KTH Computer Science
and Communication

Mining Speech Sounds

Machine Learning Methods for Automatic Speech Recognition and Analysis

GIAMPIERO SALVI

Doctoral Thesis
Stockholm, Sweden 2006

TRITA-CSC-A-2006:12
ISSN 1653-5723
ISRN KTH/CSC/A--06/12--SE
ISBN 91-7178-446-2

KTH School of Computer Science and Communication
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i datalogi fredagen den 6 oktober 2006 klockan 13.00 i F3, Sing-Sing, Kungl Tekniska högskolan, Lindstedtsvägen 26, Stockholm.

giampi@kth.se

Tryck: Universitetsservice US AB

Abstract

This thesis collects studies on machine learning methods applied to speech technology and speech research problems. The six research papers included in this thesis are organised in three main areas.

The first group of studies were carried out within the European project Synface. The aim was to develop a low latency phonetic recogniser to drive the articulatory movements of a computer-generated virtual face from the acoustic speech signal. The visual information provided by the face is used as a hearing aid for telephone users.

Paper A compares two solutions based on regression and classification techniques that address the problem of mapping acoustic to visual information. Recurrent Neural Networks are used to perform regression whereas Hidden Markov Models are used for the classification task. In the second case, the visual information needed to drive the synthetic face is obtained by interpolation between target values for each acoustic class. The evaluation is based on listening tests with hearing-impaired subjects, where the intelligibility of sentence material is compared in different conditions: audio alone, audio and natural face, and audio and synthetic face driven by the different methods.

Paper B analyses the behaviour, in low latency conditions, of a phonetic recogniser based on a hybrid of recurrent neural networks (RNNs) and hidden Markov models (HMMs). The focus is on the interaction between the time evolution model learned by the RNNs and the one imposed by the HMMs.

Paper C investigates the possibility of using the entropy of the posterior probabilities estimated by a phoneme classification neural network as a feature for phonetic boundary detection. The entropy and its time evolution are analysed with respect to the identity of the phonetic segment and the distance frombased on regression and classification techniques a reference phonetic boundary.

In the second group of studies, the aim was to provide tools for analysing a large amounts of speech data in order to study geographical variations in pronunciation (accent analysis).

Paper D and Paper E use Hidden Markov Models and Agglomerative Hierarchical Clustering to analyse a data set of about 100 millions data points (5000 speakers, 270 hours of speech recordings). In Paper E, Linear Discriminant Analysis was used to determine the features that most concisely describe the groupings obtained with the clustering procedure.

The third group belongs to studies carried out within the international project MILLE (Modelling Language Learning), which that aims at investigating and modelling the language acquisition process in infants.

Paper F proposes the use of an incremental form of Model-Based Clustering to describe the unsupervised emergence of phonetic classes in the first stages of language acquisition. The experiments were carried out on child-directed speech expressly collected for the purposes of the project.

Papers Included in the Thesis

The papers will be referred to by letters A through F.

Paper A:
Öhman, T. and Salvi, G. (1999) Using HMMs and ANNs for mapping acoustic to visual speech. *TMH-QPSR*, 1-2:45–50.

Paper B:
Salvi, G. 2006 Dynamic behaviour of connectionist speech recognition with strong latency constraints. *Speech Communication*, 48(7):802–818.

Paper C:
Salvi, G. (2006) Segment boundaries in low latency phonetic recognition. *Lecture Notes in Computer Science*, 3817:267–276.

Paper D:
Salvi, G. (2003) Accent clustering in Swedish using the Bhattacharyya distance. *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, 1149–1152.

Paper E:
Salvi, G. (2005) Advances in regional accent clustering in Swedish. *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, 2841–2844.

Paper F:
Salvi, G. (2005) Ecological language acquisition via incremental model-based clustering. *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, 1181–1184.

Author's Contribution to the Papers

Paper A:

T. Öhman developed the ANN method and performed the statistical analysis of the results, G. Salvi developed the HMM method, both authors participated in writing the manuscript.

Papers B, C, D, E, F:

The work was carried out entirely by the author, G. Salvi.

Other Related Publications by the Author

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Salvi, G., Spens, K.-E., and Öhman, T. (1999a). A synthetic face as a lip-reading support for hearing impaired telephone users - problems and positive results. In *Proceedings of the 4th European Conference on Audiology*, Oulo, Finland.

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Salvi, G., Spens, K.-E., and Öhman, T. (1999b). Two methods for visual parameter extraction in the Teleface project. In *Proceedings of Fonetik*, Gothenburg, Sweden.

Agelfors, E., Beskow, J., Granström, B., Lundeberg, M., Salvi, G., Spens, K.-E., and Öhman, T. (1999c). Synthetic visual speech driven from auditory speech. In *Proceedings of Audio-Visual Speech Processing (AVSP)*, Santa Cruz, USA.

Agelfors, E., Beskow, J., Karlsson, I., Kewley, J., Salvi, G., and Thomas, N. (2006). User evaluation of the synface talking head telephone. *Lecture Notes in Computer Science*, 4061:579–586.

Beskow, J., Karlsson, I., Kewley, J., and Salvi, G. (2004). SYNFACE - A Talking Head Telephone for the Hearing-impaired. In Miesenberger, K., Klaus, J., Zagler, W., and Burger, D., editors, *Proceedings of International on Conference Computers Helping People with Special Needs*.

Johansen, F. T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G. (2000a). The COST 249 SpeechDat multilingual reference recogniser. In *Proceedings of XLDB Workshop on Very Large Telephone Speech Databases*, Athens, Greece.

Johansen, F. T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G. (2000b). The COST 249 SpeechDat multilingual reference recogniser. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Karlsson, I., Faulkner, A., and Salvi, G. (2003). SYNFACE - a talking face telephone. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 1297–1300.

Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G. (2000). A noise robust multilingual reference recogniser based on SpeechDat(II). In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Salvi, G. (1998a). Developing acoustic models for automatic speech recognition. Master's thesis, TMH, KTH, Stockholm, Sweden.

Salvi, G. (1998b). Developing acoustic models for automatic speech recognition in Swedish. *The European Student Journal of Language and Speech*.

Salvi, G. (2003a). Truncation error and dynamics in very low latency phonetic recognition. In *Proceedings of Non Linear Speech Processing (NOLISP)*, Le Croisic, France.

Salvi, G. (2003b). Using accent information in ASR models for Swedish. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 2677–2680.

Salvi, G. (2005). Segment boundaries in low latency phonetic recognition. In *Proceedings of Non Linear Speech Processing (NOLISP)*.

Salvi, G. (2006). Segment boundary detection via class entropy measurements in connectionist phoneme recognition. *Speech Communication*. in press.

Siciliano, C., Williams, G., Faulkner, A., and Salvi, G. (2004). Intelligibility of an ASR-controlled synthetic talking face (abstract). *Journal of the Acoustical Society of America*, 115(5):2428.

Spens, K.-E., Agelfors, E., Beskow, J., Granström, B., Karlsson, I., and Salvi, G. (2004). SYNFACE, a talking head telephone for the hearing impaired. In *IFHOH 7th World Congress*.

Acknowledgements

This work has been carried out at the Centre for Speech Technology, supported by Vinnova (The Swedish Governmental Agency for Innovation Systems), KTH and participating Swedish companies and organisations. The work has also been supported by the Swedish Transport and Communications Research Board (now Vinnova) through the project Teleface, by the European Union through the project Synface,[1] and by The Bank of Sweden Tercentenary Foundation through the project MILLE (Modelling Language Learning). The Ragnar and Astrid Signeuls Foundation and the COST 277 Action (Nonlinear Speech Processing) have contributed to some of the travelling and conference expenses.

I would like to thank my supervisor *Björn Granström* for supporting my work and giving me the opportunity to focus my studies in the direction I was most interested in. For the generous feedback on my work, I am indebted to my co-supervisor *Kjell Elenius* and to *Mats Blomberg*. Their help and support have been decisive for the development of this work. *Rolf Carlson* and *Maria-Gabriella Di Benedetto* are responsible for having introduced me to the field of speech research, and to the warm atmosphere at the Department of Speech, Music and Hearing (TMH). I am indebted to *Bastiaan Kleijn* and *Arne Leijon* for stimulating discussions and courses, especially when they were still part of TMH.

A number of people at the department have contributed actively to my work and education. I am grateful to *Jonas Beskow* for being such an easy person to collaborate with, and for constant advice on Tcl/Tk programming; to *Kåre Sjölander* for sound and speech processing advice; to *Jens Edlund* for being my personal Regular Expression guru. *Håkan Melin* has always been ready to help with all sorts of questions, no matter if dealing with speech technology or the Unix file system. My admiration goes to *Kjell Gustafson* for his deep knowledge and devotion to language and for patiently answering all my naïve questions about phonetics. I wish to thank *Per-Anders Jande* and *Botond Pakucs* for sharing my search for the best way of doing things in LaTeX, and for the late nights at the department.

It has been most enjoyable to work with the Teleface and Synface project members, among which I wish to thank in particular, *Inger Karlsson*, *Eva Agelfors*, *Karl-Erik Spens*, and *Tobias Öhman*, at TMH, *Geoff Williams*, and *Cathrine Siciliano*, at University College London, and *Jo Kewley* and *Neil Thomas* at the Royal National Institute for Deaf People, United Kingdom.

Within the project MILLE, I am most grateful to *Francisco Lacerda* and *Björn Lindblom* at the Department of Linguistics at Stockholm University for their enthu-

[1] IST-2001-33327 `http://www.speech.kth.se/synface`

siasm and devotion to the study of language, and for useful comments and warm encouragement on my work. I wish to thank also the rest of the group at the Department of Linguistics, and in particular *Lisa Gustavsson*, *Eeva Klintfors*, and *Ellen Marklund* for great discussions in a relaxed and friendly atmosphere.

A special thanks goes to *Rebecca Hincks* and *Steven Muir* for proofreading my writings. If there are parts of this manuscript that sound closer to Italian than English, it is most probably because I did not give them a chance to comment on them. Thanks to *Christine Englund* for showing me how unstructured my writing often is. If many parts of this thesis are readable, it is to a large extent due to her precious comments.

I feel privileged for the chance I have been given to experience the atmosphere at the Department of Speech, Music and Hearing. I wish to thank all the members of the department who contribute to this stimulating and relaxed environment, and in particular *Roberto Bresin* for being a big brother to me, and helping with any kind of advice when I still was a disoriented student in a foreign country; *Sofia Dahl* for her never-ending patience and support when I was trying to speak Swedish, and for all the nice memories together; *Kjetil Falkenberg Hansen* for the coffee pauses, for the lumberjack activities in the Stockholm archipelago, and for introducing me to the music of Samla Mammas Manna, among a thousand other things. *Svante Granqvist* for his brilliant pedagogic skills and for always having a pen with him at lunch breaks to explain the most obscure signal processing concepts on a paper napkin. *Loredana Cerrato Sundberg* for her passion for the Italian national football team and for bringing a warm breeze from southern Italy all the way up to Scandinavia. My last two office-mates *Anna Hjalmarsson* and *Preben Wik* for creating a lively and stimulating environment in our working hours. No activity at the department could be possible without the valuable work of *Cathrin Dunger*, *Caroline Bergling*, *Markku Haapakorpi*, and *Niclas Horney*.

A number of guests have visited the department while I was here. I wish to warmly remember *Leonardo Fuks* for trying to teach me play trumpet and for turning the lunch room at the department into a Brazilian stadium during the World Cup final 1998; *Philippe Langlais* for great advice when I was new to speech research and I needed advice most; *Carlo Drioli* for the time together spent trying to solve course assignments, and for his enthusiasm for Frank Zappa's music; *Werner Goebl* for playing the piano on my late evenings at work, for his hospitality in Vienna, and for organising a trip to Öland when I was in deep need of a vacation; *Bruno L. Giordano* for great discussions on statistics and for his "gnocchi". Sento che un giorno saremo di nuovo amici.

Free software[2] has been central to my work: all experiments and writing have been performed on GNU[3] Linux systems. The pieces of software that have constituted my daily tools are: Emacs,[4] for anything that involves typing (from writing

[2] `http://www.fsf.org/`
[3] `http://www.gnu.org/`
[4] `http://www.gnu.org/software/emacs/`

e-mails to programming); LaTeX,[5] what you see is what you mean; XFig,[6] for all sorts of schematics and illustrations; Gimp,[7] when you really cannot avoid bitmaps; Perl,[8] text processing has never been this easy; Tcl-Tk,[9] from small scripts to full-featured applications; R,[10] for the statistical analysis, and for being in many respects a great substitute to Matlab®. I am really grateful to all the people and communities that have worked at developing this software. Thanks to *Jonas Beskow* and *Kåre Sjölander* for WaveSurfer[11] and Snack[12] that have been valuable tools for visualising and processing speech. I would like to also thank *Børge Lindberg*, *Finn Tore Johansen*, and the other members of the COST 249 Action for the SpeechDat reference recogniser software,[13] that has simplified my work in many respects. *Nikko Ströms* has done an outstanding job in developing the NICO Toolkit.[14] I am grateful for his recent decision to release NICO under a BSD licence. This work would have taken a much longer time without the Hidden Markov Model Toolkit (HTK).[15]

I would like to thank the co-founders of SynFace® AB, *Jonas Beskow*, *Per Junesand*, and *Pål Ljungberger*, for their enthusiasm, and for shouldering most of the company work while I was finishing this thesis.

More on a personal basis, there are a (large) number of people I would like to thank. Even though they have not directly contributed to the writing of this thesis (but there are exceptions), they have been as a big family to me, making my time in Sweden most enjoyable. These people have left a permanent sign in my memories: *Addisu*, *Anna*, *Beatriz*, *Brindusa*, *Catalina*, *Christoph*, *Francesco*, *Hal*, *Henry*, *Jessica*, *Johan*, *Lill-Ann*, *Maria*, *Miriam*, *Pablo*, *Paola*, *Pavla*, *Roope*, *Sanjoo*, *Shane*, *Shirin*, *Simone*, *Steven*, *Taneli*, *Tom*, *Veera*, and *Yolanda*.

The final period of my studies would have certainly been unbearable without *Christine*'s loving help and support. Thank you for being such an understanding, sweet and smart person.

Finally, I would like to thank my family: my parents Angelo and Fiorenza, my sister Raffaella, and my nephews Giovanni and Federica, for having supported me, and for having borne the distance. Dubito di essere stato in grado di esprimere quanto mi siete mancati in questi anni.

[5] http://www.latex-project.org/
[6] http://www.xfig.org/
[7] http://www.gimp.org/
[8] http://www.perl.com/
[9] http://www.tcl.tk/
[10] http://www.r-project.org/
[11] http://www.speech.kth.se/wavesurfer/
[12] http://www.speech.kth.se/snack/
[13] http://www.telenor.no/fou/prosjekter/taletek/refrec/
[14] http://nico.sourceforge.net/
[15] http://htk.eng.cam.ac.uk/

Symbols

$\mathbb{R}$ the field of real numbers

$\mathbb{R}^N$ Nth dimensional space on the field of real numbers

$P(A)$ probability of the event A

$p(x)$ probability density function (PDF) of the variable x

$D(x) = \int_{-\infty}^{x} p(t)dt$ cumulative probability distribution of the variable x

μ vector of means

Σ covariance matrix

$\mathbf{x}^T$ transpose of the vector $\mathbf{x}$

Abbreviations

ANN Artificial Neural Network

ASR Automatic Speech Recognition

BIC Bayes Information Criterion

EM Expectation Maximisation

GMM Gaussian Mixture Model

HMM Hidden Markov Model

HFCC Human Factor Cepstral Coefficients

IPA International Phonetic Alphabet

LDA Linear Discriminant Analysis

LP Linear Prediction

LPA Linear Prediction Analysis

LPC Linear Prediction Coefficients

MFCC Mel Frequency Cepstral Coefficients

MLP Multi-Layer Perceptron

NN Neural Network

PDF Probability Density Function

PLP Perceptual Linear Prediction

PME Perceptual Magnet Effect

RNN Recurrent Neural Network

SAMPA Speech Assessment Methods Phonetic Alphabet

TD(A)NN Time-Delayed (Artificial) Neural Network

Contents

Part I

Introduction

Chapter 1

Outline

Spoken language is, in many situations, the most natural and effective means of communication between humans. All aspects of human activity can be conveniently encoded into spoken utterances and interpreted, often effortlessly, by the listener. Several concurrent messages can be transmitted through a number of different channels. The messages can be linguistic, where a concept is formulated into a sequence of utterances, or paralinguistic, where additional information related to the feelings or intentions of the speaker is encoded using a complex mixture of acoustic and visual cues. The channels can be acoustic, such as the phonetic and prosodic channel, or visual, including speech-related movements and gestures.

The natural ease with which we carry out spoken conversations masks the complexity of language. Diversity in language emerges from many factors: *geographical*, with thousands of languages and dialects; *cultural*, because the level of education strongly influences the speaking style; *physical*, because everyone's speech organs have slightly different shapes; *psychological*, because each person can assume different speaking styles depending on her attitude, emotional state, and intention. This complexity has attracted researchers for centuries, and numerous aspects of language and speech have been described in detail. When trying to build computational models of speech communication, however, many questions are still unanswered. Automatic speech recognition and speech synthesis are still far from imitating the capabilities of humans.

The availability, in the last decades, of large data collections of spoken and written language has opened new opportunities for the speech and language communities, but, at the same time, has implied a challenge, as the task has shifted from the analysis of few examples collected in laboratory conditions, to the study of how language is used in the real world. Machine-learning methods provide tools for coping with the complexity of these data collections.

There are three main areas of applications of these methods. In applied research, they can be used to develop models of astounding complexity that can perform reasonably well in speech recognition and synthesis tasks, despite our incomplete

understanding of the human speech perception and production mechanisms. Machine learning can also be used as a complement to standard statistics to extract knowledge from multivariate data collections, where the number of variables, the size (number of data points), and the quality of the data (missing data, inaccurate transcriptions) would make standard analysis methods ineffective. Finally these methods can be used to model and simulate the processes that take place in the human brain during speech perception and production.

This thesis contains empirical investigation that pursue the study of speech-related problems, from all the above perspectives.

The first group of studies (Paper A–C), was motivated by a practical application and was carried out within the Swedish project Teleface and, subsequently, within the European project Synface. The aim was to develop a low-latency phonetic recogniser to drive the articulatory movements of a computer-generated virtual face from the acoustic speech signal. The visual information provided by the face is used as hearing aid for people using the telephone.

In the second group of studies (Paper D and E), the aim was to provide tools for analysing large amounts of speech data in order to study geographical variations in pronunciation (accent analysis).

The third and last group (Paper F) was carried out within the international project MILLE (Modelling Language Learning), which aims at studying and modelling the language acquisition process in infants.

More in detail, Paper A compares two conceptually different methodologies for mapping acoustic to visual speech. The first attempts to solve the regression problem of mapping the acoustic features to time-continuous visual parameter trajectories. The second classifies consecutive acoustic feature vectors into acoustic categories that are successively converted into visual parameter trajectories by a system of rules.

A working prototype of the system was later developed making use of speech recognition techniques that were adapted to the characteristics of the task. Recurrent multilayer neural networks were combined with hidden Markov modes to perform phoneme recognition in low-latency conditions. This combination of methods gave rise to the studies in Papers B and C. Paper B analyses the interaction between the model of time evolution learned by the recurrent and time-delayed connections in the neural network, and the one imposed by the hidden Markov model. Paper C investigates the properties of the entropy of the posterior probabilities, as estimated by the neural network of Paper B, with respect to the proximity to a phonetic boundary.

Paper D and Paper E contain two variants of a semi-supervised procedure to analyse the pronunciation variation in a Swedish telephone database containing more than 270 hours of recordings from 5000 speakers. Automatic speech recognition methods are used to collect statistics of the acoustic features of each phoneme's accent-dependent allophones. Hierarchical clustering is used to display the differences of the resulting statistical models. Similar allophones are grouped together according to an information theoretical measure of dissimilarity between probab-

ility distributions. Linear discriminant analysis is used to determine the features that best explain the resulting groupings.

Paper F contains a preliminary study that explores the possibility of using unsupervised techniques for modelling the emergence of speech categories from the typical acoustic environment infants are exposed to.

The thesis is organised as follows. Two short introductory chapters summarise the aspects of speech research (Chapter 2) and machine learning (Chapter 3) that are relevant to the thesis. These were included given the multidisciplinary nature of the studies, but are necessarily incomplete and simplified descriptions of the fields. Many references are provided for readers that want to further their understanding of the subjects. Chapter 4 describes the problem of mapping acoustic to visual information in speech, and the results obtained in Paper A–C. Chapter 5 describes the analysis of pronunciation variation in a Swedish database, and the results obtained in Paper D and E. Chapter 6 describes an attempt to model the emergence of acoustic speech categories in infants (Paper F). Finally, a discussion and summary of the thesis is presented in Chapter 7.

Chapter 2

The Speech Chain

This chapter is a brief introduction to speech production, perception, and analysis. It is intended for those with a background in signal processing that are not familiar with speech and voice research and it is limited to those concepts that are relevant to this thesis. The reader is referred to the literature for a more complete review.

An illustration from Denes and Pinson (1993) provides a good map to help navigate through the speech communication mechanism. Figure 2.1 shows a slightly modified version of the picture. The figure depicts two persons, a speaker and a

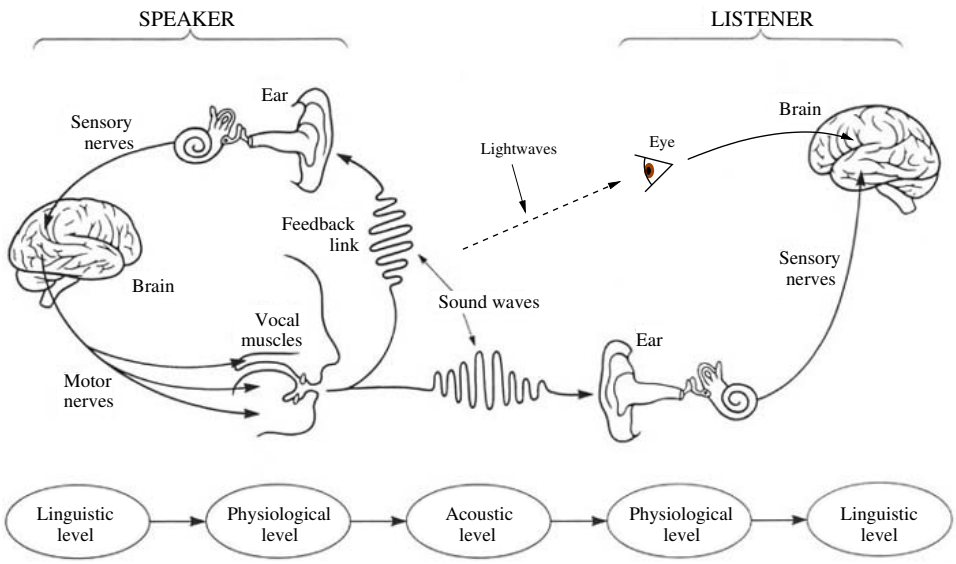


Figure 2.1: The Speech Chain from Denes and Pinson (1993), modified to include the visual path.

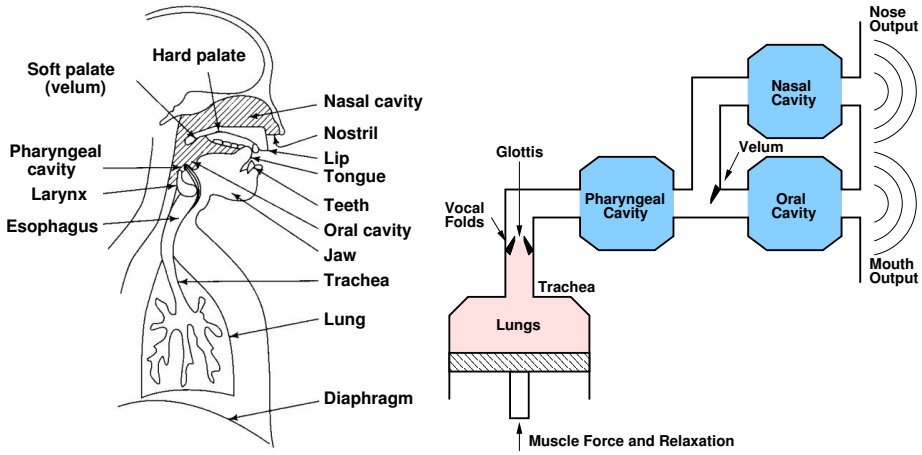


Figure 2.2: Speech production organs and a block diagram depicting their function. From Rabiner and Juang (1993).

listener. We limit the description to the *one-way* speech communication path. However, it is worth noting that many aspects of human-to-human interaction concern the way the speaker's and listener's roles are exchanged, and this picture would be misleading if the focus was on dialogue and dialogue systems research.

At the first level (linguistic), the speaker formulates ideas into language and translates the internal representation of the corresponding utterances into neural signals that control the muscles involved with speech production through the *motor nerves*. The resulting movements (physiological level) of the speech organs produce and modulate the air flow from the lungs to the lips. The modulation of the air flow produces sound waves (acoustic level) that propagate through the air and eventually reach the ears of both the listener and the speaker. At this point, an inverse process is performed. The ear transduces the acoustic signal into neural signals that travel to the brain through sensory nerves. The brain reconstructs the linguistic information with processes that are to a large extent still unknown. In the speaker this information is used as a feedback path to tune the production mechanism.

A visual path has been added to the original figure, to account for the multimodal nature of speech communication, an aspect relevant to the thesis. A number of visual events are used by the listener as a complement to the acoustic evidence during speech understanding, such as facial expressions and gestures of various kinds. The perhaps most important visual information in face-to-face communication comes from articulator movements. Lip reading has been shown to constitute a strong aid to hearing and provides information that is sometimes orthogonal to the acoustic information.

2.1 Speech Production

The processes involved in transforming the neural signals into acoustic sounds are described in this section. The left part of Figure 2.2 shows the organs involved in speech production. The right part of the same figure is a block diagram depicting the functional aspects of each part of the vocal system.

An extensive description of the physiology of speech production can be found in Titze (1994); Rabiner and Schafer (1978); Quatieri (2002). Here it suffices to say that, from the functional point of view, the speaker has control, among other organs, over:

- the air pressure in the lungs
- tension and extension of the vocal folds
- opening of the velum
- position and configuration of the tongue
- opening of the jaw
- protrusion and rounding of the lips

With the exception of click sounds typical of some African languages, speech is produced by inducing a pressure in the lungs, which excites a constriction at the glottis or along the vocal tract. The oral cavities modify the sounds produced at the constriction point by enhancing the contributions close to their resonance frequencies and damping the rest of the spectrum. Depending on the configuration of the articulators along the vocal tract, the resonance frequencies can be varied.

The Source/Filter Model

A simplified but accurate model of these processes is the Source/Filter model (Fant, 1960). The sound generation at the constriction point, and the modulation by the vocal cavities are considered to be independent (Figure 2.3).

The source assumes different states (voiced, fricative, plosive) depending on the place and kind of constriction. These are depicted in the left part of Figure 2.3.

In *voiced* sounds, the constriction is formed by the vocal folds that, due to their elastic characteristics, the lung pressure, and aerodynamic forces, start oscillating and produce a pulse-shaped air flow signal. The dynamics of vocal fold oscillation is a complex phenomenon and have generated a whole field of research (voice research). A popular model of the glottal airflow signal (Fant et al., 1985) models the derivative of the airflow (thus incorporating the effect of radiation at the lips) with a piecewise function controlled by four parameters. An example of flow obtained with the so called Liljencrants-Fant (LF) model is shown in Figure 2.4, where both the glottal flow and its derivative are plotted for three periods of oscillations. Usually, the source signal for voiced sounds is modelled by a Dirac impulse train, followed by a

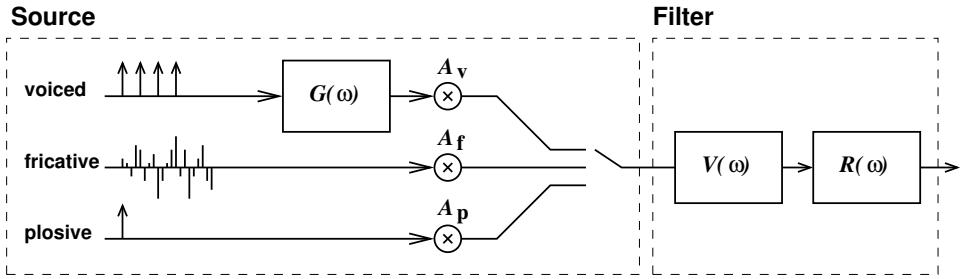


Figure 2.3: Source-Filter model of speech production

filter $G(\omega)$ that reproduces the time evolution of the glottal flow (Figure 2.3). Al vowels are voiced sounds. Additionally, vocal fold oscillations are present in voiced consonants, *e.g.* /v/ (compared to /f/) or /b/ (compared to /p/).

Fricative sounds are produced when the constriction is small enough to produce turbulence in the airflow, but the tissues that compose the constriction do not have the physical characteristics to produce oscillations. For example, whispering sounds are produced with a constriction at the glottis, but with a glottal configuration that does not induce vocal fold oscillations. More commonly, fricative sounds are produced with constrictions along the vocal tract. The source, in fricative sounds, is noise-like. Examples of these sounds are /f/ as in "fin", /s/ as in "sin", /ʃ/ as in "shin".

Finally, the constriction can be complete causing pressure to build up in the preceding cavities. At the release a *plosive* sound, such as /p/ or /b/, is produced.

Note that there may be more than one source: in voiced fricatives, for example, there is both turbulence noise at the place of articulation, and excitation from the vocal folds. This explains also why some of the examples above appear in two different categories. For example /b/ is both a voiced sound and a plosive, and /v/ is both a voiced sound and a fricative. Examples of speech sounds in the different categories for Swedish and British English are given in Appendix X.

In each state the source is characterised by a number of parameters: in the case of voiced sounds, for example, the frequency at which the vocal folds oscillate, denominated f_0, is a function of time.

The filter parameters are determined by the configuration of the vocal cavities, *i.e.*, essentially by the velum, the configuration of the tongue, the position of the jaw, and the configuration of the lips. Another factor that determines the filter is the position of the source in the mouth: for example, the resonances in the oral cavity introduce poles in the transfer function for back fricatives excited near the velum, while they act as energy sinks and introduce zeros for front fricatives that are generated near the lips.

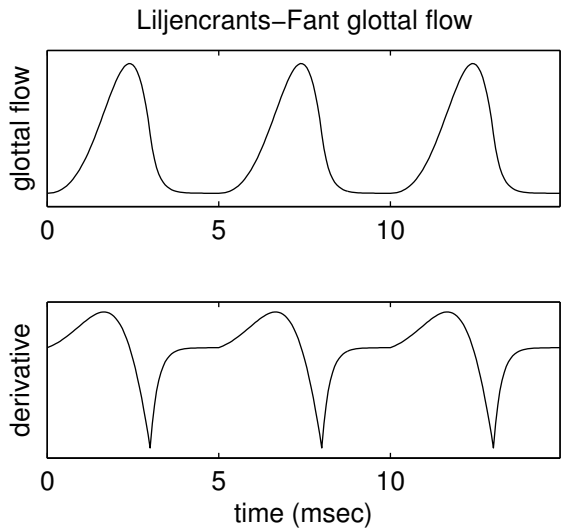


Figure 2.4: Glottal airflow and its derivative obtained with the Liljencrants-Fant model with the following parameters: period 5 msec (fundamental frequency $f_0 =$ 200 Hz), open phase = 0.6, +ve/-ve slope ratio = 0.1, closure time constant/close phase = 0.2.

Solving the wave equation in the vocal tract with a number of simplifying assumptions results in a linear system with transfer function $V(\omega)$ having a number of poles and zeros corresponding to the resonance frequencies of the cavities that are, respectively, in series or in parallel to the path from the source to the radiation point. An additional zero corresponds to the closing condition at the lips, usually referred to as radiation impedance $R(\omega)$.[1]

Both the source and the filter parameters vary continuously in time. The rate of variation is usually considered to be slow when compared to the fluctuations of the acoustic signal. In practise, when modelling the speech signal in statistical terms, short segments of speech are considered to be drawn from a stationary process.

Finally, Figure 2.5 shows an example of a voiced sound at different points in the Source/Filter model. Both the waveform and the spectrum are shown. The impulse train $i[n]$ is filtered successively by the functions $G(\omega)$, $R(\omega)$, and $V(\omega)$. The glottal flow filter $G(\omega)$ is obtained from the LF model. The vocal tract filter $V(\omega)$ is an all-pole model of order 8, obtained with typical values for the first four formants and bandwidths of the Swedish vowel [ɛ].

[1]The contribution of the radiation at the lips is often encompassed in the glottal filter $G(\omega)$, as already noted with regard to the LF model of glottal flow.

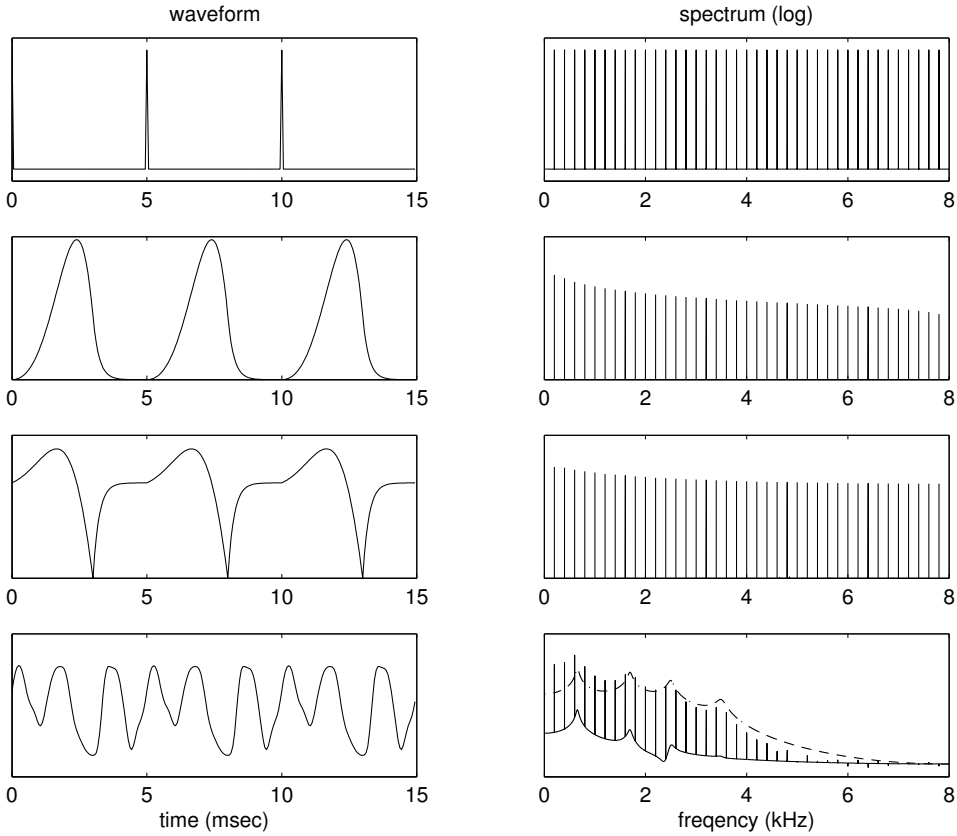


Figure 2.5: Example of a voiced sound at different points in the Source/Filter model. The left hand plots show three periods of the signal, right hand plots show the corresponding log spectrum. From above: 1) impulse train $i[n]$, 2) glottal flow $i[n] \star g[n]$, obtained when filtering $i[n]$ with the glottal shape filter $G(\omega)$ 3) derivative of the glottal flow, can be seen as $i[n] \star g[n] \star r[n]$, *i.e.*, the convolution with the radiation function $R(\omega)$, and 4) after the vocal tract filter: $i[n] \star g[n] \star r[n] \star v[n]$. Source parameters as in Figure 2.4. $V(\omega)$ is an all-pole model of order 8 with formants (bandwidths) in Hz at 654 (50), 1694 (75), 2500 (100), 3500 (150) simulating the vowel [ɛ]. The transfer function of $V(\omega)$ is indicated in the last plot by a dashed line.

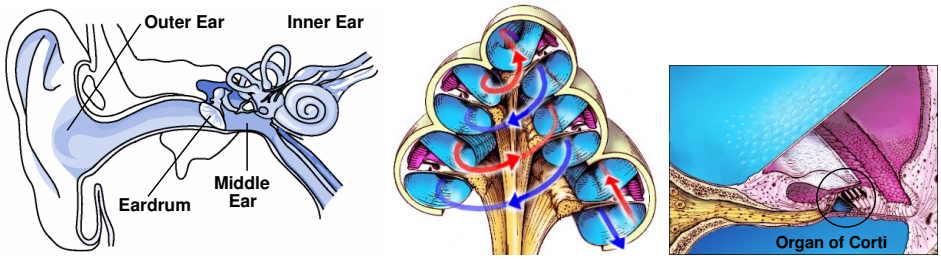


Figure 2.6: Hearing organs, drawings by S. Blatrix (Pujol, 2004), Left: outer, middle and inner ear. Middle: section of the cochlea. Right: detail of a section of the cochlea with organ of Corti.

2.2 Speech Perception

Moving forward along the Speech Chain, the ear of the listener transduces the acoustic wave into neural signals. Many details of this process are still not known. Again, we will describe briefly the results that are relevant to this thesis.

The outer and the middle ear (Figure 2.6, left) work as an impedance adaptor between the acoustic wave travelling in the air and the wave travelling in the inner ear. The inner ear consists of a spiral-formed organ called *cochlea* (Figure 2.6, middle and right), which acts as a pressure-to-neural activity transducer.

Leaving aside nomenclature and details, the cochlea is divided longitudinally into three areas by two membranes (Figure 2.6, right). The two outer areas are filled with incompressible fluid, whereas the inner area hosts the hair cells that stimulate the auditory nerves.

Stimulation of the cochlea by the small bones in the middle ear (ossicles) produces a wave on the thicker of the two membranes (basilar membrane) that propagates along the spiral. Different locations along the basilar membrane have different resonance frequencies, *i.e.*, a sinusoidal stimulus at a certain frequency corresponds to an oscillation that is spatially localised along the membrane. This behaviour is the basis for the ability of the ear to perform a spectral analysis of the incoming signals.

The discrimination of two pure tones is related to the *critical bands* that correspond, for a given frequency f^*, to the range of frequencies around f^* that activate the same part of the basilar membrane. These bands correspond to the same geometrical length in the cochlea, but are nonlinearly mapped in frequency. At high frequencies the discrimination is poorer than at low frequencies. This phenomenon is reflected into the perception of pitch, where test subjects assign the same perceived distance to tones that are further apart at higher frequencies. Stevens et al. (1937) propose a scale based on the perceived pitch called mel scale. The value of 1000 mel is arbitrarily assigned to 1000 Hz. For lower frequencies, the mel and frequency scales coincide. Above 1000 Hz the relationship between mel and fre-

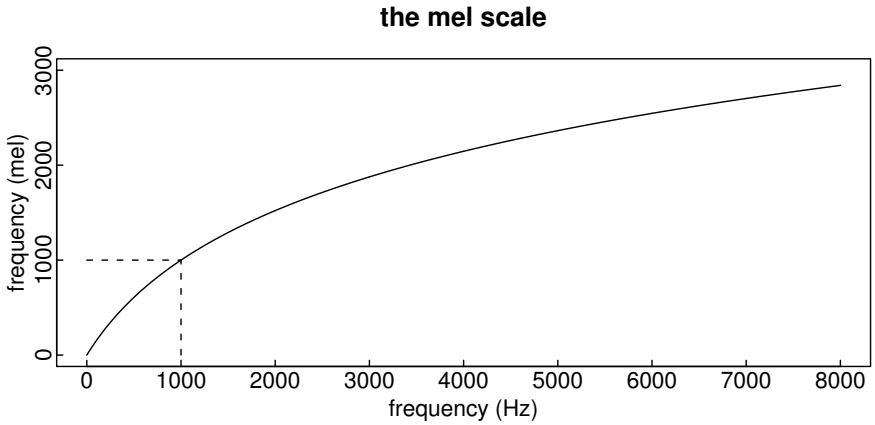


Figure 2.7: The mel scale

quency scale is logarithmic. A common approximation of the mel scale is defined by Stevens and Volkman (1940) as:[2]

$$\text{Mel}(f) = 2595.0 \log_{10}(1 + \frac{f}{700})$$

where f is the frequency in Hz. The mel scale is plotted in Figure 2.7.

The acoustic stimuli that reach the different locations along the basilar membrane are transduced into neural activities by the so called *air cells* in the organ of Corti (see right of Figure 2.6). The air cells fire random neural impulses at a rate that is proportional to the (rectified) pressure signal.

The signals that reach the brain contain both information about the frequency of the acoustic stimulus (carried by the group of nerve fibres that are activated) and temporal information (carried by the time evolution of the density of firings in each nerve fibre).

This gives rise to two groups of models of sound perception. For the first group, called *place models*, sound analysis is entirely based on the position along the basilar membrane from which the neural activities are originated. In practise it states that the brain performs a spectral analysis of the sounds. The second group, called *temporal models*, instead, states that the brain uses the temporal evolution of the firing patterns in each auditory fibre, to characterise sounds.

Most of the processes involved with speech perception are, however, not entirely understood. It is not known, for example, whether and how the sensitivity of the neural transducers along the basilar membrane can be actively controlled, and what

[2]The multiplicative constant should be changed to 1127.0 if the natural logarithm is used.

information from the auditory nerves (place, temporal or both) is used to form the percepts in the higher areas of the auditory cortex.

What most models seem to agree on is that the amplitude spectrum is central for mono-aural perception, whereas, in binaural mode, phase information may be used as well, for example to extract the relative direction between the listener and the sound source.

The Visual Channel

As mentioned in the introduction, the visual channel is a path in the speech communication chain that is relevant to this thesis. When possible, the visual information related to the speaker is integrated with the acoustic information to form a more robust percept. The listener uses, *e.g.*, facial expressions and other gestures to extract paralinguistic information, such as the mood/intention of the speaker, and signals related to turn taking (*i.e.*, when the other person can suitably be interrupted). Information that is more directly coupled to the linguistic content of the speech act is contained in the movements of the visible speech organs: essentially the lip and tongue tip movements and the jaw opening.

In the same way that the acoustic signal is categorised into phonetically relevant classes called phonemes, the visual signal is classified into *visemes*. In speech perception, the information from the acoustic and visual channels is integrated to support one of the possible hypotheses, given the context. This process is exemplified by the McGurk effect (McGurk and MacDonald, 1976) that takes place when inconsistent visual and auditory information is presented to the listener. If, *e.g.*, an auditory /ba/ is combined with a visual /ga/, a /da/ is often heard, suggesting that, among the possible hypotheses, the one that is most consistent with both the auditory and visual information, is chosen.

2.3 Speech Analysis and Acoustic Features

The methods for speech signal analysis are strongly adapted to the concepts of speech production and perception described above.

If we describe the speech signal with an all-pole model (*i.e.*, disregarding the zeros coming from the physics of the production mechanism described above), a powerful analysis tool is *Linear Prediction* (LP). The idea is that, if the system is all poles, the output signal $s[n]$ can be computed as the input signal $u[n]$ plus a weighted sum $\tilde{s}[n]$ of the past samples of the output signal. For voiced sounds, the input signal is modelled as a train of Dirac pulses.[3] Plosive sounds have a single pulse and fricatives have white noise as source signal. In all cases, the model parameters are obtained by minimising the energy of the difference between $s[n]$ and $\tilde{s}[n]$.

[3]provided that the form of the glottal pulse is considered to be part of the filter transfer function.

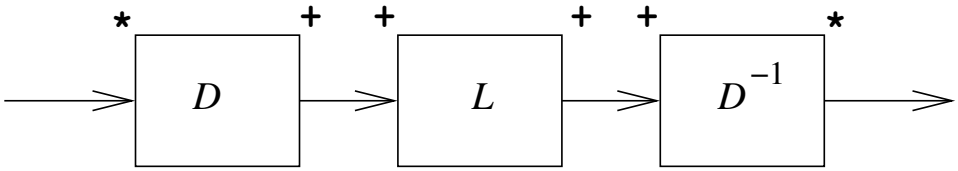


Figure 2.8: Motivation for Cepstral analysis: homomorphic filtering

A limit of Linear Prediction analysis is that zeros in the transfer function cannot be directly modelled, only approximated by a number of poles. Increasing the number of poles reduces the approximation error, but also the accuracy in the model parameter estimation.

A different way to look at the problem is to observe that the signal we try to analyse is a convolution of a number of terms corresponding to the excitation, the glottal shaping filter, and the vocal and nasal tract filtering. An analysis method that can perform deconvolution and does not suffer from the zero-pole limitation of Linear Prediction is Cepstral Analysis. This method is based on the concept of homomorphic filtering.

Homomorphic systems use nonlinear transformations D (see Figure 2.8) to transform the convolution operation into a sum. The convolutive components of the signal correspond, in the transformed domain, to additive components. In the transformed domain, we can use standard linear methods, such as spectral analysis, to analyse the different components of the signal. Moreover, linear filters L can be used to separate the components if these occupy different spectral intervals in the transformexyd domain. This operation corresponds to a deconvolution in the original domain.

We can find a homomorphic transformation when we notice that the spectrum of the convolution of two signals $x_1[n]$ and $x_2[n]$ is the product of the respective spectra, and that the logarithm of a product of two terms is a sum of the logarithm of the respective terms. We have:

$$\begin{array}{l} D \downarrow \\ \\ \\ \end{array} \left| \begin{array}{lcl} x[n] & = & x_1[n] \star x_2[n] \\ X(\omega) & = & X_1(\omega) X_2(\omega) \\ \log X(\omega) & = & \log X_1(\omega) + \log X_2(\omega) \end{array} \right| \begin{array}{l} \uparrow \\ \\ \\ D^{-1} \end{array}$$

If the aim is to eliminate a convolutive term, *e.g.*, $x_2[n]$, we can use a filter in this domain and then use the inverse transform D^{-1} to obtain an estimation $\hat{x}_1[n]$ of the signal $x_1[n]$ alone. If the aim is to analyse the different convolutive components, we do not need to reconstruct the time varying signal, but simply perform a Fourier analysis in the transformed domain.

original	derived
spectrum	cepstrum
frequency	quefrency
harmonics	rahmonics
magnitude	gamnitude
phase	saphe
filter	lifter
low-pass filter	short-pass lifter
high-pass filter	long-pass lifter

Table 2.1: Terminology in the frequency and quefrency domains

This is the definition of the Complex Cepstrum: the inverse Fourier transform of the logarithm of the Fourier transform of the signal.

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log(X(\omega)) e^{j\omega n} d\omega$$

If we are not interested in the phase term, we can simplify the complex spectrum $X(\omega)$ in the formula with its modulus $|X(\omega)|$. The result is called Real Cepstrum. Because the function $|X(\omega)|$ is even for real signals, and the logarithm does not change this property, the inverse Fourier transform can be simplified to the cosine transform. Disregarding the phase information is in agreement with models of mono-aural perception, where the amplitude spectrum seems to play the most important role.

The cepstrum has been used in many applications in which the deconvolution of a signal is necessary, *e.g.*, the recovery of old recordings by eliminating the characteristic of the recording channel. To stress the fact that, in the transformed domain, the independent variable does not correspond to time, a new terminology has been introduced, following the way the term cepstrum was coined, *i.e.*, by reversing the first syllable, examples of the terms used in the cepstrum domain are in Table 2.1.

The most commonly used features for speech recognition are the Mel Frequency Cepstrum Coefficients (MFCC, Davis and Mermelstein, 1980). The MFCCs are obtained by applying the cosine transform to the log energies of the outputs of a filterbank with filters regularly positioned along the mel frequency scale. The resulting coefficients can be liftered in order to equalise their range that can vary from low to high order coefficients. The principal advantage of the cepstral coefficients for speech recognition is that they are in general decorrelated, allowing the use of simpler statistical models.

Alternatives to the MFCCs exist that include more knowledge about speech perception. An example is Perceptual Linear Prediction (PLP, Hermansky, 1990; Junqua et al., 1993), which weights the output of the filterbanks by an equal-loudness curve, estimated from perceptual experiments. Human Factor Cepstral

Coefficients (HFCC, Skowronski and Harris, 2004), instead, make use of the known relationship between centre frequency and critical bandwidth obtained from human psychoacoustics, to define a more realistic filterbank. Many studies that compare different features for speech recognition can be found in the literature (Jankowski et al., 1995; Schukat-Talamazzini et al., 1995; Eisele et al., 1996; Nicholson et al., 1997; Batlle et al., 1998; Saon et al., 2000)

2.4 Speech Data

Speech data is usually organised in databases. The recordings are generally carried out in sessions during which a single speaker reads sentences or produces more spontaneous utterances. Short and often unconnected utterances are recorded in a sequence. Depending on the task the database is designed for, more or less realistic scenarios are simulated in the recording sessions.

Databases intended for automatic speech recognition applications are usually divided into different sets of recordings. A *training set* is devoted to building classification models, while a *test set* is dedicated to verifying their quality. The difficulty of the classification task is directly related to the degree of mismatch between the training and test data. The identity of the speakers is one of the factors that usually differentiates training and test set, at least in databases built for speaker independent speech recognition. The orthographic content is also varied for each session of recordings. However, factors as the kind of utterances, the modality of speech production (usually read speech) and the recording conditions (channel and background noise), are often homogeneous withing each database. In these respects, the degree of variation within a database is, therefore, small if compared to the richness of language.

The collection of vast recordings of spontaneous everyday conversations has demonstrated the limit of recognition systems based on words represented as strings of phonemes. Many other aspects, such as the richness in pronunciation variation for each words, the frequent use of corrections and the use of prosody and non linguistic sounds to convey meaning, have begun to be the focus of speech research.

In this thesis, most of the studies are based on data from the Swedish SpeechDat FDB5000 telephone speech database (Elenius, 2000). The database contains utterances spoken by 5000 speakers recorded over the fixed telephone network. All utterances are labelled at the lexical level and a pronunciation lexicon is provided. The database also contains information about each speaker, including *gender*, *age*, and *accent*, and more technical information about the recording, for example the type of telephone set used by the caller.

Additional experiments, often not described in the publications included in the thesis, have been carried out with SpeechDat databases in other languages (English, Flemish) and on the TIMIT American English database. TIMIT and SpeechDat are different in many respects: a) the language (British/American English), b) the quality of the recordings (8kHz, A-law for SpeechDat and 16kHz, 16bit linear for

TIMIT), c) the sentence material, and, perhaps most importantly, d) the detail of the transcriptions. In SpeechDat, only orthographic transcriptions are available, whereas TIMIT provides detailed phonetic transcriptions. This makes the phonetic classification task very different in the two cases: phonetic models built on SpeechDat with the help of a pronunciation dictionary are rich with allophonic variants, and thus closer to phonemic models. In TIMIT, each model represents acoustic realisations (phones) that are more homogeneous. These models are, however, more difficult to use in a word recognition task, because they need accurate pronunciation models to be specified for each phrase.

Another kind of material used in our studies has been collected within the project MILLE (Lacerda et al., 2004a). These recordings are specifically designed to study the interaction between parents and children in their first months. Most of the data consists of child-directed speech by a number of mothers. The specific way that voice and language are used in this situation constitutes the distinctive feature of these recordings. As a reference, the parents are also recorded when speaking with adults.

Chapter 3

Machine Learning

In the last decades, two phenomena have determined the emergence of a new research field. First, the drop of costs involved in electronically collecting and storing observations of the world has brought the need for sophisticated methods to handle the resulting data collections. Both the dimensionality of the data and the number of data points have challenged traditional statistics. Secondly, the study of human behaviour and of human perception (vision, hearing...) has shifted from a descriptive to a quantitative and functional perspective, and researchers have engaged themselves in a challenge to reproduce these abilities in artificial systems.

Data Mining aims at dealing with the first issue by extracting a summarised description of large sets of multivariate observations, to allow for meaningful interpretations of the data.

The second issue is mostly addressed by studies in *Pattern Classification* or *Recognition*, which aim to classify a set of data points into a number of classes. The classes may be predefined according to prior information, or inferred from the data itself.

These two fields are widely overlapping and share the same theoretical foundations. In the following the term Pattern Classification will be used to refer to both areas of investigations, while it should be kept in mind that the *unsupervised learning* methods described in this chapter, are often referred to in the literature as Data Mining methods.

Pattern Classification belongs to a subfield of Artificial Intelligence called *Machine Learning*, that focuses on finding methods for automatically extracting knowledge from sets of observations (so called *inductive learning*). Machine Learning covers a wide range of subjects from theoretical studies aimed at finding general philosophical implications of the concept of learning to more practical studies that try to solve specific learning problems. In all cases, however, the aim is the development of artificial systems (or algorithms) that can improve automatically through experience.

This chapter describes the concepts and the methods that are relevant to this thesis; for a complete account of Machine Learning, the reader is referred to the following books: Duda et al. (2001); Cristianini and Shawe-Taylor (2001); Schölkopf and Smola (2002); Arabie et al. (1996); Gordon (1999); Gurney (1997)

3.1 The Observations

The primary source of knowledge in inductive learning is a set of *observations*. An observation in Machine Learning can be any set of attributes associated to an outcome of a measurement of a certain phenomenon. These attributes can be of many kinds, ranging from *categorical* where the attributes assumes one of a finite number of values, to *ordinal* where the values are ordered, to *interval* where a metric defines the distance between any pair of values of the attribute.

Here we will focus on the cases in which the observations are continuous quantities represented by real numbers. An observation in the following is a set of n measurements (features) that can be represented as a vector $\mathbf{x}$ in $\mathbb{R}^n$ (feature space).

The choice of features is a central problem in pattern recognition. If the learning methods are often independent of the particular domain (*e.g.*, gene classification, speech recognition, or vision), the selection of the features that concisely and robustly convey information about the specific problem often requires deep understanding of the processes involved in it. This is exemplified by noting that, even in human perception, feature extraction is performed by organs (ear, eye, ...) that are hard coded in our genetic heritage, and have been developed through evolution. The classification and recognition tasks are, on the other hand, performed by the same kind of structure (biological neural networks) and are learned in the first years of life.[1]

The studies in this thesis utilise standard features for speech representation, which are described and motivated in detail in Chapter 2.

3.2 The Knowledge

As already stated, machine-learning methods aim at finding structural patterns in a set of observations. The specific learning task is dependent on the kind of knowledge the method aims at extracting from the data. While the observations are the input to the learning procedure, the knowledge is its output. Knowledge can, in this context, be represented in different ways. The learning procedure could, *e.g.*, extract a set of rules that relate different attributes in the observations contained in the data set. It could also assign every observation to a class (*classification*), or relate each observation to a continuous variable (*regression*). Other kind of

[1]The fact that the physiology of the brain is itself a result of evolution, is not in conflict with the point made in the example, as the physiology of the brain corresponds to a particular machine-learning method, while the learning process corresponds to the training a child goes through in the first years of life.

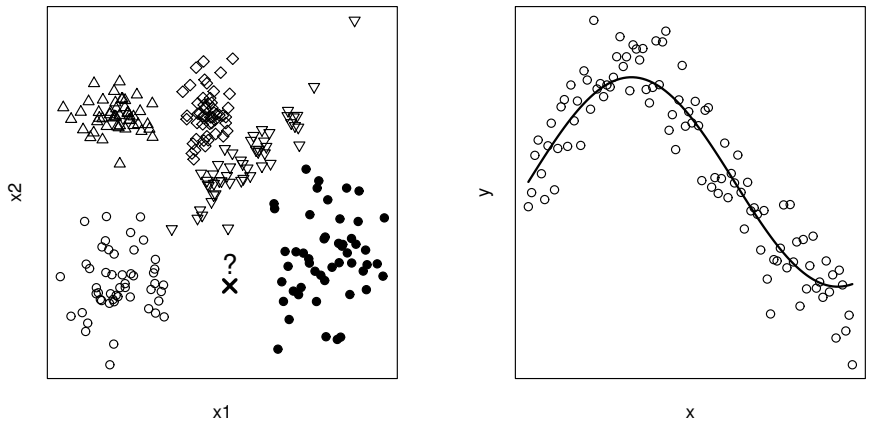


Figure 3.1: Supervised learning: the classification problem (left) and the regression problem (right)

knowledge are graphical models that relate observations with respect to some binary relation (for example a similarity criterion). This is the case of some *clustering* procedures. Methods for classification, regression and clustering will be described in the following.

3.3 Supervised vs Unsupervised Learning

Another central point in machine learning, is whether the knowledge associated with each observation (*i.e.* the right answer) is available to the method during learning. In *supervised learning*, the learning method has access to the right answer for each observation. In *unsupervised learning*, only the set of observations is available, and the learning method has to make sense of them in some way. In most learning situations that we experience, however, both these conditions are extreme cases. *Reinforcement learning* tries to describe a more natural learning scenario, where the learning method is provided with limited feedback.

Supervised Learning

Two possible tasks in supervised learning are supervised classification and regression. These can be seen as two aspects of the same problem. In the first case, we want to assign a new observation to one of C classes that are defined by previously annotated observations, as illustrated in the left part of Figure 3.1. In the second case, we want to assign to any point $\mathbf{x}$ a continuous value for the variable y, having observed a number of examples (right part of Figure 3.1). In both cases, we have

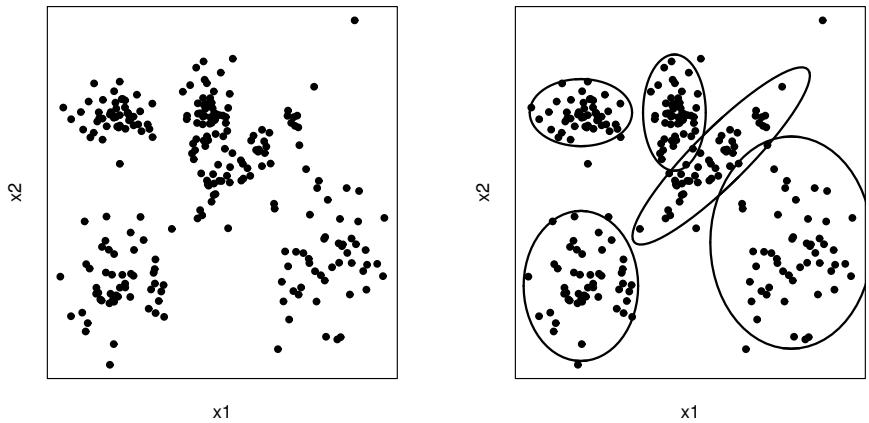


Figure 3.2: Unsupervised learning: the data points are not labelled (left). One of the tasks is, in this case, to find "natural" groupings of the data (right)

a training set that consists of a number of observations $\mathbf{x}_i$ in the input (feature) space X and the corresponding outputs y in the output space Y. X corresponds to $\mathbb{R}^n$ in both cases, whereas Y is a finite set in the case of classification, and the field of real numbers $\mathbb{R}$ in the case of regression.

Unsupervised Learning

When information about the output variable is missing during the training phase (see left plot in Figure 3.2), we talk about unsupervised learning. In this case the possible tasks may be:

- finding the optimal way, with regard to some optimality criterion, to partition the training examples into C classes (see right plot in Figure 3.2),
- finding and displaying relationships between partitions of the data with different cardinality C,
- assigning a new observation to one of the classes in a partition, or
- estimating the optimal number of *natural* classes in a data set.

These tasks are usually referred to as *Clustering* problems in the engineering literature, and simply *Classification* problems in more statistically oriented publications.

Reinforcement Learning

For completeness we give some information about reinforcement learning. This is the learning process that takes place when the learning system has access to

a reward (or penalty) that is somehow related to a series of events. The main difference compared to supervised learning is that the feedback available to the system is not directly related to a specific observation, but rather to a history of events happening in the environment the system lives in, and to the actions that the system has taken as a consequence of these events. It is up to the learning system to interpret the feedback and relate it to its state and to the state of the environment. This kind of task has not been considered in this thesis, but it may be central to one of the problems that will be described in Chapter 6, namely modelling the language acquisition process in an infant.

3.4 Theory behind Pattern Classification

Regardless of the methods used in the pattern classification or regression problem, a number of assumptions about the data generation help to obtain theoretical results of general value, by embedding the problem in a probabilistic framework. In the case of classification, it is assumed that the data is generated by a double stochastic process: first, one of a finite number of states ω_i (called in the following the *state of nature*) is chosen with *a priori* probability $P(\omega_i)$; then, one observation $\mathbf{x}$ is emitted with probability density function $p(\mathbf{x}|\omega_i)$, given that state. In the case of regression, the same model can be used, provided that the discrete states ω_i are substituted by a continuous variable ω with *a priori* density function $p(\omega)$.

Bayes Decision Theory

Bayes decision theory is based on the probabilistic model described in the previous section. If both the *a priori* probabilities $P(\omega_i)$ of each state of nature ω_i and the conditional densities $p(\mathbf{x}|\omega_i)$ of the observations, given the state, are known, using Bayes formula we can derive the *posterior probabilities* $P(\omega_i|\mathbf{x})$ that nature was in state ω_i when we have observed $\mathbf{x}$:[2]

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}$$

The *a priori* density function of the observation $p(\mathbf{x})$ can be computed by summing over all possible states of nature:

$$p(\mathbf{x}) = \sum_{i=1}^{C} p(\mathbf{x}|\omega_i)P(\omega_i)$$

where C is the number of states.

If the stochastic model that has generated the data is fully known, then, Bayes decision theory can be used to find the optimal solution to a decision task given an

[2] If fact that we derive a probability from a density function should not surprise the reader, as the fraction on the right side of the equality could be seen as a limit: $\lim_{dx \to 0} \frac{p(\mathbf{x}|\omega_i)dxP(\omega_i)}{p(\mathbf{x})dx}$

optimality criterion. The criterion can be defined by means of a *loss* function that assigns a penalty $\lambda(\alpha_i, \omega_j)$ to a decision α_i given the state of nature ω_j. Accordingly, a different weight can be assigned to different kinds of errors. The probabilistic model can be used to define a conditional *risk* function $R(\alpha_i|\mathbf{x})$ that relates the risk of taking decision α_i to the observation $\mathbf{x}$, rather than to the state of nature that is, in general, not known. This function is obtained by integrating the loss function over all possible states of nature ω_i, weighted by the posterior probability of ω_i given the observation $\mathbf{x}$:

$$R(\alpha_i|\mathbf{x}) = \sum_{i=1}^{C} \lambda(\alpha_i, \omega_j) P(\omega_i|\mathbf{x})$$

In practical cases, the model of data generation is not completely known. Machine-learning methods attempt, in this case, to estimate the optimal decisions from the available observations.

3.5 Classification Methods

Depending on how the probabilistic description of the data generation is used in the solution of the learning problem we talk about *parametric* and *non-parametric* methods.

Parametric methods rely on a probabilistic model of the process of generating the observations in which probability distributions are described in functional (parametric) form. Learning is, in this case, the process of estimating the model parameters on the basis of the available observations.

There are two kinds of non-parametric methods. The first tries to determine the probabilistic model that has generated the data, but does not assume a functional description of this model. Typical examples are histogram-based methods. Methods of the second kind are based on heuristics and do not directly attempt to estimate a model for the data generation, but, rather, attempt to minimise a criterion that is dependent on the task. Linear discriminant functions, neural networks (or multi-layer perceptrons), and support vector machines are examples of this kind. It should be noted that most of these methods can also be studied and interpreted in probabilistic terms.

Parameter Estimation

Parameter estimation aims at fitting the available probabilistic model to a set of observations, by deriving the optimal set of model parameters. When considering supervised classification, and assuming that observations related to a certain class ω_i give no information about the model for a different class ω_j, the problem can be split into N independent subproblems (N number of classes).

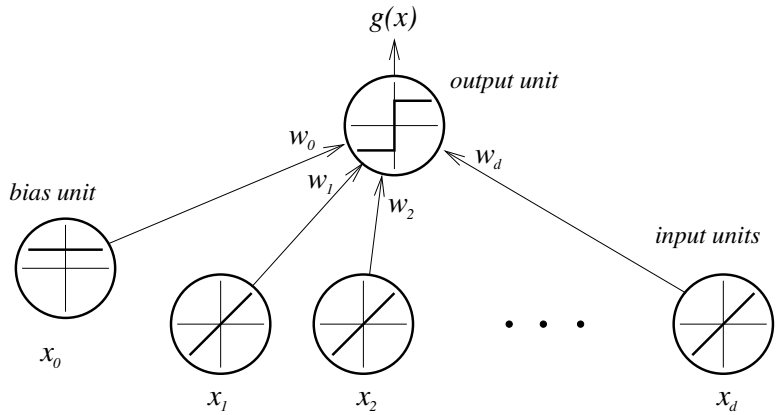


Figure 3.3: A simple linear classifier with $g(\mathbf{x}) = \sum_{k=0}^{d} a_k x_k$. This model can be generalised by using arbitrary functions of the observation $\mathbf{x}$.

Two philosophies can be followed in order to find a solution to this problem. Maximum Likelihood estimation attempts to find the model parameters that maximise the model fit to the data with no *a priori* information. Bayesian estimation, on the other hand, optimises the posterior probability of a certain class given an observation (and the training data). The model parameters are in this case stochastic variables that are characterised by probability distributions. The *a priori* distribution of the model parameters is necessary to apply Bayes theorem and compute the posterior distributions from the likelihoods (that can be computed from the data).

The advantages and disadvantages of these two paradigms are the subject of active discussion in the statistics community, and are not discussed here.

Linear Discriminant Functions

Linear discriminant functions belong to the class of nonparametric methods in the sense that the form of the probability models underlying the data generation need not be known. In reality, the learning paradigm is similar in both parametric and nonparametric models, because, also in the second case, we assume a known form of the discriminant functions (*e.g.*, linear, quadratic, polynomial,...), and we estimate the model parameters from the data. In general, the functions are in the form (often referred to as generalised discriminant functions):

$$g_i(\mathbf{x}) = \sum_{k=1}^{F} a_{ik} y_k(\mathbf{x})$$

where a_{ik} are the weights (or model parameters) that can be trained from the data for each class ω_i, and $y_k(\mathbf{x})$ are F arbitrary functions of $\mathbf{x}$. The fact that these

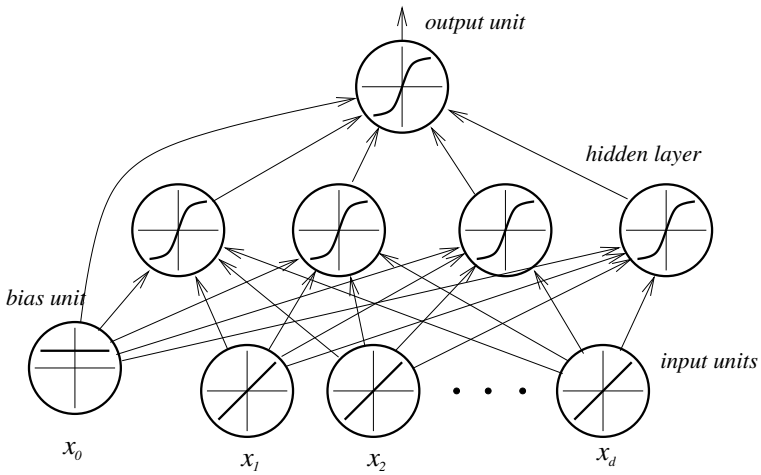


Figure 3.4: Illustration of a multi-layer perceptron.

possibly nonlinear functions are combined linearly is the origin of the name. The functions are used by assigning a point $\mathbf{x}$ to the class ω_i iff $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$.

The training procedure is aimed at optimising a criterion (*e.g.*, the training error) and is usually performed with gradient descent procedures. These procedures iteratively move in the model parameter space following the steepest direction with respect to the criterion function. Eventually the procedure finds a local minimum of the criterion function and the iterations are terminated.

Multi-Layer Perceptrons

One of the problems with linear discriminant functions is that they shift the complexity of the pattern recognition task onto the choice of proper discriminant functions, *i.e.*, the transformation of the feature space that makes the classes linearly separable.

One way to let the classification method decide the internal representation of the data is to add a (hidden) layer in the network depicted in Figure 3.3. The result is the multi-layer perceptron shown in Figure 3.4. The weights from the input to the hidden layer (symbolised by arrows in the figure) perform a linear data transformation. The activation functions of each unit in the hidden layer, are usually non-linear functions (s-shaped curves in the figure). The activities at each unit in the hidden layer, obtained by passing the weighted sum of the inputs through the non-linear activation function, can be thought of as an internal convenient representation of the patterns observed by the input layer. The output

layer makes a decision in a manner similar to linear discriminant functions.

The complication of this model, as far as the training procedure is concerned, is that the units in the hidden layer are not directly observable. However, if the so called activation functions are invertible, the error observed at the output node can be propagated to the previous layers using the current estimation of the connection weights. This is the principle of the Back Propagation algorithm, which is the most commonly used training procedure for these models.

3.6 Clustering Methods

Clustering, or unsupervised classification methods, can be divided into a number of classes. The first distinction is based on the use of probability distributions to characterise the clusters, as in Model Based Clustering, or on the use of a pairwise similarity measure, *i.e.*, a relationship between each pair of data points, as in K-means (MacQueen, 1967). It should be noted that, in particular cases, the two approaches coincide, *e.g.*, Gaussian distributions with diagonal covariance matrix in the form λI define spherical shapes equivalent to the ones obtained with the euclidean distance between data points.

Another distinction is between *relocation* and *hierarchical* methods. In the first case, starting from a desired number of clusters, the data points are iteratively relocated between clusters until an optimal partition is found. In the second case the procedure successively merges or splits clusters (agglomerative or divisive clustering), and the result is a tree structure called *dendrogram* that represents the hierarchical relationships between partitions of different orders.

Clustering is a very active research field and many new methods and theories are continuously being developed. For example, Graph Theory (Gaertler, 2002), Self Organising Maps (Kohonen, 1982), and Support Vectors (Ben-Hur et al., 2001) have been applied to the clustering problem. In the following only the methods used in this thesis will be described.

Model-Based Clustering and Gaussian Mixture Estimation

When using parametric models to perform unsupervised classification, we cannot split the problem into a number of subproblems (one for each class), as in the supervised case. The reason is that the assignment of each data point to a certain state of nature (class) that has generated it, is unknown. The parameter estimation procedure must in this case consider all the distributions simultaneously, *i.e.*, it must fit a mixture of distributions to the data.

A solution to this problem is achieved by describing the membership of each data point to each class in probabilistic terms. The union of the original data points, called *incomplete data*, and of the vector of memberships to the classes are called the *complete data*. The Expectation Maximisation (EM) algorithm (Dempster et al., 1977) can be used to compute the model parameter estimate that

Σ_k	Distribution	Volume	Shape	Orientation
λI	Spherical	Equal	Equal	N/A
$\lambda_k I$	Spherical	Variable	Equal	N/A
λDAD	Ellipsoidal	Equal	Equal	Equal
$\lambda_k D_k A_k D_k$	Ellipsoidal	Variable	Variable	Variable
$\lambda D_k A D_k$	Ellipsoidal	Equal	Equal	Variable
$\lambda_k D_k A D_k$	Ellipsoidal	Variable	Equal	Variable

Table 3.1: Parametrisation of the covariance matrix Σ_k in the Gaussian model based on the eigenvalue decomposition, and their geometric interpretation. From Fraley and Raftery (1998).

maximises the likelihood of the model on the incomplete data, by maximising the complete data likelihood. The EM algorithm iteratively finds an estimate of the membership function given the current model parameters (expectation), and the new optimal values for the model parameters given the new membership function (maximisation).

Compared to more traditional clustering methods, such as K-means, which rely on a fixed metric in the feature space, the parametric solution to the clustering problem has the advantage that the form of the clusters can be easily controlled by the shape of the distributions used. For example, in the case of Gaussian distributions, the class of the covariance matrix determines the shape of the distributions as indicated in Table 3.1.

As we will see in the following, this framework also simplifies the task of determining the most plausible number of clusters in the data (Banfield and Raftery, 1993; Dasgupta and Raftery, 1998; Fraley, 1999; Fraley and Raftery, 2003, 2002; Fraley et al., 2003) and the task of comparing different partitions of the same data set (Meilă, 2002).

Hierarchical Clustering

In their classical formulation, hierarchical methods are entirely based on the binary relation between the entities that are to be clustered. In the case of data points spanning a vectorial space, the relation is given by the definition of a metric in that space.

The procedure operates by recursively merging or splitting the clusters according to some optimality criterion. The different variants of this method are dependent on the criterion used to measure the distance between two clusters $d(c_1, c_2)$, given the pairwise distances of their members $d(\mathbf{x}_m, \mathbf{x}_n)$. Examples are given in Table 3.2.

Other formulations operate directly on the data-points. For example Ward's method clusters groups that result in the minimum increase of "information loss", defined in this case as the sum of squared errors from the member average. Hierarchical clustering has also been implemented in the framework of Model-Based

method	distance
single linkage	$d(c_1, c_2) = \min_{\mathbf{x}_m \in c_1, \mathbf{x}_n \in c_2} d(\mathbf{x}_m, \mathbf{x}_n)$
complete linkage	$d(c_1, c_2) = \max_{\mathbf{x}_m \in c_1, \mathbf{x}_n \in c_2} d(\mathbf{x}_m, \mathbf{x}_n)$
average linkage	$d(c_1, c_2) = \frac{1}{\lvert c_1 \rvert \lvert c_2 \rvert} \sum_{\mathbf{x}_m \in c_1, \mathbf{x}_n \in c_2} d(\mathbf{x}_m, \mathbf{x}_n)$
average group linkage	$d(c_1, c_2) = \frac{1}{(\lvert c_1 \rvert + \lvert c_2 \rvert)^2} \sum_{\mathbf{x}_m, \mathbf{x}_n \in (c_1 \cup c_2)} d(\mathbf{x}_m, \mathbf{x}_n)$

Table 3.2: Merging criteria for hierarchical clustering. $|.|$ indicates the cardinality of the set, and $\cup$ the union

Clustering (Fraley, 1999). Note that for diagonal covariance matrices with equal eigenvalues ($\Sigma_k = kI$), Model-Based Hierarchical Clustering is equivalent to Ward's method.

Estimating the Number of Clusters

Guessing the number of *natural groups* from a data set has long been a central problem in statistics and data mining. Here the concept of natural groups or clusters should be considered in a statistical sense.

Intuition suggests that the number of groups in the data is strongly dependent on the degree of detail the analysis is aimed at. For example, when analysing natural language, one could be interested in finding: a) the groups belonging to the same *linguistic family*, b) the *single languages* as defined by political factors, c) *dialectal differences* within the same language, d) personal speaking styles (so called *ideolects*), or even e) variation in *speaking style* of the same person in different contexts. Hence, the problem of finding the *true* number of clusters is strongly influenced by the final goal of the analysis.

However, if we consider the problem in statistical terms, the aim is to establish a methodology that allows an objective interpretation of the data in the same way that hypothesis testing can tell if there are significant differences between the group means in two data samples.

Several methods have been developed by different authors in order to predict the number of clusters. Milligan and Cooper (1985) compared 30 different indexes on a number of synthetic data sets containing from 1 to 5 well-differentiated clusters. They showed how some indexes perform reasonably well under these artificial conditions. Many of these indexes are based on the comparison between the so called scatter matrices. If we call $\mathbf{m}$ the total mean of the data set $\mathcal{D}$, $\mathbf{m}_i$ the mean of the data points belonging to one of k clusters $\mathcal{D}_i$ of size n_i, and $\mathbf{x}$ a generic data point, the scatter matrix of the ith cluster is defined as (T indicates the transpose

operation):

$$S_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$$

the within-cluster scatter matrix as:

$$S_W = \sum_{i=1}^{k} S_i$$

the between-cluster scatter matrix as:

$$S_B = \sum_{i=1}^{k} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

and the total scatter matrix as:

$$S_T = \sum_{\mathbf{x} \in D} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T$$

Most of the indexes studied in Milligan and Cooper (1985) use the scatter matrices as a measure of the spread of data points. They usually tend to reduce the within-cluster scatter and increase the between-cluster scatter, in order to produce compact and well-separated groups. Often the indexes have a similar definition, and are, therefore, highly correlated.

Another class of methods to estimate the number of clusters makes explicit use of the probabilistic model used in model-based clustering. These are, *e.g.*, the likelihood and the Bayes Information Criterion (BIC).

The likelihood $l_M(\mathcal{D}, \theta)$ is the evaluation of the density function defined by the clustering procedure on the actual data $\mathcal{D}$. This is seen as a function of the model parameters θ when the data $\mathcal{D}$ is fixed. Given the assumption of independence of each data point $\mathbf{x} \in \mathcal{D}$, the likelihood function can be evaluated as a product of the values of the density function $p(\mathbf{x}|\theta)$ evaluated at each $\mathbf{x}$. For this reason, it is more common to consider the log likelihood that turns the product into a sum, simplifying the computations and avoiding numerical problems. A drawback of the likelihood is that it increases monotonically with the model complexity: more complex models always fit the data equally or better in terms of the likelihood.

The Bayes Information Criterion is an approximation of the Bayes factor, and by definition takes into account the number of parameters and the amount of data that is available for parameter estimation, as well as the model fit to the data. It is defined as

$$BIC \equiv 2 l_M(\mathcal{D}, \theta) - m_M \log(n)$$

where $l_M(\mathcal{D}, \theta)$ is the likelihood of the data $\mathcal{D}$, given the model parameter θ, m_M is the number of free parameters in the model, and n the number of data points.

3.7 Learning Variable Length Sequences

In most machine learning problems, the observations, or data points $\mathbf{x}_i$, are considered to be independently drawn from the probability model described above. As a consequence, the order in which the set of data points is analysed does not matter.

In speech recognition and analysis, and in other areas such as gene sequence classification, on the other hand, the evolution of the sequence of data points carries most of the information. One of the key issues that make these problems exceptional when compared to more traditional pattern classification tasks, is the fact that the sequences may be of variable length. If we could assume sequences of the same length, say S, as might be possible with time sequences in economy, for example, that often follow predefined calendar time intervals, the problem could be folded back to a standard pattern classification. The whole sequence of S points $\mathbf{x}_i \in \mathbb{R}^n$ could in fact be considered, at least in principle, as one point in a space of larger dimensionality $\mathbb{R}^{n \times S}$. The fixed-length assumption would guarantee that any sequence is a point that lies in the same space $\mathbb{R}^{n \times S}$ and thus standard pattern classification methods could be used.

With variable length sequences this is not possible and more sophisticated methods must be used.

Hidden Markov Models

One way to model time-varying processes is with Markov chains. A Markov chain is a process $\{X_n, n = 0, 1, 2, \dots\}$ that assumes one of a finite or countable number of possible values ω_i (called states) at each time step. Whenever the processes is in state ω_i, there is a fixed probability P_{ij} of making a transition to any other state ω_j, *i.e.*

$$P(X_{n+1} = \omega_j | X_n = \omega_i, X_{n-1}, \dots, X_1, X_0) = P_{ij}$$

Another way to express this property is to state that the conditional probability of being in any state at time $n + 1$, given the past states at times $0, 1, \dots, n - 1$ and the present state at time n, is not dependent on past states, but only on the present state. A graphical representation of such a model is shown in Figure 3.5. Each oriented arc in the figure represents a transition probability between the states that the arc connects.

If the sequence of states is not directly observable, but rather we observe the emissions o of a phenomenon that is controlled by the current state, the model can be augmented with conditional *emitting* probabilities $P(o|\omega_i)$, *i.e.*, the probability of emitting a particular observation o when the model was in state ω_i. Because the state is not directly observable, these models are called Hidden Markov Models (HMMs).

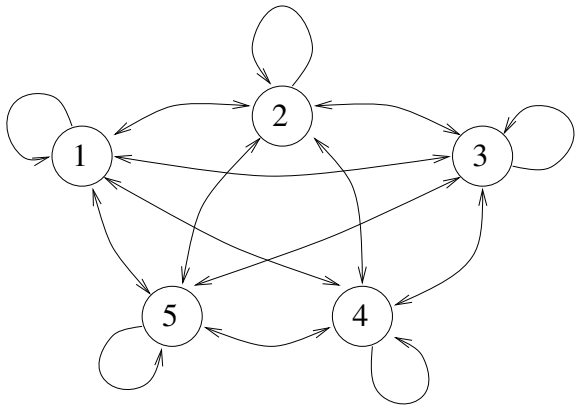


Figure 3.5: A five-state ergodic Markov model

The main tasks when using these models are:

- given a sequence of observations, find the optimal parameters of the model that has generated them (training),
- given a model and a sequence of observations, compute the likelihood that the model has produced the observations (evaluation),
- given a model and a sequence of observations, compute the most likely path through the model that has produced the observations (decoding).

The first of the three tasks is conceptually similar to the task of fitting a mixture of distributions to the data set. The similarity is in the fact that the data is incomplete. In the mixture of distributions case, the missing data is the assignment of each data point to one of the distributions. In the HMM case, the missing data is the sequence of states. Here, a solution to parameter estimation is the forward-backward or Baum-Welch algorithm, which is an instance of the Expectation Maximisation algorithm.

A conceptually simple solution to the *evaluation* problem is performed in two steps: the likelihood can be computed for each possible path through the model by multiplying the corresponding transition and emission probabilities; the total likelihood is then computed by adding the path-specific likelihoods over all possible paths. However, this is prohibitive in most practical applications, because the number of possible paths increases exponentially with the length of the sequence. The forward (or backward) algorithm solves this problem by computing the probabilities iteratively, thus keeping at each time step a number of alternatives equal to the number of states.

A fast algorithm for *decoding* is the Viterbi algorithm (Viterbi, 1967). In this case, at each time step and for each state, we keep track of the probability of the

best path that has taken us to that particular state and retain a pointer to the previous state in the path. When the end of the sequence is reached, the full path can be reconstructed by starting from the last best state, and recursively following the previous best states.

Clustering Variable-Length Sequences

A task that is receiving increasing attention, especially in the field of text processing and gene analysis, is the unsupervised classification of variable-length sequences. As in the supervised case, the most popular models are HMMs.

Oates et al. (1999), for example, apply Dynamic Time Warping to extract information about the possible number of clusters. This information is then used to train one HMM for each hypothesised cluster. The assignment between time series and HMMs can vary in the training phase, based on likelihood measurements.

In Bar-Hillel et al. (2005), the data is first modelled as a mixture of Gaussian distributions and then the parameters of a hidden Markov model are computed as transitions between candidate mixture terms.

Porikli (2004) approaches the problem by training a comprehensive HMM on the data, and then clustering the model parameters by eigenvector decomposition.

Finally, Li and Biswas (1999, 2000, 2002) perform a sequential search, optimising (i) the number of clusters in a partition, (ii) the assignment of each data object to a cluster given the size of the partition, (iii) the HMM size for each cluster, and (iv) the HMM parameters for the individual cluster.

Part II

Contributions of the Present Work

Chapter 4

Mapping Acoustic to Visual Speech

A large part of the work behind this thesis has been carried out within the Swedish project Teleface and the subsequent European project Synface. The aim was to develop a computer-animated talking face that could be used as a hearing aid for persons using the telephone.

As depicted in Figure 4.1, the only input to the system is the acoustic signal received at the hearing-impaired person's end of the telephone line. This paradigm gives a number of advantages, compared to, *e.g.*, video telephony: the telephone

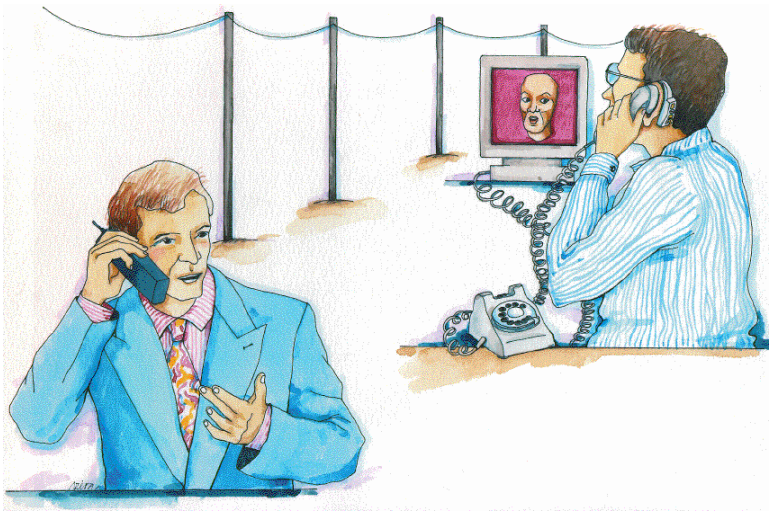


Figure 4.1: The Teleface vision and Synface reality. Illustration by Viera Larsson

line does not need to be changed, the person on the other end of the line needs only a standard telephone, and the hearing-impaired person does not need to make the impairment manifest. On the other hand, this solution requires that the facial movements, and especially the lip movements, be derived from the acoustic signal alone.

The mapping from acoustic to visual information is the focus of this part of the thesis (Papers A–C). The challenge is to produce sufficiently accurate movements, in order to convey useful information to the listener in a *real-time* system with *low latency*.

The real-time requirement arises from the intention to run the system on home computers with no extraordinary processing power. This limits the amount of calculations that can be performed per time unit without overloading the machine, and translates into a constraint on the complexity of the methods employed.

The low-latency requirement is more fundamental and originates from the need to preserve the natural flow of the conversation. Two arguments are responsible for this constraint. Firstly, it is essential that the face movement be synchronous with the acoustic signal. According to, *e.g.*, Grant and Greenberg (2001), the asynchrony between visual and acoustic presentations causes a rapid drop in intelligibility for persons relying on lip reading, especially when the audio leads the video. Secondly, studies (*e.g.*, Kitawaki and Itoh, 1991) have shown that in a communication channel, long round-trip transmission delays (ca 500 milliseconds) cause the natural turn-taking mechanism in a conversation to become troublesome. An everyday example is given by the satellite video connections often employed in TV news programmes, where the delay is in the order of seconds.[1] The first argument requires that the acoustic signal be delayed while the acoustic to visual mapping and the face animation are performed, in order to synchronise audio and video. The second argument imposes a limit to the length of the delay and, therefore, to the latency of the system. As a consequence, the mapping from acoustic to visual information at a certain time may be based on a limited amount of look-ahead, *i.e.*, on a limited amount of future acoustic evidence.

4.1 The Face Model

The face model was developed by Beskow (2003). This section describes the aspects of the model that are useful to understand the acoustic to visual mapping problem addressed in this part of the thesis.

The face model is a 3D graphic model comprised of a large number of polygons controlled by a set of continuous articulatory parameters with values varying in the range $[0, 1]$ (Table 4.1). The time evolution of the value of each parameter

[1] The transmission delay due to wave propagation is about 240 msec for geostationary satellites orbiting at 36,000 km above the earth. Delay due to the transmission protocols and to signal processing must be added.

parameter	description
V0	jaw rotation
V3	labiodental occlusion
V4	lip rounding
V5	bilabial occlusion
V6	tongue tip
V7	tongue length
V8	mouth width
V9	lip protrusion

Table 4.1: Parameters used to control the facial movements

defines movements at a conceptually high level, influencing the trajectory of groups of points in the 3D space.

An alternative higher level control system exists (Beskow, 1995) that is compatible with text-to-speech synthesis engines (*e.g.*, Carlson et al., 1982). This control system accepts as input a string of phonemes/visemes (see Chapter 2) with the corresponding time stamps. For each viseme, a set of target values for the facial parameters is defined. Some of the parameters may be unspecified for a certain viseme when this does not affect the corresponding articulator. In this case, the context must be used to determine the optimal target value. A rule-based system takes as inputs the target values and the time stamps, and performs a linear or spline interpolation, attempting to reach the targets as close as possible while preserving a smooth and consistent movement of the speech organs.

Figure 4.2 displays a subset of the parameters in Table 4.1 for the word "Synface" (/sʏnfɛis/). For each parameter, the target values (when specified) are plotted together with the final parameter trajectories obtained by spline interpolation. In the Swedish-English pronunciation of the word, the first vowel (/ʏ/) is front rounded and activates both the lip rounding and protrusion. The tongue tip is activated by the first and last /s/ and by the /n/. The labiodental occlusion is activated by the /f/, and finally the jaw rotation is moderately active during the whole utterance.

4.2 Regression vs Classification

The first question regarding the solution of the acoustic to visual parameter mapping is whether this map should be achieved directly solving a regression problem, or if the acoustic signal should first be classified into phoneme/viseme classes, after which the face parameters should be computed by means of the rule system described above.

In the first case, the map is a continuous function from the acoustic parameter space $S_p = \mathbb{R}^N$ to the face parameter space $F_p = \mathbb{R}^M$. In the second case, the map

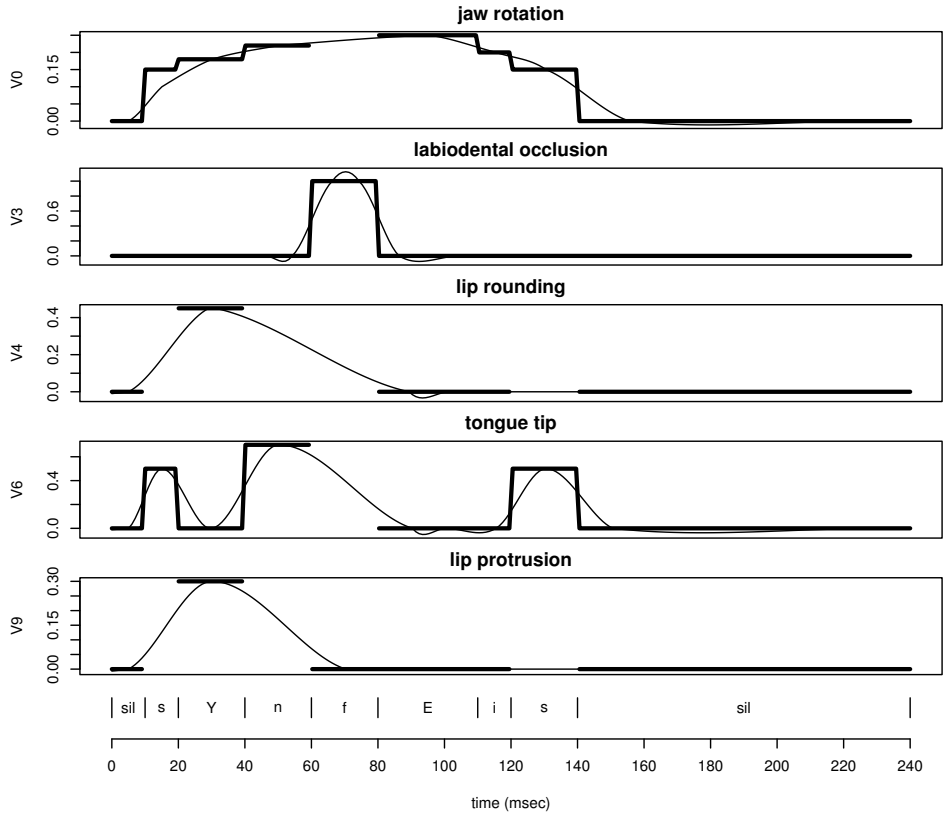


Figure 4.2: Some face parameter trajectories for the word "Synface" (/sʏnfɛis/). Target values (when specified) are indicated by the thicker line. The final parameter trajectories are obtained in this case by spline interpolation.

is a discrete function from the acoustic parameter space to a finite set of symbols $V = \{v_i\}$.

$$S_p = \mathbb{R}^N \longrightarrow F_p = \mathbb{R}^M$$
$$\searrow \quad V = \{v_i\} \quad \nearrow$$

This question was the basis of Paper A. Here, two methods are compared, one aiming at achieving regression by means of a neural network and the other classifying the speech signal into a phoneme string with Gaussian mixture models combined with hidden Markov models.

The neural network is a three-layer perceptron with 13 mel frequency cepstral coefficients (see Chapter 2) as input and the face parameters as output. The network includes both time-delayed and recurrent connections, in an attempt to grasp the time evolution of both the input and the output sequences.

A number of methods were investigated to perform the classification task. The standard scheme uses hidden Markov models with Gaussian distributions as a state-to-output probability model. The different factors that distinguish the methods are the amount of context used in the phonetic models, and whether phonemes are clustered into viseme classes already in the modelling phase or after decoding. The training procedure and the topology of the phonetic models is described in Salvi (1998).

The phonetic models were connected to each other in a free loop. An attempt has been made to define a syllable based topology of the recognition network, but this did not lead to any improvement.

Gender-dependent models have also been tried resulting in a slight improvement when the gender of the model and the test speaker agree; a marked degradation is observed in the case of gender mismatch.

4.3 Evaluation Method

The performance of the different methods can be measured at different points along the processing chain. As the aim of the system is to improve intelligibility in a conversation, the most reliable measure of performance is obtained by performing listening tests with hearing-impaired subjects. Alternatively, normal hearing subjects can be used if the acoustic signal is degraded so as to simulate a hearing impairment. This kind of test usually employs short sentences containing a number of keywords that the subject should detect. The performance is given by the percentage of correctly transcribed keywords. Paper A contains an experiment in which 10 hearing-impaired subjects were asked to repeat the short sentences they were presented with, which contained three keywords each.

Moving backward in the processing chain, an alternative evaluation method is to measure the distance between the target trajectory and the predicted trajectory for each face parameter. This allows for uniform evaluation of both regression and classification methods, but it presents a number of difficulties because the deviations from the target trajectories should be weighted with the potential effect they have on visual perception. For example, in the case of bilabial occlusion, it is essential that the corresponding parameter reaches the maximum value in order to give to the listener the impression that /b/, /p/ or /m/ has been uttered. However, for other visemes, the value of this parameter is less important. In Paper A, only visual inspection has been performed on the parameter trajectories obtained with the different methods.

Finally, classifications methods have a natural evaluation criterion: the number of correctly classified symbols (phonemes or visemes in this case). This evaluation

Example 1

```
ref: AAABBBBCCCCCCCDDDDDEEEEEEE -> ABCDE
rec: AAAAAAAAAABBBBCDEEEEEEEEEE -> ABCDE
res:                             ccccc
fbf: ccc----------c---ccccccc
```

accuracy = 100, % correct = 100%, frame-by-frame = $11/26 \times 100 = 42.3\%$

Example 2

```
ref: AAABBBBCCCCCCCDDDDDEEEEEEE -> A    BC      D E
rec: AFGHIBKDSORFKDKELFDSKDLFID -> AFGHIBKDSORFKDKELFDSKDLFID
res:                             ciiiicsiiiiiiciciiiiiiiii
fbf: c----c----------c-------
```

accuracy = $(4 - 21)/26 \times 100 = -65.4$, % correct = 100%, frame-by-frame = $3/26 \times 100 = 11.5\%$

Figure 4.3: Scoring examples. Capital letters are result symbols for each frame (observation vector) on the left and as a sequence on the right. `ref` is the reference, `rec` is the recognition output, `res` is the result after alignment (`c` = correct, `i` = insertion, `s` = substitution), and `fbf` is the frame-by-frame result.

criterion is complicated by the fact that classification results come in sequences where both the correctness of the symbol and the time alignment are important. Moreover, given the low-pass filter characteristics of the rule-based visual parameter generation, short insertions do not cause severe degradation to the final facial movements.

A number of standard evaluation methods have been used in Papers A–C. The scores are illustrated in Figure 4.3, where some of their limitations are exemplified. Two scoring methods (*accuracy* and *% correct*) are computed after aligning the sequence of recognised phonemes/visemes to the sequence contained in the reference transcription. The alignment is performed at the symbolic level, with dynamic programming. Although the position of each symbol in the recognition sequence is correlated to its time alignment with the sequence of observations, the above methods disregard the time information contained in the transcriptions. The % correct score is defined as the ratio between the number H of matching symbols after alignment and the total number N of symbols. Accuracy is defined as $\frac{H-I}{N} \times 100$ where I is the number of insertions emerged from the alignment procedure. An additional scoring method (*frame-by-frame correct rate*) simply computes the ratio between number of correctly classified and total number of frames (observation vectors). This method tends to underestimate the importance of short insertions.

As shown in Figure 4.3, these scores can give misleading results in case of phonetic recognition. In the first example, accuracy and % correct give no information about the time misalignment of the recognition sequence and reference sequence in time. The second example shows how % correct can give high scores for random recognition sequences.

4.4 Results of Paper A

The main result of Paper A is that the methods solving the classification problem, referred to as HMM methods in the paper, outperform the ones performing regression (the ANN method). The main reason for this is that the trajectories estimated by the neural network are only able to follow the trend of the target trajectories, but are not sufficiently accurate. The errors committed by the HMM methods are more drastic, because they produce a completely incorrect trajectory. However, these errors are less frequent and more acceptable when considering that, in case of correct classification, trajectories close to the optimum are obtained.

The regression methods have the advantage that the facial parameters can be estimated independently of one another from the acoustic features, and that different articulatory gestures (*e.g.*, with different degrees of articulation) could, in principle, be determined from the sounds they produce. In this application, however, the aim is to achieve stereotypical movements, rather than reproducing realistically the actual movements that produced the sound. In this respect, methods based on classification are more suitable, because they make use of a discrete phonetic representation that standardises the results.

The listening tests show how the synthetic face driven by the correct parameters gives a remarkable improvement over the audio-alone condition, even if it does not help as much as a natural face. The synthetic face driven by the HMM method gives an improvement that is significant compared to the audio-alone condition, whereas the ANN method gives no improvement. The tests also show that these results are highly dependent on the listener.

4.5 Real-Time and Low-Latency Implementation

After the studies described in Paper A, the solution based on regression was discarded and further development, mainly carried out during the Synface project, was focused on solving the classification problem described above in a real-time, low-latency system.

This research has involved both the development of new classification methods and the implementation of a software library that could be integrated with the existing face animation and control module.

The classification methods employed in the Synface prototype use a hybrid of hidden Markov models (HMMs) and artificial neural networks (ANNs). The hidden Markov models define a number of states, corresponding to subparts of

each phoneme and the way these states are traversed in time (transition model). The neural networks are trained to estimate the *a posteriori* probability of a state given an observation (Ström, 1997).

Different topologies for the ANNs, and consequently for the HMMs, have been tested in an initial phase of the project. The simplest models are feed-forward networks trained with one output unit for each phoneme. The neural network is in this case a static model; all the information about time evolution is coded into the HMM in the form of a loop of three-state left-to-right models representing a phoneme each.

A more complex model performs the same task with a time-delayed neural network (TDNN). In this case, several input vectors are considered at every time step by means of delayed connections between the input and the hidden layer. Delayed connections are also present between the hidden and the output layer. All delays are positive, making sure that only the current and past frames are necessary to produce a result at a certain time.

Another model that was trained during this phase extends the output targets in the neural network to context-dependent units. Statistics on the training data were used to estimate which phonemes are more affected by context. This was done with a methodology commonly used in speech recognition based on HMMs and Gaussian mixture distribution models, called phonetic tree clustering (Young, 1996): phonetic knowledge and statistics collected during the training phase are used to build a clustering tree, in which the leaves contain groups of similar states. The procedure can be terminated either based on the likelihood of the data given the model or when the desired number of clusters has been reached. This procedure was used to determine the number of context-dependent output units to be used in the neural network. An example of the distribution of number of output units for each phoneme is given in Figure 4.4, with a total of 376 output units. These models have been trained but never fully tested.

Finally, recurrent neural networks (RNNs) with one output unit per phoneme were trained. These models, thanks to the time-delayed recurrent connections in the hidden layer, can learn more complex time dependencies than TDNNs (Salvi, 2003a). Papers B and C describe experiments based on RNN models.

A real-time modified version of the Viterbi decoder was implemented. This truncates the back-tracking phase to a configurable number of time steps, thus allowing incremental results to be evaluated. The effects of the truncation compared to the standard Viterbi decoder were investigated in (Salvi, 2003a). As expected, the performance drops when the look-ahead length, *i.e.*, the amount of right context, is reduced. Results stabilise for look-ahead lengths greater than 100 milliseconds, for this specific task.

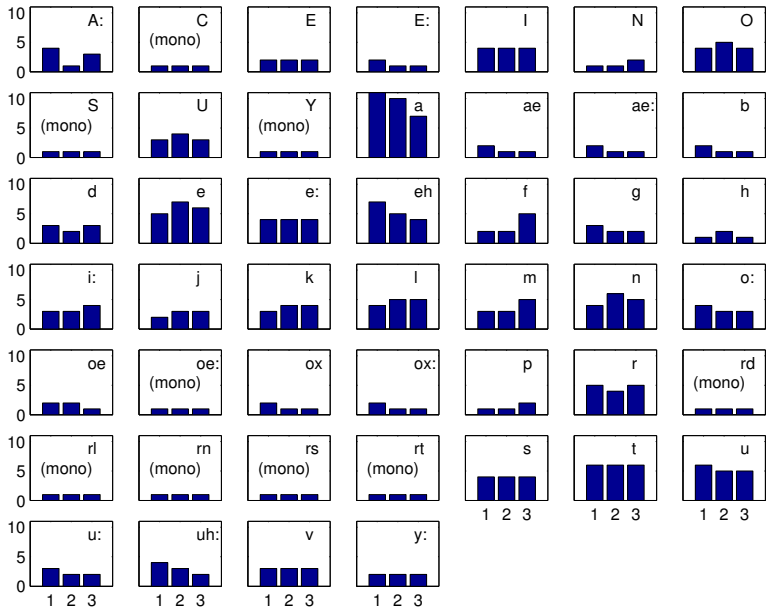


Figure 4.4: The distribution of context-dependent output units of each phoneme and segment position (initial, middle and final), indicated by the numbers 1, 2, 3. The total number of output units is in this case 376. The phonemes with only one output unit per segment are context independent (mono).

4.6 Interaction between the RNN's and HMM's Dynamic Model

As described in detail in Paper B, the use of recurrent and time-delayed neural networks violates the first order Markov chain assumption: each observation is not only dependent on the current HMM state. In other words, both the NNs and the HMMs impose a model of time evolution. Depending on the task, these two models may be conflicting, *e.g.*, if the HMM contains a sequence of states that was not present in the network's training data. The interaction between the two models can be expected to have stronger effects when the time dependencies are longer (RNNs with many delayed connections and HMMs defining word-like sequences rather than a simple loop of phonemes).

Standard analysis tools may not be used to characterise the temporal properties of these models, due to the presence of nonlinearities. A number of studies have been devoted to describing the dynamic properties of non-linear systems analytically (*e.g.*, Koop et al., 1996; Hirasawa et al., 2001). In our case, however, the use of a decoder with limited look-ahead complicates this task. Is, for example, the

truncation error larger when the time dependencies in the model are longer?

Paper B tries to answer some of these questions with empirical experiments. Three factors are considered:

1. the length of the time dependencies in the neural network,
2. the length of the time dependencies in the hidden Markov model,
3. the look ahead length.

To vary the first factor, three neural networks were compared: a feed-forward NN (with no time dependencies) and two recurrent networks with different complexities.

The second factor was varied in two ways. In one experiment (denominated word-length test), the topology of the HMM was gradually varied from a free loop of phonemes to a word loop, with words of increasing length in terms of the number of phonemes (up to 7). To achieve this, the words were artificially created by concatenating the phonemes in the transcriptions of each test utterance. In the second experiment (denominated alpha test), the HMM topology was defined as a mixture of a phoneme loop and a forced alignment model, *i.e.*, a model in which the only possible output sequence is the phoneme string from the transcription of each utterance. A parameter sets the relative weight of the two models in the mixture. Note that the alpha test with alpha set to 0 coincides with the word length test with word length equal to 1 because, in both cases, the HMM defines a loop of phonemes.

The third factor is varied by setting the look-ahead length in the decoder from 1 to 20 frames (10 to 200 milliseconds).

4.7 Results of Paper B

The results in Paper B can be summarised as follows:

- In all conditions, longer look-ahead lengths correspond to better performance.
- The word-length test shows this phenomenon in a clearer way than the alpha test.
- The degree of improvement depends on the amount of time-dependent information in the HMM and in the NN.
- When the time dependencies in the HMM are short (word length close to 1 or alpha less then 0.5), the recurrent networks take more advantage of longer look-ahead lengths than do the feed-forward networks.
- When the time dependencies in the HMM are long, the feed-forward networks seem to take more advantage of longer look-ahead lengths, but this outcome is conditioned by the fact that the results obtained with the more complex recurrent networks tend to saturate (approach a 100% correct classification rate).

4.8 A Method for Phonetic Boundary Detection

During the experiments in Paper B, the class entropy, easily estimated from the outputs of the neural networks, was considered as a possible confidence measure for the classification performance. A closer inspection of the time evolution of the entropy suggested another use of this measure. The entropy is a measure of uncertainty, and thus increases when the neural network presents activity in more than one output unit. This usually corresponds to an input that is ambiguous compared to the patterns learned during training. There are many factors that generate ambiguous patterns, *e.g.*, a voice that is particularly hard to recognise or a phoneme in a context underrepresented in the training data. One of the factors that seem to systematically affect the class entropy is the proximity to a phonetic boundary. This can be explained by the fact that the properties of the speech signal vary continuously from one segment to the next, assuming values that are intermediate between the two typical patterns. Even in the case of abrupt transitions, as for plosive sounds, the fact that the signal analysis is preformed over overlapping windows of a certain length causes the corresponding feature vectors to vary gradually.

This phenomenon suggests the use of the class entropy as a segment boundary detector. Paper C shows a preliminary analysis aimed at verifying this possibility.

4.9 Results of Paper C

The results in Paper C can be summarised as follows:

- The entropy of the output activities of the phoneme classification neural network has local maxima corresponding to phonetic boundaries.
- The absolute value of the entropy is strongly dependent on the phonetic content of the segment.
- Averaged over all the speech frames, the absolute value of the entropy and the value of the entropy normalised to the average for each phonetic segment increase close to phonetic boundaries.
- Both the absolute and normalised entropy are nevertheless not sufficient to detect phonetic boundaries efficiently.
- Dynamic entropy measures (first and second derivatives) also carry information on the proximity to a phonetic boundary.
- These preliminary results suggest that peak-picking techniques could be used to accurately detect the boundary positions from the entropy measurements.

Later results (Salvi, 2006), confirm that the phonetic boundaries can be predicted within 20 msec, with 86.4% precision and 76.2% recall, based only on entropy measurements. To be able to compare these results with the literature, it is important to

note that the task of detecting phonetic boundaries is related, but not equivalent, to the task of aligning the speech signal to a reference transcription (*e.g.* Hosom, 2002). In the first case, no information is given about the phonetic content or about the number of boundaries for the speech utterances.

Chapter 5

Accent Analysis

Papers D and E describe studies on pronunciation variation in Swedish due to the speaker's regional accent. The distinctive characteristics of these studies, when compared to more traditional phonetic investigations, is the use of large data sets containing vast populations of subjects (speakers).

Most studies in phonetics use traditional statistics to analyse a limited number of observations collected in well-controlled laboratory conditions. The value of these studies is not questioned, because they provide means to isolate the effect of the variables of interest and eliminate or control disturbing effects. However, the drawback of these studies is the limited amount of observations that can be collected and analysed under controlled conditions, mainly due to the need to carefully annotate the material.

Given the large number of physical, linguistic, and psychological variables that affect speech production, there is a risk that the observations collected in a laboratory are biased towards the phenomena the researcher is investigating. In some cases, the study might reveal, with utmost precision, phenomena that are only valid for the small population of subjects considered and are hard to generalise.

The machine-learning methods used in automatic speech recognition (ASR) represent an alternative analysis framework that helps overcome these limitations. These methods are inherently developed for large data sets that are not accurately annotated. In spite of a reduced control over the experimental parameters, caused by the lower quality of the data, these methods allow the analysis of large populations of subjects and thus lead to results that are more general and representative of the entire population for a certain language. Moreover, if the data set is sufficiently large, the effects of the parameters over which we lack control can be assumed to be randomly distributed and, thus, cancel out in the experiments.

The aim is not to substitute the more traditional methods for speech analysis, but rather to verify results obtained within controlled experiments, and to extend and generalise them to larger populations.

From the point of view of ASR applications, these studies can also throw light on aspects and limits of the models and lead to more "knowledge aware" development, as for example in Salvi (2003b).

5.1 Regional Accent vs Dialect

When talking about geographical variability of language, many linguistic aspects can be taken into account. The hierarchical classification into language families, languages, and dialects is based on a complex mixture of factors such as word forms, grammatical rules, phonetic inventory, and prosodic features.

The only linguistic aspects considered in this thesis are in regard to pronunciation variation within a particular language. The term *accent* will be used to describe a pronunciation variety, according to Crystal (1997, ch. 8, p. 24):

> Dialect or accent?
>
> It is important to keep these terms apart, when discussing someone's linguistic origins. *Accent* refers only to distinctive pronunciation, whereas *dialect* refers to grammar and vocabulary as well. [...]

To avoid confusion, a further distinction needs to be made between the use of the term accent in this thesis and its use in prosodic studies: in the latter case, the focus is on specific patterns of suprasegmental stress.

5.2 Method

The speech signal is represented with regularly spaced feature vectors that constitute data points in the feature space. Each instance of a phoneme (phone) is thus characterised by a variable number of feature vectors or data points, depending on its length in time. This contrasts with the common practise in phonetic experiments of measuring a set of features at suitable points in time chosen by the researcher, or of averaging them over previously annotated segments. As mentioned above, another difference is that, given the size of the recordings we are considering, the number of data points to be analysed can easily reach the order of hundreds of millions.

Both these observations suggest the need for an intermediate level in the analysis, *i.e.*, the analysis is not performed directly on the observations, but on a parametric (statistical) model of the observations. This allows us both to represent phonemes with a fixed-length set of parameters and to drastically reduce the size of the set of observations to be analysed.

Each phoneme, as spoken by each population of speakers from a certain accent area, is represented by a three-state left-to-right hidden Markov model. A multivariate unimodal Gaussian distribution is associated with each state of the model, representing the statistics (means μ and covariances Σ) of the data points associated with it. This is common practise in ASR and guarantees that the model

splits each phone, *i.e.*, each acoustic realisation of each phoneme, into three consecutive segments (initial, middle, and final), thus taking into account the non stationary nature of speech segments. The procedure to initialise and estimate the model parameters is also standard practise in ASR and makes recursive use of the Baum-Welsh algorithm, based on the Expectation-Maximisation paradigm. The input to this procedure, in addition to the current estimates of the model parameters, is simply the sequence of phonemes contained in each utterance, abolishing the need for time-annotated material.

The analysis of differences in the pronunciation of each phoneme is then performed in the model parameter space, rather than in the feature space. This is done either by defining a global distance between model states, *i.e.*, between Gaussian distributions, or by analysing in detail the individual model parameters. In the first case, the metric is used in conjunction with hierarchical clustering to build trees of relationships that show how the model states naturally group depending on their mutual similarity. In the second case, given two groups of states, discriminant analysis is used to rank the model parameters that best explain the separation between the groups.

The Metric

Let $p_1(x)$ and $p_2(x)$ be the two density functions we want to compare; the metric used is the *Bhattacharyya distance* (Bhattacharyya, 1943) defined as:

$$\mathcal{B}(p_1, p_2) = -\ln \int p_1(x)^{\frac{1}{2}} p_2(x)^{\frac{1}{2}} dx$$

This is an example of the Ali-Silvey class of information-theoretic distance measures (Ali and Silvey, 1966), of which the more common Kullback-Leibler and Chernoff distances are members. All these distances carry informations about the classification error of a classifier based on the two distributions considered. The Kullback-Leibler distance, defined as

$$\mathcal{D}(p_1, p_2) = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} dx$$

has the drawback of being asymmetric, *i.e.*, $\mathcal{D}(p_2, p_1) \neq \mathcal{D}(p_1, p_2)$. The Chernoff distance is a generalisation of the Bhattacharyya distance:

$$\mathcal{C}(p_1, p_2) = \max_{0 \leq t \leq 1} -\ln \int p_1(x)^{1-t} p_2(x)^t dx$$

The Chernoff distance provides a tighter bound on the classification error than the Bhattacharyya distance, but it implies an optimisation.

In the case that $p_1(x)$ and $p_2(x)$ are multivariate Gaussian densities, the Bhattacharyya distance can be computed in terms of the means μ_1, μ_2 and the covariance

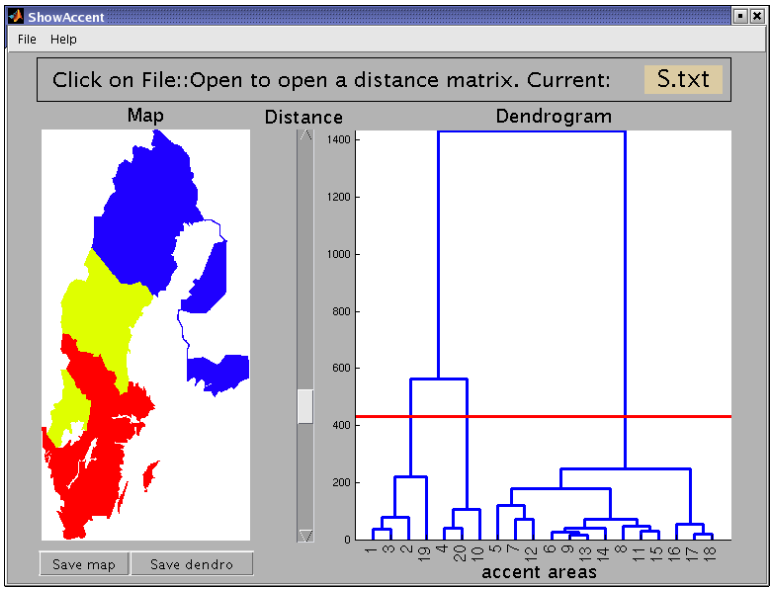


Figure 5.1: The ShowAccent interface. The map on the left displays the geographical location of the clusters for a distance level specified by the slider. On the right, the *dendrogram* is a compact representation of the clustering procedure.

matrices Σ_1, Σ_2:

$$\begin{aligned}\mathcal{B}(\mu_1, \Sigma_1, \mu_2, \Sigma_2) &= \frac{1}{8}(\mu_2 - \mu_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (\mu_2 - \mu_1) \\ &\quad + \frac{1}{2} \ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}\end{aligned}$$

The Analysis

The analyses performed in Papers D and E are based on agglomerative hierarchical clustering. The pairwise distances between the distributions corresponding to each HMM state are used to build a dendrogram that represents the hierarchical grouping of the corresponding allophones. The *complete linkage* method has been used to guarantee maximum separation between the clusters.

In Paper D, each phoneme is analysed independently and the distance between two allophones is averaged over the three HMM states. The analysis was performed by visual inspection; a Matlab tool was developed to simplify this task (Figure 5.1). For each phoneme, the tool shows the dendrogram on the right pane, a map on the left pane, and a slider in the middle. The slider can be used to select levels at

Region	Subregions	Subjects
I	15,16,17,18	1044
II	10,11,12,13,14	1098
III	8,9	1332
IV	7	76
V	5,6	307
VI	1,2,3,4	975
VII	19	25

Table 5.1: The number of subjects and the subdivision of the seven main accent areas in Sweden and part of Finland.

which the dendrogram can be cut. The map shows the geographical distribution of the groups of regions obtained cutting the dendrogram at that particular level. No attempt was made in this paper to establish automatically the most plausible number of clusters for each case.

In Paper E, the complete pool of states is clustered simultaneously, allowing distributions belonging to different phonemes to form groups of similarity. In Paper E, moreover, for each split in the binary tree obtained by the clustering procedure, linear discriminant analysis (LDA) is used to rank the model parameters that best explain the split. A number of indexes have been investigated in order to predict the optimal number of clusters from the data. Many of the indexes from Milligan and Cooper (1985) are highly correlated, which Milligan also remarks. Additionally, the Bayes Information Criterion was calculated.

5.3 Data

As in other studies in this thesis, the data is extracted from the Swedish SpeechDat database (Elenius, 2000). Other collections of recordings in Swedish might be suitable for similar investigations. The project SweDia (Bruce et al., 1999), *e.g.*, has collected a remarkable set of recordings with the aim of studying genuine dialects of Swedish. As already explained, the focus in this thesis is rather on accent variation than on dialect analysis. Both the rich phonetic knowledge and the data collected in the SweDia project are, however, precious resources for this kind of studies.

The SpeechDat database contains annotations of each speaker's accent region, as defined by Elert (1995) and illustrated in Figure 5.2. The number of speakers for each accent area (see Table 5.1) is representative of the total population for that region. Other variables, such as gender and age, are well balanced across accent areas. A detailed description of the pronunciation variation for each region is given in Paper D (Figure D.2, page D6).

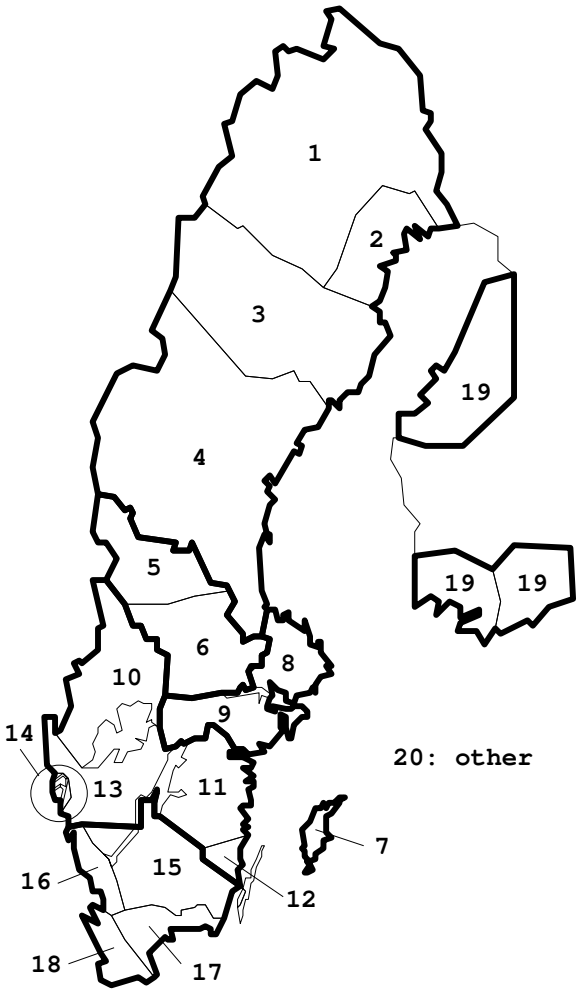


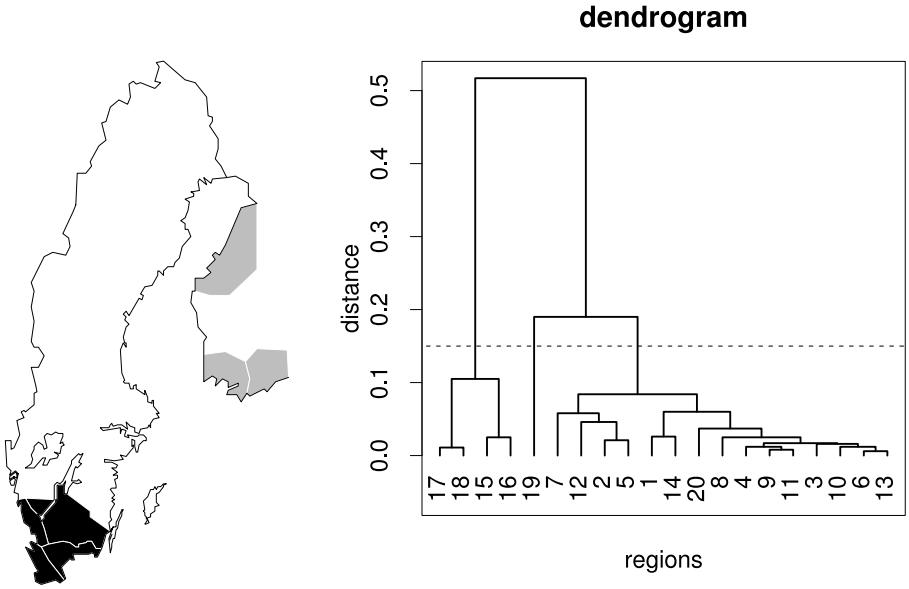
Figure 5.2: Geographic representation of the twenty accent areas in Sweden and part of Finland. The thick borders represent the seven main accent areas.

The official SpeechDat training set was used in the experiments for a total of 4500 speakers and 270 hours of speech. This results in approximately 97 million feature vectors computed at 10 millisecond intervals containing 13 Mel cepstrum coefficients $\{c_0$–$c_{12}\}$, and their first and second order differences $\{d_0$–d_{12}, a_0–$a_{12}\}$.

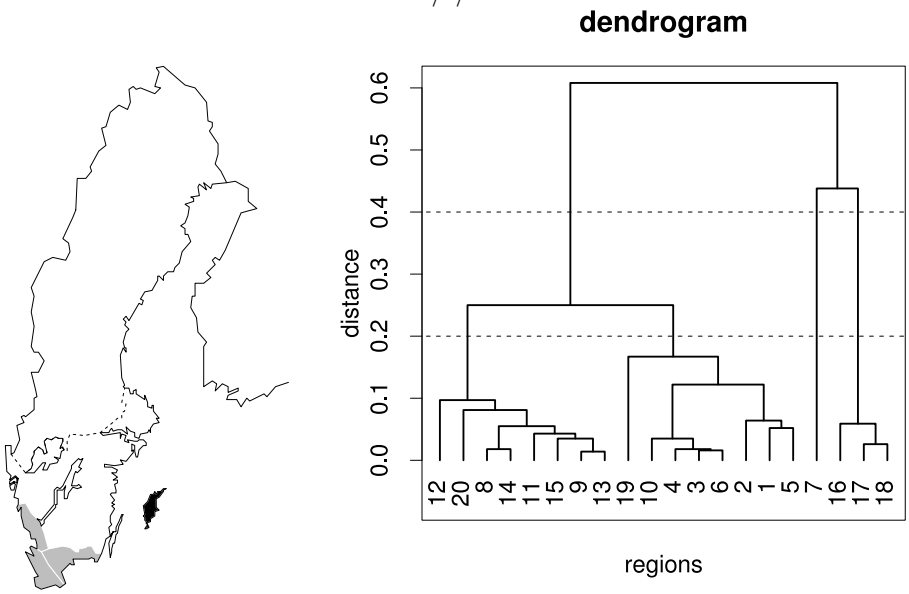
During training, 49 three-state HMM models (46 phonemes, 1 silence, 2 noise models) were trained for each of the 20 regions. This resulted in a total of 2940 states (or distributions). In the analysis of Paper E, the noise models and the retroflex allophone [ɭ] were removed (the last for lack of data points in certain regions). Thus, a total of 2760 states (distributions) were used for the clustering procedure.

5.4 Results of Paper D

Some of the results are reported in Figure 5.3 and can be summarised as follows. In all cases except for the fricative /ç/, the dendrogram shows two or three distinct clusters. In the case of /r/, the retracted pronunciation [ʁ] of the southern regions (15–18) corresponds to a very distinctive cluster in the analysis. Also, the Finnish variant (region 19) could be separated from the more standard pronunciation [ɹ]. For the phoneme /u:/, cutting the dendrogram at distance 0.4, for example, we get three clusters corresponding to the southern regions (16–18) where the vowel is diphthongised, to Gotland (region 7) where it is pronounced as [o:] (Elert, 1995), and to the rest of Sweden and Finland. An alternative partitioning, also plausible considering the dendrogram at distance 0.2, would split the third group with a border indicated by the dashed line on the map. For the fricative /ç/, the dendrogram does not show clearly distinctive clusters; however, a possible partition is the one shown on the map with a group corresponding to the affricate variant as spoken in Finland. Finally, the two clear clusters in the case of the fricative /ʃ/ correspond to the standard velar pronunciation in the central and south part of Sweden [ɧ] and to the more frontal pronunciation in the north and in Finland.

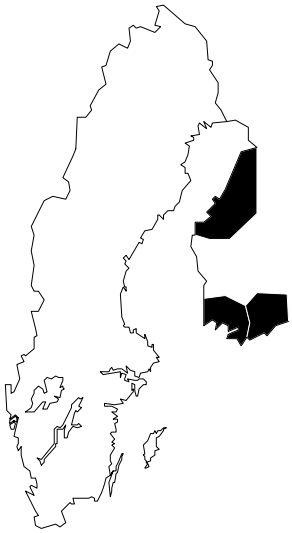


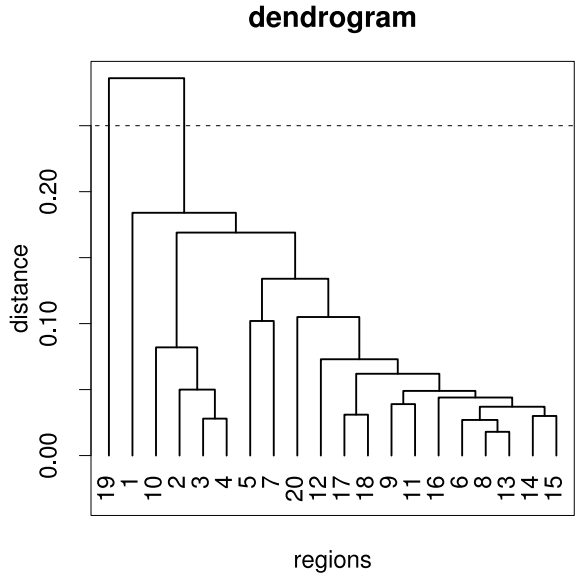
/r/



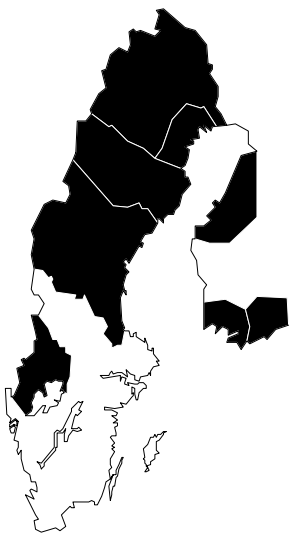
/uː/

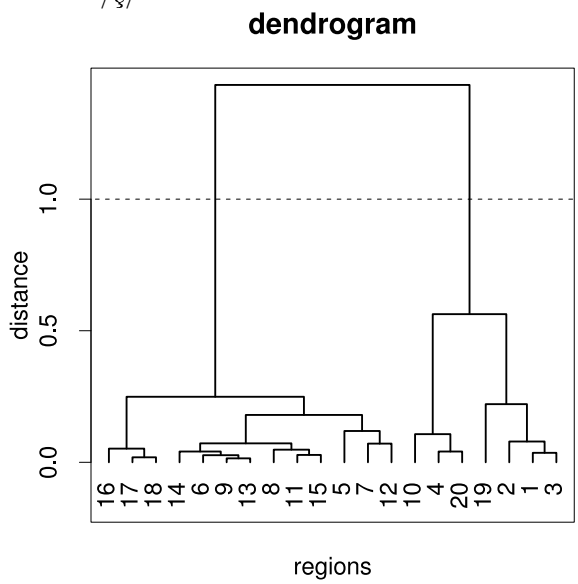
Figure 5.3: Four examples of pronunciation variation across Sweden and part of Finland. White, black, and grey regions represent clusters where the acoustic features are homogeneous. The dashed lines in the dendrograms represent the possible distance levels that produce the clusterings displayed in the maps.





/ç/





/ʃ/

Figure 5.3: (continued)

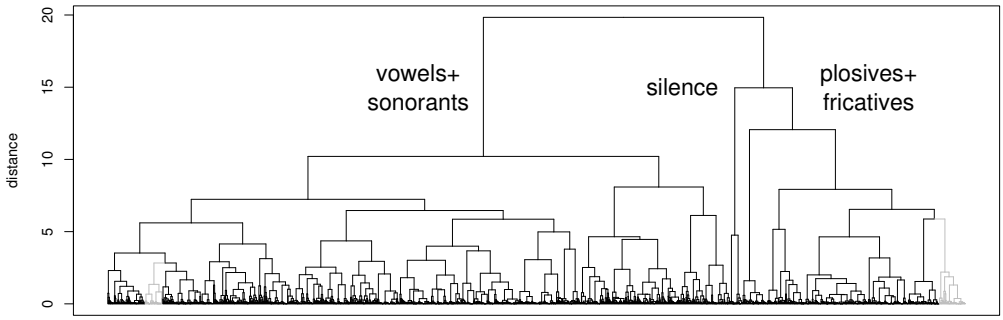


Figure 5.4: Dendrogram of the full clustering tree. The y-axis shows the dissimilarity level; the x-axis shows the states in the form phoneme-segment-region. Given the number of states, it is not possible to display each label. Broad classes are also shown in the picture. An enlargement of the two greyed-out subtrees is given in Figure 5.5.

5.5 Results of Paper E

The complete clustering tree for the 2760 distributions is shown in Figure 5.4. The reader is referred to the paper for a detailed description of the results that can be summarised as follows:

- Given the large number of distributions, it is not possible to visualise all the details simultaneously; one way to eliminate this problem is to split the analysis in two phases by approaching the tree from the top or bottom.
- Starting from the top of the tree we can observe the emergence of broad classes such as vowel/sonorant, plosive/fricative and silence.
- LDA shows that the first cepstral coefficients are mainly responsible for these groupings (mainly related to the energy and spectral tilt).
- Starting from the bottom of the tree, we can observe that the state position (initial, middle and final) and the identity of the phoneme are more prominent variables than the accent region. This means that states in different positions of a segment and belonging to different phonemes are split first, whereas the effect of the accent region comes last. In this case, the subtrees referring to the same phoneme and position can be used to observe relationships between different accent variants in a similar way as in Paper D (see the case of /r/ in Paper E).
- There are exceptions to this regularity that are particularly interesting, because they show the cases in which the pronunciation of one phoneme in the

language is assimilated to the pronunciation of another phoneme that is also part of the phonetic inventory of that language. Paper E gives a number of examples with fricatives and vowels. More examples are given in the following discussion and in Figure 5.5.

- Most of the indexes computed in order to estimate the optimal number of clusters are a monotone function of the number of clusters. This shows that the number of parameters used in the accent-dependent HMM sets is low if compared to the complexity of the problem and to the amount of data available for parameter estimation. This is in agreement with the fact that ASR models are usually some orders of magnitude more complex, by using both context-dependent models and mixtures of up to tens of Gaussian distributions for each state.

Although several interesting cases have been observed in Paper E, these results are not exhaustive and the process of analysing and interpreting the tree in Figure 5.4 continues. In Figure 5.5, we give two examples that did not appear in the papers and that are indicative of some of the above points.

The two dendrograms are extracted from the full tree of Figure 5.4 in which they are greyed out. The left dendrogram shows the clusters of the first states (s1) of three back vowels: two long /oː/, /uː/ and one short /ɔ/. The three vowels appear next to each other if we consider the full tree, but they form distinctive clusters, as the dendrogram illustrates. Exceptions to this are the allophones of /oː/ from the southern regions of Sweden (r16–r18) and the allophone of /uː/ from Gotland (r07); both are more open if compared to the more standard pronunciations, and therefore cluster with /ɔ/.

Similarly, for the right dendrogram in Figure 5.5, the fricatives /s/, /f/, and /v/ form distinctive clusters. However, the pronunciation of /ʃ/ is split into two main allophones: more retracted in middle and southern Sweden and more frontal in the north of Sweden and in Finland. This second allophone clusters with the southern allophone of the phoneme /ʂ/, and successively with the other front fricatives /s/, /f/, and /v/. Note also that the two clusters of the phoneme /ʃ/ are identical to the ones obtained in Paper D (Figure 5.3), where each phoneme was analysed separately, confirming the robustness of the method.

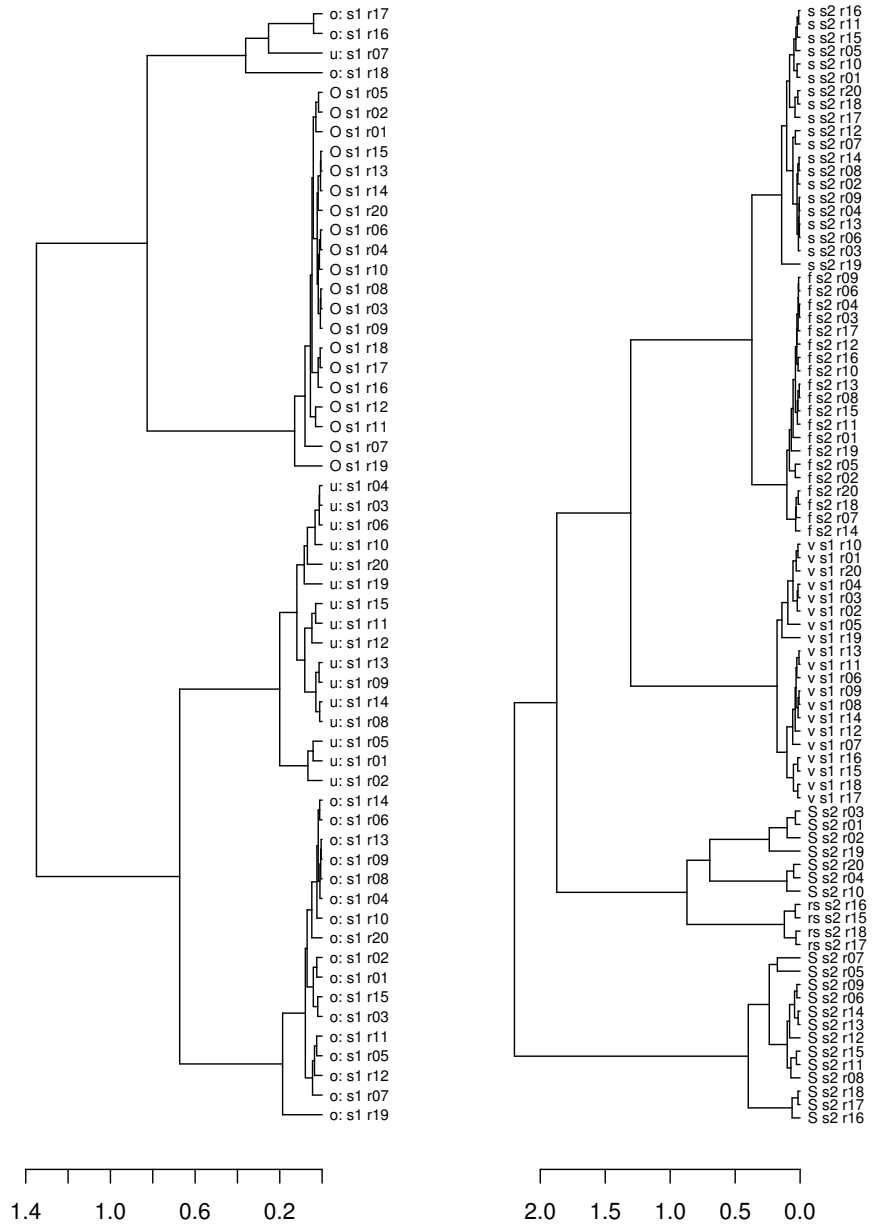


Figure 5.5: Two examples extracted from the complete clustering tree (Figure 5.4), with back vowels (left) and fricatives (right). The labels indicate phoneme (SAMPA), state (s1, s2 or s3), and region (r01–r20). The SAMPA symbols that graphically deviate from IPA are: O → ɔ, S → ʃ and rs → ʂ.

Chapter 6

Modelling Language Acquisition

When we described the speech communication chain in Chapter 2, we assumed that the two interlocutors were familiar with the language they were speaking. This chapter describes a study carried out within the project MILLE (Lacerda et al., 2004b) that aims at studying the phenomena that occur in an infant's attempt to acquire the skills necessary to engage in spoken interaction.

The theoretical standpoint of the project MILLE are the Ecological Theory of Language Acquisition (Lacerda et al., 2004a) and Lindblom's Emergent Phonology (Lindblom, 1999) which claim that very little, possibly only the physiology of the speech and hearing organs and of the brain, is transmitted genetically from generation to generation. The skills to control these organs and to interpret sensory information are not inherited and must be learned from the environment. This is in line with Gibson's Ecological Psychology (Gibson, 1963) that was initially developed for visual perception, and first stressed the importance of the environment in understanding human behaviour.

Figure 6.1 is a modified version of Figure 2.1 (page 7) that takes into account language learning aspects of the speech chain. Here, speaker and listener are replaced by child and parent. The child is exposed both to the sounds she has produced and to the physical environment, which consists mainly of acoustic, visual and tactile stimuli that the parent generates by talking, making gestures, and showing objects to the child. The loop on the left side of the figure, *i.e.*, from the brain of the child/speaker to her vocal muscles, to the speech sounds, to her ear, and back to her brain, is important both for skilled speakers and for speakers taking their first steps in the use of a language.

For the skilled speaker, it can be seen as a control loop in which the speaker's perception of the sounds she has produced is compared to a system of acoustic categories in order to fine-tune her production mechanism (for example, correcting a *tongue slip*). In this case, both the motor control necessary to drive the speech organs and the perceptual acoustic categories are considered to be well established in the speaker's brain.

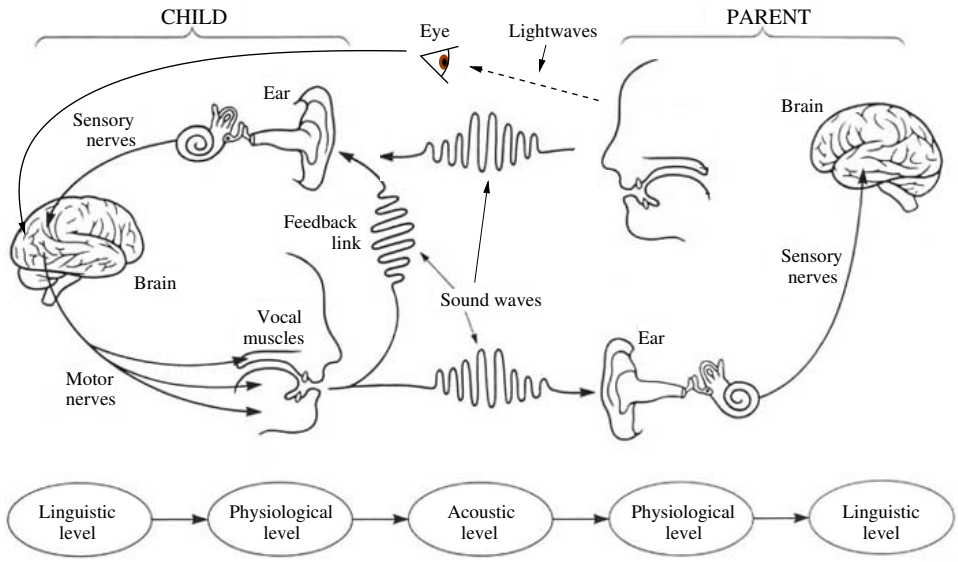


Figure 6.1: The Speech Chain after Denes and Pinson (1993), modified to account for language learning.

In the case of a child learning how to produce speech sounds, this feedback mechanism is equally important. During babbling, infants explore the articulatory and acoustic spaces in an attempt to gain familiarity with their production and sensory organs. An attempt to model this behaviour is Guenther's DIVA model of speech production. In its original formulation (Guenther, 1994), the model was an attempt to show how a child learns the motor control of the speech production organs. The model randomly generates configurations of the vocal tract and adjusts them in order to minimise the distance between the resulting sounds and some reference speech categories. Because the focus of this model is on production, the speech categories that constitute the perceptual reference are assumed to be already established in the child and to correspond to the specific language's phonemes. This is supported by evidence that, by the age of 10–12 months, children already show well-developed phonetic categorisation (Werker and Tees, 1984).

A more realistic model, when focusing on the first stages of language development, considers both the production and perception mechanisms to be in continuous evolution in the infant. Referring to Figure 6.1, the infant at the same time:

- develops acoustic categories (perhaps on a statistical basis),
- compares parent's production to own production, in order to bias her originally random babbling towards language-specific events (*e.g.* syllables), and

- correlates acoustic to visual (and other sensory) information in an attempt to associate linguistic events to a representation of her environment (semantics).

The theoretical standpoint of ecological theories of language acquisition is the assumption that each of these tasks are performed on the basis of very few evolutionary principles, based on statistics and energetics. The latter is supported by a vast body of literature focusing on locomotion in various species, which shows how the locomotive strategies in both animals and humans are optimised in order to minimise energy consumption.

The three points listed above are likely to interact with each other. For example, the phonetic inventory (acoustic categorisation) in a language depends on the contrastive use the language makes of each speech sound. The emergence of acoustic categories is, therefore, guided, not only by the statistical occurrence of each sound (first point), but also by the attempt to relate the acoustic evidence to meaning (third point). This interaction was ignored in the study described in Paper F that only focuses on the first of the three points.

6.1 The Emergence of Speech Categories

The study described in Paper F is an attempt to model the way acoustic categories emerge on the basis of the child's exposure to the linguistic environment mainly produced by the parent's voice. The main constraints to the learning process are:

- learning should be unsupervised because we do not want to assume innate information about speech sounds,
- learning should be incremental because the categorical perception improves with the child's increasing exposure to language, and
- the learning method should be compatible with the psycho-neurological processes that are likely to take place in the child's brain.

With regard to the last constraint, it is important to state the level of detail at which the modelling is aimed. Detailed computational models of neural systems are available for both supervised and unsupervised learning. Guenther and Bohland (2002) simulate the emergence of a phenomenon called Perceptual Magnet Effect (PME), first introduced by Kuhl (1991), in categorical perception of vowels with a self-organising map that closely resembles the functionality of a neural system; Sandberg et al. (2003) simulate in detail the functionality of the working memory, whereas Johansson et al. (2006) model the neurons in the mammalian visual cortex, just to give some examples.

In this phase of the project MILLE, however, the focus is on modelling the phenomenon *per se*, rather than its neurological implementation. Moreover, we are interested in modelling the phenomenon in its entirety rather than focusing on specific details. For this reason, and with the future aim of integrating this computational model into a larger system that would take into account other aspects

of language learning, the methods used are purely statistical. However, both the data representation and the properties of the algorithms are compatible with the real learning processes in many respects.

A problem that is in some respect similar to modelling the emergence of speech categories has been addressed in the last two decades in the field of automatic speech recognition (ASR). This is the data-driven optimisation of the sub-word units that constitute the building blocks of the ASR acoustic models (Holter and Svendsen, 1997; Deligne and Bimbot, 1997; Singh et al., 2002). In these studies, even though the sub-word units are derived from the data, rather than from phonetic knowledge, the target words are defined by orthographic transcriptions, making the methods supervised, from our perspective.

As discussed elsewhere in this thesis, when representing speech with regularly spaced acoustic feature measurements, the resulting feature vectors must be considered as sequences, rather than independent outcomes of some stochastic process. Furthermore, any attempt to relate these observations to meaning (*e.g.*, to build a phonological system out of a set of acoustic features), results in a variable-length sequence matching problem. In spite of this, the study presented in Paper F is limited to the classification of feature vectors in the feature space, and does not consider the time aspects of the problem. The focus here is on incremental, unsupervised classification.

A number of methods for clustering variable-length sequences have been introduced in Chapter 3. Most of the methods have been developed focusing on discrete observations, and are particularly appropriate for applications such as text clustering or gene analysis. None of the methods consider the problem of learning incrementally. Yet, they propose interesting concepts that should be further investigated when the emergence of syllable- and word-like sequences is taken into account.

6.2 Method

The speech signal is represented, as in the other studies in this thesis, by mel frequency cepstral coefficients (see Chapter 2). These are loosely related to the properties of the human ear (Davis and Mermelstein, 1980). Parameterisations that are more closely related to psychoacoustic phenomena exist (*e.g.* Hermansky, 1990; Skowronski and Harris, 2004) and may be considered in future studies. Given the nature of the study in Paper F, the conclusions obtained are likely to generalise to psychoacoustically-based features as well.

The method is based on Model-Based Clustering (Fraley and Raftery, 1998), described in Section 3.5. The algorithm has been modified according to Fraley et al. (2003), in order to account for incremental learning, and can be summarised in the following steps:

1. Start with a Gaussian model.

2. Get new data.
3. Adjust current model parameters to the new data.
4. Divide the new data into well-modelled and poorly-modelled points.
5. Try a more complex model adding a distribution for the poorly modelled points.
6. Choose the best model according to the Bayes Information Criterion (BIC). If the more complex model is best, go to 4; otherwise, go to 2.

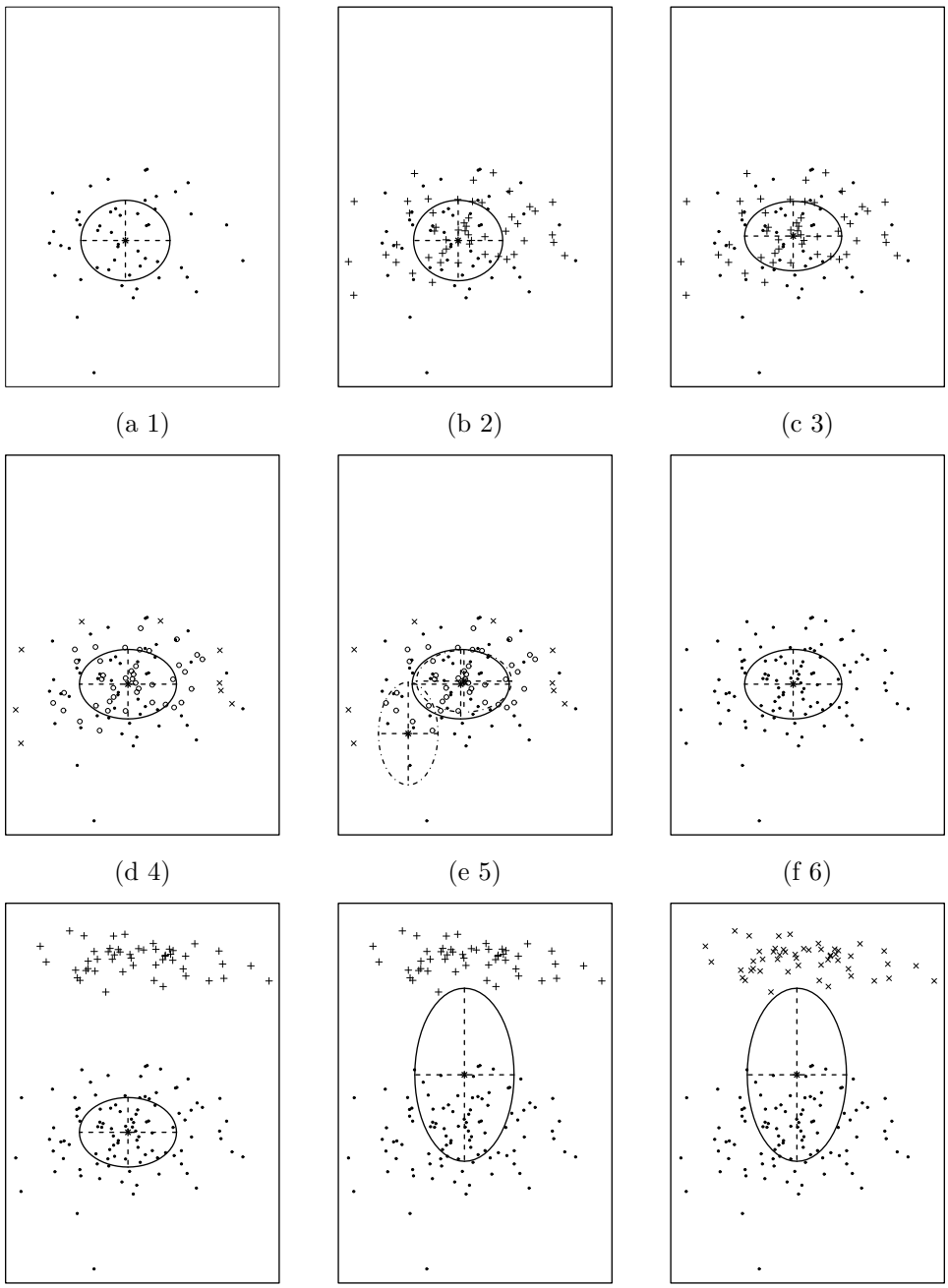
Figure 6.2 illustrates the algorithm with an example. Synthetic data points are drawn from two-dimensional distributions. Each plot is indexed in time by a letter (a–r). An additional number (1–6) refers, for each plot, to the corresponding step in the list above.

New data points are indicated by a "+", well-modelled points by "∘", and poorly-modelled points by "×". The current best model is indicated by one or more ellipses corresponding to the standard deviation of the bivariate multimodal Gaussian distributions. The alternative more complex model introduced by step 5 in the algorithm is represented by a dashed line. The sequence shown in the figures illustrates how the complexity of the best model incrementally follows the characteristics of the data.

6.3 Data

The data used in the experiments is a subset of the recordings done within the project MILLE. The interactions between infants and their parents (predominantly their mothers) are recorded using both the acoustic and visual channel. Conversations between the parents and adults have also been recorded as reference. Only the child-directed acoustic channel has been used in our experiments.

Using this kind of data, as opposed to traditional speech databases, ensures that the method is exposed to a similar environment as the child during learning. This is necessary because child-directed speech has peculiar characteristics if compared to adult-directed speech. The number of repetitions and the similarity of subsequent repetitions in child-directed speech seem to be optimised for facilitating learning. Also, the pseudo words that the parent makes up in their custom language seem to be an attempt to steer the child's random babbling towards syllable-like sequences and to facilitate the association between words and objects through onomatopoeia. An example is the utterance "brummeli, brummeli", which is often used in our material in association with a car toy.

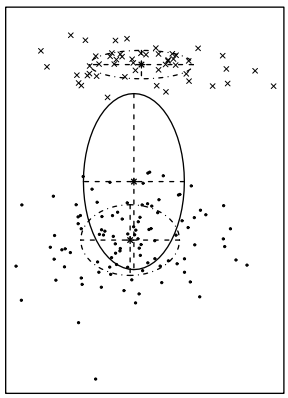


(a 1) (b 2) (c 3)

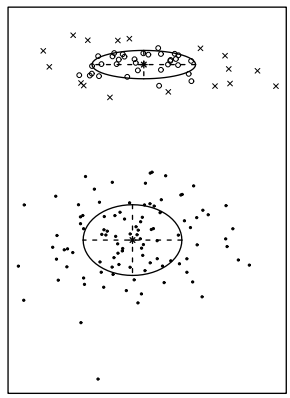
(d 4) (e 5) (f 6)

(g 2) (h 3) (i 4)

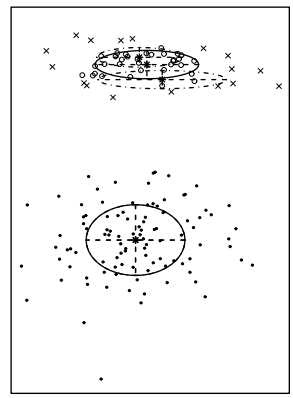
Figure 6.2: Illustration of the incremental clustering algorithm on synthetic bi-dimensional data. See the text for a thorough description.



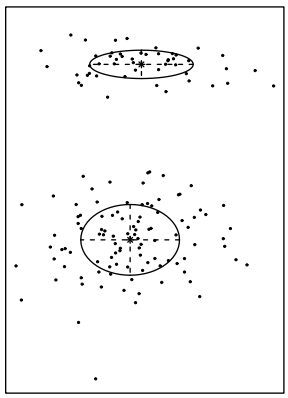
(j 5)



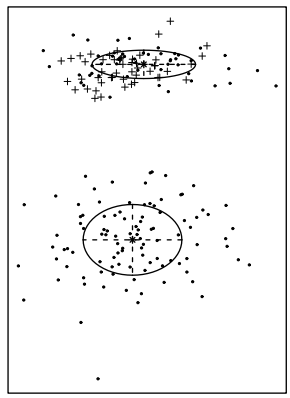
(k 4)



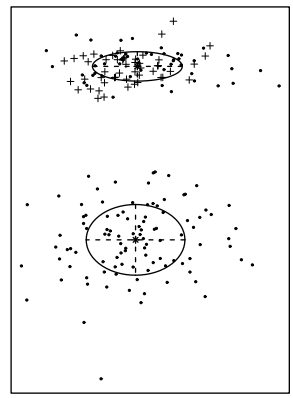
(l 5)



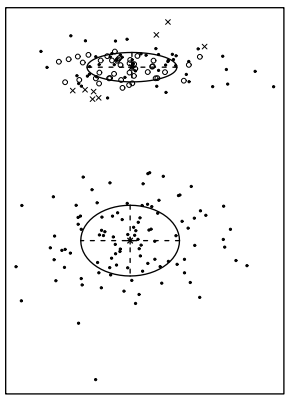
(m 6)



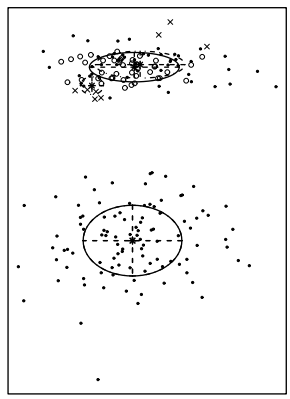
(n 2)



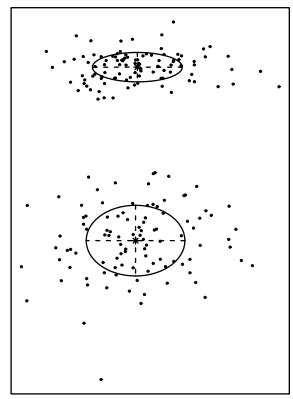
(o 3)



(p 4)



(q 5)



(r 6)

Figure 6.2: (continued)

6.4 Experimental Factors

The effects of two experimental factors have been investigated in the paper:

- The size of the data chunks that are incrementally presented to the algorithm ("frame length" in the paper).
- The size of the feature vectors that constitute a single observation ("number of coefficients" in the paper).

The first factor is related to the length of the auditory store that is available to the child for processing. This is an interesting factor in the light of the various theories on the existence of different auditory "buffers" with different sizes and properties (Cowan, 1984).

The second factor can be related to the amount of information carried by the auditory nerves from the cochlea to the auditory cortex. It is reasonable to assume that the child initially has a coarse perception of the sounds, mainly based on energy levels (or amplitude). This allows her to extract information that is mainly related to the prosody of the language. Later, the perception is refined and an increasing number of "coefficients" are adopted, allowing for a finer classification of the speech sounds.

6.5 Results of Paper F

Paper F is an exploratory study aimed mainly at observing the properties of incremental-model-based clustering on child-directed speech. Given the nature of the data and the task, the results should be compared to perceptual studies on the emergence of speech categories in children, in order to establish the model's predictive ability. The choice made in this study is to measure performance in a relative way by comparing the partitions obtained by the method when the experimental factors are varied. The measure of similarity between two partitions is based on information theoretical concepts described in Meilă (2002). This relative measure can validate the method in terms of stability, with regard to the experimental factors; however, it does not provide information on the method's ability to predict the perceptual data.

The simulations in Paper F show that although the size of the data chunks presented to the clustering algorithm has a relatively little effect on the clustering results, the number of coefficients play a fundamental role.

Another observation is that, in spite of the method not making use of time evolution information, the clusters obtained are stable in time.

The classifications obtained in different conditions, *e.g.*, with different sizes of the data chunks, are also in good agreement, in spite of the fact that the resulting partitions correspond to different total numbers of clusters.

Chapter 7

Discussion and Conclusions

7.1 General Discussion

Two main goals have been pursued in all cases. The first is the characterisation of speech sounds in statistical terms. The purpose of this characterisation has varied from speech recognition applications (Paper A), to the analysis of pronunciation variation (Papers D and E), to simulating the emergence of speech categories in infants (Paper F). The second goal that pervades these studies is the empirical analysis of machine-learning methods on real data over a wide range of problems, including supervised and unsupervised classification and regression.

In all cases, the continuous speech signal has been represented by equally spaced vectors of acoustic features,[1] which has highlighted one of the central problems in speech processing; that is, the need for modelling variable-length sequences.[2] Paper B addresses one aspect of this problem by analysing the behaviour of models that are commonly used in variable-length sequence classification, in specific conditions. Paper C contributes to the studies aimed at finding landmark points in the speech signal (in this case phonetic boundaries).

A common criticism of statistical machine-learning methods is that they provide "black box" solutions to practical problems, without adding any knowledge about the phenomena at hand. The studies described in part of this thesis (Papers D, E, and F) may contribute to showing that this is not necessarily the case, and that models obtained with machine-learning methods can be a valuable source of knowledge.

In the following, each paper is briefly summarised and discussed; see Chapters 4, 5, and 6 for more details.

[1] The fact that the same kind of features has been used consistently throughout this thesis is also a unifying aspect, but of secondary importance, because the methods employed are to a great extent independent of the features.

[2] In this respect, speech research shows similarities with, *e.g.*, gene classification research.

7.2 Paper A

The problem of mapping the acoustic speech signal onto a set of visible articulatory parameters is addressed in this paper. The articulatory parameters are used to animate an avatar that provides a visual aid to hearing-impaired persons using the telephone.

Two methods for acoustic to visual mapping have been investigated. Because both the acoustic signal and the articulatory parameters are represented by vectors of continuous measurements, the task of estimating each articulatory parameter is solved by the first method as a regression problem. The second method classifies the acoustic parameter vectors into visemes, *i.e.*, groups of sounds that share the same target values for the visual articulatory parameters. The parameter trajectories are, in this case, obtained by interpolation of the target values. These two strategies have been implemented by means of recurrent neural networks for regression and hidden Markov models for classification.

The classification method gave a number of advantages over the regression method, mainly because of the lower number of degrees of freedom in this task: in case of correct classification, somewhat stereotypical movements are produced, which simplifies the task of lip reading by the listener; the target values of some critical parameters (such as bilabial occlusion) are fully reached, reducing ambiguity of the movements; the interpolation procedure produces smooth trajectories that are pleasant to see.

An improvement to the regression method could be obtained by imposing constraints on the possible outcome of the mapping function and by smoothing the outputs of the neural network.

7.3 Paper B

Hidden Markov models are widely used in speech recognition. The Markov chain model specifies the time evolution of the speech production process, whereas the state-to-output probability models specify the relationship between states and acoustic observations (feature vectors). When using recurrent neural networks (RNNs) to estimate the state-to-output probabilities, a potential conflict emerges due to the fact that the time evolution model learned by the RNN can be in contrast with the Markov chain structure.

Paper B analyses this phenomenon in a phoneme recogniser with low-latency constrains. The results reveal an interaction between the two dynamic models. The degree of interaction depends both on the complexity of the RNNs and on the length of time dependencies in the Markov chain.

7.4 Paper C

Paper C analyses the phonetic boundaries obtained by the SynFace phoneme recogniser under low-latency constraints. A neural network estimates the posterior probabilities of a phonetic class given the acoustic feature vector (observation). The entropy of the probability estimates is studied in relation to the proximity to a phonetic boundary. Visual investigation shows that the entropy as a function of time assumes local maxima at phonetic boundaries. Results over a number of test sentences confirm that the entropy tends to be higher close to a boundary, even if variation due to other factors is large. The first and second derivatives of the entropy also carry information about the proximity to a boundary. Later studies, not included in this publication, show that the phonetic boundaries can be predicted within 20 msec, with 86.4% precision and 76.2% recall based only on entropy measurements.

7.5 Paper D

Pronunciation variation related to geographical factors is the topic of Paper D. A method is proposed to analyse large amounts of speech data that have not been transcribed at the phonetic level. Automatic speech recognition (ASR) techniques are used to fit a statistical model of the spectral features for each phoneme to the data containing recordings from 5000 speakers. The analysis is then performed on the model parameters, rather than on the data points, with the help of a measure of dissimilarity between probability distributions. Agglomerative hierarchical clustering is used to visualise the results and analyse the groups of similarities that emerge from the data. A few examples are shown where the clusters emerging from this procedure clearly correspond to well-known phenomena of pronunciation variation in Swedish.

7.6 Paper E

Paper E is an extension of Paper D in several respects. The statistics for three segments of each phoneme (initial, middle and final) are considered independently to take into account the dynamic properties of each phonetic segment. The agglomerative clustering procedure is performed on the full pool of distributions, allowing allophones of a certain phoneme to group with allophones of other phonemes, based on their acoustic similarity. An attempt to establish the optimal number of clusters in the data is discussed in the paper. Cutting the clustering tree (dendrogram) at any point results in two groups of distributions. Linear Discriminant Analysis is used to find the spectral characteristics that best explain acoustic differences between the groups so obtained. The results are, in most cases, in agreement with those in Paper D. They also show that the accent variable has usually a weaker influence on the acoustic feature that the phoneme identity and the position within

a phoneme. A few exceptions to this correspond to the cases where the allophonic variation *within* a particular phoneme exceed the variation *between* different phonemes.

7.7 Paper F

An incremental version of Model-Based Clustering has been used to simulate the unsupervised emergence of acoustic speech categories in the early stages of language acquisition in an infant. Preliminary experiments are performed using recordings from a mother talking to her child. The effect of two parameters are considered: the dimensionality of the acoustic features used as observations and the number of data vectors that are presented to the method at each step of the incremental learning procedure. The results are analysed in terms of number of acoustic categories as a function of time and by comparing the classifications obtained in different conditions with an information theoretical criterion. The results show that the acoustic categories emerging from the data are to a high degree consistent across the different experimental conditions. The classes are usually stable in time, *i.e.*, consecutive frame vectors belong often to the same class. The total number of classes is strongly dependent on the dimensionality of the acoustic feature frames, but only weakly dependent on the size of the data chunks incrementally presented to the method.

Bibliography

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of teh Royal Statistical Society*, 28:131–142.

Arabie, P., Hubert, L. J., and De Soete, G. (1996). *Clustering and Classification*. World Scientific.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821.

Bar-Hillel, A., Spiro, A., and Stark, E. (2005). Spike sorting: Bayesian clustering of non-stationary data. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17, pages 105–112. MIT Press, Cambridge, MA.

Batlle, E., Nadeu, C., and Fonollosa, J. A. R. (1998). Feature decorrelation methods in speech recognition. a comparative study. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 951–954.

Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. (2001). Support vector clustering. *Journal of Machine Learning Research*, 2:125–137.

Beskow, J. (1995). Rule-based visual speech synthesis. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 299–302, Madrid, Spain,.

Beskow, J. (2003). *Talking Heads - Models and Applications for Multimodal Speech Synthesis*. PhD thesis, KTH, Speech, Music and Hearing.

Beskow, J. (2004). Trainable articulatory control models for visual speech synthesis. *Journal of Speech Technology*, 7(4):335–349.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109.

Bruce, G., Elert, C.-C., Engstrand, O., and Wretling, P. (1999). Phonetics and phonology of the swedish dialects - a project presentation and a database demonstrator. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pages 321–324, San Francisco, CA.

Carlson, R., Granström, B., and Hunnicutt, S. (1982). A multi-language text-to-speech module. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1604–1607.

Cowan, N. (1984). On short and long auditory stores. *Psychological Buletin*, 96(2):341–370.

Cristianini, N. and Shawe-Taylor, J. (2001). *An Introduction to Support Vector Machine and other Kernel-Based Methods*. Cambridge University Press.

Crystal, D. (1997). *The Cambridge encyclopedia of language*. Cambridge university press, second edition.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with cluster via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Deligne, S. and Bimbot, F. (1997). Inference of variable-length acoustic units for continuous speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1731–1734.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Denes, P. B. and Pinson, E. N. (1993). *The Speech Chain: Physics and Biology of Spoken Language*. W. H. Freeman.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley-Interscience. John Wiley & Sons, INC.

Eisele, T., Haeb-Umbach, R., and Langmann, D. (1996). A comparative study of linear feature transformation techniques for automatic speech recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 252–255.

Elenius, K. (2000). Experience from collecting two Swedish telephone speech databases. *International Journal of Speech Technology*, 3(2):119–127.

Elert, C.-C. (1995). *Allmän och svensk fonetik*. Norstedts Förlag, 7th edition.

Fant, G. (1960). *The Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G., Liljencrants, J., and guang Lin, Q. (1985). A four parameter model of glottal flow. *QPSR*, 26(4):1–13.

Fisher, C. G. (1968). Confusions among visually percieved consonants. *Journal of Speech and Hearing Research*, 11(4):796–804.

Fraley, C. (1999). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281.

Fraley, C. and Raftery, A. (2003). MCLUST: Software for model-based clustering, density estimation and discriminant analysis. Technical Report 415, Department of Statistics, University of Washington.

Fraley, C., Raftery, A., and Wehrensy, R. (2003). Incremental model-based clustering for large datasets with small clusters. Technical Report 439, Department of Statistics, University of Washington.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41(8):578–588.

Fraley, C. and Raftery, A. E. (2002). Model based clustering, discriminant analysis, and density estimation. *Journal of American Statistical Association*, 97(458):611–631.

Gaertler, M. (2002). Clustering with spectral methods. Master's thesis, Universität Konstanz Fachbereich Mathematik und Statistik, Fachbereich Informatik und Informationswissenschaft.

Gibson, J. J. (1963). The useful dimensions of sensitivity. *American Psychologist*, 18(1):1–15.

Gordon, A. D. (1999). *Classification*. Chapman & Hall/CRC, 2nd edition.

Grant, K. and Greenberg, S. (2001). Speech intelligibility derived from asynchrounous processing of auditory-visual information. In *Proceedings of Audio-Visual Speech Processing (AVSP)*, Scheelsminde, Denmark.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, 72(1):43–53.

Guenther, F. H. and Bohland, J. W. (2002). Learning sound categories: A neural model and supporting experiments. *Acoustical Science and Technology*, 23(4):213–220.

Gurney, K. (1997). *An Introduction to Neural Networks*. UCL Press.

Hermansky, H. (1990). Perceptual linear prediction (PLP) analysis for speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hirasawa, K., Hu, J., Murata, J., and Jin, C. (2001). A new control method of nonlinear systems based on impulse responses of universal learning networks. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 31(3):362–372.

Holter, T. and Svendsen, T. (1997). Combined optimisation of baseforms and model parameters in speech recognition based on acoustic subword units. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 199–206.

Hosom, J. P. (2002). Automatic phoneme alignment based on acoustic-phonetic modeling. In *International Conference on Spoken Language Processing (ICSLP)*, volume I, pages 357–360.

Jankowski, C. J., Vo, H.-D., and Lippmann, R. (1995). A comparison of signal processing front ends for automatic word recognition. *IEEE Transactions on Speech and Audio Processing*, 3(4):286–293.

Johansson, C., Rehn, M., and Lansner, A. (2006). Attractor neural networks with patchy connectivity. *Neurocomputing*, 69(7–9):627–633.

Junqua, J.-C., Wakita, H., and Hermansky, H. (1993). Evaluation and optimization of perceptually-based ASR front-end. *IEEE Transactions on Speech and Audio Processing*, 1(1):39–48.

Kitawaki, N. and Itoh, K. (1991). Pure delay effects on speech quality in telecommunications. *IEEE Journal on Selected Areas in Communications*, 9(4):586–593.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics*, 74(1):119–147.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, 50(2):93–107.

Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, L., Marklund, E., and Sundberg, U. (2004a). Ecological theory of language acquisition. In *Proceedings of the Fourth International Worksop on Epigenetic Robotics*, pages 147–148.

Lacerda, F., Sundberg, U., Carlson, R., and Holt, L. (2004b). Modelling interactive language learning: Project presentation. In *Proceedings of Fonetik*, pages 60–63.

Li, C. and Biswas, G. (1999). Temporal pattern generation using hidden Markov model based unsupervised classification. In *Advances in Intelligent Data Analysis: Third International Symposium*, volume 1642, pages 245–256.

Li, C. and Biswas, G. (2000). A Bayesian approach to temporal data clustering using hidden Markov models. In *International Conference on Machine Learning*, pages 543–550, Stanford, California.

Li, C. and Biswas, G. (2002). Applying the hidden Markov model methodology for unsupervised learning of temporal data. *International Journal of Knowledge-based Intelligent Engineering Systems*, 6(3):152–160.

Lindblom, B. (1999). Emergent phonology. In *Proceedings of the Twenty-fifth Annual Meeting of the Berkeley Linguistics Society*.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, California.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Meilă, M. (2002). Comparing clusterings. Technical Report 418, Department of Statistics, University of Washington.

Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179.

Nicholson, S., Milner, B., and Cox, S. (1997). evaluating feature set performance using the F-ratio and J-measures. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 413–416.

Oates, T., Firoiu, L., and Cohen, P. R. (1999). Clustering time series with hidden Markov models and dynamic time warping. In *IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*, pages 17–21.

Porikli, F. (2004). Clustering variable length sequences by eigenvector decomposition using HMM. *Lecture Notes in Computer Science*, 3138:352–360.

Pujol, R. (2004). Promenade around the cochlea. http://www.cochlea.org.

Quatieri, T. F. (2002). *Discrete-Time Speech Signal Processing, Principles and Practice*. Prentice Hall.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Signal Processing. Prentice Hall.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice Hall.

Salvi, G. (1998). Developing acoustic models for automatic speech recognition. Master's thesis, TMH, KTH, Stockholm, Sweden.

Salvi, G. (2003a). Truncation error and dynamics in very low latency phonetic recognition. In *Proceedings of Non Linear Speech Processing (NOLISP)*, Le Croisic, France.

Salvi, G. (2003b). Using accent information in ASR models for Swedish. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 2677–2680.

Salvi, G. (2006). Segment boundary detection via class entropy measurements in connectionist phoneme recognition. *Speech Communication*. in press.

Sandberg, A., Tegnér, J., and Lansner, A. (2003). A working memory model based on fast Hebbian learning. *Network: Computation in Neural Systems*, 14(4):798–802.

Saon, G., Padmanabhan, M., Gopinath, R., and Chen, S. (2000). Maximum likelihood discriminant feature spaces. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1129–1132.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels, Support Vector Machines, Optimization and Beyond*. The MIT Press.

Schukat-Talamazzini, E., Hornegger, J., and Niemann, H. (1995). Optimal linear feature transformations for semi-continuous hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 361–364.

Siciliano, C., Williams, G., Faulkner, A., and Salvi, G. (2004). Intelligibility of an ASR-controlled synthetic talking face (abstract). *Journal of the Acoustical Society of America*, 115(5):2428.

Singh, R., Raj, B., and Stern, R. M. (2002). Automatic generation of subword units for speech recognition systems. *IEEE Transactions on Speech and Audio Processing*, 10(2):89–99.

Skowronski, M. D. and Harris, J. G. (2004). Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. *Journal of the Acoustical Society of America*, 116(3):1774–1780.

Stevens, S. and Volkman, J. (1940). The relation of pitch to frequency. *American Journal of Psychology*, 53(3):329–353.

Stevens, S. S., Volkmann, J. E., and Newmann, E. B. (1937). A scale for the measurement of a psychological magnitude: Pitch. *Journal of the Acoustical Society of America*, 8(1):185–190.

Ström, N. (1997). Phoneme probability estimation with dynamic sparsely connected artificial neural networks. *Free Speech Journal*, 5.

Titze, I. R. (1994). *Principles of Voice Production*. Prentice Hall.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymtotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13:260–269.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behaviour and Development*, 7(1):49–63.

Young, S. J. (1996). Large vocabulary continuous speech recognition: A review. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 3–28, Snowbird, Utah.

Appendix X

Phonetic and Viseme Symbols

In the papers included in this thesis different phonetic symbols are used depending on the focus of the work and on problems related to graphic software. This appendix lists IPA (International Phonetic Alphabet) and SAMPA (computer readable phonetic alphabet) symbols for Swedish and British English with examples. A slightly modified version of SAMPA, here called HTK SAMPA, has been used during the experiments to overcome the limitations of the string definition in the HTK software package.

The definition of the visemic groups used in the Synface project are also given here using HTK SAMPA symbols. Different viseme classifications can be found in the literature with varying degrees of details. Fisher (1968), *e.g.*, defines broad classes based on confusions in perceptual studies. The definitions given here are from Beskow (2004) for Swedish and Siciliano et al. (2004) for English. They were developed in order to control the visual synthesis in the Synface avatar, and are therefore more detailed (22 classes for Swedish and 26 for English, including silence). Broader viseme definitions can be obtained by merging some of the groups presented here.

X.1 Swedish

Plosives, fricatives and sonorants

IPA	HTK SAMPA (SAMPA)	Example	Transcription	
			IPA	SAMPA
6 plosives				
p	p	pil	pi:l	pi:l
b	b	bil	bi:l	bi:l
t	t	tal	tɑ:l	tA:l
d	d	dal	dɑ:l	dA:l
k	k	kal	kɑ:l	kA:l
g	g	gås	go:s	go:s
6 fricatives				
f	f	fil	fi:l	fi:l
v	v	vår	vo:ɾ	vo:r
s	s	sil	si:l	si:l
ʃ	S	sjuk	ʃʉ:k	Suh:k, S}:k
h	h	hal	hɑ:l	hA:l
ç	C	tjock	çɔk	COk
6 sonorants				
m	m	mil	mi:l	mi:l
n	n	nål	no:l	no:l
ŋ	N	ring	ɾɪŋ	rIN
r	r	ris	ri:s	ri:s
l	l	lös	lø:s	lox:s, l2:s
j	j	jag	jɑ:g	jA:g

Vowels

IPA	HTK SAMPA (SAMPA)	Example	Transcription	
			IPA	SAMPA
9 long vowels				
i:	i:	vit	vi:t	vi:t
e:	e:	vet	ve:t	ve:t
ɛ:	E:	säl	sɛ:l	sE:l
y:	y:	syl	sy:l	sy:l
ʉ:	uh: (}:)	hus	hʉ:s	huh:s, h}:s
ø:	ox: (2:)	föl	fø:l	fox:l, f2:l
u:	u:	sol	su:l	su:l
o:	o:	hål	ho:l	ho:l
ɑ:	A:	hal	hɑ:l	hA:l
9 short vowels				
ɪ	I	vitt	vɪt	vIt
e	e	vett	vet	vet
ɛ	E	rätt	ɾɛt	rEt
ʏ	Y	bytt	bʏt	bYt
ɵ	u0	buss	bɵs	bu0s
œ	ox (2)	föll	fœl	foxl, f2l
ʊ	U	bott	bʊt	bUt
ɔ	O	håll	hɔl	hOl
a	a	hall	hal	hal

Allophones

IPA	HTK SAMPA (SAMPA)	Example	Transcription	
			IPA	SAMPA
important allophones				
æː	ae: ({:)	här	hæːɹ	hae:r, h{:r
œ̞ː	oe: (9:)	för	fœ̞ːɹ	foe:r, f9:r
æ	ae ({)	herr	hæɹ	haer, h{r
œ̞	oe (9)	förr	fœ̞ɹ	foer, f9r
ə	eh (@)	pojken	pɔjkən	pOjkehn, pOjk@n
ʂ	rs	fors	fɔʂ	fOrs
less frequent allophones				
ʈ	rt	hjort	jʉʈ	jUrt
ɖ	rd	bord	buːɖ	bu:rd
ɳ	rn	barn	bɑːɳ	bA:rn
ɭ	rl	karl	kɑːɭ	kA:rl

Visemes

name	phonemes
sil	sil fil sp spk
C	C j
A:	A:
e	E e E: eh
e:	e:
i:	I i:
oe	oe ox: ox oe:
u0	u0
ae	ae ae:
g	N g h k S
O	O

name	phonemes
rd	rd rn rs rt
U	U u:
Y	Y y:
a	a
o:	o:
b	b m p
d	d n t
f	f v
uh:	uh:
l	l rl r
s	s

X.2 British English

Plosives, affricates, fricatives and sonorants

IPA	HTK SAMPA (SAMPA)	Example	Transcription	
			IPA	SAMPA
6 plosives				
p	p	pin	pɪn	pIn
b	b	bin	bɪn	bIn
t	t	tin	tɪn	tIn
d	d	din	dɪn	dIn
k	k	kin	kɪn	kIn
g	g	give	gɪv	gIv
2 phonemic affricates				
tʃ	tS	chin	tʃɪn	tSIn
dʒ	dZ	gin	dʒɪn	dZIn
9 fricatives				
f	f	fin	fɪn	fIn
v	v	vim	vɪm	vIm
θ	T	thin	θɪn	TIn
ð	D	this	ðɪs	DIs
s	s	sin	sɪn	sIn
z	z	zing	zɪŋ	zIN
ʃ	S	shin	ʃɪn	SIn
ʒ	Z	measure	meʒə	mezeh, meZ@
h	h	hit	hɪt	hIt
7 sonorants				
m	m	mock	mɒk	mQk
n	n	knock	nɒk	nQk
ŋ	N	thing	θɪŋ	TIN
r	r	wrong	rɒŋ	rQN
l	l	long	lɒŋ	lQN
w	w	wasp	wɒsp	wQsp
j	j	yacht	jɒt	jQt

Vowels

IPA	HTK SAMPA (SAMPA)	Example	Transcription	
			IPA	SAMPA
6 "checked" vowels				
ɪ	I	pit	pɪt	pIt
e	e	pet	pet	pet
æ	ae ({)	pat	pæt	pae t, p{t
ɒ	Q	pot	pɒt	pQt
ʌ	V	cut	kʌt	kVt
ʊ	U	put	pʊt	pUt
1 central vowel				
ə	eh (@)	another	ənʌðə	ehnVDeh, @nVD@
13 "free" vowels				
iː	i:	ease	iːz	i:z
eɪ	eI	raise	reɪz	reIz
aɪ	aI	rise	raɪz	raIz
ɔɪ	OI	noise	nɔɪz	nOIz
uː	u:	lose	luːz	lu:z
əʊ	ehU (@U)	nose	nəʊz	nehUz, n@Uz
aʊ	aU	rouse	raʊz	raUz
ɜː	Eh: (3:)	furs	fɜːz	fEh:z, f3:z
ɑː	A:	stars	stɑːz	stA:z
ɔː	O:	cause	kɔːz	kO:z
ɪə	Ieh (I@)	fears	fɪəz	fIehz, fI@z
eə	eeh (e@)	stairs	steəz	steehz, ste@z
ʊə	Ueh (U@)	cures	kjʊəz	kjUehz, kjU@z

Visemes

name	phonemes
sil	sil fil sp spk
O:	O: Ueh
A:	A: aq
D	D T
I	I ih
ae	ae V ax e eh
g	N g h k
Ieh	Ieh
oh	oh Q
OI	OI
S	S Z dZ tS j
U	U
eI	eI

name	phonemes
u:	u:
b	b m p
d	d l n r t
f	f v
eeh	eeh
i:	i:
ehU	ehU
aI	aI
s	s
Eh:	Eh:
w	w
z	z
aU	aU