OULU 2014
C 487

ACTA UNIVERSITATIS OULUENSIS

Eero Väyrynen

EMOTION RECOGNITION FROM SPEECH USING PROSODIC FEATURES

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING,
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING;
INFOTECH OULU

ACTA UNIVERSITATIS OULUENSIS
C Technica 487

EERO VÄYRYNEN

EMOTION RECOGNITION FROM SPEECH USING PROSODIC FEATURES

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Technology and Natural Sciences of the University of Oulu for public defence in OP-sali (Auditorium L10), Linnanmaa, on 9 May 2014, at 12 noon

UNIVERSITY OF OULU, OULU 2014

Acta Univ. Oul. C 487, 2014

Supervised by
Professor Tapio Seppänen

Reviewed by
Professor Tom Bäckström
Professor Klára Vicsi

Opponent
Professor Paavo Alku

ISBN 978-952-62-0403-1 (Paperback)
ISBN 978-952-62-0404-8 (PDF)

ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2014

Väyrynen, Eero, Emotion recognition from speech using prosodic features.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering, Department of Computer Science and Engineering; Infotech Oulu
Acta Univ. Oul. C 487, 2014
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Emotion recognition, a key step of affective computing, is the process of decoding an embedded emotional message from human communication signals, e.g. visual, audio, and/or other physiological cues. It is well-known that speech is the main channel for human communication and thus vital in the signalling of emotion and semantic cues for the correct interpretation of contexts. In the verbal channel, the emotional content is largely conveyed as constant paralinguistic information signals, from which prosody is the most important component. The lack of evaluation of affect and emotional states in human machine interaction is, however, currently limiting the potential behaviour and user experience of technological devices.

In this thesis, speech prosody and related acoustic features of speech are used for the recognition of emotion from spoken Finnish. More specifically, methods for emotion recognition from speech relying on long-term global prosodic parameters are developed. An information fusion method is developed for short segment emotion recognition using local prosodic features and vocal source features. A framework for emotional speech data visualisation is presented for prosodic features.

Emotion recognition in Finnish comparable to the human reference is demonstrated using a small set of basic emotional categories (neutral, sad, happy, and angry). A recognition rate for Finnish was found comparable with those reported in the western language groups. Increased emotion recognition is shown for short segment emotion recognition using fusion techniques. Visualisation of emotional data congruent with the dimensional models of emotion is demonstrated utilising supervised nonlinear manifold modelling techniques. The low dimensional visualisation of emotion is shown to retain the topological structure of the emotional categories, as well as the emotional intensity of speech samples.

The thesis provides pattern recognition methods and technology for the recognition of emotion from speech using long speech samples, as well as short stressed words. The framework for the visualisation and classification of emotional speech data developed here can also be used to represent speech data from other semantic viewpoints by using alternative semantic labellings if available.

Keywords: affective computing, data visualisation, emotion recognition, machine learning, speech prosody

Väyrynen, Eero, Emootiontunnistus puheen prosodisten piirteiden avulla.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta, Tietotekniikan osasto; Infotech Oulu
Acta Univ. Oul. C 487, 2014
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Emootiontunnistus on affektiivisen laskennan keskeinen osa-alue. Siinä pyritään ihmisen kommunikaatioon sisältyvien emotionaalisten viestien selvittämiseen, esim. visuaalisten, auditiivisten ja/tai fysiologisten vihjeiden avulla. Puhe on ihmisten tärkein tapa kommunikoida ja on siten ensiarvoisen tärkeässä roolissa viestinnän oikean semanttisen ja emotionaalisen tulkinnan kannalta. Emotionaalinen tieto välittyy puheessa paljolti jatkuvana paralingvistisenä viestintänä, jonka tärkein komponentti on prosodia. Tämän affektiivisen ja emotionaalisen tulkinnan vajaavaisuus ihminen-kone – interaktioissa rajoittaa kuitenkin vielä nykyisellään teknologisten laitteiden toimintaa ja niiden käyttökokemusta.

Tässä väitöstyössä on käytetty puheen prosodisia ja akustisia piirteitä puhutun suomen emotionaalisen sisällön tunnistamiseksi. Työssä on kehitetty pitkien puhenäytteiden prosodisiin piirteisiin perustuvia emootiontunnistusmenetelmiä. Lyhyiden puheenpätkien emotionaalisen sisällön tunnistamiseksi on taas kehitetty informaatiofuusioon perustuva menetelmä käyttäen prosodian sekä äänilähteen laadullisten piirteiden yhdistelmää. Lisäksi on kehitetty teknologinen viitekehys emotionaalisen puheen visualisoimiseksi prosodisten piirteiden avulla.

Tutkimuksessa saavutettiin ihmisten tunnistuskykyyn verrattava automaattisen emootiontunnistuksen taso käytettäessä suppeaa perusemootioiden joukkoa (neutraali, surullinen, iloinen ja vihainen). Emootiontunnistuksen suorituskyky puhutulle suomelle havaittiin olevan verrannollinen länsieurooppalaisten kielten kanssa. Lyhyiden puheenpätkien emotionaalisen sisällön tunnistamisessa saavutettiin taas parempi suorituskyky käytettäessä fuusiomenetelmää. Emotionaalisen puheen visualisoimiseksi kehitetyllä opetettavalla epälineaarisella manifoldimallinnustekniikalla pystyttiin tuottamaan aineistolle emootion dimensionaalisen mallin kaltainen visuaalinen rakenne. Mataladimensionaalisen kuvauksen voitiin edelleen osoittaa säilyttävän sekä tutkimusaineiston emotionaalisten luokkien että emotionaalisen intensiteetin topologisia rakenteita.

Tässä väitöksessä kehitettiin hahmontunnistusmenetelmiin perustuvaa teknologiaa emotionaalisen puheen tunnistamiseksi käytettäessä sekä pitkiä että lyhyitä puhenäytteitä. Emotionaalisen aineiston visualisointiin ja luokitteluun kehitettyä teknologista kehysmenetelmää käyttäen voidaan myös esittää puheaineistoa muidenkin semanttisten rakenteiden mukaisesti.

Asiasanat: affektiivinen laskenta, emootiontunnistus, koneoppiminen, prosodiikka, tiedon visualisointi

Acknowledgements

This work was carried out during the years 2004–2014 at the Language and Audio Technology team of the MediaTeam Oulu research group in the Department of Computer Science and Engineering, University of Oulu, Finland.

First of all, I would like to thank my supervisor, Professor Tapio Seppänen, for supervising my thesis and providing guidance throughout my postgraduate studies. I also wish to thank Dr. Juhani Toivanen for his indispensable role in advising me, as well as for his major contributions to the research and the original publications. The work of the other co-authors, that is, M.Sc. Heikki Keränen, and Dr. Jukka Kortelainen is also gratefully acknowledged for their contributions and our discussion moments always so helpful. I would also like to give thanks to all the MediaTeam and Biosignal Processing Team personnel for the inspiring and positive work environment.

I would like to thank the official reviewers, Professor Tom Bäckström and Professor Klára Vicsi, for their work. Thanks go also to Dr. Pertti Väyrynen for the proofreading of the thesis manuscript.

For the financial support, I am grateful to Infotech Oulu Graduate School, Walter Ahlström Foundation, Tauno Tönning Foundation, and Oulu University Scholarship Foundation.

I am forever grateful to my parents for everything they have done for me. Final thanks go to my wife Maria-Melina, in particular, for her unwavering love and support, but also to all my friends for the much needed support and escape from the stress of the academic work which they provided.

Eero Väyrynen
Oulu, 2014

Abbreviations

AQ	*Amplitude Quotient*
ASR	*Automatic Speech Recognition*
AvRE	*Average Relative Error*
CART	*Classification and Regression Tree*
ClQ	*Closing Quotient*
CMS	*Cepstral Mean Subtraction*
CV	*Cross-Validation*
F_0	*Fundamental frequency of a speech signal*
FFT	*Fast Fourier Transformation*
GA	*Genetic Algorithm*
GRNN	*General Regression Neural Network*
HMM	*Hidden Markov Model*
HNR	*Harmonics-to-Noise Ratio*
HRTF	*Head Related Transfer Function*
IAIF	*Iterative Adaptive Inverse Filtering*
KF	*Kalman Filter*
kNN	*k-Nearest Neighbour*
KPCA	*Kernel Principal Component Analysis*
LDA	*Linear Discriminant Analysis*
LDC	*Linear Discriminant Classifier*
LFE	*Low Frequency Energy*
LLE	*Local Linear Embedding*
LOOCV	*Leave-One-Out Cross-Validation*
LPC	*Linear Predictive Coding*
MFCC	*Mel-Frequency Cepstral Coefficient*
NAQ	*Normalised Amplitude Quotient*
NB	*Naive Bayesian*
NFL	*No Free Lunch*
NN	*Neural Network*
OQ	*Open Quotient*
PCA	*Principal Component Analysis*

PDA	*Pitch Determination Algorithm*
PFS	*Promising First Selection*
PLP	*Perceptual Linear Predictive*
QOQ	*Quasi Open Quotient*
RMS	*Root Mean Square*
SBS	*Sequential Backward Search*
SD	*Standard Deviation*
SFFS	*Sequential Forward Floating Search*
SFS	*Sequential Forward Search*
SPL	*Sound Pressure Level*
SQ	*Speed Quotient*
STE	*Short Time Energy*
SVM	*Support Vector Machine*
V/UV	*Voiced/Unvoiced*
ZCR	*Zero Crossing Rate*

List of original articles

This thesis is based on the following four publications, referred to in the text by their Roman numerals (I–IV):

I Toivanen J, Väyrynen E & Seppänen T (2004) Automatic discrimination of emotion from spoken Finnish. Language and Speech 47(4): 383–412.

II Väyrynen E, Keränen H, Seppänen T & Toivanen J (2005) Performance of F0Tool - A new speech analysis software for analyzing large speech data sets. Proceedings of the 2nd Baltic Conference on Human Language Technologies: 353–358.

III Väyrynen E, Toivanen J & Seppänen T (2011) Classification of emotion in spoken Finnish using vowel-length segments: Increasing reliability with a fusion technique. Speech Communication 53(3): 269–282.

IV Väyrynen E, Kortelainen J & Seppänen T (2013) Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody. IEEE Transaction on Affective Computing 4(1): 47–56.

Contents

Abstract
Tiivistelmä
Acknowledgements **7**
Abbreviations **9**
List of original articles **11**
Contents **13**
1 Introduction **15**
1.1 Background 15
1.2 Research problem and objectives 16
1.3 Research scope and approach 17
1.4 Original publications and authors' contributions 18
2 Literature review **21**
2.1 Affective computing 21
2.1.1 Emotion recognition 23
2.2 Human voice and speech 24
2.2.1 Speech production 25
2.2.2 Source filter model 26
2.2.3 Auditory perception 27
2.3 Extraction of speech features 29
2.3.1 Acoustic features 30
2.3.2 Prosodic features 32
2.4 Vocal expression of emotion 39
2.4.1 Emotional speech models 39
2.4.2 Emotional speech databases 44
2.5 Emotion recognition from speech 45
2.5.1 Machine learning and pattern recognition 45
3 Research contributions **55**
3.1 Emotion recognition for Finnish speech 55
3.1.1 Emotion recognition using global prosodic features 56
3.1.2 Performance testing of prosodic feature extraction 60
3.2 Fusion technique for multi-modal classification of emotion 62

3.3 Visualisation of emotional speech data 66
3.3.1 Isomap based framework for semantic speech content visualisation 66
3.4 Summary of research contributions 72
4 Discussion **73**
4.1 Significance of results 73
4.1.1 Emotion recognition for Finnish speech 73
4.1.2 Fusion technique for multi-modal classification of emotion 74
4.1.3 Visualisation of emotional speech data 74
4.2 Limitations and generalizability 75
5 Summary and conclusions **77**
References **79**
Original articles **87**

1 Introduction

1.1 Background

Emotion recognition is a key process of an affective computer. In emotion recognition, the decoding of emotional content from visual, audio, and/or physiological signals is attempted. The understanding of emotional relevance is seen as an essential capability for artificial computing to interpret the semantic meaning of signals.

A thirst for understanding emotion and related phenomena in the human interaction is as old as the civilisation itself. Written evidences of a strong understanding of emotional relevance to semantics can be traced back to ancient times. For example, the Greek philosopher Aristotle taught in his great treatise *Rhetoric* about the concept of using emotion in public speaking as a tool for appealing to the crowds' innate emotional biases for a desired effect (i.e. ethos and pathos). During the early modern era in 1649, the philosopher Descartes in his last published work, the *Passions of the Soul*, identified emotions as simple and primitive passions from where all other forms of passion arise. Although Descartes acknowledged the mind-body union of a human being, he separated the two and placed the passions in the body from where they interfere with the decisions of the rational mind. Now, it is generally understood that this total division of mind and body was the so-called Descartes' error, greatly affecting the thinking of philosophers and scientists afterwards. From studying people with brain injuries, it can be argued that having emotions is an integral and essential component of conscious decision making and thus entwined with the reason itself (Damasio 1994).

Scientific study to model and understand emotion can also be traced back by more than two centuries. The current understanding of affects has evolved immensely from the early treatises on emotion, e.g. Bell (1806). Darwin (1872/1965) studied emotional expressions and made observations that link some emotional behaviour to evolutionary processes. During the last few decades, a popular concept of basic emotions (Izard 1992, Ekman 1992, 1999) has been found successful, especially when considering facial expressions. However, in the search of a more primitive, fundamental representation of affects, the old views that attribute emotions, and expressions thereof, to lists of monopolar emotional labels have been augmented by modern dimensional approaches (Russell 2003, Fontaine *et al.* 2007). Although emotional labels are still a relevant and

useful concept, a bipolar dimensional circumplex model (Russell 1980) can be seen as an essential representation of affects. Dimensional models can be used to represent the strict emotional labels in a looser, more malleable continuous structure.

Affective computers and affective technology, or the lack of, have been a staple theme in science fiction for decades. Engineering emotionally intelligent computers and agents is, however, no more just a good story. Affective computing has become a recognised and important field of technological study (Picard 1997). Affective computing is a multidisciplinary field bridging knowledge together from multiple fields such as cognitive sciences, psychology, computer sciences, natural language processing, phonetics, and linguistics. Advantages that affective computing can bring, range from better usability and user experience of human computer interfaces to more independent context aware autonomous technological agents. Ultimately, perhaps, a better understanding of semantics and affects can lead to a point when the Holy Grail of computer science, a fully conscious, self-aware artificial intelligence, becomes a reality.

1.2 Research problem and objectives

The overall research problem of this thesis is: How to automatically recognise communicated emotions from speech? Due to the vast size of the field of emotion recognition, even if nothing but all paralinguistic communication modalities (e.g. facial expressions, postures, gestures, speech, or other physiological signals) would be considered, the research problem had to be still narrowed. Speech is generally identified as an important channel for the communication of emotion, yet it has been relatively little studied, compared with facial expressions, in the context of emotion.

Consequently, in this thesis, the problem of emotion recognition is approached using only speech signals. More specifically, the most important paralinguistic component of speech, i.e. prosody, is primarily considered. The prosody is chosen because the prosody of emotional continuous spoken Finnish has not been extensively studied and no automatic emotion recognition technology exists yet for continuous Finnish speech using long-term global prosodic features. To accommodate for the narrowed approach, a fusion of different modalities is nevertheless considered during the research in order to retain some generalizability of results with the overall field of emotion recognition. The visualisation of data is a closely related problem. The nonlinearity of semantic concepts, e.g. emotion, requires more effective data representation techniques to generate semantically meaningful visualisations. An aim is to study nonlinear modelling

techniques in order to produce visualisations of emotional speech data that are consistent with a semantic evaluation. Specific sub-questions for this research can be formulated as:

1. Is emotion recognition from speech possible from Finnish using long-term global prosodic features?
2. What kind of technology is needed to combine different signal modalities and/or feature sets for multimodal emotion recognition?
3. Is it possible to produce visualisations of the intrinsic emotional data structures consistent with a dimensional model of emotion?

The technological goal was to develop methods for the automatic recognition of emotions from speech prosody. The papers published on the topic aim at three distinct main objectives. The first objective is to establish state-of-the-art emotion recognition for continuous Finnish speech using pattern recognition techniques. The second objective was to develop an information fusion based technology for the recognition of emotional speech when using multimodal features. The third objective was to develop a supervised method for emotional speech data visualisation.

1.3 Research scope and approach

The objectives of this thesis are approached using short- and long-term prosodic features and other related acoustic features in the context of Finnish speech. The selected approach moves the scope distinctively away from the standard Automatic Speech Recognition (ASR) field. By using speech prosody, the analysis is now concentrated on the semantic quality of the speech signal rather than the lexical information content of natural language. The intonation and acoustic quality of speech is a property that is continuously present in any speech signals and has a profoundly important meaning and impact on speech communication, yet it is still a quite poorly understood phenomenon. Technical tools for the study of this implicit speech information are needed.

The emotional corpus data used in the original publications was recorded in 2003 and 2006, using high quality condenser microphones and a controlled environment. The speech material was produced using recruited native Finnish professional actors with theatrical acting backgrounds. The recorded corpus contains hours of emotional speech material. From the material, a set of 280 around 10 second length basic emotion renditions was selected for the emotion recognition and the visualisation of emotion

studies. A set of 450 long [a:] vowel samples was selected for the short segment fusion study.

The simulated acted emotional content was chosen because one of the goals of the study was to develop emotional data visualisation techniques. The fact that an immense amount of available recorded material, e.g. television programs, are all acted material support this choice of an approach for technical analysis method development. The consequence of this approach is that for spontaneous speech many of the results of this study must be separately verified, however. A separate testing data containing classified radio communications between fighter jet pilots during a war exercise was used in the robustness testing of the feature extraction. The flight communication contains high noise and both high physical and psychological stress situations that can be used to test the extraction procedures at extreme conditions.

In this work, statistical pattern recognition techniques are utilised in the development of emotion recognition technology. First, statistical features are developed that can be automatically extracted from speech prosody. A classifier technology is used to enable a robust machine learning of emotional speech. Next, fusion techniques are investigated to enable the use of multimodal information ensembles for the classification of emotion in very short time frames. Finally, nonlinear modelling and classifier based optimisation techniques are developed for the visualisation of high-dimensional speech feature manifolds in a low dimensional semantically relevant space.

1.4 Original publications and authors' contributions

Paper I focuses on the classification of Finnish emotional speech content. In the paper, a standard pattern recognition approach to emotion recognition task is investigated using prosodic features extracted with a fully automatic method. In Paper II, the performance and quality of the used automatic prosodic feature extraction method is investigated. In Paper III, fusion techniques to combine prosodic features with voice quality features are investigated to enhance the emotion classification performance of short vowel length utterances. Finally, in Paper IV, a framework for semantic speech visualisation is presented in the context of emotional speech using a nonlinear manifold estimation technique, Isomap.

In all four original publications, the author was primarily responsible for the study design, development and implementation of algorithms and analysis of results. The author was also the principal manuscript writer of papers II–IV. In papers I–III, Dr.

Juhani Toivanen provided phonetic and language expertise, along with being the responsible principal manuscript writer for Paper I. Dr. Toivanen was also the primary contributor responsible for the design and collection of the emotional speech corpus used throughout the research (the author contributed technical assistance and supervision only). M.Sc. Heikki Keränen contributed to the data analysis design, provided manual reference estimates and annotations for the algorithm and speech error analysis, and contributed to the manuscript writing of Paper II. Dr. Jukka Kortelainen contributed to the study and algorithm design, as well as to the manuscript writing and revision of Paper IV. Professor Tapio Seppänen contributed to all publications as a supervisor and participated in the revision of the manuscripts.

2 Literature review

This chapter provides a review of the relevant literature and terminology related to the topic of the thesis. The review is divided into five sections. Section 2.1 reviews emotion recognition technology in the affective computing scope. In Section 2.2, the fundamental characteristics of the human speech production and perception systems are reviewed. In Section 2.3, the taxonomy and extraction of different features of speech are reviewed. Section 2.4 concentrates on the vocal expression of emotion in the context of natural languages. In particular, the role and importance of speech prosody, emotional models, and emotional corpus data are discussed. Finally, in Section 2.5, a review of current emotion recognition research and methodologies used in the context of emotional speech is provided.

2.1 Affective computing

"... emotion and feelings may not be intruders in the bastion of reason at all: they may be enmeshed in its networks, for worse and for better."
Antonio R. Damasio

Affective computing is a multidisciplinary field of computer science, psychology, and cognitive sciences related to emotion and the implications thereof to computing. Affective computing as a scientific technical field is still in its infancy. The defining work of Picard (1995, 1997) is a very good introduction to the general and technological problems of affects.

Before continuing, first, a few concepts have to be discussed. Emotion and affect are key terms frequently used in the field of affective computing, sometimes inconsistently. The terms are hard to define precisely and many definitions exist (Russell 2003). *Emotion* can be characterised as a brief, conscious and self-evident, state of mind that is accompanied by subjective feelings and tendencies for physiological, behavioural, neural, and/or expressive responses. An affect refers to a non-conscious change of state. An *affect* is the experience arising from a change in an affective state (e.g. emotion, mood, or feeling) due to stimuli, external or internal. Affect and emotion are routinely used in the literature synonymously. However, affect is a more general encompassing term that includes emotions, feelings and more general longer-term moods. In this

thesis, emotion is used to refer to distinct prototypical states in the affective dimension and affect, in its turn, is used when referring to the underlying broader context.

As the field of affective computing is young, many questions have not been robustly answered yet. The concept of emotion and cognition, as well as the interconnection of the two, in decision-making is also still an active field of study with large open questions, for a review, see Blanchette & Richards (2010). Many theories for cognitive appraisal have been presented, usually focusing on the conscious mind, e.g. by Oatley & Johnson-laird (1987), Ortony *et al.* (1990), and not so much on the sentic modulations (i.e. physiological signals). Nevertheless, the concept of body-mind interactions of affects is generally accepted. The general consensus currently is that emotions are both cognitive and physical. Thinking and feeling are neither separate processes nor the same; rather, they are interconnected and entwined together complementing each other. Affects can be triggered, reinforced, or suppressed by external stimuli, sentic modulations, and/or from within the conscious processes themselves (e.g. social rules), complicating any recognition, expression, and/or modelling of emotion.

To somewhat help sort the multifaceted aspect of emotions, Damasio (1994) distinguished between primary and secondary emotions. The primary emotions can be described as the more primitive, immediate emotional effects (typically accompanied by unintentional spontaneous sentic modulations). The secondary emotions, in their turn, are a product of cognition that might not be accompanied with any readily observable reactions (but can also include intentional sentic modulations). The notion of basic emotions (e.g. sad, anger, happy, disgust, fear, and surprise) argued to have explicit facial expressions (Ekman 1992) can be tied to the primary emotions, while secondary emotions relate to more subtle emotions typical in social interactions (e.g. shame or guilt). However, any kind of lists of emotional labels or hard taxonomical structures for emotions derived from natural languages are still problematic (Ortony & Turner 1990).

For a computer to have an affective dimension, it needs to account for at least some emotional components. The most primitive requirement for an affective computer is the ability to recognise emotional signals. The recognition of emotion is typically connected with some reasoning to understand the context of the detected signals but no more. The ability to express emotions is an important but straightforward expansion of an affective computer. The expression of emotion is dependent on some internal or external reasoning, but does not require any deeper level of understanding other than proper contexts. Both recognition and expression can be seen more or less as simple pattern recognition and/or signal processing tasks. Finally, for a computer to actually have

emotions, it needs to have a whole range of emotional capabilities. Picard (1997) lists the components needed for an emotional system as: emergent emotions, fast primary emotions, cognitive emotions, emotional experience, and body-mind interactions.

2.1.1 Emotion recognition

Emotion recognition is in principle based on detecting and unscrambling sentic modulations. Picard (1997) divides the possible sentic modulations into signals `apparent` and `less apparent` to others (see Table 1).

Table 1. Sentic modulations.

Apparent	Less-apparent
Facial expressions	Respiration
Voice	Heart rate/pulse
Gesture/movement	Temperature
Posture	Electro-dermal response/perspiration
Pupillary dilation	Muscle action potentials
	Blood pressure

Picard (1997) also identifies criteria for emotion recognition, summarised briefly in Table 2. The criteria for emotion recognition can be seen as relevant steps in a classic pattern recognition and machine learning problem. The inputs in the context of this thesis are audio speech recordings. Prosodic and acoustic features can be extracted from the recordings to produce feature sets used for pattern recognition. The reasoning and learning steps can be included in machine learning using feature selection and

Table 2. Criteria for emotion recognition.

Criteria	Short description of criteria
Input	Receiving of input signals (i.e. raw sentic modulation data).
Pattern recognition	Feature extraction and classification (i.e. relevant emotional features and structures for input signals).
Reasoning	Prediction of emotion based on knowledge about emotion generation and expression, i.e. reasoning about situations, goals, preferences, social rules, and other perceived context.
Learning	Learning of person dependent factors and updating of rules used in future reasoning based on new information and reasoning.
Bias	Optional bias in recognition to account for internal emotional states, if emotionally induced behaviour is defined.
Output	Descriptions of recognised emotions and expressions (e.g. probabilities for current and predicted emotions).

other supervised learning approaches. The bias information can be seen as an option for recognition. Emotional bias is relevant if an inner emotional state is defined and emotionally influenced behaviour is sought for an affective system. The emotional bias can be included in machine learning using supervision. For classifier decisions, the emotional bias can also be implemented at a later stage if the classifier outputs are defined more robustly than just the most likely class (e.g. probability estimates are given to all prospective classes).

Multi-modality

The numerous different sentic modulations suggest that the underlying affective state can be approximated using multiple sources of information. This multi-modal nature of emotional expression is well known and explored (Brown 2005). From the apparent sentic modulations, facial expressions and voice are the best known and robust sources for emotion recognition (Russell 2003). These different modal sources are, however, not exclusive to emotional expressions. Sentic modulations are convoluted with other signals, both physiological functions (e.g. pupil adaptation to light changes) and other semantic signals (e.g. gestures to point at objects). Perception of semantic information is routinely done using as many clues as possible. For example, humans interpret speech content using bimodal audio-visual information (Brown 2005). The multi-modal nature of emotion requires robust recognition modalities. However, using classifier fusion techniques this problem can be approached by a divide and conquer tactic. The overall multi-modal recognition problem is first reduced to single modality problems and then an affective system can use fusion to make multi-modal decisions based on available single modal experts.

2.2 Human voice and speech

The voice has a fundamental meaning for us human beings. Without the voice, no language, speech, or a culture of information exchange like ours would exist. Even the essence of our conscious thinking could be fundamentally different. It is thus not surprising that the human voice and speech are one of the most interesting and studied phenomena in our world. This chapter is a brief overview of the human speech production system, perception, and related acoustic basics. For a more thorough treatment of the human auditory system and acoustic properties thereof, see the *Encyclopedia of*

Language & Linguistics (Brown 2005), the works of a pioneer of speech G. Fant (Fant 1960, 2004), or any good book on the subject, e.g. Kent & Read (1992), Titze (1994). For an in depth review of speech from the ASR viewpoint, see Rabiner & Juang (1993).

2.2.1 Speech production

The human speech production system (Fig. 1) is a complex system capable of producing a theoretically infinite number of distinct sounds. The system can be divided into three major subsystems that contribute to specific operations in speech production: respiratory subsystem (diaphragm, lungs), laryngeal subsystem (larynx) and articulatory subsystem (oral/nasal cavities, soft/hard palate, tongue, jaw, lips and teeth) (Kent & Read 1992). It is important to notice that the respiratory and laryngeal subsystems, providing the source signals of speech, are in effect separate from the articulation processes of the vocal tract. Therefore, the *fundamental frequency* (F_0), i.e. the quasi-periodic cycle frequency of voiced speech, and the glottal waveforms differ only a little across the vowels. The F_0

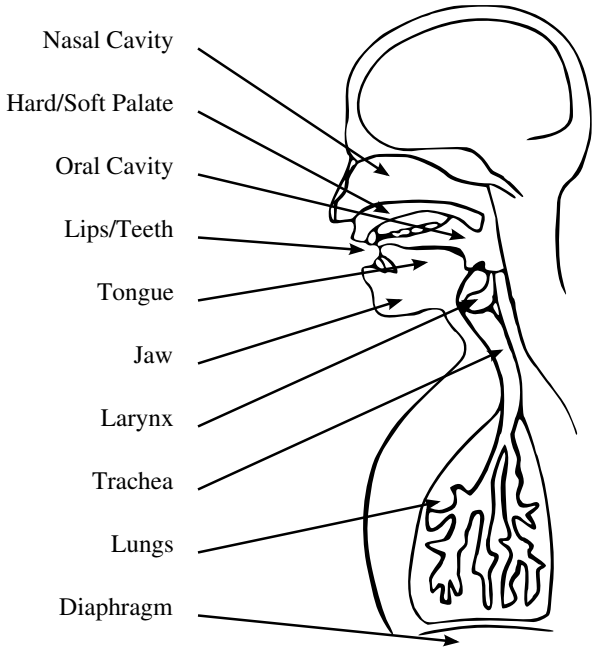


Fig. 1. An overview of the human speech production system.

and the patterns thereof, called *intonation*, as well as the *intensity* of speech, i.e. the power of a speech signal per unit area, are mainly the functions of pulmonary pressure from the respiratory subsystem and the larynx (Kent & Read 1992, Titze 1994). The articulatory subsystem is then responsible for the final articulation of different phones such as vowels (e.g. '[a]'), fricatives (e.g. '[f]') and plosives (e.g. '[p]').

2.2.2 Source filter model

The finding that laryngeal and articulatory processes are in effect separate is the basis of one of the most fundamental models used in speech technology. The assumption allows a total separation of the complete speech production process into a linear convolution of a source signal, i.e. a glottal waveform or a noise source in normal speech, and a vocal tract filter. This linearized model is used widely in various fields of speech technology and is commonly referred to as the source filter model (see Fig. 2).

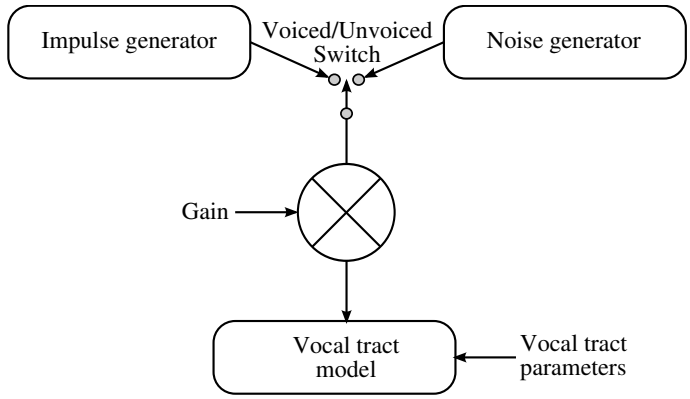
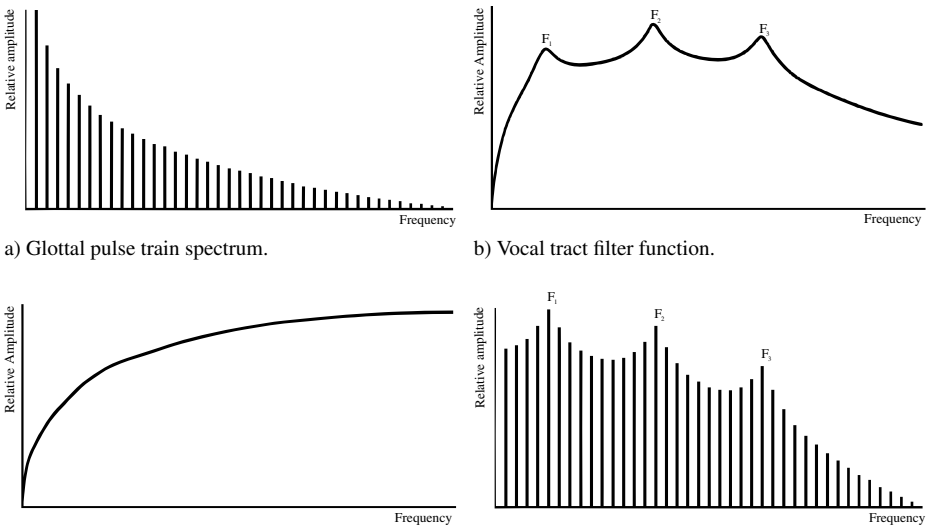


Fig. 2. Source-filter model of speech production.

The main parameters in the source filter model are the cycle period, gain and vocal tract parameters. The switch operating the voiced and unvoiced input selection can be seen as a slider that also allows a mixture of both input signals and controls the harmonics-to-noise ratio of speech. The cycle period and the gain parameters are directly related to the F_0 and intensity of speech, respectively. Vocal tract parameters include the formant frequencies (F_1, F_2 and F_3) that describe the vowel produced for voiced speech. (Kent & Read 1992, Titze 1994)



a) Glottal pulse train spectrum.

b) Vocal tract filter function.

c) Lips and teeth radiation function.

d) Radiated sound pressure spectrum of voiced speech.

Fig. 3. Illustration of source signal and filter functions for voiced speech production.

The equation for the linear model in the frequency domain is:

$$\mathrm{P}(f) = \mathrm{U}(f)\mathrm{T}(f)\mathrm{R}(f), \tag{1}$$

where, assuming voiced speech, $\mathrm{P}(f)$ is the radiated sound pressure spectrum of speech (Fig. 3d), $\mathrm{U}(f)$ is the glottal source function (Fig. 3a), $\mathrm{T}(f)$ is the vocal tract function (Fig. 3b), $\mathrm{R}(f)$ is a radiation function of lips (Fig. 3c) and f is frequency.

2.2.3 Auditory perception

The influence of perception cannot be disregarded when addressing speech production. The human auditory system is an inseparable combination and adaptation of both hearing and voice. Implications of this signal adaptation are profound in speech technology. The way the human ear and auditory system processes sound has many consequent qualities defined as psychoacoustics (for a throughout treatment of psychoacoustics, see Fastl & Zwicker (2007)). Most importantly, the psychological perception of auditory signals is in no way linearly correlated with the corresponding physical acoustic scales. Even

basic properties like the bandwidth of hearing is subject to changes, for example age is a well-known factor.

Frequency perception

Pitch is the perceived frequency of an auditory signal. The perceived pitch of an acoustic signal can differ from the actual frequency of the signal considerably. The perceptual correspondence to the physical measures can be estimated with a scaling such as the mel-scale (Stevens *et al.* 1937) that is designed to represent the perceptual pitch in relation to the actual frequency. The bark-scale (Zwicker 1961) introduces a concept of critical bands where the pitch perception range is divided into filter banks according to the frequency sensitivity of the cochlea in the inner ear. The concept of critical bands is closely related to the basic psychoacoustic phenomena of frequency masking. Masking, in general, is the ability of another signal, close in time or frequency, to modify, suppress, or otherwise hide the perception of a signal. To make matters more complicated, the masking properties of auditory perception are also dependent on the spectral and temporal composition of the acoustic stimuli (e.g. noises, narrowband tones, or complex harmonic sounds) leading to a multitude of tuning curves. For speech, the impact of masking is also seen in the formant and vowel perception, key elements of speech signals, where the vowel formant frequencies integrate at 3.5 bark distances (Beddor & Hawkins 1990). The frequency masking of formant frequencies has a tremendous impact on any semantic evaluation of speech analysis (e.g. formant analysis).

Intensity perception

The perceived intensity, i.e. *loudness*, of audio signals, as a concept is also at least equally complex as the perception of frequency. The human interpretations of intensity for different audio stimuli can be modelled by exponential functions (Stevens 1957). The perceptual loudness of an audio signal is measured in the units of phon for pure tonal signals. The loudness perception as a function of frequency is a nonlinear function as well. Typically, the perception of loudness as a function of frequency is represented with equal-loudness contours (Fletcher & Munson 1933, Robinson & Dadson 1956). As is the case with the pitch perception, a multitude of different scalings exists dependent on the signal properties. For tonal signals, the A-scale, representing roughly the tonal

response of the 40-phon equal-loudness contour, is often used as a simplified solution for loudness scaling. A more thorough measurement of the tonal signal equal-loudness contour is defined in the ISO 226:2003 standard. The ITU-R 468 standard presents a perceptual loudness scale for noise signals. The equal-loudness contours are, however, presented for frontal audio sources only. When the highly individual Head Related Transfer Functions (HRTF) are considered, the perception of loudness of direction free sound sources clearly becomes even more complicated. An HRTF is an omnidirectional estimate of the impact that the various tissues of the head, especially the pinnae, have on the auditory signal.

2.3 Extraction of speech features

The taxonomy of speech derived features is not straightforward. In general, the inputs to a pattern recognition system are called just features. However, a higher level of taxonomy is useful to assess the information expected in a feature. The categorization of features can be seen as a kind of preselection of expectedly efficient features. Speech conveys not only linguistic messages, but also includes a major paralinguistic component, prosody. Prosody is used to convey information that augments the linguistic message, e.g. cues of speakers' emotional state.

It is generally defined in linguistics, e.g. by Brown (2005), that prosodics comprises the following suprasegmental perceived speech properties/acoustic correlates: pitch/F_0, loudness/intensity, and rhythm/duration. *Suprasegmental* property is defined in this linguistic context as a property that spans over identifiable discrete phonetic and linguistic unit borders (e.g. phonemes and utterances). The taxonomy of speech properties into suprasegmental and segmental components is not to be confused with any technical segmentation of a speech signal during the extraction of features. Features that describe prosody are referred to as *prosodic features*. Features that are not designed to describe prosody are called *acoustic features*.

However, there is a plurality of extraction techniques for features that do not intuitively fall under the prosodic or acoustic categories. For example, many linguistic and phonetic correlates (e.g. stress) are strictly segmental in nature, i.e. as defined in phonetic and linguistic sciences. Features describing segmental units are often referred to as acoustic features in the literature even if such a segmental unit is signalled using prosody (e.g. stress using pitch). This division into suprasegmental prosodic and segmental acoustic features is still even more problematic as there are many features

that are attributed to acoustic features in the literature but estimate some suprasegmental property (e.g. voice quality) and thus are intuitively close to prosody. Finally, features that attempt to preserve the whole speech signal, i.e. transformation based features, contain also prosody but are traditionally included in acoustic features. Transformation based features are also often referred to more precisely by using the name of the transformation, e.g. MFCC features.

The taxonomy of the speech features in the literature thus seems to be heavily influenced by the originating field of research. In this thesis, the main taxonomy of speech features is defined as suprasegmental prosodic features and other acoustic features to retain resemblance to the literature. Prosodic features are, however, extended to include the various voice quality features.

2.3.1 Acoustic features

Acoustic features often attempt to parameterise the whole acoustic speech signal in an efficient way. Traditional acoustic features are in many times engineering motivated and used to efficiently solve specific practical problems. Feature extraction is routinely based on basic methods found robust in other engineering approaches. The Fast Fourier Transform (FFT) is a great example of such a method and consequently is widely employed in acoustic feature extraction (e.g. FFT is also the basis of cepstral coefficient extraction). The use of FFT based approaches also explains the fact that common acoustic features are almost invariably using short segment based extraction ideology. As a consequence, while accomplishing a robust efficient algorithmic description, the connection to linguistic and/or physiological representations or interpretations may be lost. Some acoustic features, however, are more closely connected to a linguistic and phonetic background. For example, different acoustic features have been designed to quantify distinct phoneme information (e.g. features derived from speech formant estimation) and thus are strongly identifiable as segmental phonetic/linguistic features.

Typically, acoustic features are used in speech diarisation applications, for a review, see Moattar & Homayounpour (2012). For speech technological applications where individual speaker qualities and long-term properties are also relevant (e.g. affective sensing and analysis or speaker recognition), the additional use of longer term features has been found improving performance (Shriberg 2007). For the aforementioned purposes, a number of longer term acoustic correlates have been developed and adopted. These typically include non-linguistically motivated suprasegmental features, e.g. the

Zero-Crossing Rate (ZCR), Short Time Energy (STE), Harmonics-to-Noise Ratio (HNR), and long-term average spectrum.

Cepstral coefficients

A classic acoustic feature set commonly used is the Mel-Frequency Cepstral Coefficients (MFCC), attributed often to (Bridle & Brown 1974, Mermelstein 1976, Davis & Mermelstein 1980), and its variants (e.g. log power coefficients, delta MFCC, and delta-delta MFCC). The MFCC features have been found extremely efficient in various speech technologies. Other spectral derivations and LPC based features such as Perceptual Linear Prediction (PLP) coefficients (Hermansky 1990) are also often used.

The basic extraction process for cepstral coefficients includes a frequency domain transform of the speech signal (typically an FFT). In the frequency domain, a filter bank with a number of pre-selected frequency bands (the number of filters corresponds to the number of cepstral coefficients extracted) is used to filter the resulting power spectra. The filter bank is usually selected using a perceptually motivated logarithmically spaced scale, e.g. mel-scale (Stevens *et al.* 1937) for MFCC (see Fig. 4), but sometimes, the bark-scale (Zwicker 1961) is used as is the case for PLP. Triangular windowing functions are commonly applied. Alternative filtering functions, e.g. rectangular or Gaussian windows, are also used. Intensity weighting relative to frequency, e.g. using an equal-loudness contour, and/or nonlinear intensity-loudness compression can be employed to produce perceptually motivated features that more closely model the human auditory system.

In the case of MFCC, the logarithms of the resulting filtered frequency banks are then transformed back into the time domain where the amplitudes of the filters are then interpreted as the cepstral coefficients. Typically, segmental difference features (delta, and delta-delta) are also collected, as well as the total log power (i.e. the zero coefficient). The coefficients are often also decorrelated using the Cepstal Mean Subtraction (CMS) method where the long-term means of the cepstral coefficients are subtracted from individual cepstral coefficients resulting in zero mean signals depicting cepstrum variability. For LPC based approaches, the transform to final features is performed typically by an all-pole autoregressive modelling of the spectral envelopes.

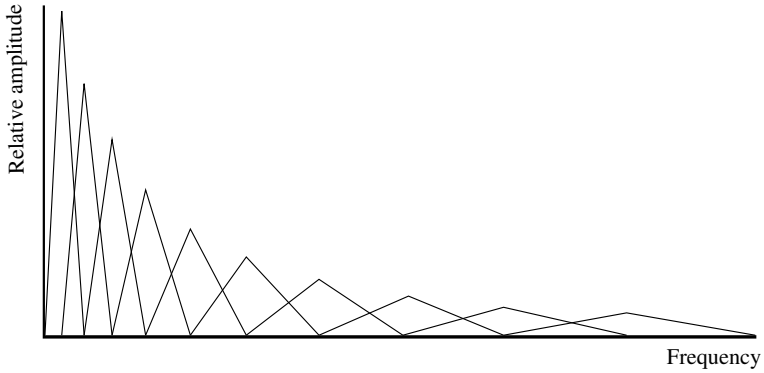


Fig. 4. Normalised mel-frequency filter bank.

Speech formants

Speech formant structures code the phonemes of spoken languages. As features, the speech formants describe the articulation of speech and are estimates for the parameters of the vocal tract. As the main function of formats is to define single phoneme length segments, they are usually considered as segmental features from a linguistics perspective. However, formant structures can have acoustic properties that are carried over the phoneme boundaries, thus giving them also some suprasegmental qualities. Formants and features derived from formant data are therefore many times used as voice quality features and hence can also be argued to have some prosodic qualities.

Formant analysis is typically based on autoregressive modelling techniques, e.g. LPC (Atal & Hanauer 1971, Makhoul 1973). The spectrum of a voiced speech sample or segment is analysed for peaks in the spectrum envelope. The peaks are then interpreted as discrete formant frequencies enumerated in ascending frequency order (typically denoted as $F_1, F_2, F_3, ...$). The resulting voiced samples can then be projected on a space with axes defined by the formant frequencies (typically in bark scaling using the first two formants only). Fig. 5 illustrates a typical formant map for Finnish vowels.

2.3.2 Prosodic features

Prosody of speech is defined in the linguistic literature as the suprasegmental properties of speech. Traditionally, the definition of prosody is thought to include the pitch/F_0, loudness/intensity, and rhythm/duration aspects of speech (Brown 2005). Purely from a

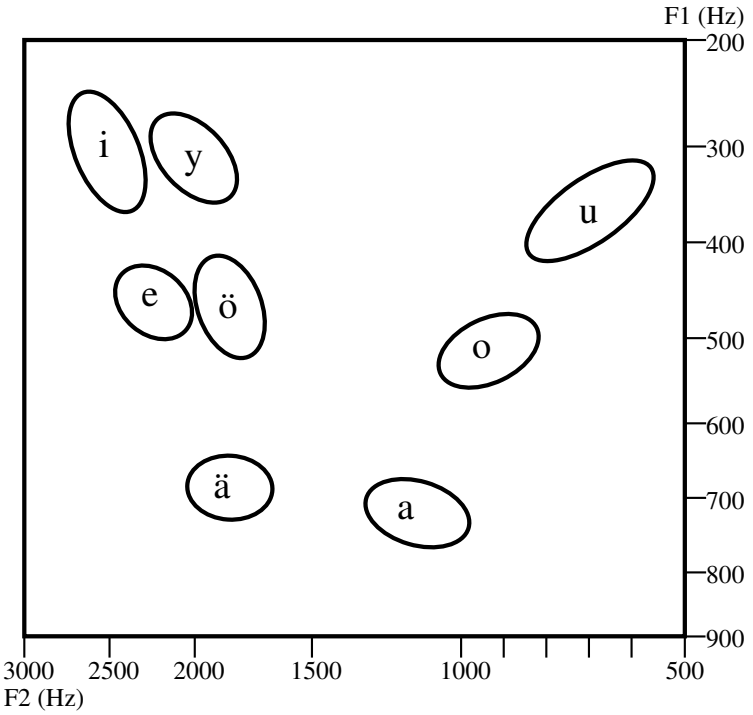


Fig. 5. Sketch of Finnish vowel chart (bark scaled axes).

technical point of view, there are no features that are intrinsically prosodic, i.e. prosodic features can be estimated from segmentals but also audio features or variants thereof can be calculated over multiple segmentals to describe suprasegmental properties. The features are called prosodic features mostly as they have phonetic and linguistic model etymology and aim to model the suprasegmental properties of speech. Therefore, many other features that have a different background can also be seen as prosodic features if they are seen capturing the suprasegmental aspects of speech. As a consequence, there does not exist any technical clear boundary between acoustic and prosodic features.

Many longer time domain acoustic features (e.g. long-term spectral measures) also describe the suprasegmental properties of speech. Especially, the voice quality aspects of speech are generally seen forming the most obvious group of speech properties that has not been traditionally viewed as a part of prosody. Unlike with the other prosodic speech properties, voice quality cannot be fundamentally associated with a single perceptual property and its corresponding acoustic correlate. However, the voice quality

is suprasegmental in nature and thus convincing arguments have been made to include voice quality under prosody (Campbell & Mokhtari 2003). In this thesis, voice quality has been included taxonomically under prosody together with the three other prosodic primes. The definition of prosody used in this thesis when classifying features is as follows: pitch/fundamental frequency, loudness/intensity, rhythm/duration, and voice quality.

Pitch/Fundamental frequency

Pitch is the most important prosodic property of speech. The pitch contour, a perceptual property, is directly related to the fundamental frequency F_0 contour, an acoustic correlate, formed by the larynx during the phonation of speech. The F_0 contour is extracted from speech using a Pitch Detection Algorithm (PDA). Various PDA approaches have been studied extensively, see Hess (1983) for an overview of pitch determination. Current state-of-the-art methods are typically constructed using cepstrum based PDA algorithms (Noll 1967, 1970, Schroeder 1968), or autocorrelation based methods (Boersma 1993). Pitch determination in the context of human speech is still an active research area. However, in the context of this thesis, the performances of PDAs are at a sufficiently high level (cf. Rabiner *et al.* (1976), and Bagshaw *et al.* (1993)) and, consequently, the performance of algorithms does not seriously limit the quality of long-term prosodic feature extraction.

Extracted features for the F_0 contour typically include distribution parameters such as the mean, median, range, variance, skewness, and kurtosis of contour values. However, the F_0 contour values can be extracted in a wide domain. F_0 can be measured from small or large segmental units such as phonemes or utterances, but also from whole paragraphs of spoken material. The analysis segment size chosen for feature extraction greatly influences the information captured in the extracted features. Therefore, a short segment F_0 measure can also be a single value only. Additionally, a range of derivative data and distribution features thereof can be measured. The steepness of F_0 movements and durations of accents can be calculated. Furthermore, the F_0 contour has a tendency of declination towards the ends of sentences for normal speech. The type of language has a large impact on the properties of speech F_0 contours as well due to different functional use of pitch in languages. Between tone and pitch accent type of languages, the intonational contour structure differs enormously (Thymé-Gobbel & Hutchins 1999).

Loudness/Intensity

The importance of the intensity aspects of the speech signal for prosodic signalling is comparable to that of the F_0 contour. Technically, the actual physical quantity called the intensity of speech, i.e. the acoustic intensity, itself is not directly measured by microphones but, rather, the sound pressure is sensed. Using calibrated microphones the logarithmic relative measure called sound pressure level (SPL) can be estimated. The SPL is a quantity closely related to the intensity of speech. If certain conditions are fulfilled (i.e. recordings in an anechoic room, in the far field, and using an omnidirectional microphone capsule), the intensity of speech is proportional to the measured SPL. Therefore, the recorded pressure signal can be used as a proxy of intensity.

A prosodic intensity contour is technically trivial to extract from speech recordings, for example, directly from the time domain signal using some short segment root mean squared (RMS) routine or from the power spectrum of speech derived from the frequency domain. A STE contour can be seen as a prosodic intensity measure. Loudness features that attempt to describe the perceptual experience of a human listener are clearly more difficult to produce. The audio signals must be transformed using the various psychoacoustic corrections, e.g. the equal-loudness contours. Statistical features (Mean, median, range, variance, skewness and kurtosis) can be calculated from the derived feature contours. Psychological effects also affect the intensity of speech. *Vocal effort*, a quantity used to adapt speech according to the perceived distance to the receiver, has an important effect on the intensity of speech (Traunmüller & Eriksson 2000). Vocal effort also affects F_0 and voice quality.

In addition to personal variations, the measurement of intensity, unfortunately, has an increased sensitivity to variations in the recording conditions compared with F_0 or other frequency measures. The most inherently problematic nature of the intensity related features is the squared relationship between the distance of the sound source and the sound pressure measured at the recording microphone. Additional problems are presented by the physical dimensions of the recording environment, as well as the used microphone capsule types. The Lombard effect, i.e. the tendency of increasing vocal effort in the presence of noise, is also a factor.

Features derived from intensity contours are consequently also affected. Without knowledge about distance, the intensity measures can be unreliable on their own. The general unreliability of intensity derived features for uncalibrated recordings is one reason why comparably little research interest has been directed to intensity features

than F_0 based prosodic features. Only basic statistical features are commonly included in prosodic feature sets (Ververidis & Kotropoulos 2006).

Rhythm/Duration

The timing and duration of various speech parts forms the rhythm/duration family of prosody. The duration and timing of speech segments are usually estimated during voicing analysis performed in order to do pitch determination. Although duration properties are formed of clear segmental parts of the speech signal, the resulting derived prosodic features (e.g. speech rate) are typically suprasegmental and can be defined in a wide domain. Features for durations include ratios of voiced, unvoiced and silent segments. The classification of pauses and voiced segments can be used to calculate articulation and speech rate related correlates. The classification of segments can also be used to form ratios for different length segment usages. It needs to be taken into account, however, that speech rates, the usage of pauses and other timing related features based on preset classification criteria can be language, culture and even situation dependent (e.g. read or spontaneous speech) (Lehtonen 1985, Campione & Véronis 2002). For western languages, the distribution of silent pauses can be modelled by a bi-Gaussian distribution (Campione & Véronis 2002). Changes in read text speech rates move the bi-Gaussian distributions of pauses, but the overall model remains (Demol *et al.* 2007). The pause lengths are also relative to intonational complexity and other cognitive load factors (Krivokapić 2007).

The role of timing and duration in speech is indeed broad. The usage of different pauses and the timing together with intonations are very common as cues of the functional level of speech. This multi-role aspect of durations complicates the evaluation of the features and affects the very information the features capture when a different domain is used for extraction, i.e. a short window captures more local functional cues and a longer window may reveal more general aspects of rhythm. Furthermore, in language identification studies, e.g. by Thymé-Gobbel & Hutchins (1999), Farinas & Pellegrino (2001), the role of rhythm has been highlighted between different languages. Considerable differences in usages of durations between the syllable timed and the pitch accent language families have been reported. Segment durations have also been detected reflecting differences in the syllabic structures of languages that are more closely related.

Voice quality

Contrary to the other acoustic correlates of prosody, F_0, intensity, and duration, the quality of voice does not have an acoustic property that is easily distinguishable and measurable from a speech signal. Voice quality is composed of many aspects of the speech production. Usually, the quality of voice is characterised by qualitative terms such as hoarseness, whispering, creakiness, etc. Traditionally, the voice quality is not seen as a part of prosody. This may be due to poor understanding of the role the related acoustic parameters have on the voice quality. In any case, voice quality is definitely suprasegmental in nature and, therefore, arguments have been made to place the quality aspects of voice under prosody, e.g. by Campbell & Mokhtari (2003). Differences in quality can be seen in the spectral properties of speech, but also the subtler changes in F_0 and intensity can contribute to voice quality. The spectral features can include the spectrum tilt and ratios of energy in different frequency bands. Voice quality, however, is not prominently represented in such trivial spectral features. More specific spectral features have been developed using formant frequencies (Lugger *et al.* 2006). Many voice quality features have been researched and developed in medical voice research. Among the more well-known features, the shimmer and jitter of voiced segments, originally developed to analyse pathological voices, see e.g. Horii (1979, 1980), can be important source-related features of speech quality (Bachorowski 1999). The HNR of speech, originally used for pathological voice analysis, is also strongly related to normal voice quality as a correlate of hoarseness (Yumoto *et al.* 1982).

Research into voice source properties has also identified many parameters that are seen fundamentally associated with voice quality. To produce source features, the glottal flow function must be extracted. For a thorough review of glottal flow estimation and parameters, see Alku (2011). The glottal flow can be calculated using an iterative adaptive inverse filtering (IAIF) method (Alku 1992). Typically, the flow function is estimated using an idealised flow model, e.g. the Liljencrants-Fant (LF) model (Fant *et al.* 1985), that is fitted to the inverse filtered flow, e.g. by a nonlinear least square error minimisation or a Kalman Filter (KF) (Alku 1992, Airas 2008, Li *et al.* 2011). Features describing the voice source can then be extracted from the inverse filtered glottal flow and the derivative flow, but also from the fitted model (i.e. estimates for the model parameters can be seen as features describing voice quality). Model derived features are, however, reported as sensitive to noise. More robust methods that estimate the model parameters from the frequency domain have been proposed, e.g. by Kane *et al.* (2010).

For a good treatment of model extracted features, see Scherer *et al.* (2013).

In Fig. 6, an illustrative example of glottal flow and the corresponding derivative flow through vocal folds is depicted. Many source related parameters can be derived from the glottal flow and its derivative. The Open Quotient (OQ) and the Speed Quotient (SQ) forms an essential representation of the glottal flow. OQ is defined as the ratio of time the glottis is open in relation to the total cycle period time. The SQ is the ratio of rise and fall time of the glottal flow; thus, the ratio between the glottal opening and the closing phases reflects the asymmetry of the glottal pulse, skewness. Many times, the glottal opening is formed of two distinct phases (note the knee in Fig. 6) that represent the primary (i.e. the onset of flow) and the secondary (i.e. an abrupt increase in the flow) openings of glottis. The two-phased opening of glottis is perhaps caused by a pistonlike movement of the vocal folds due to pulmonary pressure just before the actual

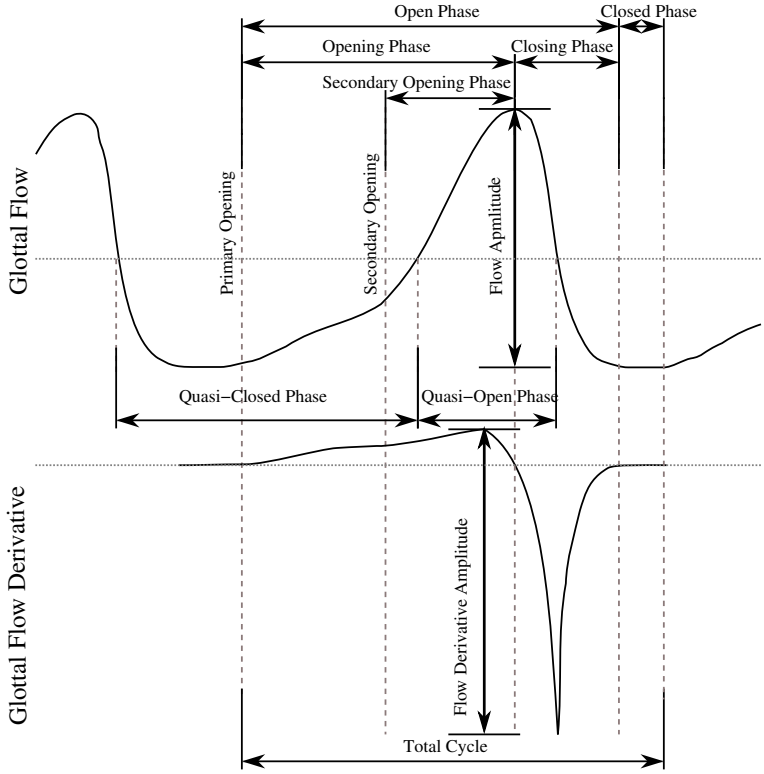


Fig. 6. Glottal flow and derivative flow waveforms (III, published by permission of Elsevier B.V.).

opening of glottis. The primary and secondary openings can be used to define open quotients corresponding to the primary opening (OQ1) and the secondary opening (OQ2). Corresponding primary and secondary speed quotients (SQ1 and SQ2) are also defined from the two openings, respectively. Due to the problems of accurate detection of the instants of the glottal openings, other variants of OQ have been also developed. In the Quasi Open Quotient (QOQ), a threshold is applied to the glottal flow to produce a more robustly evaluated quasi-open phase. The quasi-open phase now represents the opening and closing of glottis that is comparable to the open phase used in OQ. SQ and QOQ indirectly describe vocal fold vibration (Laukkanen *et al.* 1996). Another example is a variation of the open quotient (OQa) derived using the LF model (Gobl & Ní Chasaide 2003). The Closing Quotient (ClQ) can be further defined as the ratio of the closing phase in relation to the total cycle.

Because OQ , ClQ, and consequently SQ, are sensitive to errors in the estimation of glottal closures, more robust parameterisations of the glottal flow have been proposed. The Amplitude Quotient (AQ) (Alku & Vilkman 1996) defines a ratio of glottal flow peak-to-peak pulse amplitude in relation to the flow derivative peak. A normalised version, the Normalised Amplitude Quotient (NAQ) (Alku *et al.* 2002), is further defined as the AQ normalised with respect to the fundamental period time. NAQ can be seen as an estimate for glottal adduction (i.e. it is closely related to ClQ) that is more robustly measured because an accurate estimation of the glottal closure is not required. NAQ (Eq. 2) is therefore a robust parameter that is related to the pressedness of speech.

$$\mathrm{NAQ} = \frac{AQ}{\mathrm{T}} = \frac{f_{ac}}{d_{peak}\mathrm{T}}, \tag{2}$$

where f_{ac} denotes the peak-to-peak alternating flow amplitude, d_{peak} is the flow derivative amplitude, and T is the total cycle period length. The intensity of the speech signal itself can also be seen as a component of speech quality as intensity has a direct effect on the other speech quality properties (Murray *et al.* 1996, Zetterholm 1999).

2.4 Vocal expression of emotion

2.4.1 Emotional speech models

Finding an emotional model of speech has been for a long time an aspired goal for researches. The simplest discrete label based emotion model can be traced back by

more than a century (Darwin 1872/1965). The discrete emotional model has a strong ecological foundation. A group of so-called "big six" fundamental basic emotions (sadness, happiness, anger, fear, disgust, surprise) is a distinctive example of this link between evolutionary survival related processes and the chosen emotional model labels. This label based model has been found remarkably efficient and thus is still a commonly used paradigm in many emotional studies, e.g. by Dellaert *et al.* (1996), McGilloway *et al.* (2000), Oudeyer (2002), Bosch (2003).

Although the usage of emotional class labels has provided efficient techniques, the problem of a discrete model still remains. When a more accurate model is desired, the inflexibility of strict labels leads to an ever larger amount of emotional classes resulting in an unnecessarily complex model that is susceptible to the curse of dimensionality (i.e. an additional class in the model requires exponential amount of training data to fill the resulting space). More flexible dimensional models have been introduced in an attempt to reduce the emotional space into a few semantic dimensions. A circumplex model of emotion (Russell & Mehrabian 1977, Russell 1980) is the basis for these dimensional emotional models (Fig. 7).

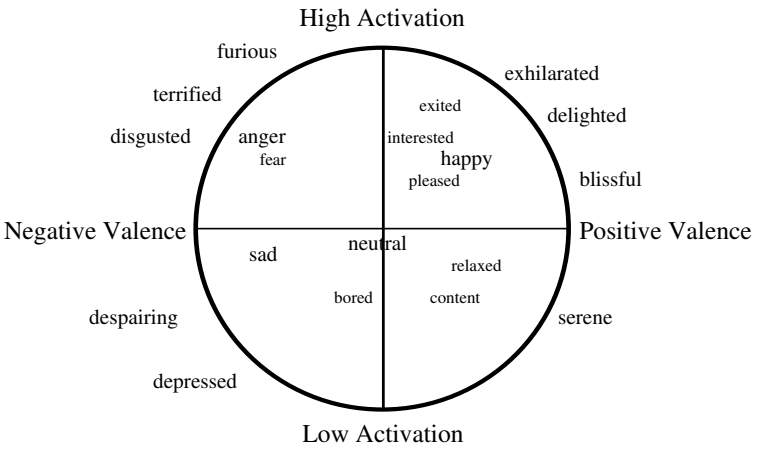


Fig. 7. A two dimensional circumplex model of emotion. Adapted from Cowie *et al.* (2000).

The dimensional model of emotion is currently a widely accepted paradigm for the representation of affect. However, the dimensions associated with models are still up for debate. The original circumplex model is constructed using only the two most

accepted emotional dimensions, activation and valence. Activation and valence as emotional dimensions are widely accepted. Russell & Feldman (1999) argues that these dimensions are the only relevant emotional components. The view is that other proposed emotional dimensions do not have any orthogonal components to activation and valence and hence should reduce to the two emotional dimensions. The validity of the two prime emotional dimensions, activation and valence, is not challenged. However, many researchers argue that other emotional dimensions exist. The most common additional emotional dimension is potency. The argument supporting potency as an independent dimension is the common observation that some negative yet fundamentally different emotions (e.g. fear and anger) are poorly differentiated in the two-dimensional approach, e.g. Russell & Mehrabian (1977), MacKinnon & Keating (1989), Fontaine *et al.* (2007). Smith & Ellsworth (1985) have argued that up to six emotional dimensions could be distinguishable. Togneri *et al.* (1992) have postulated, however, that the intrinsic dimensionality of speech signals should not be higher than four, limiting the emotional structure of speech into similar dimensional bounds. The integration of *emotional intensity*, i.e. the strength of emotion, in the dimensional models is also not a settled issue. Laukka *et al.* (2005) have modelled the emotional intensity as a fourth dimension. Daly *et al.* (1983) presented a cone-like model to include the emotional intensity in a three-dimensional structure. The two-dimensional circumplex model is also capable of modelling emotional intensity to some degree. In the circumplex models, emotional intensity is typically assumed to be the distance of emotion from the circumplex origin (Cowie *et al.* 2000).

Prosodic and acoustic features of emotion

Prosody has many other functions than emotional signalling. Intonations and speech rhythm changes can indicate other common speech related functions. The stressing of words or syllables for emphasis or other conversationally relevant functions such as turn signalling in conversations are often used regardless of any emotional state. Murray *et al.* (1996) conclude that any emotional changes to prosody must be seen in addition to the underlying normal prosodic processes. Also, acoustic features do not specify emotional information only. Even simpler acoustic contours such as ZCR, HNR, or STE measures are very sensitive to different speech properties that are not related to emotion. More sophisticated spectral feature contours such as MFCC or PLP features capture also much, or all, of the normal speech processes.

Research to find emotionally relevant prosodic components and signal features have been traditionally conducted by performing perception tests or by manually inspecting the prosodic parameters of real speech recordings to guess what kind of changes are emotionally induced (Scherer 2003). Access to modern computational resources has also made systematic data mining approaches more effective with the capability to search through vast amounts of suspected or even randomly generated acoustic and prosody derived candidate features to find emotional correlates, e.g. by Oudeyer (2002), Batliner *et al.* (1999). A more fundamental approach is to model emotional speech itself and with synthesis techniques to produce candidate samples using different model parameters. The synthesised samples are then used in perception tests to identify which parameter changes convey emotional signals (Murray & Arnott 1995, Murray *et al.* 1996, Schröder *et al.* 2001, Schröder 2001). For a review of emotionally relevant features and extraction techniques, see Cowie & Cornelius (2003) or Ververidis & Kotropoulos (2006).

It is generally accepted that the most contributing factor for emotional speech is prosody through the fundamental frequency (F_0) and variations in the F_0 contour. Significant differences in F_0 contours have been observed, especially between basic emotions (Banse & Scherer 1996, Paeschke & Sendelmeier 2000, Toivanen 2001). Other important prosodic feature categories identified are the variations in energy (intensity), variations in duration, and variations in speech quality. Energy statistics are directly related to the perceived activation level of emotions (Banse & Scherer 1996, Ehrette *et al.* 2002). The timing and duration of the different parts of an utterance as well as the overall speech rate are all emotionally sensitive features (Murray *et al.* 1996, Thymé-Gobbel & Hutchins 1999, Farinas & Pellegrino 2001, Bosch 2003). It has been further noted that pause lengths are different between read and spontaneous speech. The speech rate has an effect on articulation, e.g. segment deletions, that can be detected using formant and spectral analysis (Kienast & Sendlmeier 2000). Using synthesis techniques, voice quality changes have been shown to have a significant supporting role in the signalling of emotion, particularly among milder affective states (Gobl & Ní Chasaide 2003, Grichkovtsova *et al.* 2012).

A summary of commonly associated emotion effects in relation to normal speech is shown in Table 3, adapted from Scherer (1986), Murray & Arnott (1993), and Nwe *et al.* (2003).

Table 3. Summary of common emotional effects.

	Anger	Fear	Joy	Sadness	Disgust	Surprise
Speech rate	$>$	$\gg$	$>$ or $<$	$<$	$\lll$	$>$
Pitch average	$\ggg$	$\ggg$	$\gg$	$<$	$\lll$	$>$
Pitch range	$\gg$	$\gg$	$\gg$	$<$	$>$	$>$
Pitch changes	abrupt	normal	smooth up	down	down terminal	high
Intensity	$>$	$=$	$>$	$<$	$<$	$>$
Voice quality	breathy	irregular	breathy	resonant	grumbled	breathy
Articulation	tense	precise	normal	slurring	normal	

Note: Symbols "$>$", "$\gg$" and "$\ggg$" represent increase and symbols "$<$", "$\ll$" and "$\lll$" decrease in a relevant category. A "$=$" symbol indicates no perceived change. The emotional effects represent changes found in the context of western languages.

Perception of emotional speech

The perception of emotion from actor-produced simulated emotions has been mostly studied using listening tests with randomly ordered recordings and answer choices limited to forced choice categories, typically a set of basic emotions. In this kind of discrimination study, a success rate of 50%–60% has commonly been achieved for the six basic emotions and neutral. The problem of guessing the correct answer, or using one category (e.g. neutral) as a default choice, however, is not easily compensated. Criticism has been made against the forced choice methodology (Russell 1994). When using a wider list of common emotions for the possible answers, or even not giving any fixed alternatives to choose from, a more accurate estimate for recognition might be possible. A large meta-analysis of cross-cultural emotion recognition by Elfenbein & Ambady (2002) could not, however, find any significant difference between methodologies due to a low amount of studies using alternative answering methods. The study did, on the other hand, find that emotion perception is universal. A dimensional annotation of emotion is also an option, e.g. Cowie *et al.* (2000). Common emotional labels can be argued to be more familiar to naive listeners; however, in Laukka *et al.* (2005), both professional and amateur listeners were shown to be able to perform equally well when annotating emotional expressions using typical dimensional scales (i.e. activation, valence, potency, and emotional intensity). Table 4 (Scherer 2003) shows the recognition rate for eleven western countries and one non-western country.

Table 4. Perception of simulated vocal expressions of emotion.

	Neutral	Anger	Fear	Joy	Sadness	Disgust	Surprise	Mean
Western	74%	77%	61%	57%	71%	31%		62%
Non-western	70%	64%	38%	28%	58%			52%

Note: The perception of surprise has not been commonly studied for vocal expressions and hence is missing in the table.

Using fewer emotional categories has the expected tendency of inflating recognition rates. Statistically, it is obvious that a smaller set of classes leads to a higher percentage of correct choices due to increasing chance of guessing the correct class. However, due to different semantic evaluation of available class choices (e.g. some choices such as neutral are selected more frequently when unsure of the correct choice), it can be difficult to compare recognition rates obtained using different sets of classes. For recognition, an average rate of 66% has also been reported by Scherer (2000) for the first four basic emotions and neutral. In a Spanish discrimination investigation for the first three basic emotions and neutral, a rate of 87% was recorded Montero *et al.* (1998).

2.4.2 Emotional speech databases

Emotional databases have evolved much during the last decade. Older studies on emotion recognition have relied almost exclusively on small and heterogeneous databases collected by the researchers for their own personal use, for a review, see Ververidis & Kotropoulos (2006). Typically, the contents and used annotations in the small databases have a large degree of variance making the comparison or fusion of databases hard. Douglas-Cowie *et al.* (2003) presented a guideline for future databases to address these issues and guide the database collection processes. The main considerations for a database development have been identified to be as follows: scope, naturalness, context, descriptors and accessibility. For a more thorough discussion about the identified considerations, see Douglas-Cowie *et al.* (2003).

The more modern databases have not only adopted the presented guidelines, but also incorporated many advances in emotional models, e.g. dimensional model annotations. Furthermore, the modern databases have logically included the multi-modal aspects of emotion. Thus, the more modern larger databases, e.g. the *HUMAINE Database* (Douglas-Cowie *et al.* 2007), or *MAHNOB-HCI* (Soleymani *et al.* 2012), have begun including not only speech, but also a full range of multi-modal signals such as audio-

visual recordings from audio-visual stimuli, gestures, and physiological biosignals (i.e. ECG, skin conductance, respiration, eye tracking, etc.). The database used in this thesis is of the older variety, limiting its usability in the future, due to the fact that Finnish emotional speech data has not yet been extensively collected using the more modern collection guidelines and multimodal scope. Nevertheless, the *MediaTeam Emotional Speech Corpus* (Seppänen *et al.* 2003) is still currently the largest corpus of emotional Finnish speech.

2.5 Emotion recognition from speech

Emotion recognition from speech using pattern recognition techniques is a relatively recent approach that has become possible with the increase in computational resources. Traditional label based classification solutions have been developed using multiple state-of-the-art classifier techniques. Reasonably good results around 50%–80% correct classification for 4–6 basic emotions (e.g. neutral, sad, angry, happy, fear, disgust) have been attained (Dellaert *et al.* 1996, McGilloway *et al.* 2000, Oudeyer 2002, Bosch 2003). Label based classification attempts are, however, prone to semantic confusion in the truth data and other problems in data collection (e.g. methodological issues in obtaining some types of emotion) (Douglas-Cowie *et al.* 2003) limiting their performance. More robust approaches using dimensional models of emotion, e.g. Zeng *et al.* (2005), Nicolaou *et al.* (2011), are very recent. For a state-of-the-art review of emotion recognition and classification techniques from traditional approaches to recent advances, see Cowie *et al.* (2001), Ververidis & Kotropoulos (2006), Zeng *et al.* (2009), and/or Calvo & D'Mello (2010).

2.5.1 Machine learning and pattern recognition

Machine learning and pattern recognition are not standard procedures although an overview of the required steps in the process can be presented. Rather, machine learning and pattern recognition offer tools that can be used to solve problems by utilising knowledge (Duda *et al.* 2001). No Free Lunch (NFL) theories (Wolpert & Macready 1997) have been proven, indicating that no single optimisation or search strategy can be universally better than another, with certain exceptions for co-evolutionary self-game scenarios (Wolpert & Macready 2005). A generic classifier, feature transformation, or supervised learning scheme therefore cannot be expected *a priori* to work better

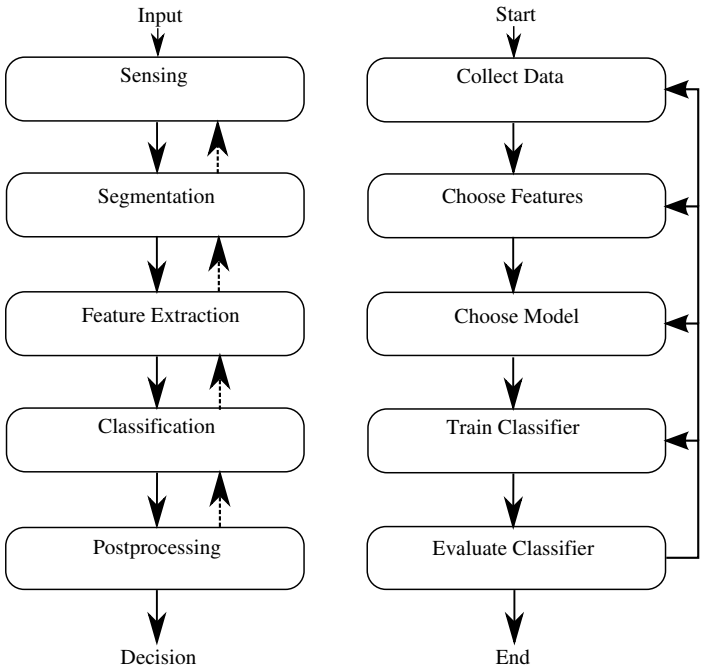


Fig. 8. General overview of a pattern recognition system and design process. Adapted from Duda *et al.* (2001).

than the alternatives on an arbitrary problem. The key to well performing solutions is the understanding of the problem at hand and identifying methods that align well with the problem characteristics. It is, however, useful to look for working solutions for similar problems. It is also important to note that, since many machine learning and pattern recognition approaches are closely related, typically, no one approach is significantly superior to another in performance. In Fig. 8, the general structure of a pattern recognition system is illustrated as a series of steps from input to decision (with optional feedback to previous steps indicated by the dashed arrows). The design process follows a similar structure from start to end, involving various choices and tasks in each step (with the possibility of going back to previous steps if necessary, indicated by the backwards arrows).

Feature selection and transformations

Feature selection and transformation techniques are used to find, or remodel, features that are relevant to the pattern recognition problem. In selection techniques, to improve model performance, the relevant and efficient features are kept while the irrelevant and noisy features are excluded. Feature selection is a typical way to implement supervised learning. Optimal feature selection can be performed by brute force, i.e. an exhaustive search of all alternatives. Methods using cost functions that guarantee global optimum, e.g. Branch and Bound style methods, can also be implemented to produce optimal solutions at the expense of computational complexity (Duda *et al.* 2001). For practical applications, however, optimal solutions are often computationally too inefficient and suboptimal search strategies, i.e. heuristics, are used instead. Many methods have been attempted in emotion recognition research. Simple sequential selection algorithms, e.g. Promising First Selection (PFS) or Sequential Forward Search (SFS) (Dellaert *et al.* 1996, Lee *et al.* 2001), have been found effective. Very similar Sequential Backward Search (SBS) (Oudeyer 2002) has also been used successfully. A more generalized Sequential Forward Floating Search (SFFS) (Pudil *et al.* 1994) is therefore a good candidate. Evolutionary computing motivated Genetic Algorithms (GA) (McGilloway *et al.* 2000) can be used, but they are not ideal for decision type objective functions, nor when using a low amount of features, i.e. around 50–200, typical in emotional speech analysis.

In transformations, the existing features are transformed into a new set of features. The transformed set of features typically satisfies some desired property. For example, the new transformed feature set is an orthogonal set of features, which is the case in Principal Component Analysis (PCA). The target dimensionality of the transformations is almost always a lower dimension than the starting feature space to counteract the curse of dimensionality. Although transformations are usually defined as unsupervised methods, they can also implement supervised learning, e.g. Linear Discriminant Analysis (LDA). Nonlinear transformations are often used in supervised learning. Using the kernel trick (Aizerman *et al.* 1964), an extension of PCA to nonlinear spaces, i.e. Kernel PCA (KPCA), is straightforward. KPCA can be seen as a general framework for nonlinear transformations and thus many nonlinear techniques can be reduced to special cases of KPCA (Maaten *et al.* 2009). Neural Networks (NN) (Haykin 1994) can also be used to learn nonlinear transformations, e.g. using a General Regression Neural Network (GRNN) method. Manifold modelling techniques can also be seen as a type of

feature transformations. Manifold modelling techniques are discussed in more detail in Section 2.5.1.

Both selection and transformation methods can be used at the same time. It is important to note that neither approach gives guaranteed increases in performance as implied by the NFL theories. Transformations and selection are, however, great tools for implementing desired models in feature extraction to simplify subsequent classifier and decision methods.

Manifold modelling

Manifold modelling techniques are a group of specialized data transformation methods. The taxonomy of manifold modelling methods is typically divided into linear and nonlinear methods, convex and non-convex optimisation techniques, or full and sparse spectral groups. Many of these methods share considerable similarities, however, even when the methods can be taxonomically placed in seemingly different branches. For a throughout review of methods and general taxonomy, see Maaten *et al.* (2009).

Classic data transformation methods, e.g. linear PCA, can be included in the taxonomy of manifold modelling techniques, even though the methods are not specific to any spatial structures implied by the assumption of manifold data. The more advanced convex manifold modelling techniques, however, are based on the use of some spatial data properties (e.g. neighbourhood information) and a suitable distance measure to model data structures as manifolds. The metric choice for the distance measurement is typically some kind of formulation of path information, e.g. geodesic distances, to effectively model nonlinear manifolds by mapping the path distances to the desired, typically lower, dimension. An effective nonlinear manifold modelling technique that uses the geodesic distance mapping of a local neighbourhood linked data for a global solution is Isomap (Tenenbaum *et al.* 2000). Another successful method that aims to retain the local neighbourhood structure of data manifolds for a sparse generalized solution is Local Linear Embedding (LLE) (Roweis & Saul 2000). The base methods of manifold modelling techniques are typically unsupervised but can be modified for supervised learning.

Nonlinear manifold modelling is a strong tool and a clear advance in learning and representation of complex convex data structures. The nonlinear manifold modelling techniques still have some problematic properties. Many assumptions made in the methods can lead to serious problems if violated. Many methods, but especially the local

linearity based sparse methods, e.g. LLE, suffer badly from the curse of dimensionality, local linearity violations, gaps in the data or outliers, and numerical problems (Maaten *et al.* 2009). The general problems increase the requirements for the data, both in quality and quantity. The neighbourhood linking that is used to produce data structures as a graph tree exposes the global methods, in particular, e.g. Isomap, to short-circuiting errors (Balasubramanian & Schwartz 2002). A short-circuit error usually leads to a major deformation of the manifold structures resulting in a poor performing and ultimately incorrect model. The most important shortcoming for most nonlinear manifold learning methods is, however, the lack of generalization property (i.e. the mapping between the feature and manifold spaces is lost during the process). Consequently, new data cannot be readily mapped to the generated manifold model, and the embedding must be recalculated to add new data to the model. The need to recalculate the embeddings for supervised training and the inclusion of many additional parameter optimisations greatly increases the overall computational complexity of nonlinear manifold learning. The computational complexity is thus a seriously limiting factor for the practical implementation of the nonlinear manifold learning methods for large databases.

Methods have been developed to, at least partially, overcome problems such as the short-circuiting and the lack of generalization property. The short-circuiting can be addressed by finding major short-circuits and removing them from the graph, e.g. by identifying unusually large flow nodes (Choi & Choi 2007). Another solution is to use supervised learning by weighting known structures differently in the data (e.g. class structures) to lessen the chance of short-circuits (Vlachos *et al.* 2002). The lack of generalization is currently a more difficult problem to solve. The mapping of new data to the generated manifold can be constructed again by using another nonlinear method, e.g. GRNN (Geng *et al.* 2005). Mapping can also be established using the kernel trick (Zhang *et al.* 2010a). However, the use of the kernel trick in supervised learning is still somewhat limited because correct class weights are needed in the projection and they cannot be specified without first knowing the class information of the new unknown samples. A method using average weight for an unknown class has been suggested by Gu & Xu (2007).

Emotion recognition has been attempted successfully using nonlinear manifold modelling with prosodic features. The approaches (You *et al.* 2006, Zhang *et al.* 2010a,b) have been more aligned with a traditional concept of feature reduction during the feature extraction process. The amount of dimensionality reduction used, although considerable, has not been aimed at a sufficiently low dimensionality to be very suitable

for visualisation purposes. Reduction attempts at a sufficiently low dimensionality for visualisation have also been made, but the studies, e.g. by Jain & Saul (2004) and Kim *et al.* (2010), have been restricted to short vowel or word length samples and use spectral features.

Classifiers

A classifier is a key element of machine learning. A multitude of classifiers has been developed. Linear classifiers, e.g. Naive Bayesian (NB), Linear Discriminant Classifier (LDC), or Perceptron (Duda *et al.* 2001), are typical examples of classic statistical classification methods. Further advances in statistical classification have made nonlinear classification methods available. Typical methods capable of nonlinear classification are k-Nearest Neighbour (kNN) and various advanced Neural Network (NN) methods. The linear classifiers can also be extended to nonlinear classification, e.g. using the kernel trick (Aizerman *et al.* 1964). Using the kernel trick, a linear Support Vector Machine (SVM) (Vapnik & Lerner 1963) is generalized to nonlinear classification (Boser *et al.* 1992) making SVM a very effective framework for classification.

Many classifiers have been found effective in emotion recognition from speech, typically using mostly prosodic features and a label based class approach (typically 4–6 basic emotions). LDC approaches have been found effective with feature selection (McGilloway *et al.* 2000, Lee *et al.* 2001). Classification And Regression Trees (CART) performed well for the classification of frustration in Ang *et al.* (2002). Oudeyer (2002) found the CART approach also efficient when combined with a meta-optimisation method. Simple NB classifiers exhibited poor performance, but when combined with a good feature selection method, good performance was reached (Dellaert *et al.* 1996, Oudeyer 2002). Various kNN classifiers were all found performing well when a suitable feature selection method was utilised (Dellaert *et al.* 1996, Lee *et al.* 2001, Yu *et al.* 2001, Oudeyer 2002). The NN type of classifiers has not been found effective. In McGilloway *et al.* (2000), a generative vector quantisation approach led to over learning problems, which resulted in a poor performance. In Oudeyer (2002), radial basis function neural networks and voting perceptron type of classifier setups did not perform well, either. Linear SVMs were found performing poorly, but when using polynomial kernels, a good performance was observed (McGilloway *et al.* 2000, Oudeyer 2002). In Nwe *et al.* (2003), a Hidden Markov Model (HMM) based solution was found very effective.

It can be seen from the literature that emotion recognition from speech has the

typical dualistic solution property common in classification tasks. A simple classifier can be effective when the features are selected and the input nonlinearities are handled first with a more robust method (e.g. manifold modelling or nonlinear transformation). Another solution is to use a classifier that is particularly well suited for the problem. For example, one can use an SVM that has the feature space problem solved by using the correct kernel or use a solution that is intrinsically well aligned with the information structure of the input features, e.g. an HMM that robustly incorporates the segmental nature of speech.

Validation and bias in supervised learning

One of the cornerstones of machine learning is the ability to evaluate or validate the results of a supervised learning method. Objective estimates for classifier performances are essential for any attempt to generalize the learning results for new data. In supervised learning, there is an inherent risk in over learning the training and testing data. A method for validation is thus necessary for an objective assessment of any such results. The problem of a separate validation data, e.g. a typical choice of 30% data that is left for validation, is that now a significant portion of data cannot be used for training. Almost always it would be advantageous to have more data. Another problem is that the validation set is still quite small as there is the need to have as large as possible training and testing sets as well, due to the curse of dimensionality, for instance. The use of a small validation set leads to a larger variance of performance estimates. (Duda *et al.* 2001)

Cross-Validation (CV) (Stone 1974) techniques attempt to mitigate the problems of using separate validation data. In CV, a fraction of data is held out in turn as a validation set (e.g. 10% for 10-fold CV), while the other part is used for training and testing. Then, an estimate of generalized performance is formed using all folds of the validation procedure (e.g. mean performance of the CV folds). The performance estimate gained from CV is an almost unbiased estimate of the true generalization error. An extreme case of cross-validation is the Leave-One-Out Cross-Validation (LOOCV). In LOOCV, each sample is rotated in turn out of the training and testing sets and used for performance estimation. LOOCV offers the maximal use of data for training and still retains generalization. The downside of CV methods is the increased computational cost.

The use of CV has become very common in pattern recognition, especially in model

selection, feature selection, and hyper-parameter optimisation. For example, in emotion recognition from speech, Oudeyer (2002) used AdaBoost meta-optimisation to produce the best performance. There is, however, an insidious problem when using CV in incremental wrapper methods, sometimes called metaheuristics. The use of CV in wrapper methods violates the assumption of statistical independence of the validation because the results of validation are used to tune the parameters (i.e. the validation data "bleeds" information to the feature vector or hyper-parameters). The results are prone to over-learning because the CV error is no longer unbiased for finite data due to the estimation variance (Cawley & Talbot 2010). Other possible meta-optimisation algorithms, e.g. bootstrapping or jackknifing (Efron & Tibshirani 1994), are also equally effected. A solution is to use a separate validation data for the generalization testing, but unfortunately, this solution leads to the obvious problems of using separate validation data discussed earlier. Another solution is to stop the training early (Qi *et al.* 2004) to avoid complex models (i.e. a simple realisation of Occam's razor). A nested CV can be used to generate effectively unbiased performance estimates for wrapper methods (Varma & Simon 2006). In a nested CV, each trained CV wrapper method folds (i.e. inner CVs) are tested using another outer CV loop. The outer CV layer holds out a fraction of data that is unused in the inner CV layers. A nested CV, however, can lead to different optimisation results within each inner CV fold making the learning results ambiguous.

Fusion techniques

The fundamental idea behind fusion techniques (Kittler 2000) is the integration of information from multi-modal sources. Typically, the modalities are highly differentiated in feature and analysis techniques (e.g. facial expressions and speech signal). Fusion can be achieved at multiple stages in pattern recognition systems. In feature fusion, an attempt to add multi-modal information at the feature level is made by including features, or derived features (i.e. transformed features using both modal feature sources), from two or more modal sources. In the simplest form, the feature sets are just united into a new larger set of features and the problem of fusing the information is left for the classifier to learn. Another way of performing fusion is to fuse the information at a later stage (e.g. after classifier decisions) by treating the outputs of multiple single modal experts as features or inputs to the fusion. Commonly, this late stage fusion is referred to as score or decision level fusion. In the most straightforward case where the classifier

outputs can be defined as probability distributions, the fusion is a trivial product of the expert outputs. However, due to the presence of noise and the fact that classifiers are invariably limited by finite data, the use of multiplication in fusion is problematic (e.g. multiplication by zero from one bad expert nullifies the outputs of other experts). It is found that a sum fusion of outputs has more robust properties in many real application cases (Kittler *et al.* 1998). Vote fusion between expert outputs has also been investigated but found inferior to sum fusion in real application cases (Kittler & Alkoot 2003). In emotion recognition from speech, Dellaert *et al.* (1996) used a voting technique similar to fusion to produce an ensemble of majority voting specialists that produced the best performance reported.

3 Research contributions

The research contributions of the original Publications I–IV are presented here. In Paper I, a state-of-the-art emotion recognition method is presented with results for Finnish acted emotional speech. The classifier and feature selection methods developed in Paper I are also used in Papers III–IV. In Paper II, the performance testing of the prior developed PDA algorithm is described. In Paper III, short vowel length segments are studied in order to develop robust multimodal fusion enhanced classification of emotion for short word length inputs. Paper IV, in its turn, proposes a framework for the visualisation of semantic speech data. In the paper, classifier optimised non-linear manifold modelling techniques are used to create a visualisation of the emotional content of the speech database used in this research.

In papers I and IV, the passage length emotional monologs of the MediaTeam Emotional Speech corpus are used. In Paper II, separate, classified, very challenging and noisy radio communication data between Finnish F-18 Hornet fighter pilots recorded during a war exercise is used. In Paper III, the second larger multiple rendition extension to the *MediaTeam Emotional Speech Corpus* is used.

3.1 Emotion recognition for Finnish speech

As outlined in Sections 2.3 and 2.4, the expression of continuous emotional states in speech is conveyed primarily via prosody. Emotion recognition from speech can be fundamentally seen as a pattern recognition problem (Section 2.5). In Paper I, a state-of-the-art emotion recognition technology is presented and the performance of fully automatic feature extraction techniques for prosody analysis is demonstrated for Finnish emotional speech samples. In Paper II, the performance of the used feature extraction technique is further tested in a very demanding physical and psychical stress environment. The relevance of developed prosodic features in a challenging environment is important for future application implementations.

3.1.1 Emotion recognition using global prosodic features

Previously, the emotional properties of the long-term prosody of continuous spoken Finnish have been very little studied. Research on the vocal parameters of emotional Finnish speech has often been focused on very short speech units, e.g. syllables and vowels. The authors are not aware of any prior automatic classification studies of continuous Finnish emotional speech using long-term global prosody. The primary technological goal in Paper I was first to establish an automatic recognition of emotion using the prosodic features of speech. The automatic classification of continuous emotional Finnish speech was then attempted and the results compared with a human listening test reference.

The data set used in Paper I is the passage length emotional monologues of the MediaTeam Emotional Speech corpus (Seppänen *et al.* 2003) in core basic emotions, i.e. four emotional categories, including neutral. To generate the data, 14 Finnish speaking actors (eight male, six female, aged between 26 and 50) were recruited. Around one minute length renditions of emotional speech in each emotional class (neutral, sad, happy, and angry) were produced by each actor by reading an otherwise emotionally neutral text passage, for a total of one hour of speech material. The speech samples were recorded in an anechoic chamber with a high quality condenser microphone and a digital audio tape (DAT) recorder using a lossless 44kHz 16bit audio format.

Each of the resulting passages was then divided into five segments to produce 70 samples of around ten seconds in length in each emotional category for a total of 280 samples. The division of raw data was required to produce enough samples for classification. The sample length of around ten seconds was chosen to guarantee at least one full utterance and the pauses between utterances in each sample (there are eleven utterances in the full text). A shorter sample length would have been more desirable for classification purposes, i.e. the more samples, the better. However, as the optimal sample length for global prosody extraction is not known, a larger sample was preferred to avoid oversensitivity to local phenomena. In order to retain a fully automatic capability, the segmentation was made using a simple division of the recordings. The samples were labelled according to speaker identity, gender, and emotional categories.

The F0Tool feature extraction software, developed by the author in Väyrynen (2005), was then used to calculate the long-term global prosodic features and related acoustic features used in the study. Features used in Paper I are all automatically calculable from the segmented raw speech recordings, for an overview of the calculated features see

Table 5. The definitions for each feature are presented in detail in Väyrynen (2005). The same extracted global features were later on also used in Paper IV. In Paper II and III, the feature contours are derived using the same PDA architecture (see, Väyrynen (2005) for a detailed description), but raw feature contours (Paper II) or short time statistical parameters of the contours (Paper III) are used instead of the global features.

The pattern recognition approach used for emotion classification in Paper I consists of a CV-based metaheuristic search algorithm that utilises nested kNN classifiers and an SFFS feature selection method (see Fig. 9). The approach was selected to retain maximal use of the still rather small amount of data available, i.e. all data can be used for training and testing. A person independent partitioning of the data was adopted for the CV loop. Each person's set of samples, i.e. 20 samples at a time, was rotated in and out of the training set, i.e. 260 samples in each fold, in turn to produce fourteen fold CV performance estimates for the feature selection algorithm. The problem of over fitting is not easily compensated for when using meta-search algorithms (see, Section 2.5). In this approach, the structure of floating search does not allow the feature vector to change too heavily during the search, however, shielding the process from finding totally random combinations of features. Over-learning is primarily handled by limiting the search dimension to a low number of initial features and the CV performance testing. The kNN

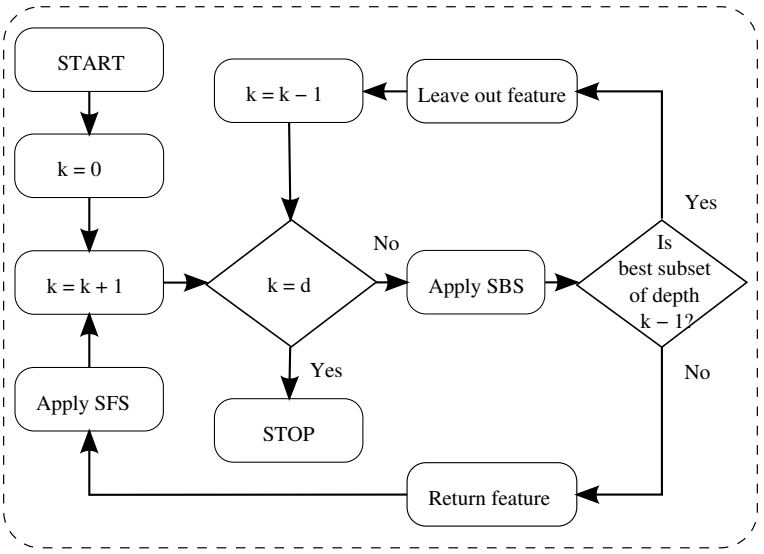


Fig. 9. Sequentual forward floating search flowchart.

Table 5. Global prosodic features and other related acoustic features.

Feature name	Short description of feature
Mean	Mean F_0 frequency (Hz)
Median	Median F_0 frequency (Hz)
Max	99% value of F_0 frequency (Hz)
Min	1% value of F_0 frequency (Hz)
fracMax	95% value of F_0 frequency (Hz)
fracMin	5% value of F_0 frequency (Hz)
fracRange	5–95% F_0 frequency range (Hz)
Range	1–99% F_0 frequency range (Hz)
F0Var	F_0 variance
Shimmer	Mean proportional random intensity perturbation
Jitter	Trend corrected mean proportional random F_0 perturbation
LFE1000	Proportion of Low Frequency Energy under 1000Hz
LFE500	Proportion of Low Frequency Energy under 500Hz
GDposav	Average F_0 rise during cont. voiced segment (Hz)
GDnegav	Average F_0 fall during cont. voiced segment (Hz)
GDriseav	Average F_0 rise steepness (Hz/cycle)
GDfallav	Average F_0 fall steepness (Hz/cycle)
GDrisemax	Max rise of F_0 during continuous voiced segment (Hz)
GDfallmin	Max fall of F_0 during continuous voiced segment (Hz)
GDmax	Max steepness of F_0 rise (Hz/cycle)
GDmin	Max steepness of F_0 fall (Hz/cycle)
MeanInt	Mean RMS intensity (abs., dB)
MedianInt	Median RMS intensity (abs., dB)
MaxInt	Max RMS intensity (abs., dB)
MinInt	Min RMS intensity (abs., dB)
IntRange	Intensity range (abs., dB)
fracMaxInt	95% value of intensity (abs., dB)
fracMinInt	5% value of intensity (abs., dB)
fracIntRange	5–95% intensity range (abs., dB)
IntVar	Intensity variance (abs., dB)
mavlngth	Average length of voiced runs
manlngth	Average length of unvoiced segments shorter than 300 ms
maslngth	Average length of silence segments shorter than 250 ms
mlnlngth	Average length of unvoiced segments longer than 300 ms
mlslngth	Average length of silence segments longer than 250 ms
max_vlngth	Max length of voiced segments
max_nlngth	Max length of unvoiced segments
max_slngth	Max length of silence segments
perc_short	Percentage of pauses shorter than 50 ms
perc_mid	Percentage of 50–250 ms pauses
perc_long	Percentage of 250–700 ms pauses
spratio	Ratio of speech against unvoiced segments >300ms
vratio	Ratio of voicing against unvoiced segments
sratio	Ratio of silence against speech
norm_intvar	Normalised segment intensity distribution width
norm_freqvar	Normalised segment frequency distribution width

Table 6. Results of global prosody based emotion recognition for Finnish.

	Scenario 1	Scenario 2
Instruction mode	85.7%	71.4%
Human vote mode	81.8%	71.1%

classifier is also resistant to inflated performance readings when larger k values are used.

Experiments were made using two scenarios. In Scenario 1, a person dependent approach was used. In Scenario 2, a more general setting of an unknown speaker was adopted. In both cases, also the selection of the ground truth data was evaluated. First, the original emotions requested from the speech actors were used as the labels (i.e. instruction mode). Next, labels generated via a majority vote from the answers provided by the human listening tests were used (i.e. human vote mode). Both used labellings produced very similar results. This could be expected as both labellings were largely congruent. Since the instruction mode yielded slightly better results, the human voting based labels were not used in the subsequent papers. The emotion recognition performance for basic emotions achieved using long-term global prosodic features (see Table 6) approached the used human reference of 76.9%.

Feature vectors corresponding to the best achieved results are presented in Table 7. In the speaker dependent Scenario 1, long-term spectral features (i.e. LFE500 and

Table 7. Selected feature vectors for global emotion classification.

Instruction mode		Human vote mode	
Scenario 1	Scenario 2	Scenario 1	Scenario 2
Mean	fracRange	Mean	fracMax
Median	Shimmer	Median	fracRange
fracMax	Jitter	fracMin	F0Var
fracMin	GDrisemax	fracRange	Shimmer
LFE1000	IntRange	Shimmer	Jitter
LFE500	IntVar	LFE1000	GDfallav
GDnegav	sratio	LFE500	MaxInt
GDfallav	norm_intvar	GDposav	MinInt
GDrisemax		GDfallav	IntRange
MedianInt		GDfallmin	fracMaxInt
MinInt		GDmax	fracIntRange
vratio		MeanInt	IntVar
sratio		MinInt	vratio
		spratio	
		norm_intvar	

LFE1000) and mean values for important prosodic parameters (e.g. Mean, Median) were among the discriminative features. In the speaker independent Scenario 2, the selected features are typically those that capture the dynamic variation of prosody (e.g. fracRange, IntRange, and Jitter). The difference in the selected features between the scenarios can be explained by the ability of the spectral and mean values of prosodic contours to identify individual speakers. The classifiers in Scenario 1 have learned the individual representations of emotions. In Scenario 2, the classifiers are more aligned to identify the speaker independent patterns of emotions.

3.1.2 Performance testing of prosodic feature extraction

In Paper II, the performance of the feature extraction method was tested using speech data recorded in a very challenging environment. The data was obtained by recording actual F-18 Hornet fighter pilot communications inside the aircraft cabin during an in-flight war exercise. The high noise environment is mainly composed of broad band noises produced by the jet engines, the aerodynamic flow, and cockpit equipment (e.g. pressurisation). The recordings were made using a microphone inside the pilots' oxygen mask with a DAT recorder attached to the pilots' combat gear. A lossless 44.1kHz 16 bit audio format was utilised. A set of 40 communication utterances (ten random utterances from each of the four pilots) was then extracted from the total material and used for the performance evaluation.

The performance measurement was based on a semi-automatic waveform matching procedure (a tool programmed in Java, see Fig. 10) where a human operator constructed the correct reference F0-contours by manually identifying each F0 cycle. A least squared error minimisation with quadratic interpolation was then applied to fine tune the identified peaks and to extract the reference contour values. The resulting reference F0 values were then inspected again by the operator and corrected if necessary. A total of 6,501 cycles were individually identified and tagged. Each reference contour was then compared with a corresponding contour provided by the fully automatic extraction algorithm (F0Tool) to produce performance measurements.

The performance measurements (see Table 8) revealed that the extracted basic prosodic parameters are quite robust to noise environment. The errors were found to be unbiased. The total average relative error (AvRE) of F0 estimation was at 1.0% level. The total SD of errors was also found to be very acceptable at 3.4%. The errors were mainly explained by fine pitch determination errors. Only 0.56% of the errors (0.22%

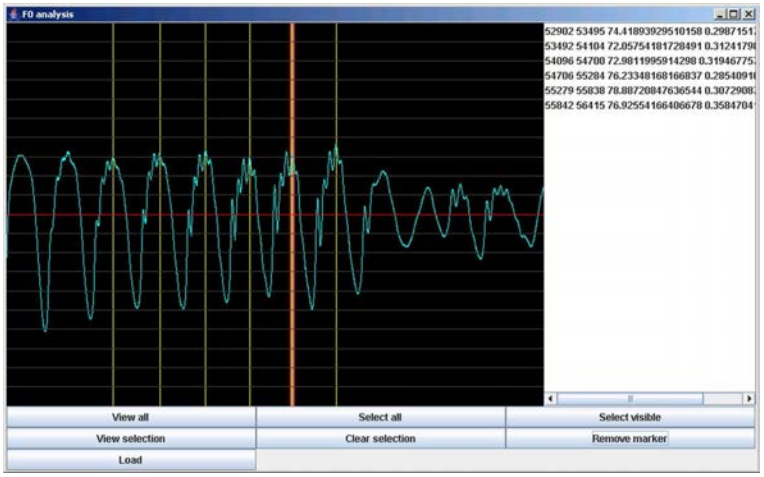


Fig. 10. UI of the reference contour generation application.

high and 0.34% low) were defined as gross errors (i.e. an error of more than 20%). The AvRE of fine errors was 0.8% with an SD of 2.2%.

The performance level does not adversely impact the final long-term (i.e. multiple second) prosodic features calculated using multiple F0 measurements, however, because the mean error becomes practically zero very soon. Features relying on a few measures would typically incur an error of no more than a few percentages. The Perturbation contour based feature estimates (e.g. Shimmer and Jitter), on the other hand, are more affected by the noise environment rendering them unreliable for high noise recordings.

Table 8. Results of performance measurement.

Fine errors		Total errors		Gross error rate		
AvRE	SD	AvRE	SD	high	low	total
0.8%	2.2%	1.0%	3.4%	0.22%	0.34%	0.56%

Note: Fine errors include F0 measures deviating < 1ms from the reference contour. Total errors include all measurements. Gross error rate is the percentage of measures with an error exceeding $\pm 20\%$ from the reference (+ for high, - for low errors).

3.2 Fusion technique for multi-modal classification of emotion

The aim in Paper III was to investigate the influence of emotion on very short stressed words. Short stressed words are routinely used in voice commands and when acknowledging messages with voice. In the paper, a fusion classifier based emotion recognition of short speech segments using F0Tool (Väyrynen 2005) local prosodic features and Aparat (Airas 2008) vocal source features, describing mostly voice quality, is proposed.

A set of 450 samples of the MediaTeam Emotional Speech Corpus data was used in the study. The data was produced in an anechoic room by nine Finnish speaking actors (five men and four women aged between 26 and 45) in five emotional categories, including neutral. Ten renditions of speech produced by reading a passage length text was recorded using a random sequence where no single category (neutral, sad, happy, angry, and tender) was repeated immediately in the next rendition. The text was constructed in such a way that in the middle of the passage a specifically constructed target utterance "Taakkahan se vain on" 'It's only a burden' was placed to create a strong and consistently emphasised position on the first syllable of the word "Taakkahan" 'Only a burden'. The passage frame is intended to guarantee that emotional speech is consistently portrayed when the target utterance is reached by the naïve speakers. The first long [a:] vowel (about 100–250ms in length) situated in a primary stressed position was then extracted from the target utterance. An open long [a:] vowel was needed in order to robustly extract the vocal source features. The extracted emotional vowel sample is used as a proxy of a short independent discoursal unit such as an acknowledgement or a short command.

Local prosodic features and vocal source features were extracted from the vowel samples using F0Tool and Aparat software, respectively. A sum fusion based leveraging method using CV meta learning was developed (see Fig. 11) in order to merge the two feature sets at decision level. The fusion method was implemented using groups of four odd k (3, 5, 7, and 9) valued kNN classifiers in the leveraging process in order to marginalise over the parameter k. A base classifier group using local prosodic features is first trained using SFFS feature search. The base classifier group is then used with sum fusion to search with SFFS for a vocal source feature based classifier group that maximises the performance of the fusion classifier. During the sum fusion, weights were given to the prosodic feature and vocal source feature classifier groups according to their relative separate performances. The sum fusion (Eq. 3) is calculated using estimates of

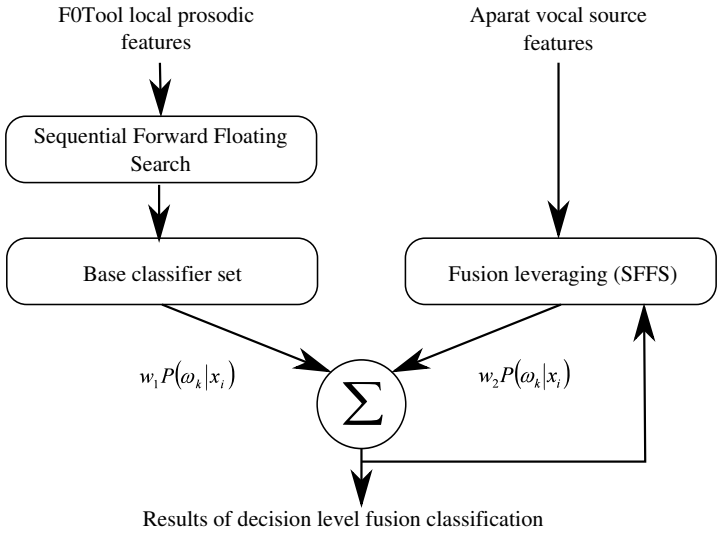


Fig. 11. Leveraging of fusion classifiers using meta learning (III, published by permission of Elsevier B.V.).

the a-posteriori probabilities of the kNN classifier groups (Eq. 4).

$$\begin{aligned} &assign \qquad Z \rightarrow c_j \qquad if \\ &(1-R)P(c_j) + \sum_{i=1}^{R} w_i P(c_j|x_i) \\ &= \max_{k=1}^{m} \left[(1-R)P(c_k) + \sum_{i=1}^{R} w_i P(c_k|x_i) \right], \end{aligned} \tag{3}$$

where Z is the pattern to be assigned to class c_j among m number of classes $(c_1, \ldots, c_m)$ using R classifiers that represent patterns given by sample vectors x_i, $i = 1, \ldots, R$. The sum of weights is normalised $\sum_{i=1}^{m} w_i = 1$. The a-posteriori probabilities were defined as:

$$P(c_i|x) = \frac{p(x|c_i)P(c_i)}{\sum_{j=1}^{m} p(x|c_j)P(c_j)} \widehat{=} \frac{k_i}{k}, \tag{4}$$

where c_i denotes the i:th class of the m total number of classes, x is a sample vector, and k_i is the number of k-neighbourhood prototypes with class label c_i.

The results of short segment emotion recognition (see Table 9 for average classification performance results) reflects the knowledge that voice quality features are

Table 9. Results of fusion based emotion recognition of short speech segments.

Human	Aparat	F0Tool	Fusion
42.4%	32.9%	40.9%	45.3%

Note: Average classification results are provided for neutral, sad, happy, angry, and tender emotions.

effective in describing emotions where the speech quality is more pressed (e.g. anger). The observed low performance of 32.9% correct, on average, when using vocal source features only (Aparat) can therefore be explained largely to be due to the incapability of classifying whole emotional categories. The vocal source features were effective when classifying neutral, happy, and angry samples. Intonation level features (F0Tool) are clearly emotionally relevant, also in the case of stressed short segments of speech, indicated by the robust recognition performance (40.9%). The proposed fusion method was capable of including the information in both the vocal source features and the local prosodic features to reach the best performance of 45.3% correct, on average, even surpassing the performance of the human reference (42.4%). Feature level fusion was also investigated. However, the decision level fusion based approach invariably resulted in a better performing classifier structure that combined both intonation describing prosodic features, as well as voice quality describing source features.

The best performing feature vectors used in the fusion study are collected into Table 10. The best identified vocal source features for emotion recognition were the features sensitive to the pressedness of speech (i.e. features derived from ClQ, AQ, and NAQ). Local prosodic features effective in emotion recognition were, expectedly, the mean values for pitch and amplitude together with features describing the amplitude dynamics of the stressed short segments. The fusion leveraging did not result in a noticeably different set of vocal source features.

The results of the fusion classification suggest that local prosodic features and vocal source features could be used effectively to evaluate the emotional content of short discoursal units such as command words and acknowledgements. Unfortunately, as the used Aparat vocal source features are currently limited by the IAIF technique, i.e. a good clean quality open vowel is needed for the analysis, the results are not immediately usable for practical applications. The main intent of the study is, however, to demonstrate the expected utility when combining the feature sets using a fusion classification architecture. When a robust feature extraction method becomes available, the proposed fusion architecture can be immediately used to incorporate the new features for the classification of emotion. The leveraging process further enables a parallel

Table 10. Selected feature vectors for short segment fusion emotion classification (III, published by permission of Elsevier B.V.).

Aparat vocal source features only	F0Tool local prosodic features only	Aparat vocal source features in the fusion
Mean of OQ2	Mean of F0 frequency	Mean of OQ2
Mean of ClQ	Mean of cycle peak amplitude	Mean of ClQ
Median of ClQ	Median deviation from the median of cycle peak amplitude	Median of NAQ
Median of SQ2	Median deviation from the median of Shimmer	Median of SQ2
Median deviation from the median of OQ1		Median deviation from the median of OQ1
Median deviation from the median of NAQ		Median deviation from the median of NAQ
Median deviation from the median of AQ		Median deviation from the median of AQ
Median deviation from the median of QOQ		Median deviation from the median of OQa
		Median deviation from the median of QOQ
		Median deviation from the median of SQ1

approach to the final trained fusion classifier structure that allows the use of voice quality parameters if they are available by a simple sum operation (Fig. 12). The extraction of features can be handled in separate processes and if one expert's feature extraction fails, the missing expert can be easily handled in the sum fusion by simply zeroing the output of the expert.

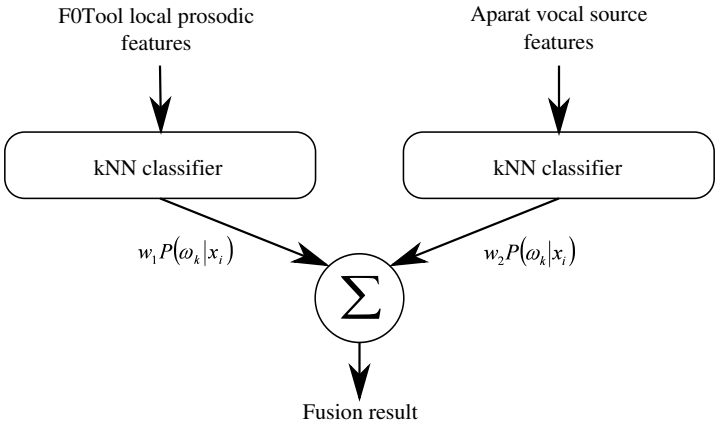


Fig. 12. Weighted sum fusion of experts (III, published by permission of Elsevier B.V.).

3.3 Visualisation of emotional speech data

Visualisation of different semantic qualities in speech recordings is a highly sought capability for speech researchers. In Paper IV, the visualisation of speech data is studied. As semantic concepts (e.g. emotion) in speech are typically highly nonlinear phenomena, it is very hard to gain a deep insight from data via simple visualisations such as plots of different raw feature values. A machine learning based data mining approach is required in order to access the nonlinear statistical patterns that are implicitly sampled in the recorded speech materials. Traditional label based classifications and statistical modelling of class properties are often incapable of learning complex patterns and are prone to errors in the labelling of data. The correct labelling of semantic content cannot be guaranteed to be flawless, and consequently, the annotated labellings often contain a large amount of error.

In order to compensate the labelling errors, it can be helpful to relax the rigid labelling. A continuous dimensional modelling of a semantic aspect, e.g. annotating with Feeltrace (Cowie *et al.* 2000), or another continuous measure known to be directly related to the semantic aspect in question, can be used to provide a more flexible annotation of data. A regression analysis can then be used with a continuous annotation to produce clear measures and visualisations for semantic properties, e.g. for language fluency, by using the number of annotated speech errors as a dependent variable (Väyrynen *et al.* 2009). However, many semantic aspects are also notoriously hard to define on continuous dimensions (cf. dimensional models of emotion). A solution can be to use modelling techniques that use the intrinsic data structure (e.g. geodesic distance) to provide an alternate mapping of the labelled data. The mapping of data into a new supervised structure provides a data driven version of the labelling space that is not as strict as the original labels and, consequently, is not as much influenced by the labelling errors (i.e. data similarity is emphasised).

3.3.1 Isomap based framework for semantic speech content visualisation

In Paper IV, an Isomap based framework for the nonlinear visualisation of speech data is presented that uses a classifier based supervised learning of data. The data used in Paper IV is the same data which was used in Paper I (Four emotional categories: neutral, sad, angry, and happy) with exactly the same sampling and extraction of features. The

original labelling of samples was used in the training and testing of models, i.e. the instruction mode labels.

An SFFS-based feature selection procedure and a CV-based estimation of model performance are used to produce low dimensional nonlinear embeddings of high dimensional prosodic and acoustic features. The classifier learning procedure enables an efficient way to emphasise the desired semantic property to be visualised (e.g. emotional speech content) and yet still retains the topological structure of the data in the visualisation (see Fig. 13, for a sketch of 3D model of emotion for four emotional labels).

In the training, using the SFFS feature selection method, for each prospective feature vector, first, the Isomap embedding procedure is performed (Fig. 14). The resulting model is then evaluated in the embedding space using a kNN classifier and a CV performance estimation. The Isomap and other model parameters (i.e. the Isomap linking k, the kNN classifier k, dissimilarity function parameters, and GRNN parameters) are searched during the training using an exhaustive grid search within reasonable parameter ranges.

A PCA based global linear methods (using ROBPCA, Hubert *et al.* (2005), and a similar supervised classifier optimisation setup as was used in the Isomap methods) is

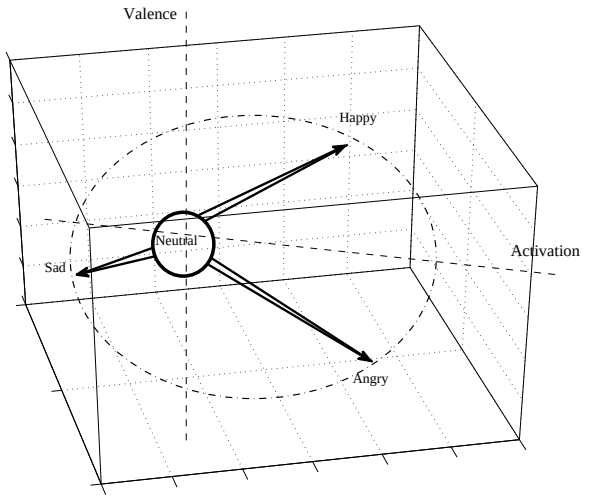


Fig. 13. Sketch of a 3D model of emotion.

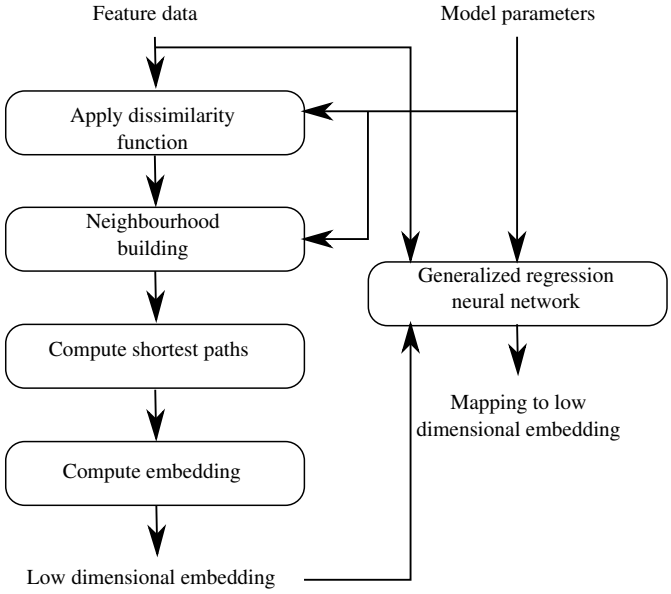


Fig. 14. Flowchart of the Isomap embedding procedure.

also provided as a reference and to highlight the need for nonlinear learning in order to build usable visualisations. The PCA based embedding (see Fig. 15), although effective in classification performance, does not provide a very good visualisation of the data.

In the paper, two different dissimilarity functions were used to produce the nonlinear manifold embeddings, a linear weighted (Vlachos *et al.* 2002) version (W-Isomap) and a nonlinear version (S-Isomap) using the dissimilarity function (Eq. 5) from Geng *et al.* (2005):

$$D(\boldsymbol{x}_i,\boldsymbol{x}_j) = \begin{cases} \sqrt{1-\exp^{\frac{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\beta}}} & y_i = y_j \\ \sqrt{\exp^{\frac{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\beta}}} - \alpha & y_i \neq y_j \end{cases}, \tag{5}$$

where $d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is the squared Euclidean distance between vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with class labels y_i and y_j, respectively. β was chosen to be the average Euclidean distance of the data set and α is optimised in feature selection process.

The impact of different target embedding dimensionality was also evaluated during training. The best classification performances were consistently obtained when using

three dimensional target embedding space. Lower dimensions (i.e. one and two dimensional spaces) were clearly not as effective embedding targets, resulting in worse performance. Higher dimensions, although providing performances comparable to that of the three-dimensional space, would require more complex representations of the embeddings (e.g. they would require multiple visualisations projected along different dimensions to properly illustrate the structures). The classifier learning procedure enables comparison between the classification performances of the various approaches.

The average classification performances for a direct high dimensional classification using a kNN classifier, a linear PCA based mapping, the two nonlinear Isomap based methods W-Isomap and S-Isomap, and a human listening test are presented in Table 11. The performances of the two nonlinear embeddings compare well with the direct kNN classifier and the linear PCA reference methods. Furthermore, the performances of the nonlinear methods are also very close to the human reference. The best performing nonlinear embeddings were also observed, subjectively, to provide the most visually pleasing low dimensional representations of the speech database (Fig. 16 and Fig. 17). Feature vectors corresponding to the best achieved results are presented in Table 12. All feature vectors are composed of features from each of the major prosodic feature categories. Emotionally discriminative features seem to be predominantly describing the range dynamics of the F_0, intensity, and duration aspects of speech. Only a few

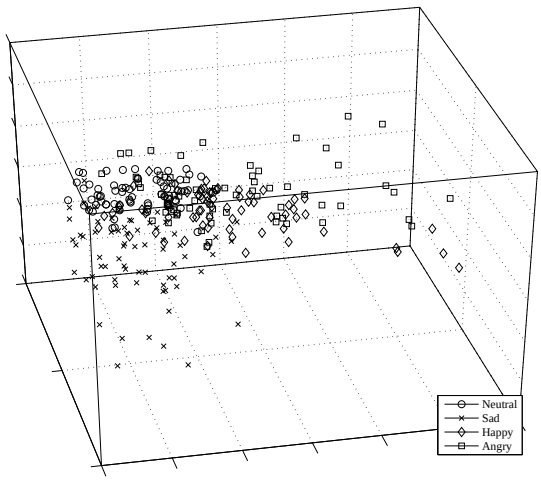


Fig. 15. Linear PCA embedding of a speech database.

Table 11. Classification results of low dimensional embedding methods and reference methods.

kNN	PCA	W-Isomap	S-Isomap	Human
71.4%	73.2%	75.4%	76.1%	76.9%

spectral features (i.e. LFE500 and LFE1000) are present in the feature vectors of Isomap embedding based approaches.

The visualisation offering the best classification performance, and also, via subjective evaluation, the most pleasing structure, was then further studied for emotional intensity structures by means of listening tests. The hypothesis was that the used weighting and the supervision that relies on geodesic data structures is also robust in preserving other topological structures of the data. It was therefore expected that the emotional manifold structures should contain to some degree the variations in emotional intensities although no such information was used in the training. Speech samples selected along the emotional manifolds were annotated in two emotional intensity groups (high and low) by the human listeners. Wilcoxon rank-sum tests and logistic regression were then used to test the tagged groups for statistically significant structures (Table 13). Significant median distances between the high and low emotional intensity groups were observed in each emotional category. The listeners' annotations of emotional intensity

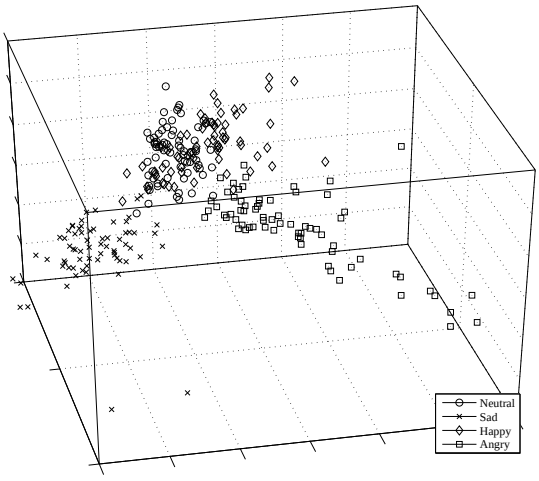


Fig. 16. Linear weighted W-Isomap based visualisation of an emotional speech database.

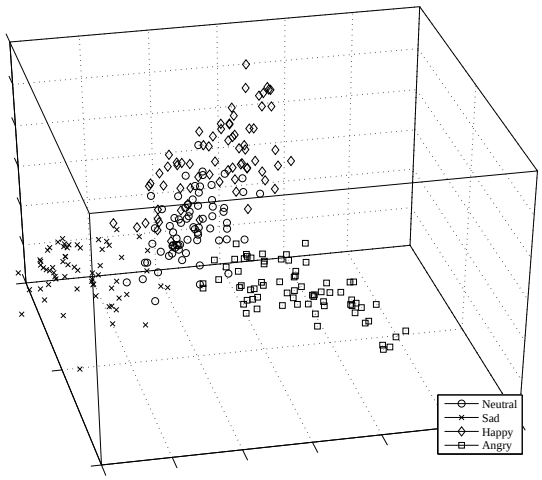


Fig. 17. Nonlinear weighted S-Isomap based visualisation of an emotional speech database.

were briefly compared and found consistent across all categories. The logistic regression tests further revealed that the emotional intensity structures were significantly predicted by the embedding coordinates. The low intensity areas of emotions were as expected closer to neutral samples and the high intensity samples resided farther away along the corresponding emotional sub-manifolds.

Table 12. Feature vectors corresponding to the best achieved results (IV, published by permission of IEEE).

S-Isomap	W-Isomap	PCA	kNN
fracRange	fracRange	Mean	fracRange
Jitter	F0Var	Max	Shimmer
LFE1000	Shimmer	fracMax	Jitter
GDrisemax	Jitter	fracRange	GDrisemax
GDfallmin	LFE500	Shimmer	IntRange
MeanInt	GDrisemax	Jitter	IntVar
fracMaxInt	fracMaxInt	GDposav	sratio
fracIntRange	fracIntRange	GDfallav	norm_intvar
mlnlngth	mlslngth	GDmin	
mlslngth	spratio	fracMaxInt	
spratio	norm_intvar	vratio	
vratio	norm_freqvar		
norm_intvar			
norm_freqvar			

Table 13. Significance testing results of S-Isomap emotional sub-manifold structures (IV, published by permission of IEEE).

	Sad	Angry	Happy
Listener agreement	85%	85%	86.5%
Median distance	1.0599**	2.3024***	1.9745***
Correct count (Dist)	62.0%	80.0%***	74.5%***
Correct count (3D)	73.0%***	84.0%***	73.5%***

Note: Agreement ratios of the listeners are shown in percentages. The median distances of the low/high emotional intensity groups in different emotions (arbitrary units) are shown with the associated p-values from Wilcoxon rank-sum tests. The logistic regression correct counts in percentages are shown with the associated p-values from Likelihood Ratio (LR) tests. ** indicates p-values < 0.01 and *** indicates p-values < 0.001

3.4 Summary of research contributions

The contributions of this thesis can be divided into three parts. First, a technology for the recognition of emotion for Finnish speech was developed. The Finnish long-term global prosody of emotions has been very little studied. Often, the studies of the vocal parameters of emotional Finnish speech have been concentrated on very short speech units. Furthermore, no automatic classification approaches for continuous emotional Finnish speech has been previously performed.

Second, a novel fusion technique was developed to enhance the emotion classification of short speech segments. An increased classification performance was achieved by using vocal source derived features and more traditional intonation describing local prosodic features.

Third, a technological framework for the visualisation of emotional speech data was developed. A novel classifier based optimisation approach was used for the supervised Isomap embedding of high dimensional emotional data into a low dimensional manifold. An efficient and topologically consistent visualisation of emotional data was achieved using long-term global prosodic features.

4 Discussion

4.1 Significance of results

4.1.1 Emotion recognition for Finnish speech

In this thesis, the automatic recognition of emotion has been studied using Finnish emotional speech recordings. The technology developed in Paper I for emotion recognition is based on features describing primarily global long-term prosody. Some related acoustic features of speech are also included. The feature extraction and pattern recognition techniques have been developed aiming at a fully automatic mode of operation and a robust noise performance (Paper II).

Using long samples (i.e. multiple seconds in length), a robust emotion recognition for the basic emotions is possible. A long-term emotion recognition capability can be useful in making experiences with different interfaces (e.g. robot behaviours, adapted user interface flows, or automatic assessment of communications) more familiar and natural for the users. As a result, the operational effectiveness of such interfaces can also be expected to increase.

The recognition results for emotion recognition for Finnish speech were found to be consistent with prior literature findings on emotion recognition in the western language context. A level of classification performance around two thirds correct to over 80% correct can be seen as an attainable goal for a limited basic emotion recognition task depending on the amount of emotional categories used (Dellaert *et al.* 1996, McGilloway *et al.* 2000, Oudeyer 2002, Bosch 2003). In Finnish speech, the activation dimension of affect seems to be more prominently signalled in the voice, a finding common for vocal expression of emotion (Bachorowski 1999, Scherer 2003). The recognition rates for Finnish listeners were found to be typical and consistent with the results reported in western language contexts, see e.g. Elfenbein & Ambady (2002), Scherer (2003). Feature vectors describing the learned models suggest that a similar prosody based signalling of emotion is used also in Finnish speech.

4.1.2 Fusion technique for multi-modal classification of emotion

Short emphasised vowels are often used in commands and acknowledging purposes in conversations carrying important information that can be utilised to detect frustration or other emotional behaviour that possibly indicates unwanted functionality when using speech user interfaces. A robust recognition of emotion can thus be a desirable enhancement for voice command based user interfaces, e.g. automatic call centres.

In Paper III, meta learning algorithms were developed for the learning of emotional models and to combine relevant information from multiple sources with fusion techniques. The study of fusion techniques indicates that a decision based fusion method is preferable for the recognition of emotion when combining short term voice quality features with traditional prosodic, and related often used acoustic features. Furthermore, the decision level sum fusion allows for an efficient technological realisation for multimodal operation.

For the short signals, the detection of emotion is the most effective in high activation emotions when the speech quality is more pressed (e.g. Anger). For short segments, the voice quality features describing the vocal source characteristics were mostly capable in describing neutral, happy, and angry emotions. This was expected as many of the voice quality describing features are sensitive to the pressedness of the voice, e.g. NAQ (Alku *et al.* 2002). Using fusion, recognition results were increased accordingly, as expected. For this reason, applications attempting to analyse frustration and possible erroneous operations in human computer interactions where short vocal commands are utilised are the most likely to benefit from short speech segment based emotion recognition. However, as the performance of the used IAIF based vocal source extraction is currently limited, a more robust feature extraction technique for the vocal source features is most likely required for most practical applications. The use of the proposed fusion architecture with alternative features is straightforward.

4.1.3 Visualisation of emotional speech data

In Paper IV, the presented method for emotional speech data visualisation relies on the structural properties of the data. Supervised learning of manifolds not only improves on the classification performance of emotional speech over standard classification methods, but also provides an effective representation of speech. A visualisation of intrinsic data

manifolds helps in interpreting the emotional speech data and provides a means for further exploration.

Using nonlinear weighting, the S-Isomap approach produced both the best classification rate and qualitatively the best visualisation of emotional Finnish speech data. The topology of the produced visualisation was also inspected using listening tests to verify if the perceived emotional intensity would be distributed in a meaningful way along the emotional sub-manifolds. The distribution of emotional intensity was found correlating strongly with the manifold coordinates. The presence of clear manifold structures that was not guided by the supervised learning process can be seen as a strong indication of robustness of the presented visualisation framework in retaining the original space manifold topology after a considerable reduction in dimensionality. Furthermore, the resulting visualisations of emotional speech data were found strongly resembling the dimensional emotional models presented in the literature, e.g. by Russell & Mehrabian (1977), Daly *et al.* (1983). During the optimisation process, the strong performance of the very low dimensional representations (3–5) is also a strong support for the assumption of low intrinsic dimensionality of speech and emotion therein (Togneri *et al.* 1992, Kim *et al.* 2010).

The proposed manifold modelling technology is not specifically designed for emotion visualisation only. The presented framework should thus be generalizable for other semantic concept classification and visualisation tasks (e.g. speech style or fluency). The supervision can be changed to incorporate arbitrary class labellings as long as the required prior assumptions for continuous connected manifold structures are not violated. It is also possible to incorporate other continuous functions to focus the supervision on more specific structures. For example, it might be useful to incorporate the emotional intensity estimates gained from human experts into the manifold dissimilarity functions to produce more highlighted visualisation of the emotional intensity manifolds.

4.2 Limitations and generalizability

This study is based on Finnish emotional speech data only. Thus, the results are only to be interpreted in the context of Finnish speech. Especially, in Paper III, the used short speech segment results are critically based on the fact that in Finnish the first word of utterances, short or long, is always stressed. The emotional behaviour and the prosody of short verbal commands can be markedly different in other languages and the study must therefore be replicated before any generalizations of the results over different

languages can be made.

Also, it should be noted that the used emotional data only contains acted emotional content in limited basic emotions. However, the vast majority of recorded data available is actually acted. Hence, the results are likely to be applicable for at least applications using such data (e.g. the analysis of emotional plays and other entertainment productions). However, spontaneous emotional speech might not be signalled using similar, perhaps stereotypical, ways naturally limiting the utility of the presented results in a real environment application. The used technology should in general be applicable to natural emotional speech data also, but the presented results of feature selections and meta learnings will likely be different with a spontaneous data and should be reanalysed accordingly. The inclusion of more emotions can be done without changes in the proposed methods. A lowered overall performance when working with a larger scale of emotion or more complex emotions is naturally expected.

The IAIF based extraction of vocal source features used in Paper III to demonstrate the performance of fusion classification is currently limited to high quality clean speech recordings of an open vowel. It is likely that a more robust feature extraction method is required for most practical applications. However, the proposed fusion classification architecture itself is not limited by the used features and can be directly used with features provided by alternative extraction methods.

We can expect that the presented method for nonlinear manifold learning and visualisation of speech data (Paper IV) should be generalizable for other languages, spontaneous speech, or semantic concepts other than emotion. Adding more emotions or categories and therefore more data samples, or hyper-parameters in the training, can be problematic, however, due to computational requirements. The high computational complexity of the presented method is an apparent limitation in learning and finding complex manifold embeddings from large databases. A final learned embedding, on the other hand, can be used effectively in classification and further analysis, making the computational requirements problematic only during the supervised learning.

5 Summary and conclusions

In this thesis, emotion recognition from speech is studied using long and short term prosodic and related acoustic features. Finnish acted emotional speech data is used to produce a state-of-the-art classification of spoken emotional content in basic emotional categories. The work is divided into three main parts:

1. Emotion recognition from speech using long-term prosodic and acoustic features
2. Emotion recognition using the fusion of short term prosodic and voice quality features
3. Visualisation of emotional speech data using nonlinear manifold modelling.

In the first part, a state-of-the-art level basic emotion recognition from Finnish speech was developed using global prosodic and acoustic features. The robustness and performance of the feature extraction were investigated. The aim was to enable automatic operation using raw audio recordings with little or no user defined parameters. An automatic pattern recognition method was developed for emotional speech recognition using a meta learning feature selection algorithm. An emotion classification rate approaching the human reference was obtained using a limited set of basic emotions.

In the second part, short term prosody and voice source features were investigated to enable the emotion classification of short stressed words. Fusion technology was developed to combine the feature sets in order to enhance the classification performance. An increased emotion classification performance was demonstrated.

In the third part, the visualisation of emotional speech data was investigated. Nonlinear manifold modelling techniques were used with classifier based supervision to enable a low dimensional visualisation of high dimensional prosodic and related acoustic speech features. A framework for semantic content visualisation for speech data was developed. An increased classification and a low dimensional visualisation of emotional speech data were achieved. Large similarities with the dimensional emotional models and the produced visualisations were observed.

The developed emotion recognition technology can be used to enhance user interfaces and devices to be more emotionally responsive and natural. Measuring the emotional responses of user experiences can be used to further develop the interfaces. The presented framework for emotional speech visualisation can be used for speech research and data representation purposes. The method can also be used for other semantic content visualisations and the meta-analysis of speech data.

References

Airas M (2008) TKK Aparat: An environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocology 33(1): 49–64.

Aizerman MA, Braverman EA & Rozonoer L (1964) Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25: 821–837.

Alku P (1992) Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11: 109–118.

Alku P (2011) Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana 36(5): 623–650.

Alku P, Bäckström T & Vilkman E (2002) Normalized amplitude quotient for parameterization of the glottal flow. Journal of the Acoustical Society of America 112(2): 701–710.

Alku P & Vilkman E (1996) Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. Speech Communication 18: 131–138.

Ang J, Dhillon R, Krupski A, Shriberg E & Stolcke A (2002) Prosody-based automatic detection of annoyance and frustration in human-computer dialog. Proc. 7th International Confrence on Spoken Language Processing, Denver, CO, USA, 2037–2040.

Atal BS & Hanauer SL (1971) Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America 50(2B): 637–655.

Bachorowski JA (1999) Vocal expression and perception of emotion. Current Directions in Psychological Science 8(2): 53–57.

Bagshaw PC, Hiller SM & Jack MA (1993) Enhanced pitch tracking and the processing of F_0 contours for computer aided intonation teaching. Proc. 3rd European Conference on Speech Communication and Technology, Berlin, Germany, 1003–1006.

Balasubramanian M & Schwartz EL (2002) The isomap algorithm and topological stability. Science 295(5552): 7.

Banse R & Scherer KR (1996) Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70(3): 614–636.

Batliner A, Buckow J, Huber R, Warnke V, Nöth E & Niemann H (1999) Prosodic feature evaluation: Brute force or well designed? Proc. 14th International Congress of Phonetic Sciences, San Francisco, CA, USA, 3: 2315–2318.

Beddor PS & Hawkins S (1990) The influence of spectral prominence on perceived vowel quality. The Journal of the Acoustical Society of America 87(6): 2684–2704.

Bell C (1806) Essays on the anatomy of expression in painting. London: Longman, Hurst, Rees, and Orme.

Blanchette I & Richards A (2010) The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. Cognition & Emotion 24(4): 561–595.

Boersma P (1993) Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. Institute of Phonetic Sciences, 17: 97–110.

Bosch L ten (2003) Emotions, speech and the ASR framework. Speech Communication 40(1-2): 213–225.

Boser BE, Guyon IM & Vapnik VN (1992) A training algorithm for optimal margin classifiers. Proc. 5th Annual ACM Workshop on Computational Learning Theory, ACM Press, New York, NY, USA, 144–152.

Bridle JS & Brown MD (1974) An experimental automatic word recognition system. Tech. Rep. JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.

Brown K (ed) (2005) Encyclopedia of language and linguistics, 2nd edition. Elsevier Ltd, Oxford.

Calvo RA & D'Mello S (2010) Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing 1: 18–37.

Campbell N & Mokhtari P (2003) Voice quality, the 4th prosodic dimension. Proc. 15th International Congress of Phonetic Sciences (ICPhS2003), Barcelona, 2417–2420.

Campione E & Véronis J (2002) A large scale multilingual study of silent pause duration. Proc. Speech Prosody an International Conference, Aix-en-Provence, France.

Cawley GC & Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11: 2079–2107.

Choi H & Choi S (2007) Robust kernel isomap. Pattern Recognition 40: 2007.

Cowie R & Cornelius RR (2003) Describing the emotional states that are expressed in speech. Speech Communication 40(1-2): 5–32.

Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M & Schröder M (2000) 'FEELTRACE': an instrument for recording perceived emotion in real time. Proc. ISCA Tutorial and Research Workshop on Speechand Emotion, Newcastle, Northern Ireland, UK, 19–24.

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

Daly EM, Lancee WJ & Polivy J (1983) A conical model for the taxonomy of emotional experience. Journal of Personality and Social Psychology 45: 443–457.

Damasio AR (1994) Descartes' error: Emotion, reason, and the human brain. Putnam's Sons, New York, NY.

Darwin C (1872/1965) The expression of the emotions in man and animals. University of Chicago Press, Chicago, IL.

Davis SB & Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing 28: 357–366.

Dellaert F, Polzin TS & Waibel A (1996) Recognizing emotion in speech. Proc. International Congress on Spoken Language Processing, Philadelphia, PA, 3: 1970–1973.

Demol M, Verhelst W & Verhoeve P (2007) The duration of speech pauses in a multilingual environment. Proc. 8th Annual Conference of the International Speech Communication Association, INTERSPEECH–2007, 990–993.

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

Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin JC, Devillers L, Abrilian S, Batliner A, Amir N & Karpouzis K (2007) The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In: Paiva A, Prada R & Picard R (eds) Affective Computing and Intelligent Interaction, volume 4738 of *Lecture Notes in Computer Science*, 488–500. Springer Berlin / Heidelberg.

Duda RO, Hart PE & Stork DG (2001) Pattern classification (2nd edition). John Wiley & Sons Inc,

New York, NY.
Efron B & Tibshirani RJ (1994) Introduction to the bootstrap. Monographs on Statistics and Applied Probability. Chapman & Hall.
Ehrette T, Chateau N, d'Alessandro C & Maffiolo V (2002) Prosodic parameters of perceived emotions in vocal server voices. Proc. Speech Prosody an International Conference, Aix-en-Provence, France.
Ekman P (1992) An argument for basic emotions. Cognition & Emotion 6: 169–200.
Ekman P (1999) Basic emotions. In: Dalgleish T & Power M (eds) Handbook of Cognition and Emotion, 45–60. John Wiley & Sons Inc, New York, NY.
Elfenbein H & Ambady N (2002) On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological Bulletin 128: 203–235.
Fant G (1960) Acoustic theory of speech production. Mouton, The Hague.
Fant G (2004) Speech acoustics and phonetics: Selected writings. Text, Speech and Language Technology. Springer.
Fant G, Liljencrants J & Lin Q (1985) A four- parameter model of glottal flow. STL-QPSR 4: 1–13.
Farinas J & Pellegrino F (2001) Automatic rhythm modeling for language identification. Proc. 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, 1: 2539–2541.
Fastl H & Zwicker E (2007) Psychoacoustics: Facts and models. Springer Series in Information Sciences. Springer-Verlag, 3 edition.
Fletcher H & Munson WA (1933) Loudness, its definition, measurement and calculation. Journal of the Acoustical Society of America 5: 82–108.
Fontaine JR, Scherer KR, Roesch EB & Ellsworth PC (2007) The world of emotions is not two-dimensional. Psychological Science 18(12): 1050–1057.
Geng X, Zhan DC & Zhou ZH (2005) Supervised nonlinear dimensionality reduction for visualization and classification. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 35(6): 1098–1107.
Gobl C & Ní Chasaide A (2003) Amplitude-based source parameters for measuring voice quality. Proc. ISCA Tutorial and Research Workshop VOQUAL'03 on Voice Quality: Functions, Analysis and Synthesis, ISCA, ISCA Archive, 151–156.
Grichkovtsova I, Morel M & Lacheret A (2012) The role of voice quality and prosodic contour in affective speech perception. Speech Communication 54(3): 414–429.
Gu Rj & Xu Wb (2007) Weighted kernel Isomap for data visualization and pattern classification. In: Wang Y, Cheung Ym & Liu H (eds) Computational Intelligence and Security, volume 4456 of *Lecture Notes in Computer Science*, 1050–1057. Springer Berlin / Heidelberg.
Haykin S (1994) Neural networks: A comprehensive foundation. Macmillan Publishing Company.
Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87(4): 1738–1752.
Hess WJ (1983) Pitch determination of speech signals: Algorithms and devices. Springer-Verlag, Berlin, Germany.
Horii Y (1979) Fundamental frequency perturbation observed in sustained phonation. Journal of Speech and Hearing Research 22: 5–19.
Horii Y (1980) Vocal shimmer in sustained phonation. Journal of Speech and Hearing Research 23: 202–209.
Hubert M, Rousseeuw PJ & Branden KV (2005) ROBPCA: a new approach to robust principal

component analysis. Technometrics 47: 64–79.
Izard CE (1992) Basic emotions, relations among emotions, and emotion-cognition relations. Psychological Review 99: 561–565.
Jain V & Saul L (2004) Exploratory analysis and visualization of speech and music by locally linear embedding. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004 (ICASSP'04), 3: iii–984.
Kane J, Kane M & Gobl C (2010) A spectral lf model based approach to voice source parameterisation. Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH–2010, ISCA, Makuhari, Chiba, Japan, 2606–2609.
Kent RD & Read C (1992) The acoustic analysis of speech. Singular Publishing Group, Inc., San Diego, CA.
Kienast M & Sendlmeier W (2000) Acoustical analysis of spectral and temporal changes in emotional speech. Proc. Proceedings of the ISCA Tutorial and Research Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, 92–97.
Kim J, Lee S & Narayanan S (2010) An exploratory study of manifolds of emotional speech. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2010 (ICASSP'10), 5142–5145.
Kittler J (2000) A framework for classifier fusion: Is it still needed? In: Advances in Pattern Recognition, 45–56. Springer Berlin / Heidelberg.
Kittler J & Alkoot F (2003) Sum versus vote fusion in multiple classifier sysytems. IEEE Transaction on Pattern Analysis and Machine Intelligence 25(1): 110–115.
Kittler J, Hatef M, Duin R & Matas J (1998) On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3): 226–239.
Krivokapić J (2007) Prosodic planning: Effects of phrasal length and complexity on pause duration. Journal of Phonetics 35(2): 162–179.
Laukka P, Juslin P & Bresin R (2005) A dimensional approach to vocal expression of emotion. Cognition & Emotion 19(5): 633–653.
Laukkanen AM, Vilkman E, Alku P & Oksanen H (1996) Physical variations related to stress and emotional state: a preliminary study. Journal of Phonetics 24: 313–335.
Lee CM, Narayanan S & Pieraccini R (2001) Recognition of negative emotions from the speech signal. Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio Trento, Italy.
Lehtonen J (1985) Speech rate in Finnish. In: Hurme P (ed) Papers in Speech Research, volume 6, 16–27. University of Jyväskylä, Jyväskylä, Finland.
Li H, Scaife R & O'Brien D (2011) LF model based glottal source parameter estimation by extended Kalman filtering. Proc. 22nd IET Irish Signals and Systems Conference.
Lugger M, Yang B & Wokurek W (2006) Robust estimation of voice quality parameters under realworld disturbances. Proc. IEEEInternational Conference on Acoustics, Speech, and Signal Processing, 2006 (ICASSP'06), 1(I): 1097–1100.
Maaten LJP van der, Postma EO & Herik HJ van den (2009) Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University.
MacKinnon NJ & Keating LJ (1989) The structure of emotions: Canada-United States comparisons. Social Psychology Quarterly 52(1): 70–83.
Makhoul J (1973) Spectral analysis of speech by linear prediction. Audio and Electroacoustics, IEEE Transactions on 21(3): 140–148.
McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M & Stroeve S (2000)

Approaching automatic recognition of emotion from voice: A rough benchmark. Proc. ISCA Tutorial and Research Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, 207–212.

Mermelstein P (1976) Distance measures for speech recognition - psychological and instrumental. In: Chen CH (ed) Pattern Recognition and Artificial Intelligence, 374–388. Academic Press, Inc.

Moattar M & Homayounpour M (2012) A review on speaker diarization systems and approaches. Speech Communication 54(10): 1065–1103.

Montero JM, Gutierrez-Arriola J, Palazuelos S, Enriquez E, Aguilera S & Pardo JM (1998) Emotional speech synthesis: From speech database to TTS. Proc. 5th International Conference on Speech and Language Processing, Sydney, Australia, 923–926.

Murray IR & Arnott JL (1993) Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America 93(2): 1097–1108.

Murray IR & Arnott JL (1995) Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication 16: 369–390.

Murray IR, Arnott JL & Rohwer EA (1996) Emotional stress in synthetic speech: Progress and future directions. Speech Communication 20: 85–91.

Nicolaou MA, Gunes H & Pantic M (2011) Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing 2: 92–105.

Noll AM (1967) Cepstrum pitch determination. Journal of Acoustical Society of America 41(2.): 293–309.

Noll AM (1970) Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. Proc. Symposium on Computer Processing in Communication, Brooklyn, NY, 19: 779–797.

Nwe T, Foo S & De Silva L (2003) Speech emotion recognition using hidden markov models. Speech Communication 41: 603–623.

Oatley K & Johnson-laird PN (1987) Towards a cognitive theory of emotions. Cognition & Emotion 1(1): 29–50.

Ortony A, Clore G & Collins A (1990) The cognitive structure of emotions. Cambridge University Press.

Ortony A & Turner TJ (1990) What's basic about basic emotions? Psychological Review 97: 315–331.

Oudeyer P (2002) Novel useful features and algorithms for the recognition of emotions in human speech. Proc. Speech Prosody an International Conference, Aix-en-Provence, France.

Paeschke A & Sendelmeier WF (2000) Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. Proc. ISCA Tutorial and Research Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, 75–80.

Picard RW (1995) Affective computing. Tech. Rep. Perceptual Computing Section Technical Report No. 321, MIT Media Laboratory, 20 Ames St., Cambridge, MA 02139.

Picard RW (1997) Affective computing. MIT Press, Cambridge, MA, USA.

Pudil P, Novovičová J & Kittler J (1994) Floating search methods in feature selection. Pattern Recognition Letters 15: 1119–1125.

Qi Y, Minka TP, Picard RW & Ghahramani Z (2004) Predictive automatic relevance determination by expectation propagation. Proc. Twenty First International Conference on Machine Learning

(ICML-04), 671–678.
Rabiner LR, Cheng MJ, Rosenberg AE & McGonegal CA (1976) A comparative study of several pitch detection algorithms. IEEE transactions on Acoustics, Speech, and Signal Processing 24: 399–413.
Rabiner LR & Juang BH (1993) Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ.
Robinson DW & Dadson RS (1956) A re-determination of the equal-loudness relations for pure tones. British Journal of Applied Physics 7(5): 166–181.
Roweis ST & Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500): 2323–2326.
Russell JA (1980) A circumplex model of affect. Journal of Personality and Social Psychology 39: 1161–1178.
Russell JA (1994) Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychological Bulletin 115: 102–141.
Russell JA (2003) Core affect and the psychological construction of emotion. Psychological review 110(1): 145.
Russell JA & Feldman BL (1999) Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology 76: 805–819.
Russell JA & Mehrabian A (1977) Evidence for a three-factor theory of emotions. Journal of Research in Personality 11(3): 273–294.
Scherer KR (1986) Vocal affect expression: A review and a model for future research. Psychological Bulletin 99(2): 143–165.
Scherer KR (2000) A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. Proc. International Confrence on Spoken Language Processing, Beijing, China, 2: 379–382.
Scherer KR (2003) Vocal communication of emotion: A review of research paradigms. Speech Communication 40(1-2): 227–256.
Scherer S, Kane J, Gobl C & Schwenker F (2013) Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Computer Speech & Language 27(1): 263–287.
Schröder M (2001) Emotional speech synthesis: A review. Proc. 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, 1: 561–564.
Schröder M, Cowie R, Douglas-Cowie E, Westerdijk M & Gielen S (2001) Acoustic correlates of emotion dimensions in view of speech synthesis. Proc. 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, 1: 87–90.
Schroeder MR (1968) Period histogram and product spectrum: New methods for fundamental frequency measurement. Journal of Acoustical Society of America 43(4.): 829–834.
Seppänen T, Toivanen J & Väyrynen E (2003) MediaTeam Speech Corpus: A first large Finnish emotional speech database. Proc. 15th International Congress of Phonetic Sciences, Barcelona, Spain, 3: 2469–2472.
Shriberg E (2007) Higher-level features in speaker recognition. In: Müller C (ed) Speaker Classification I, volume 4343 of *Lecture Notes in Artificial Intelligence*, 241–259. Springer-Verlag Berlin Heidelberg.
Smith CA & Ellsworth PC (1985) Patterns of cognitive appraisal in emotion. Journal of personality and social psychology 48(4): 813–838.

Soleymani M, Lichtenauer J, Pun T & Pantic M (2012) A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing 3(1): 42–55.
Stevens SS (1957) On the psychophysical law. Psychological Review 64(3): 153–181.
Stevens SS, Volkmann J & Newman EB (1937) A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America 8(3): 185–190.
Stone M (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 36(2): 111–147.
Tenenbaum JB, Silva Vd & Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500): 2319–2323.
Thymé-Gobbel A & Hutchins SE (1999) Prosodic features in automatic language identification reflect language typology. Proc. 14th International Congress of Phonetic Sciences, San Francisco, CA, USA, 29–32.
Titze IR (1994) Principles of voice production. Prentice Hall, Englewood Cliffs, NJ.
Togneri R, Alder M & Attikiouzel Y (1992) Dimension and structure of the speech space. IEE Proceedings I (Communications, Speech and Vision) 139(2): 123–127.
Toivanen J (2001) Perspectives on intonation: English, Finnish and English spoken by finns. Verlag Peter Lang AG, Frankfurt am Main, Germany.
Traunmüller H & Eriksson A (2000) Acoustic effects of variation in vocal effort by men, women, and children. The Journal of the Acoustical Society of America 107(6): 3438–3451.
Vapnik V & Lerner A (1963) Pattern recognition using generalized portrait method. Automation and Remote Control 24: 774–780.
Varma S & Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7(1): 91.
Väyrynen E (2005) Automatic emotion recognition from speech. Master's thesis, University of Oulu.
Väyrynen E, Keränen H, Toivanen J & Seppänen T (2009) Automatic classification of segmental second language speech quality using prosodic features. Proc. $XXII^{th}$ Swedish Phonetics Conference Fonetik 2009, 116–119.
Ververidis D & Kotropoulos C (2006) Emotional speech recognition: Resources, features, and methods. Speech Communication 48(9): 1162–1181.
Vlachos M, Domeniconi C, Gunopulos D, Kollios G & Koudas N (2002) Non-linear dimensionality reduction techniques for classification and visualization. Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 645–651.
Wolpert DH & Macready WG (1997) No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1(1): 67–82.
Wolpert DH & Macready WG (2005) Coevolutionary free lunches. IEEE Transactions on Evolutionary Computation 9(6): 721–735.
You M, Chen C, Bu J, Liu J & Tao J (2006) Emotional speech analysis on nonlinear manifold. Proc. 18th International Conference on Pattern Recognition, 2006 (ICPR 2006), 3: 91–94.
Yu F, Chang E, Xu YQ & Shum HY (2001) Emotion detection from speech to enrich multimedia content. Proc. Second IEEE Pacific Rim Conference on Multimedia, Beijing, China, 550–557.
Yumoto E, Gould WJ & Baer T (1982) Harmonics-to-noise ratio as an index of the degree of hoarseness. Acoustical Society of America Journal 71: 1544–1550.
Zeng Z, Pantic M, Roisman G & Huang T (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1): 39–58.

Zeng Z, Zhang Z, Pianfetti B, Tu J & Huang T (2005) Audio-visual affect recognition in activation-evaluation space. Proc. IEEE International Conference on Multimedia and Expo, 2005 (ICME 2005), 4 pp.

Zetterholm E (1999) Emotional speech focusing on voice quality. Proc. 12th Swedish Phonetics Conference, Götheborg, Sweden, 145–148.

Zhang S, Lei B, Chen A, Chen C & Chen Y (2010a) KIsomap-based feature extraction for spoken emotion recognition. Proc. IEEE 10th International Conference on Signal Processing (ICSP 2010), Beijing, China, 1374–1377.

Zhang S, Li L & Zhao Z (2010b) Speech emotion recognition based on supervised locally linear embedding. Proc. International Conference on Communications, Circuits and Systems, 2010 (ICCCAS), Chengdu City, China, 401–404.

Zwicker E (1961) Subdivision of the audible frequency range into critical bands (frequenzgruppen). The Journal of the Acoustical Society of America 33(2): 248–248.

Original articles

I Toivanen J, Väyrynen E & Seppänen T (2004) Automatic discrimination of emotion from spoken Finnish. Language and Speech 47(4): 383–412. DOI:10.1177/00238309040470040301

II Väyrynen E, Keränen H, Seppänen T & Toivanen J (2005) Performance of F0Tool - A new speech analysis software for analyzing large speech data sets. Proceedings of the 2nd Baltic Conference on Human Language Technologies: 353–358.

III Väyrynen E, Toivanen J & Seppänen T (2011) Classification of emotion in spoken Finnish using vowel-length segments: Increasing reliability with a fusion technique. Speech Communication 53(3): 269–282.

IV Väyrynen E, Kortelainen J & Seppänen T (2013) Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody. IEEE Transaction on Affective Computing 4(1): 47–56.

Original publications are not included in the electronic version of the dissertation.

471. Haapalainen, Mikko (2013) Dielectrophoretic mobility of a spherical particle in 2D hyperbolic quadrupole electrode geometry

472. Bene, József Gergely (2013) Pump schedule optimisation techniques for water distribution systems

473. Seelam, Prem Kumar (2013) Hydrogen production by steam reforming of bio-alcohols : the use of conventional and membrane-assisted catalytic reactors

474. Komulainen, Petri (2013) Coordinated multi-antenna techniques for cellular networks : Pilot signaling and decentralized optimization in TDD mode

475. Piltonen, Petteri (2013) Prevention of fouling on paper machine surfaces

476. Juuso, Esko (2013) Integration of intelligent systems in development of smart adaptive systems : linguistic equation approach

477. Lu, Xiaojia (2013) Resource allocation in uplink coordinated multicell MIMO-OFDM systems with 3D channel models

478. Jung, Sang-Joong (2013) Personal machine-to-machine (M2M) healthcare system with mobile device in global networks

479. Haho, Päivi (2014) Learning enablers, learning outcomes, learning paths, and their relationships in organizational learning and change

480. Ukkonen, Kaisa (2014) Improvement of recombinant protein production in shaken cultures : focus on aeration and enzyme-controlled glucose feeding

481. Peschl, Michael (2014) An architecture for flexible manufacturing systems based on task-driven agents

482. Kangas, Jani (2014) Separation process modelling : highlighting the predictive capabilities of the models and the robustness of the solving strategies

483. Kemppainen, Kalle (2014) Towards simplified deinking systems : a study of the effects of ageing, pre-wetting and alternative pulping strategy on ink behaviour in pulping

484. Mäklin, Jani (2014) Electrical and thermal applications of carbon nanotube films

485. Niemistö, Johanna (2014) Towards sustainable and efficient biofuels production : Use of pervaporation in product recovery and purification

486. Liu, Meirong (2014) Efficient super-peer-based coordinated service provision

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-952-62-0403-1 (Paperback)
ISBN 978-952-62-0404-8 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)

UNIVERSITY of OULU
OULUN YLIOPISTO