are designed for the purpose of affect sensing.

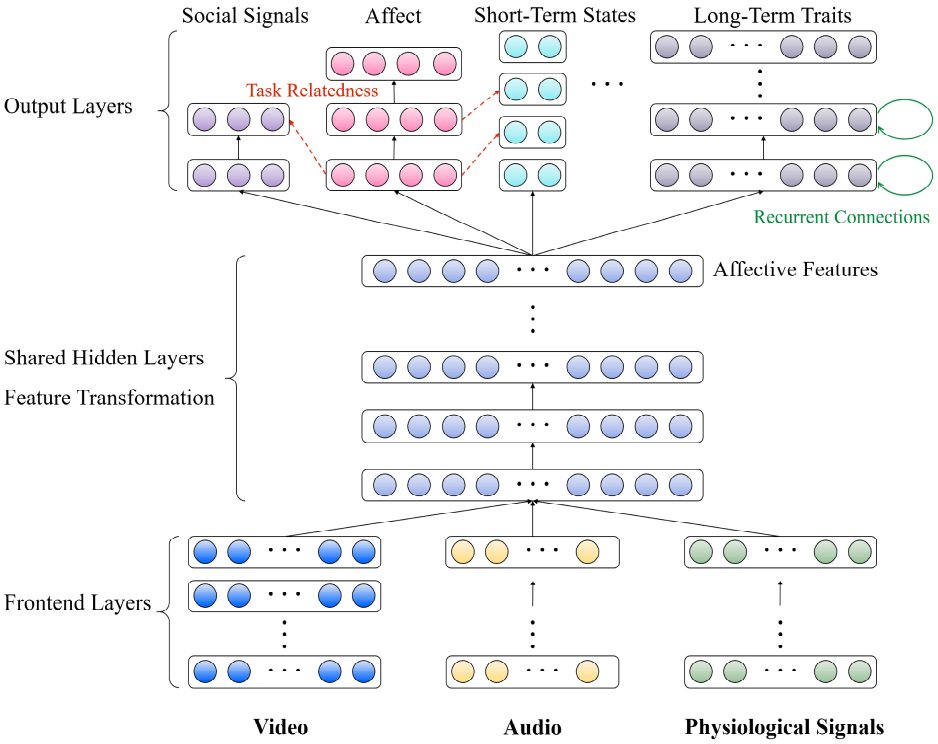


Figure 11.2: End-to-end multi-modal multi-task framework.

Figure 11.2 illustrates the architecture of the multi-modal multi-task framework. Starting from raw signals, affective features are extracted in the shared hidden layers for different modalities and tasks. Then, cross-domain transfer learning is performed to apply the affective features to a wide spectrum of human-centred phenomena, including social signals, affective and other short-term states, as well as long-term traits. Recent advances in deep learning support the feasibility of end-to-end training [228], which allows high adaptivity for the input and output layers of the DNN.

By joining forces to achieve multi-modal holistic machine perception, the united research fields in human-like computing will show the true potential of machine intelligence, and herald a new generation of AI with hitherto unseen skills to interact with humans, thereby verifying once more Aristotle's postulate "the whole is greater than the sum of its parts".

Bibliography

[1] A. Abad, E. Ribeiro, F. Kepler, R. F. Astudillo, and I. Trancoso, "Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers." in *Proc. 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2413–2417.

[2] M. Abadi, A. Agarwal, P. Barham *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," pp. 1–21, 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[3] D. Abercrombie, *Elements of general phonetics*. Edinburgh University Press, 1967.

[4] ——, "Paralanguage," *International Journal of Language & Communication Disorders*, vol. 3, no. 1, pp. 55–59, 1968.

[5] S. Alexander, "Sincerity, intonation, and apologies: A case study of Thai EFL and ESL learners," Ph.D. dissertation, Indiana University, Bloomington, IN, 2011.

[6] P. Angkititrakul and J. Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 634–646, 2006.

[7] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[8] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[9] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders," in *Proc. 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 191–194.

[10] J. L. Austin, *How to do things with words*. Oxford University Press, 1975.

[11] A. Baddeley, "Working memory," *Science*, vol. 255, no. 5044, pp. 556–559, 1992.

[12] B. Bakker and T. Heskes, "Task clustering and gating for bayesian multitask learning," *Journal of Machine Learning Research*, vol. 4, pp. 83–99, 2003.

[13] T. Bänziger, M. Mortillaro, and K. Scherer, "Introducing the Geneva Multimodal Expression Corpus for experimental research on emotion perception," *Emotion*, vol. 12, pp. 1161–1179, 2012.

[14] B. Barrett, R. Brown, M. P. Mundt, G. Thomas, S. Barlow, A. Highstrom, and M. Bahrainian, "Validation of a short form Wisconsin upper respiratory symptom survey," *Health and Quality of Life Outcomes*, vol. 7, no. 1, p. 76, 2009.

[15] M. Barrick and M. Mount, "The big five personality dimensions and job performance: A meta-analysis," *Personnel Psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[16] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, 2000.

[17] D. Behne, S. Rivera, and D. Pisoni, "Effects of alcohol on speech: Durations of isolated words, sentences, and passages in fluent speech," *Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 2311–2311, 1991.

[18] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 567–580.

[19] R. Birdwhistell, *Introduction to kinesics*. University of Louisville, 1952.

[20] M. Black, D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. V. Segbroeck, J. Kim, P. Georgiou, and S. Narayanan, "Automated evaluation of non-native English pronunciation quality: combining knowledge- and data-driven features at multiple time scales," in *Proc. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Dresden, Germany: ISCA, 2015, pp. 493–497.

[21] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. of 11th Annual Conference on Computational Learning Theory*. Madison, WI: ACM, 1998, pp. 92–100.

[22] D. Bone, M. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan, "Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors," pp. 3217–3220, 2011.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[24] A. Braun and H. Künzel, "The effect of alcohol on speech prosody," in *Proc. 15th International Congress of Phonetic Sciences (ICPhS)*, vol. 2645. Barcelona, Spain: ICPhS Organizing Committee, 2003, pp. 2645–2648.

[25] R. Buck, "The biological affects: A typology," *Psychological Review*, vol. 106, no. 2, p. 301, 1999.

[26] R. Burbidge, J. Rowland, and R. King, "Active learning for regression based on query by committee," in *Proc. 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*. Birmingham, UK: Springer, 2007, pp. 209–218.

[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. 9th European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 5. Lisbon, Portugal: ISCA, 2005, pp. 1517–1520.

[28] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proc. 7th International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta: ELRA, 2010.

[29] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, ""Would you buy a car from me?" - On the likability of telephone voices," in *Proc. 12th Conference of the International Speech Communication Association (INTERSPEECH)*. Florence, Italy: ISCA, 2011, pp. 1557–1560.

[30] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[31] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.

[32] R. P. Clapham, L. van der Molen, R. van Son, M. van den Brekel, and F. Hilgers, "NKI-CCRT corpus – speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy," in *Proc. 8th International Conference on Language Resources and Evaluation (LREC)*, vol. 4, Istanbul, Turkey, 2012, pp. 3350–3355.

[33] L. Coleman and P. Kay, "Prototype semantics: The English word lie," *Language*, vol. 57, no. 1, pp. 26–44, 1981.

[34] O. Cooney, "Acoustic analysis of the effects of alcohol on the human voice," Ph.D. dissertation, Dublin City University, 1998.

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[36] E. Coutinho, F. Hönig, Y. Zhang, S. Hantke, A. Batliner, E. Nöth, and B. Schuller, "Assessing the prosody of non-native speakers of English: Measures and feature sets," in *Proc. 10th Language Resources and Evaluation Conference (LREC)*. Portoroz, Slovenia: ELRA, 2016, pp. 1328–1332.

[37] D. Crystal, "Paralinguistics," *The Body As a Medium of Expression*, pp. 162–174, 1975.

[38] ——, *A Dictionary of Linguistics and Phonetics, 6th Edition*. John Wiley & Sons, 2008.

[39] J. Curtin, C. Patrick, A. Lang, J. Cacioppo, and N. Birbaumer, "Alcohol affects emotion through cognition," *Psychological Science*, vol. 12, no. 6, pp. 527–531, 2001.

[40] A. Damianou and N. Lawrence, "Deep Gaussian processes," in *Proc. 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 31. Scottsdale, AZ: PMLR, 2013, pp. 207–215.

[41] C. Darwin, *The expression of the emotions in man and animals.* John Murray, 1872.

[42] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[43] V. de Sa, "Learning classification with unlabeled data," in *Proc. of 7th Advances in Neural Information Processing Systems (NIPS).* Denver, Colorado: MIT Press, 1994, pp. 112–119.

[44] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Florence, Italy: IEEE, 2014, pp. 4851–4855.

[45] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Vancouver, Canada: IEEE, 2013, pp. 8599–8603.

[46] B. DePaulo, R. Rosenthal, J. Rosenkrantz, and C. R. Green, "Actual and perceived cues to deception: A closer look at speech," *Basic and Applied Social Psychology*, vol. 3, no. 4, pp. 291–312, 1982.

[47] T. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[48] P. Donmez, J. G. Carbonell, and J. Schneider, "Efficiently learning the accuracy of labeling sources for selective sampling," in *Proc. 15th ACM SIGKDD International conference on Knowledge Discovery and Data Mining.* Paris, France: ACM, 2009, pp. 259–268.

[49] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1, pp. 33–60, 2003.

[50] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *Proc.*

2nd international conference on Affective Computing and Intelligent Interaction (ACII). Springer, 2007, pp. 488–500.

[51] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. of 9th Advances in Neural Information Processing Systems*. Denver, CO: MIT Press, 1996, pp. 155–161.

[52] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, and D. Tao, "Exploring representativeness and informativeness for active learning," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 14–26, 2017.

[53] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and long-term features for emotion recognition," in *Proc. 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brighton, UK: ISCA, 2011, pp. 344–347.

[54] E. Eaton, M. desJardins, and T. Lane, "Modeling transfer relationships between learning tasks for improved inductive transfer," in *Proc. of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Antwerp, Belgium: Springer, 2008, pp. 317–332.

[55] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[56] P. Ekman, M. O'Sullivan, W. Friesen, and K. Scherer, "Face, voice, and body in detecting deceit," *Journal of Nonverbal Behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[57] I. Engberg and A. Hansen, "Documentation of the Danish emotional speech database DES," Aalborg, Denmark, Tech. Rep., 1996.

[58] J. Eriksson, "Straight talk: Conceptions of sincerity in speech," *Philosophical Studies*, vol. 153, no. 2, pp. 213–234, 2011.

[59] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, WA: ACM, 2004, pp. 109–117.

[60] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, ser. Springer Theses. Springer International Publishing, 2015.

[61] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. 18th ACM International Conference on Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[62] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE,

the Munich open-source multimedia feature extractor," in *Proc. 21st ACM International Conference on Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[63] A. Fetzer, "Sincerity and credibility in political interviews," in *Politics as Text and Talk: Analytic approaches to political discourse*. John Benjamins Publishing, 2002, pp. 173–201.

[64] K. Fish and M. Pell, "Prosodic markers of sincere compliments and white lies," in *Proc. Experimental and Theoretical Advances in Prosody 2*. Montréal, Canada: ISCA, 2011, pp. 23–25.

[65] K. Fish, K. Rothermich, and M. Pell, "The sound of (in)sincerity," *Journal of Pragmatics*, vol. 121, pp. 147–161, 2017.

[66] N. H. Frijda, P. Kuipers, and E. Ter Schure, "Relations among emotion, appraisal, and emotional action readiness," *Journal of Personality and Social Psychology*, vol. 57, no. 2, pp. 212–228, 1989.

[67] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, Cambridge, UK, 2016.

[68] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd International Conference on Machine Learning (ICML)*. New York, NY: IMLS, 2016, pp. 1050–1059.

[69] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[70] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. 17th Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 3603–3607.

[71] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT Press Cambridge, 2016, vol. 1.

[72] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using adaboost," in *Proc. 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 220–224.

[73] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," *Proc. 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3522–3526, 2017.

[74] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in Laboratory Phonology*, vol. 7, 2002.

[75] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *Proc. 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. Toulouse, France: IEEE, 2006, pp. 1033–1036.

[76] A. Graves, "Practical variational inference for neural networks," in *Proc. 24th International Conference on Neural Information Processing Systems (NIPS)*. Granada, Spain: Curran Associates, 2011, pp. 2348–2356.

[77] M. Green, "How do speech acts express psychological states?" in *John Searle's Philosophy of Language: Force, Meaning and Thought*, 2007, pp. 267–284.

[78] C. Greenwood, K. Thiemann-Bourque, D. Walker, J. Buzhardt, and J. Gilkerson, "Assessing children's home language environments using automatic speech recognition technology," *Communication Disorders Quarterly*, vol. 32, no. 2, pp. 83–92, 2011.

[79] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. 9th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. San Juan, PR: IEEE, 2005, pp. 381–385.

[80] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*. Hannover, Germany: IEEE, 2008, pp. 865–868.

[81] T. Grósz, R. Busa-Fekete, G. Gosztolya, and L. Tóth, "Assessing the degree of nativeness and Parkinson's condition using Gaussian processes and deep rectifier neural networks," in *Proc. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Dresden, Germany: ISCA, 2015, pp. 919–923.

[82] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. 1st International Workshop on Emotion Synthesis, Representation, and Analysis in Continuous Space (EmoSPACE)*. Santa Barbara, CA: IEEE, 2011, pp. 827–834.

[83] R. Gupta, K. Audhkhasi, S. Lee, and S. Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking," in *Proc. 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 173–177.

[84] R. Hahnloser, R. Sarpeshkar, M. Mahowald, R. Douglas, and S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.

[85] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato, 1999.

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

[87] A. Hanani, M. Russell, and M. Carey, "Human and computer recognition of regional accents and ethnic groups from British English speech," *Computer Speech and Language*, vol. 27, no. 1, pp. 59–74, 2013.

[88] J. Hansen and L. Arslan, "Foreign accent classification using source generator based prosodic features," in *Proc. 20th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. Detroit, MI: IEEE, 1995, pp. 836–839.

[89] J. Hansen, S. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database," in *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 4. Rhodes, Greece: ISCA, 1997, pp. 1743–1746.

[90] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. El-Desoky Mousa, and B. Schuller, "I hear you eat and speak: Automatic recognition of eating condition and food types, use-cases, and impact on ASR performance," *PLoS ONE*, vol. 11, no. 5, pp. 1–24, 2016.

[91] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[92] J. M. Hernández-Lobato and R. Adams, "Probabilistic backpropagation for scalable learning of bayesian neural networks," in *Proc. 32nd International Conference on Machine Learning (ICML)*, vol. 37. Lille, France: JMLR.org, 2015, pp. 1861–1869.

[93] A. Hill, *Introduction to linguistic structures: From sound to sentence in English.* Harcourt Brace & World, 1958.

[94] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[95] J. B. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing deceptive from non-deceptive speech," in *Proc. 9th European Conference on Speech Communication and Technology (EUROSPEECH).* Lisbon, Portugal: ISCA, 2005, pp. 805–809.

[96] H. Hollien, G. Dejong, C. Martin, R. Schwartz, and K. Liljegren, "Effects of ethanol

intoxication on speech suprasegmentals," *Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3198–3206, 2001.

[97] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of failure: Employing word accent information for pronunciation quality assessment of English L2 learners," in *Proc. International Workshop on Speech and Language Technology in Education (SLaTE)*. Wroxall, UK: ISCA, 2009, pp. 41–44.

[98] ——, "How many labellers? Modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody," in *Proc. Workshop on Speech and Language Technology in Education (SLaTE)*. Tokyo, Japan: ISCA, 2010, pp. 137–140.

[99] ——, "Automatic assessment of non-native prosody for English as L2," in *Proc. 5th International Conference on Speech Prosody*. Chicago, IL: ISCA, 2010, pp. 1–4.

[100] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody – Annotation, modelling and evaluation," in *Proc. International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*. Stockholm, Sweden: KTH Computer Science and Communication, 2012, pp. 21–30.

[101] F. Hönig, T. Bocklet, K. Riedhammer, A. Batliner, and E. Nöth, "The automatic assessment of non-native prosody: Combining classical prosodic analysis with acoustic modelling," in *Proc. 13th Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR: ISCA, 2012, pp. 823–826.

[102] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[103] D.-Y. Huang, Z. Zhang, and S. S. Ge, "Speaker state classification based on fusion of asymmetric simple partial least squares (SIMPLS) and support vector machines," *Computer Speech & Language*, vol. 28, no. 2, pp. 392–419, 2014.

[104] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, 2013, pp. 7304–7308.

[105] M. Huckvale, "Within-speaker features for native language recognition in the Interspeech 2016 computational paralinguistics challenge," in *Proc. 17th Conference of the International Speech Communication Association (INTERSPEECH)*. San Francsico, CA: ISCA, 2016, pp. 2403–2407.

[106] M. Ingre, T. Åkerstedt, B. Peters, A. Anund, G. Kecklund, and A. Pickles, "Subjective sleepiness and accident risk avoiding the ecological fallacy," *Journal of Sleep Research*, vol. 15, no. 2, pp. 142–148, 2006.

[107] A. Ivanov and X. Chen, "Modulation spectrum analysis for speaker personality trait recognition," in *Proc. 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR: ISCA, 2012, pp. 278–281.

[108] L. James, R. Demaree, and G. Wolf, "Estimating within-group interrater reliability with and without response bias," *Journal of Applied Psychology*, vol. 69, no. 1, pp. 85–98, 1984.

[109] S. Jarvis and M. Paquot, "Native language identification," in *Cambridge Handbook of Learner Corpus Research*, S. Granger, G. Gilquin, and F. Meunier, Eds. Cambridge University Press, 2015.

[110] J. H. Jeon, R. Xia, and Y. Liu, "Level of interest sensing in spoken dialog using decision-level fusion of acoustic and lexical evidence," *Computer Speech & Language*, vol. 28, no. 2, pp. 420–433, 2014.

[111] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, "Emotional audio-visual speech synthesis based on PAD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 570–582, 2011.

[112] B. Johannes, V. P. Salnitski, H.-C. Gunga, and K. Kirsch, "Voice stress monitoring in space – Possibilities and limits," *Aviation, Space, and Environmental Medicine*, vol. 71, no. 9, pp. A58–65, 2000.

[113] K. Johnson, D. Pisoni, and R. Bernacki, "Do voice recordings reveal whether a person is intoxicated? A case study," *Phonetica*, vol. 47, no. 3–4, pp. 215–237, 1990.

[114] H. Kaya and A. Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks." in *Proc. 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2046–2050.

[115] H. Kaya, T. Özkaptan, A. A. Salah, and S. F. Gürgen, "Canonical correlation analysis and local fisher discriminant analysis based multi-view acoustic feature reduction for physical load prediction," in *Proc. 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore, Singapore: ISCA, 2014, pp. 442–446.

[116] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Proc. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Dresden, Germany: ISCA, 2015, pp. 909–913.

[117] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple subsystems," in *Proc. 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR: ISCA, 2012, pp. 534–537.

[118] S. Kim, M. Filippone, F. Valente, and A. Vinciarelli, "Predicting the conflict level in television political debates: An approach based on crowdsourcing, nonverbal communication and Gaussian processes," in *Proc. 20th ACM International Conference on Multimedia*. Nara, Japan: ACM, 2012, pp. 793–796.

[119] F. Klingholz, R. Penning, and E. Liebhardt, "Recognition of low-level alcohol intoxication from speech signal," *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 929–935, 1988.

[120] M. Kockmann, L. Burget, and J. Černockỳ, "Brno university of technology system for Interspeech 2010 paralinguistic challenge," in *Proc. 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan: ISCA, 2010, pp. 2822–2825.

[121] J. Krajewski, A. Batliner, and M. Golz, "Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach," *Behavior Research Methods*, vol. 41, no. 3, pp. 795–804, 2009.

[122] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Proc. 7th International Conference on Neural Information Processing Systems (NIPS)*, vol. 7. Dencer, CO: MIT Press, 1994, pp. 231–238.

[123] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[124] M. Krzywinski and N. Altman, "Points of significance: Visualizing samples with box plots," *Nature Methods*, vol. 11, no. 2, pp. 119–120, 2014.

[125] A. Kumar and H. Daume, "Learning task grouping and overlap in multi-task learning," in *Proc. of 29th International Conference on Machine Learning*. Edinburgh, GB: Omnipress, 2012, pp. 1383–1390.

[126] J. LeBreton and J. Senter, "Answers to 20 questions about interrater reliability and interrater agreement," *Organizational Research Methods*, vol. 11, no. 4, pp. 815–852, 2008.

[127] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[128] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, "Learning a meta-level prior for feature relevance from multiple related tasks," in *Proc. of 24th International Conference on Machine Learning (ICML)*. ACM, 2007, pp. 489–496.

[129] M. Levit, R. Huber, A. Batliner, and E. Noeth, "Use of prosodic speech characteristics for automated detection of alcohol intoxication," in *Proc. ISCA Tutorial and Research*

Workshop (ITRW) on Prosody in Speech Recognition and Understanding. Red Bank, NJ: ISCA, 2001, pp. 103–106.

[130] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th International ACM SIGIR conference on Research and development in information retrieval*. Dublin, Ireland: Springer, 1994, pp. 3–12.

[131] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, 2007.

[132] W.-H. Lin and A. Hauptmann, "Vox populi annotation: Measuring intensity of ideological perspectives by aggregating group judgments," in *Proc. 6th Language Resources and Evaluation Conference (LREC)*. Marrakech, Morocco: ELRA, 2008, pp. 1–6.

[133] J. Loy, H. Rohde, and M. Corley, "Lying, in a manner of speaking," in *Proc. 8th International Conference on Speech Prosody*, vol. 8, Boston, MA, 2016, pp. 984–988.

[134] B. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia content description interface*. John Wiley & Sons, 2002, vol. 1.

[135] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd International Conference on Data Engineering, Workshop on Multimedia Databases and Data Management*. Atlanta, GA: IEEE, 2006, pp. 1–8.

[136] A. Mehrabian, *Silent messages*. Wadsworth Publishing Company, 1971, vol. 8.

[137] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *Proc. 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan: ISCA, 2010, pp. 2818–2821.

[138] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *Proc. 2nd Language Resources and Evaluation Conference (LREC)*. ELRA, 2000, pp. 957–964.

[139] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, 2017, pp. 2227–2231.

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

[141] C. Montacié and M.-J. Caraty, "Pitch and intonation contribution to speakers' traits

classification," in *Proc. 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR: ISCA, 2012, pp. 526–529.

[142] ——, "Prosodic cues and answer type detection for the deception sub-challenge." in *Proc. 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2016–2020.

[143] G. Montavon, G. B. Orr, and K.-R. Müller, Eds., *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science. Springer, 2012, vol. 7700.

[144] I. Muslea, S. Minton, and C. Knoblock, "Active learning with multiple views," *Journal of Artificial Intelligence Research*, vol. 27, pp. 203–233, 2006.

[145] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children," *Computer Assisted Language Learning*, vol. 21, no. 5, pp. 393–408, 2008.

[146] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$," in *Doklady AN SSSR (translated as Soviet. Math. Docl.)*, vol. 269, 1983, pp. 543–547.

[147] M. Newman, J. Pennebaker, D. Berry, and J. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and Social Psychology Bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[148] T. L. Nwe, H. Li, and M. Dong, "Analysis and detection of speech under sleep deprivation," in *Proc. 9th International Conference on Spoken Language Processing (ICSLP)*. Pittsburgh, PA: ISCA, 2006, pp. 1846–1849.

[149] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. 9th Language Resources and Evaluation Conference (LREC)*. ELRA, 2014, pp. 342–347.

[150] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[151] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[152] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters," in *Proc. 9th Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia: ISCA, 2008, pp. 759–762.

[153] R. Picard, *Affective Computing*. MIT Press, 1997, vol. 252.

[154] T. Piske, I. R. MacKay, and J. Flege, "Factors affecting degree of foreign accent in an L2: A review," *Journal of Phonetics*, vol. 29, no. 2, pp. 191–215, 2001.

[155] D. Pisoni and C. Martin, "Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses," *Alcoholism: Clinical and Experimental Research*, vol. 13, no. 4, pp. 577–587, 1989.

[156] J. Platt, "Fast training of support vector machines using sequential minimal optimization," *Advances in Kernel Methods*, pp. 185–208, 1999.

[157] ——, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[158] R. Plutchik, *Emotion: A psychoevolutionary synthesis*. Harper and Row, 1980.

[159] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[160] A. Ragni, E. Dakin, X. Chen, M. J. Gales, and K. Knill, "Multi-language neural network language models," in *Proc. 17th Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 3042–3046.

[161] B. Rammstedt and O. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[162] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proc. 1st International Conference on Speech Prosody*. Aix-en-Provence, France: ISCA, 2002, pp. 115–120.

[163] R. Rao, "Prosodic consequences of sarcasm versus sincerity in Mexican Spanish," *Concentric: Studies in Linguistics*, vol. 39, no. 2, pp. 33–59, 2013.

[164] O. Räsänen and J. Pohjalainen, "Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech," in *Proc. 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 210–214.

[165] C. Rasmussen, "Gaussian processes for machine learning," in *Advanced Lectures on Machine Learning*. Canberra, Australia: Springer, 2004, pp. 63–71.

[166] V. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proc. 26th International Conference on Machine Learning*. Montreal, Canada: ACM, 2009, pp. 889–896.

[167] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[168] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.

[169] S. Rigoulot, K. Fish, and M. Pell, "Neural correlates of inferring speaker sincerity from white lies: An event-related potential source localization study," *Brain Research*, vol. 1565, pp. 48–62, 2014.

[170] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language impaired children," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, pp. 1328–1342, 2011.

[171] P. Rockwell, D. B. Buller, and J. Burgoon, "The voice of deceit: Refining and expanding vocal cues to deception," *Communication Research Reports*, vol. 14, no. 4, pp. 451–459, 1997.

[172] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Proc. 7th IEEE Workshop on Applications of Computer Vision*. Breckenridge, CO: IEEE, 2005, pp. 29–36.

[173] M. Rosenstein, Z. Marx, L. P. Kaelbling, and T. Dietterich, "To transfer or not to transfer," in *Proc. of 18th Advances in Neural Information Processing Systems (NIPS), Workshop on Inductive Transfer: 10 Years Later*, vol. 898. Vancouver, Canada: MIT Press, 2005, pp. 1–4.

[174] O. Rudovic, M. Nicolaou, and V. Pavlovic, "Machine learning methods for social signal processing," in *Social Signal Processing*, J. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli, Eds. Cambridge University Press, 2017, pp. 234–254.

[175] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[176] ——, "Core affect and the psychological construction of emotion," *Psychological Review*, vol. 110, no. 1, pp. 145–172, 2003.

[177] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 100, no. 5, pp. 401–409, 1969.

[178] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proc. of 3rd International Symposium on Intelligent Data Analysis*. Cascais, Portugal: Springer, 2001, pp. 309–318.

[179] K. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal Processes in Emotion: Theory, Methods, Research*, vol. 92, no. 92–120, p. 57, 2001.

[180] F. Schiel, "Perception of alcoholic intoxication in speech," in *Proc. 12th Conference of the International Speech Communication Association (INTERSPEECH)*. Florence, Italy: ISCA, 2011, pp. 3281–3284.

[181] F. Schiel and C. Heinrich, "Laying the foundation for in-car alcohol detection by speech," in *Proc. 10th Conference of the International Speech Communication Association (INTERSPEECH)*. Brighton, UK: ISCA, 2009, pp. 983–986.

[182] F. Schiel, C. Heinrich, and V. Neumeyer, "Rhythm and formant features for automatic alcohol detection," in *Proc. 11th Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan: ISCA, 2010, pp. 458–461.

[183] F. Schiel, C. Heinrich, and S. Barfüsser, "Alcohol Language Corpus: The first public corpus of alcoholized German speech," *Language Resources and Evaluation*, vol. 46, no. 3, pp. 503–521, 2011.

[184] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, 2017.

[185] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2014.

[186] B. Schuller, M. Wimmer, D. Arsić, G. Rigoll, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2. Honolulu, HI: IEEE, 2007, pp. 733–736.

[187] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, 2009.

[188] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brighton, UK: ISCA, 2009, pp. 312–315.

[189] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. 11th Conference of the International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan: ISCA, 2010, pp. 2794–2797.

[190] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[191] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. 12th Conference of the International Speech Communication Association (INTERSPEECH)*. Florence, Italy: ISCA, 2011, pp. 3201–3204.

[192] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing*, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[193] B. Schuller, M. Wöllmer, F. Eyben, G. Rigoll, and D. Arsić, "Semantic speech tagging: Towards combined analysis of speaker traits," in *Proc. 42nd Audio Engineering Society Conference (AES)*. Ilmenau, Germany: Audio Engineering Society, 2011, pp. 89–97.

[194] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *Proc. 12th Conference of the International Speech Communication Association (INTERSPEECH)*. Florence, Italy: ISCA, 2011, pp. 1553–1556.

[195] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. 13th Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR: ISCA, 2012, pp. 254–257.

[196] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language – State-of-the-Art and the Challenge," *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013.

[197] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. 14th Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 148–152.

[198] B. Schuller, F. Friedmann, and F. Eyben, "The Munich BioVoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *Proc. 9th Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland: ELRA, 2014, pp. 1506–1510.

[199] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," in *Proc. 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore, Singapore: ISCA, 2014.

[200] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states – A review on intoxication, sleepiness and the first challenge," *Computer Speech and Language, Special Issue on Broadening the View on Speaker Analysis*, vol. 28, no. 2, pp. 346–374, 2014.

[201] B. Schuller, Y. Zhang, F. Eyben, and F. Weninger, "Intelligent systems' holistic evolving analysis of real-life universal speaker characteristics," in *Proc. 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data (ES^{3}LOD), satellite of the 9th Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland: ELRA, 2014, pp. 14–20.

[202] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Höth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition," in *Proc. 16th Conference of the International Speech Communication Association (INTERSPEECH)*. Dresden, Germany: ISCA, 2015, pp. 478–482.

[203] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Computer Speech and Language, Special Issue on Next Generation Computational Paralinguistics*, vol. 29, no. 1, pp. 100–131, 2015.

[204] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception & Sincerity," in *Proc. 17th Conference of the International Speech Communication Association (INTERSPEECH)*. San Francsico, CA: ISCA, 2016, pp. 2001–2005.

[205] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proc. 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 3442–3446.

[206] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben,

E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech and Language*, vol. 32, 2018.

[207] M. V. Segbroeck, R. Travadi, C. Vaz, J. Kim, M. Black, A. Potamianos, and S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *Proc. 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore, Singapore: ISCA, 2014, pp. 751–755.

[208] B. Settles, "Active learning literature survey," Madison, WI, Tech. Rep. TR 1648, 2010.

[209] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[210] V. Sheng, F. Provost, and P. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, 2008, pp. 614–622.

[211] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[212] J. C. Smuts, *Holism and Evolution*. Macmillan Publishers, 1926.

[213] L. Sobell, M. Sobell, and R. Coleman, "Alcohol-inducted disfluency in non-alcoholics," *Folia Phoniatrica et Logopaedica*, vol. 34, no. 6, pp. 316–323, 1982.

[214] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[215] G. Stebbins and C. Goetz, "Factor structure of the unified Parkinson's disease rating scale: Motor examination section," *Movement Disorders*, vol. 13, no. 4, pp. 633–636, 1998.

[216] S. Steidl, "Automatic classification of emotion-related user states in spontaneous children's speech," Ph.D. dissertation, University of Erlangen-Nuremberg, 2009.

[217] S. Steininger, S. Raubold, O. Dioubina, and F. Schiel, "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. 3rd Language Resources and Evaluation Conference (LREC), Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, Spain, 2002, pp. 33–37.

[218] S. S. Stevens, J. Volkmann, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[219] L. Streeter, R. Krauss, V. Geller, C. Olson, and W. Apple, "Pitch changes during attempted deception," *Journal of Personality and Social Psychology*, vol. 35, no. 5, p. 345, 1977.

[220] H. Strömfelt, Y. Zhang, and B. Schuller, "Emotion-augmented machine learning: Overview of an emerging domain," in *Proc. 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*. San Antonio, TX: IEEE, 2017, pp. 305–312.

[221] A. Stuhlsatz, J. Lippel, and T. Zielke, "Discriminative feature extraction with deep neural networks," in *Proc. International Joint Conference on Neural Networks (IJCNN)*. Barcelona, Spain: IEEE, 2010, pp. 5688–5691.

[222] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic: IEEE, 2011, pp. 5688–5691.

[223] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sönmez, "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners," in *Proc. 1st Conference of the International Speech Communication Association (INTERSPEECH)*, Beijing, P. R. China, 2000, pp. 187–190.

[224] S. Thrun and L. Pratt, *Learning to Learn*. Springer Science & Business Media, 2012.

[225] J. Torres, E. Moore, and E. Bryant, "A study of glottal waveform features for deceptive speech classification," in *Proc. 33rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Las Vegas, NV: IEEE, 2008, pp. 4489–4492.

[226] G. Trager, "Paralanguage: A first approximation," *Studies in Linguistics*, vol. 13, pp. 1–12, 1958.

[227] ——, "The typology of paralanguage," *Anthropological Linguistics*, pp. 17–21, 1961.

[228] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Shanghai, P. R. China: IEEE, 2016, pp. 5200–5204.

[229] F. Trojan and K. Kryspin-Exner, "The decay of articulation under the influence of alcohol and paraldehyde," *Folia Phoniatrica et Logopaedica*, vol. 20, no. 4, pp. 217–238, 1968.

[230] G. Tur, D. Hakkani-Tür, and R. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[231] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *Proc. Richmedia Conference*. Citeseer, 2003, pp. 109–119.

[232] G. Villar, J. Arciuli, and H. Paterson, "Vocal pitch production during lying: Beliefs about deception matter," *Psychiatry, Psychology and Law*, vol. 20, no. 1, pp. 123–132, 2013.

[233] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *Proc. 3rd International Conference on Affective Computing and Intelligent Interaction (ACII)*. Amsterdam, Netherlands: IEEE, 2009, pp. 1–4.

[234] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[235] A. Vinciarelli, S. Kim, F. Valente, and H. Salamin, "Collecting data for socially intelligent surveillance and monitoring approaches: The case of conflict in competitive conversations," in *Proc. 5th International Symposium on Communications Control and Signal Processing (ISCCSP)*. IEEE, 2012, pp. 1–4.

[236] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[237] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, "From nonverbal cues to perception: Personality and social attractiveness," *Cognitive Behavioural Systems*, pp. 60–72, 2012.

[238] A. P. Vogel, J. Fletcher, and P. Maruff, "Acoustic analysis of the effects of sustained wakefulness on speech," *Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3747–3756, 2010.

[239] A. Vrij, K. Edward, K. Roberts, and R. Bull, "Detecting deceit via analysis of verbal and nonverbal behavior," *Journal of Nonverbal Behavior*, vol. 24, no. 4, pp. 239–263, 2000.

[240] J. Wagner, T. Baur, D. Schiller, Y. Zhang, B. Schuller, M. Valstar, and E. André, "Show me what you've learned: Applying cooperative machine learning for the semi-automated annotation of social signals," in *Proc. of 2nd Workshop on Explainable Artificial Intelligence (XAI), held in conjunction with the 27th International Joint Conference on Artificial Intelligence (IJCAI) and the 23rd European Conference on Artificial Intelligence (ECAI)*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence, 2018, pp. 171–177.

[241] A. Warlaumont, M. VanDam, E. Bergelson, and A. Cristia, "HomeBank: A repository for long-form real-world audio recordings of children," *Proc. 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 815–816, 2017.

[242] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 59–69, 2009.

[243] D. Watson and L. A. Clark, "On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model," *Journal of Personality*, vol. 60, no. 2, pp. 441–476, 1992.

[244] P. Watson, I. Watson, and R. Batt, "Prediction of blood alcohol concentrations in human subjects. Updating the Widmark equation," *Journal of Studies on Alcohol*, vol. 42, no. 7, pp. 547–556, 1981.

[245] F. Weninger, E. Marchi, and B. Schuller, "Improving recognition of speaker states and traits by cumulative evidence: Intoxication, sleepiness, age and gender," in *Proc. 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Portland, OR: ISCA, 2012, pp. 1–4.

[246] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Psychology, Special Issue on Expression of Emotion in Music and Vocal Communication*, vol. 4, no. 292, pp. 1–12, 2013.

[247] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Proc. 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. Hawaii, HI: IEEE, 2007, pp. 809–812.

[248] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 2005.

[249] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes – Towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia: ISCA, 2008, pp. 597–600.

[250] Y. Yan, R. Rosales, G. Fung, and J. Dy, "Active learning from crowds," in *Proc. 28th International Conference on Machine Learning (ICML)*, vol. 11, Bellevue, WA, 2011, pp. 1161–1168.

[251] T. F. Yap, J. Epps, E. Ambikairajah, and E. Choi, "Voice source under cognitive load: Effects and classification," *Speech Communication*, vol. 72, pp. 74–95, 2015.

[252] Y. Zhang, E. Coutinho, Z. Zhang, M. Adam, and B. Schuller, "On rater reliability and agreement based dynamic active learning," in *Proc. 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII)*. Xi'an, P. R. China: IEEE, 2015, pp. 70–76.

[253] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Agreement-based dynamic active learning with least and medium certainty query strategies," in *Proc. Workshop on Advances in Active Learning: Bridging Theory and Practice, held in conjunction with the*

32nd International Conference on Machine Learning (ICML). Lille, France: IMLS, 2015, pp. 1–5.

[254] ——, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proc. 17th International Conference on Multimodal Interaction (ICMI)*. Seattle, WA: ACM, 2015, pp. 275–278.

[255] Y. Zhang, Y. Liu, and B. Schuller, "Towards human-like holistic machine perception of speaker states and traits," in *Proc. of Human-Like Computing Machine Intelligence Workshop (MI20-HLC)*. Windsor, U. K.: Springer, 2016, pp. 1–3.

[256] Y. Zhang, F. Weninger, A. Batliner, F. Hönig, and B. Schuller, "Language proficiency assessment of English L2 speakers based on joint analysis of prosody and native language," in *Proc. of 18th ACM International Conference on Multimodal Interaction (ICMI)*. Tokyo, Japan: ACM, 2016, pp. 274–278.

[257] Y. Zhang, F. Weninger, Z. Ren, and B. Schuller, "Sincerity and deception in speech: Two sides of the same coin? A transfer- and multi-task learning perspective," in *Proc. 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2041–2045.

[258] Y. Zhang, Y. Zhou, J. Shen, and B. Schuller, "Semi-autonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis," in *Proc. 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Shanghai, P. R. China: IEEE, 2016, pp. 6090–6094.

[259] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, 2017, pp. 4990–4994.

[260] Y. Zhang, F. Weninger, B. Liu, M. Schmitt, F. Eyben, and B. Schuller, "A paralinguistic approach to holistic speaker diarisation – Using age, gender, voice likability and personality traits," in *Proc. 25th ACM International Conference on Multimedia*. Mountain View, CA: ACM, 2017, pp. 387–392.

[261] Y. Zhang, F. Weninger, and B. Schuller, "Cross-domain classification of drowsiness in speech: The case of alcohol intoxication and sleep deprivation," in *Proc. 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 3152–3156.

[262] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, 2014.

[263] Z.-H. Zhou, K.-J. Chen, and H.-B. Dai, "Enhancing relevance feedback in image retrieval using unlabeled data," *ACM Transactions on Information Systems (TOIS)*, vol. 24, no. 2, pp. 219–244, 2006.

[264] J. Zhu, H. Wang, B. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1323–1331, 2010.

[265] X. Zhu, "Semi-supervised learning literature survey," Madison, WI, Tech. Rep. TR 1530, 2006.

[266] M. Zuckerman, B. DePaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," in *Advances in Experimental Social Psychology*. Elsevier, 1981, vol. 14, pp. 1–59.