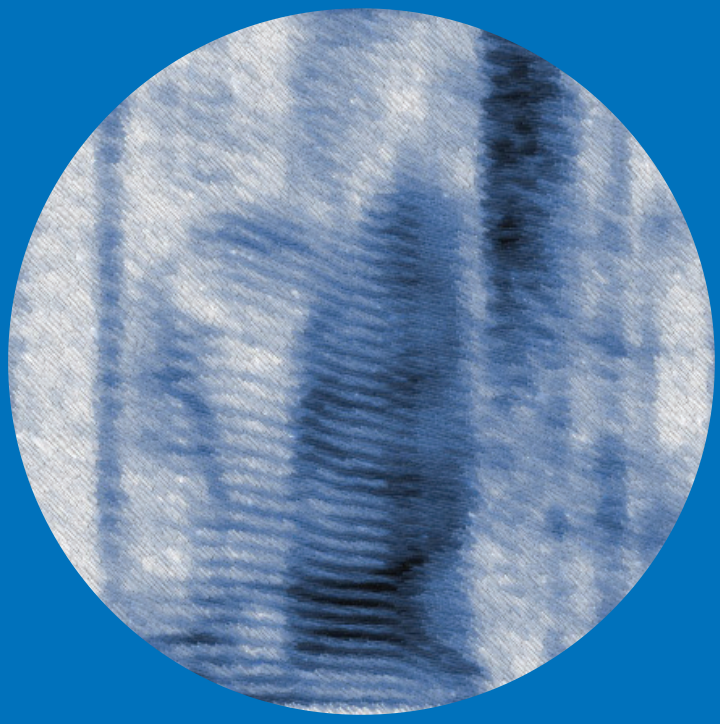
Department of Signal Processing and Acoustics

Studies on unsupervised and weakly supervised methods in computational modeling of early language acquisition

Okko Räsänen



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 55/2013

Studies on unsupervised and weakly supervised methods in computational modeling of early language acquisition

Okko Räsänen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall S1 of the school on 14 June 2013 at 12.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
Speech Technology Team

Supervising professor
Professor Unto K. Laine

Thesis advisor
Professor Unto K. Laine

Preliminary examiners
Professor Olli Aaltonen, University of Helsinki, Finland
Professor Rolf Carlson, Kungliga Tekniska högskolan (KTH), Sweden

Opponent
Professor Odette Scharenborg, Radboud University Nijmegen, Netherlands

Aalto University publication series
DOCTORAL DISSERTATIONS 55/2013

ISBN 978-952-60-5096-6 (printed)
ISBN 978-952-60-5097-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5097-3

Unigrafia Oy
Helsinki 2013

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Okko Räsänen

Name of the doctoral dissertation
Studies on unsupervised and weakly supervised methods in computational modeling of early language acquisition

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 55/2013

Field of research Language Technology

Manuscript submitted 3 December 2012 **Date of the defence** 14 June 2013

Permission to publish granted (date) 4 March 2013 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

This thesis addresses computational modeling of early language acquisition using statistical learning mechanisms. There is a constantly increasing amount of evidence from experimental psychology and brain imaging studies that human infants are sensitive to the statistical structure of sensory input and that their ability to extract statistics of speech signals plays a central role in learning of the native language. The idea of domain-general statistical learning mechanisms in language acquisition is in contrast to the nativist view of language acquisition, in which many language-specific innate factors have been traditionally assumed to exist in the human brain.

This thesis presents a series of computational studies addressing the questions of what kind of representations are learnable from speech signals and what kind of computational mechanisms are needed for the learning. The core idea is to model language acquisition from the perspective of a tabula rasa agent that does not have any advance knowledge of language or its relevant units such as phones, phonemes, syllables, or words, but simply comes into being with a number of generic statistical learning algorithms. When exposed to speech input in different experimental settings, these algorithms then start to model recurring patterns in the data and link these patterns to contextual variables such as simulated visual input associated with the speech contents. From a machine learning perspective, the studied methods correspond to unsupervised and weakly supervised machine learning algorithms, since language learning takes place without explicit supervision.

As a result of these studies, it is shown that spoken words can be learned from continuous speech based on the statistical structure of the speech input and without assuming a phonetic or other linguistically motivated intermediate representation of language. Different strategies for grounding the acoustic word patterns into their visual referents are also studied, and new methods for segmentation of speech into phone-like units and clustering of acoustic features into discrete categories are presented. Finally, it is shown that frequency characteristics of the human auditory system can also be derived from the statistics of speech signals, suggesting that distributional learning in auditory perception may not be limited to learning of linguistic representations of speech.

Keywords computational modeling, language acquisition, pattern discovery, speech processing, cognitive modeling, speech segmentation, unsupervised learning

ISBN (printed) 978-952-60-5096-6 **ISBN (pdf)** 978-952-60-5097-3

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2013

Pages 196 **urn** http://urn.fi/URN:ISBN:978-952-60-5097-3

Tekijä
Okko Räsänen

Väitöskirjan nimi
Ohjaamattomat ja heikosti ohjatut menetelmät kielenoppimisen laskennallisessa mallinnuksessa

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 55/2013

Tutkimusala Kieliteknologia

Käsikirjoituksen pvm 03.12.2012 **Väitöspäivä** 14.06.2013

Julkaisuluvan myöntämispäivä 04.03.2013 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Tämä väitöskirja käsittelee varhaisen kielenoppimisen laskennallista mallinnusta hyödyntäen tilastollisia oppimismenetelmiä. Jatkuvasti kasvava määrä kokeellisen psykologian ja aivotutkimuksen tutkimuksia on osoittanut että ihmislapset ovat herkkiä aistiärsykkeiden tilastollisille ominaisuuksille, ja että näillä tilastollisilla ominaisuuksilla on keskeinen rooli varhaisessa äidinkielen kehityksessä. Ajatus kielen omaksumisesta pelkkänä mukautumisena aistiärsykkeiden rakenteellisiin ominaisuuksiin ilman synnynnäisiä kielispesifejä oppimismekanismeja on ristiriidassa niin kutsutun perinteisen nativistisen ajattelumallin kanssa. Jälkimmäisessä synnynnäisille kielellisille mekanismeille annetaan suuri painoarvo.

Tämä väitöskirja sisältää joukon tutkimuksia jotka pyrkivät selvittämään minkälaisia tilastollisia rakenteita on opittavissa puhesignaaleista ja minkälaisilla oppimisalgoritmeilla tämä oppiminen voidaan saavuttaa. Työn ydinajatuksena on lähestyä kielenoppimista niin sanotun “tyhjän” oppivan agentin näkökulmasta. Tällä ei ole minkäänlaista ennakkokäsitystä tai -tietoa kieleen liittyvistä rakenteista, kuten äänteistä, tavuista tai sanoista. Sen sijaan agentti on varustettu tilastolliseen oppimiseen soveltuvilla algoritmeilla, jotka pyrkivät erilaisissa puhetta sisältävissä oppimistilanteissa löytämään signaaleista rakenteellisesti merkittäviä hahmoja. Koneoppimisen näkökulmasta kyseessä on ohjaamattomien ja heikosti ohjattujen hahmontunnistusmenetelmien kehitys ja soveltaminen, sillä varhainen kielenoppiminen tapahtuu poikkeuksetta ilman täsmällistä opetusta.

Tutkimuksen tuloksena voidaan osoittaa että puheessa esiintyvät sanat voidaan oppia jatkuvasta puheesta puhesignaalin tilastollisia ominaisuuksia hyödyntäen ja ilman että oppija tulkitsee puheen käyttäen ensin foneettisia tai muita lingvistisesti merkityksellisiä yksiköitä. Tutkimuksessa käydään läpi myös erilaisia oppimisstrategioita sanoja vastaavien akustisten hahmojen sekä niiden merkityksien yhdistämiseen että esitellään uudet menetelmät puheen segmentointiin äänteenkaltaisiksi yksiköiksi sekä akustisten piirteiden kategorisointiin klusteroinnin avulla. Lopuksi työssä osoitetaan, että ihmisen kuulojärjestelmän taajuusominaisuudet voidaan johtaa tilastollisella oppimismenetelmällä suoraan puhesignaalin aika-taajuus -rakenteista. Tämä viittaa siihen, että tilastollinen oppiminen ei välttämättä rajoitu kuulohavaintojen jäsentämisessä pelkästään kielellisten rakenteiden oppimiseen.

Avainsanat laskennallinen mallinnus, kielenoppiminen, hahmojen etsintä, puheenkäsittely, kognitiivinen mallinnus, puheen segmentointi, ohjaamaton oppiminen

ISBN (painettu) 978-952-60-5096-6 **ISBN (pdf)** 978-952-60-5097-3

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2013

Sivumäärä 196 **urn** http://urn.fi/URN:ISBN:978-952-60-5097-3

SATISFACTION OF ONE'S CURIOSITY IS ONE OF THE GREATEST SOURCES OF HAPPINESS IN LIFE.

- LINUS PAULING

Preface

The work related to this thesis was started in 2007 when Prof. Unto K. Laine recruited me as an undergraduate student to work on the computational modeling of language acquisition in a project called Acquisition of Communication and Recognition Skills (ACORNS). Since then, the passing years have been full of intriguing exploration and study across numerous interesting topics that have been more or less related to the initial target of my work. More than often have I found myself wandering from a topic to another due to the immense amount of interesting research problems related to speech technology, human cognition, and intelligent artificial and living systems in general.

However, due to the inspiring supervision of Prof. Laine, the work has never been truly a random walk. Therefore I would like to thank him for the immeasurable effort he has put into the supervisory work, and thank also for the hundreds of hours of stimulating and entertaining discussions we have had together related to various topics. He has also shown great patience and confidence in me by allowing me to spend enormous amounts of time on studies and experiments that I have decided to pursue without me being able to show their potential value in advance. The pre-examiners of the thesis, Prof. Rolf Carlson and Prof. Olli Aaltonen, also have my sincere gratitude for their evaluation of the manuscript.

I would like to give my special thanks to Toomas Altosaar for the close collaboration during these years. By always being eager to proofread my manuscripts, Toomas has mainly been responsible for the development of my academic language skills to a level where someone else in the community can also hopefully understand my intentions. In this context, I also thank Luis Costa for his guidance in the use of English language. I also thank Jukka P. Saarinen from Nokia for the fruitful collaboration and insightful discussions on many of our research topics and Prof. Paavo Alku for his support and advise during various occasions. I'm grateful to the other members of our research team, Seppo Fagerlund, Heikki Rasilo, Sofoklis Kakouros and Juho Knuuttila for the collaboration and positive working environment. Marko T. and Sami deserve thanks for their active

participation to the official 10.30 am (!) lunch breaks. Hannu, Jouni, Tapani, Tuomo, Emma, Magge, Mikkis, Ville P., Olli S., and many others whose name I have forgotten here have also helped with practical or not so practical things, provided useful discussions on various research topics, or kept me company during conference trips. In addition, our former laboratory engineer Martti Rahkila and our former and current secretaries Lea Söderman, Heidi Koponen, Mirja Lemetyinen, Ulla Sikander, Markku Hietala, our IT admin Jussi Hynninen, and Prof. Jorma Skyttä as the head of the department have my thanks for doing their best in enabling our daily work at the lab. I would also like to thank everyone else in the acoustics lab for the warm solidarity and friendly atmosphere, so that I have never had to feel weary when waking up early in the morning for the coming day at work.

During these years, my research projects have been supported by the following instances: EU IST Sixth Framework Programme, Nokia Research Center Tampere, the Finnish Graduate School of Language Studies, and Tekes. I have also received support in the form of personal research grants from Nokia Foundation, Jenny and Antti Wihuri Foundation, the Finnish Foundation of Technology Promotion, and the Research Foundation of Helsinki University of Technology.

Finally, I express my gratitude to my parents Tarja Räsänen and Prof. Keijo Räsänen, to my sister Annika, for being the best family one could possibly have, and to my lovely girlfriend Ida who has shown great patience and understanding towards my recurring mental absence during these several years.

Espoo, March 18, 2013,

Okko Räsänen

Table of contents

List of publications

I Okko Räsänen, Unto K. Laine and Toomas Altosaar. Blind segmentation of speech using non-linear filtering methods. In Ivo Ipsic (Eds.): *Speech Technologies*, InTech, pp. 105–124, 2011

II Okko Räsänen, Unto K. Laine and Toomas Altosaar. Computational language acquisition by statistical bottom-up processing. *Proc. Interspeech'08*, Brisbane, Australia, 1980–1983, 2008

III Okko Räsänen, Unto K. Laine and Toomas Altosaar. Self-learning vector quantization for pattern discovery from speech. *Proc. Interspeech'09*, Brighton, England, pp. 852–855, 2009

IV Okko Räsänen and Unto K. Laine. A method for noise-robust context-aware pattern discovery from categorical sequences. *Pattern Recognition*, Vol. 45, pp. 606–616, 2012

V Okko Räsänen. A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, Vol. 120, pp. 149–176, 2011

VI Okko Räsänen. Context induced merging of synonymous word models in computational modeling of early language acquisition. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12)*, pp. 5037–5040, 2012

VII Okko Räsänen and Heikki Rasilo. Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language. *Proc. 34th Annual Conference of the Cognitive Science Society (CogSci2012)*, Sapporo, Japan, pp. 887–892, 2012

VIII Okko Räsänen. Average spectrotemporal structure of continuous speech matches with the frequency resolution of human hearing. *Proc. Interspeech'2012*, Portland, Oregon, 2012

IX Okko Räsänen. Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, Vol. 54, pp. 975–997, 2012

Author's contribution

Publication I: "Blind segmentation of speech using non-linear filtering methods"

The author contributed to the development of the published algorithm and to the design of the experimental setup. Furthermore, the author implemented all versions of the studied algorithm, carried out all experiments and data analyses, and wrote the manuscript. Prof. Laine provided the original idea for the first version of the algorithm and Prof. Laine and D.Sc. Altosaar provided useful comments on all versions of the manuscript.

Publication II: "Computational language acquisition by statistical bottom-up processing"

The author of this thesis formulated the computational model reported in the publication, carried out all experiments, and wrote the manuscript. Prof. Laine provided helpful comments and general supervision of the work. D.Sc. Altosaar also provided useful comments related to the manuscript.

Publication III: "Self-learning vector quantization for pattern discovery from speech"

The author of this thesis took part in the design of the earlier versions of the algorithm together with Prof. Laine and formulated the final version of the published algorithm. In addition, the author implemented the algorithm, designed and performed evaluation experiments, and wrote the manuscript. Prof. Laine and D.Sc. Altosaar participated in the earlier versions of the algorithm and provided useful comments on the manuscript.

Publication IV: "A method for noise-robust context-aware pattern discovery from categorical sequences"

The present author carried out the development and implementation of the algorithm, in addition to designing and performing of all experiments. The author also situated the current work in relation to the previous literature and wrote the manuscript. Prof. Laine, in addition to general supervision, helped with the formulation of the mathematical notation and provided a number of useful comments and corrections to the manuscript.

Publication V: "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events"

The author was fully responsible for all the content in the publication. However, the author is grateful to D.Sc. Altosaar for extensive proofreading of the manuscript.

Publication VI: "Context induced merging of synonymous word models in computational modeling of early language acquisition"

The current author is responsible for all aspects of the publication. However, the author would like to thank D.Sc. Jukka P. Saarinen for providing the idea of using Random Indexing method in the study.

Publication VII: "Acoustic analysis supports the existence of a statistical learning mechanism in structural rule learning in artificial language acquisition"

The current author is responsible for formulating and performing the experiments, analyzing the data, drawing the conclusions, and writing the first version of the manuscript. The second author, H. Rasilo, developed the articulatory synthesizer used in the experiments. Rasilo also provided useful comments on the manuscript.

Publication VIII: "Average spectrotemporal structure of continuous speech matches with the frequency resolution of human hearing"

The current author is responsible for all aspects of this work. However, the author would like to thank Prof. Unto K. Laine, D.Sc. Daniel Aalto and other colleagues for the useful discussions related to the topic.

Publication IX: "Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions"

The author of the thesis is solely responsible for all aspects of this work.

List of abbreviations

ASR	automatic speech recognition
EEG	electroencephalography
EM	expectation maximization
FFT	fast Fourier transform
GMM	Gaussian mixture model
HAC	histogram of acoustic co-occurrences
IDS	infant-directed speech
LA	language acquisition
MFCC	Mel-frequency cepstral coefficients
NLM-e	native language magnet theory expanded
NMF	non-negative matrix factorization
OME	online mixture estimation
P&G	Park & Glass (algorithm)
PRIMIR	a developmental framework for processing rich information from multi-dimensional interactive representations
SLVQ	self-learning vector quantization
SOM	self-organizing map
STM	short-term memory
SWD	statistical word discovery (algorithm)
TOME	topographic online mixture estimation
TP	transitional probability

List of figures

List of tables

1. Introduction

Language is one of the most fundamental factors that differentiate humans from other animals. Its referential symbolic nature allows us to code, transmit, receive, and store information about the surrounding world and across the members of our society in a highly efficient manner. Importantly, language allows us to detach ourselves from our immediate surroundings in time and space, enabling the conceptualization of the state of the world in the past and in the future, and enabling the acquisition of knowledge from situations that we are not personally witnessing. In addition, inner speech serves our conscious thinking by integrating complex experience-driven mental representations of the world into a compact symbolic code that allows compositional structuring and manipulation of the contents of our thought (Vygotsky, 2012). Although it is difficult to come up with an exact definition for what language actually is, it is evident that language is ultimately realized as verbal, signed, or written communication with its characteristics shaped by the sociocultural environment of the language users, and with pragmatic purposes related to everyday social behavior (cf., Saussure, 1916; Port, 2010).

Since communication plays such an important role in our daily life, language has been an integral part of research since the early antiquity. Since then, the study of language has expanded and specialized to numerous scientific fields and disciplines (Fig. 1). Despite the inevitable fragmentation of research into different areas with more focused interests, or "scientific reductionism", integration is also required in order to form a unified understanding of the way that language is represented in our minds, how it is implemented at a neural level, and ultimately, how it is acquired and used in different communicative contexts (see, e.g., Meltzoff et al., 2009; Moore, 2007).

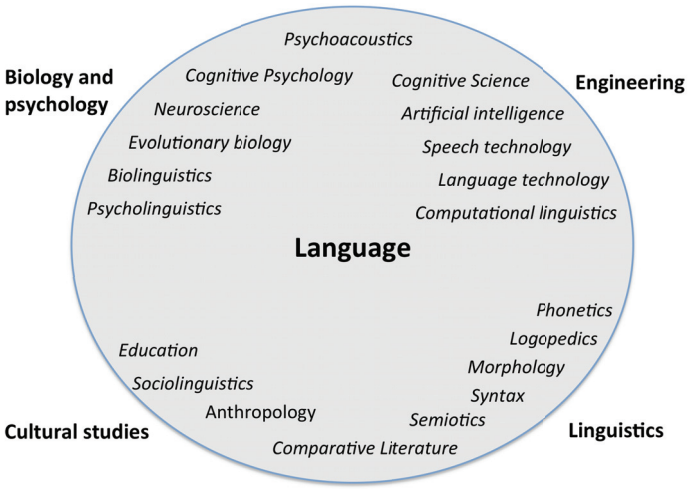


Figure 1: An example of different research fields and disciplines involved in the research of language. The listing is not meant to be comprehensive, but simply illuminates the diversity of research involved in understanding language as a phenomenon.

One central question calling for integration across disciplines in the scientific study of language is *how do human children learn to understand and produce their native language*? From an external perspective, the manner that human infants acquire their native language seems almost effortless. Instead of being explicitly taught, they learn to understand complex utterances and produce speech through everyday interaction with other people in various contexts. Due to continuous linguistic exposure, children also become capable to understand speech in adverse acoustic conditions despite different acoustic characteristics of different talkers. Moreover, they are able to fill in missing semantic and referential content of speech with the help of the context in which the communication takes place and add tens of new words to their vocabularies on a daily basis. The astonishing effectiveness of human learning becomes evident when one tries to create a comprehensive model explaining the early language acquisition (LA) process. Despite years of research, answers to questions such as how do infants discover words from continuous signals without pauses between words, how much of language capability is innate and how much is learned from experience (the so called nativism versus empiricism debate; see, e.g., McNeilage & Davis, 2005), how language becomes represented in the human brain, or what are the origins and proper rehabilitation approaches to language pathologies such as autism, dyslexia, or otherwise delayed language development, are still largely unknown. Although a plethora of important findings and candidate theories related to the above and other questions have been made, bits and pieces of information are still on the table waiting for assembly with some crucial parts still missing (Moore, 2007).

1.1 Language, speech, and their perceptual learning

Language has been traditionally characterized as a system with a property called *duality of patterning* (Hockett, 1960): all meaningful signs (words) of the system can be divided into a finite number of discrete, non-overlapping, building blocks (phones, phonemes, or syllables) that do not carry any individual meaning but can be combined sequentially in novel ways in order to form new words. Moreover, the combinatorial property of the discrete words themselves gives rise to a property of "*discrete infinity*", meaning that an infinite number of expressions can be generated from a finite set of words when they are repeatedly combined to form larger compositional structures (see, e.g., Studdert-Kennedy & Goldstein, 2003; Abler, 1989). Finally, the compositional structure of the elements is governed by the grammar of the language – a set of rules operating on the elements in order to distinguish roles, causalities, and temporal properties of the messages (Chomsky, 1965). Basically all research in mainstream linguistics assumes this type of discrete hierarchical representation of language, even though precise definitions for the sub-word elements may vary (Port, 2010; but see also Frank et al., 2012).

As tempting as the duality of patterning and discrete sequential coding of the language may seem in the theoretical sense, the question of *how do infants learn that a language consists of a sequence of meaningless units that make up meaningful words* is not currently understood. The major challenge is that the language is primarily realized as speech (or as sign language) and its physical characteristics vary notably from context to context and from talker to another. Young infants only have access to the acoustic surface structure of speech where individual phonemes or words and their boundaries are not readily perceivable, not to mention the abstract mental concepts inside the head of the talker. Instead of hearing sequences of discrete phonemic units, infants hear a continuous stream of air pressure variations that is a result of continuous movement of the articulators. The discrete sequential representation of language is already absent at the level of articulatory gestures where subsequent sounds are articulated in parallel using different articulators and causing temporally neighboring speech sounds to merge together into context-sensitive syllables and words (e.g., Studdert-Kennedy & Goldstein, 2003 or Lieberman, 2007, for an overview). The mapping from an abstract linguistic message to a physical speech signal is so complex that researchers in speech sciences have struggled for decades in order to find comprehensive descriptions for the mapping from the variable speech acoustics to the invariant discrete units of a language (see Port 2007, 2010, for reviews).

Against this background, one may ask what is actually learnable from speech and what kind of innate mechanisms and constraints are required to explain the language learning process? If speech is so different from the underlying abstract and discrete structure of language, how can one learn the latter by only having perceptual access to the former?

Naturally, in order to function at all, the spoken language must contain regularities that the talker can use to formulate a mental concept into a precisely defined series of motor commands – commands that then become converted into physical signals according to the laws of physics and are ultimately recognized as *familiar* by the listener. In this context, it has been known for a long time that humans and animals are sensitive to the statistical regularities in the sensory input and that the mammalian brain readily extracts and adapts to these regularities, or *patterns*, providing a starting point for language learning without any a priori linguistic knowledge.

In his seminal work, Hebb (1949) described a finding that the connections between biological neurons become strengthened due to their concurrent activity, explaining how brain is able to perform associative learning between neural representations (so-called *Hebbian learning*). This finding was followed by studies showing how the development of visual cortex depends on the post-natal visual experience (Wiesel & Hubel, 1963; Blakemore & Cooper, 1970), how different aspects of sensory processing are driven by learning (see Edeline, 1999 for a review), and e.g., how human infants become sensitive to the characteristics specific to native speech sounds due to linguistic exposure (Werker & Tees, 1984; Werker & Lalonde, 1988). It is now understood that the mammalian neocortex acts as a generic mechanism for learning statistical regularities in sensory input (see Mountcastle, 1978 and Hawkins & Blakeslee, 2005), and can be illustrated by the experiments where animals learn to process visual input with their auditory cortex (Sur et al., 1988) or where congenitally blind humans learn to process visual input with electrical stimulation of their tongue (Ptito et al., 2005).

However, it has been only during the last few decades when the study of early language acquisition has taken big leaps forward under a paradigm called *statistical learning* (also known as *distributional learning*). The statistical learning in language acquisition research sprung from the seminal paper of Saffran et al. (1996a) who showed that 8-month-old infants are able to segment words from an unknown language by relying solely on the statistical relationships between neighboring speech sounds. Since then, numerous studies have described how statistical learning operates already at birth (Teinonen et al., 2009), plays a role in perceptual categorization (Maye et al., 2002, 2008), word segmentation (Saffran, 2001; Newport & Aslin, 2004), acquisition of structural rules (see, e.g., Laakso & Calvo, 2011; Frank et al., 2012; Aslin & Newport, 2012), word referent mapping (Smith & Yu, 2008; Vouloumanos, 2008; Smith et al., 2011), and reading skills (Arciuli & Simpson, 2012). Furthermore, it seems that the mechanism is not limited to language but seems to be generic across all auditory and visual perception (Saffran et al., 1999; Kirkham et al., 2002; Bulf et al., 2011) and also exists in other primates (Hauser et al., 2001; Newport et al., 2004). The major implication of all these (and many other) studies is that a large proportion of the language acquisition seems to be explicable in terms of learning from the

statistics of speech and other sensory input[1] and that this learning can be accomplished without language-specific learning mechanisms or a "language module". This in contrast to the earlier poverty of the stimulus argument (e.g., Chomsky, 1975, 1980) that states that there is not enough structure in the language input to children for language to be learned from it, thereby also making the earlier nativist theories such as Universal Grammar questionable (but see also Yang, 2004 and Aslin & Newport, 2012). However, how such statistical learning mechanisms actually work, what type of language representations they can generate, and how much of language acquisition can they ultimately explain is not yet very well understood.

1.2 Computational modeling of spoken language acquisition

From an engineering point of view, the idea of statistical learning has close parallels to the fields of signal processing and machine learning. Whereas behavioral studies on statistical learning attempt to illuminate what type of signal statistics are learnable by the human brain, signal processing and machine learning researchers are devoted to understanding how different (sensory) signals can be efficiently represented in computational systems and how functionally relevant patterns can be learned from these signal representations. This is also where *computational models of early language acquisition* step in: as long as the human brain is regarded as a computational device obeying the laws of physics, any theory or model related to language acquisition should also endure computational implementation of the model so that the functionality of the model can be verified through simulations in realistic settings.

However, implementation of a theoretical model typically leads to a number of issues: first of all, a theory may be useful in understanding the LA process and in the formulation of more specific research questions, but the theory may be too vague to be implemented as an algorithm (cf. Marr's levels of analysis; Marr, 1982). Another possibility is that the theory covers the computational aspects of the process, but is not implementable given the existing limitations in the hardware. However, possibly the largest issue is that, given the complexity of the phenomenon, no single model can address all aspects of the learning problem simultaneously. This means that a large number of assumptions need to be made regarding the processes not studied in a given simulation, yielding very different results for different assumptions and having significant consequences on the ecological validity of the results.

Despite the evident challenges, computational models can provide useful knowledge regarding the LA process. First of all, they can set statistical

[1] Passive perception of speech input is naturally not enough for language acquisition, but the learning always takes place in an interactive setting between the learner, a caregiver, and a shared environmental context. However, these factors together with other cognitive development are not in conflict with the distributional learning hypothesis, but provide constraints to the learning process in the language domain.

baselines to the *learnability of data* (what, at least, can be learned from the data with the given constraints and assumptions). An estimation of *upper limits of learning* can be also attempted (e.g., Feldman et al., 2009a; Smith et al., 2006), although reaching conclusive results without notable simplification from the real world settings is usually difficult. Moreover, in addition to *replicating behavioral findings*, computational models may also help to *formulate new behavioral hypotheses* to be verified and to better predict or understand the nature of different developmental disorders related to the language faculty.

There is also another role for the computational models of human language acquisition: understanding the computational principles of the learning process also enables the development of computational devices that can acquire human-like speech processing capabilities and therefore also enables new ways of human-machine interaction. The existing state-of-the-art automatic speech recognition (ASR) systems are based on the estimation of statistical correspondences between acoustic and textual representations. This calls for expert knowledge in phonetics and huge amount of work in preparing the speech material for the estimation of the system parameters. Still, ASR systems perform well only on speech input that conforms to the acoustical and lexical content of the training material (see Lippmann, 1997). Novel words, grammatically incorrect constructions, background noise, and the paralinguistic aspects of everyday communication all cause major challenges to the system with a pre-defined set of capabilities. These shortcomings are not least due to the fact that *ASR systems do not understand speech*, but they simply convert acoustic input *directly* into textual output using the given elementary units and their estimated correspondences in both modalities. This is in contrast to human speech perception where everything is primarily based on the situated understanding of the message, and presentation of the speech as a written text is only a secondary goal achievable by literate people. The performance of the systems following the traditional ASR paradigm is still improving, but the improvement is becoming incrementally smaller and the estimated saturation level is far behind human speech perception capabilities (see, e.g., Scharenborg, 2007, and references therein). Given the complexity of language, it now seems evident that, even for machines, speech perception and production skills have to be *learned* through ever increasing experience with the linguistic environment if systems with human-like speech communication capabilities are desired in the future (Räsänen et al., 2012).

Given the above motivation, this thesis studies the computational modeling of language acquisition. More specifically, the focus is on unsupervised learning of statistical structures of continuous real speech with the methodological aim of working towards a self-learning computational agent capable of spoken language understanding. The methodological work is paralleled with the theoretical goal to understand what kind of structures are actually learnable from speech and what kind of environmental or

internal ("innate") constraints are needed for the language acquisition process to be successful.

1.3 Aims of the thesis

This thesis concentrates on models of statistical learning from spoken language input. The aim is to understand what kind of representations can be learned from speech signals without making strong a priori assumptions on the units that might be relevant for language processing (e.g., phones, phonemes, syllables, or words) and to see whether the simulated learning results are also supported by behavioral findings. Since as little advance expert knowledge is desired to be introduced to the algorithms as possible, the focus is on unsupervised and weakly supervised learning algorithms that do not require manually labeled learning data. The algorithms are evaluated in computer simulations that attempt to model different aspects of the language acquisition process.

The main research questions addressed in this thesis can be listed as follows:

- Does the speech signal contain statistical regularities that can be learned without strong supervision and assumptions of language specific innate mechanisms?
- What types of signal processing and pattern discovery algorithms are required for successful statistical learning from speech?
- What is the relationship between the automatically learnable structures and linguistically motivated units such as phones, words, or syllables?
- Does the statistical structure of speech have any implications to the manner that the human auditory system represents signal information in time and frequency?

The thesis specifically focuses on computational studies related to the learning of representations from *continuous speech* without the a priori linguistic knowledge, i.e. on the acquisition of the very first building blocks upon which later stages of language learning can rely. Since no model has been successful in acquiring a phonemic or orthographic representation of auditory speech (see also the discussion in P-V), the work related to computational models dealing with word learning or grammar induction from phonetic transcription or the orthographic layer are excluded from this work, but are discussed in depth elsewhere (e.g. in Witner, 2010; Daland & Pierrehumbert, 2011; Buttery, 2006). On the same basis, models of adult word perception such as TRACE (McClelland & Elman, 1986) or Shortlist (Norris, 1994) or language learning simulations utilizing manually trained ASR systems (e.g., Roy, 2005; Krunic et al., 2009) are not discussed since they do not explain how the representations used in the models come into

being during the early human development. Instead, the reader is recommended to see Scharenborg & Boves (2010) for an overview. Note that the aim is not to argue against conventional linguistic representations of language such as the phonemic system and duality of patterning but to understand how such systems could arise from perceptual learning. Also, the role of speech production is not much discussed since the focus of the current work is limited to the perceptual aspects of language acquisition. However, the development of speech production and its possible links to speech perception will be addressed in future work (see Rasilo et al., submitted).

1.4 Organization of the thesis

This thesis consists of two introductory sections and nine peer-reviewed publications. The first introductory section briefly describes how the computational models of language acquisition can be interpreted from the theoretical background of earlier, mainly behavioral, language learning studies. Section 3 then reviews the major findings from the computational modeling work performed so far. Since the computational modeling of language acquisition is a relatively new topic and many of the other studies have been carried out simultaneously with the work presented in this thesis, the contributions of the current work are directly integrated to the review of other work in the field. Both introductory sections are based on a more extensive review of the topic presented in P-IX.

The second part of the thesis consists of three journal articles, five papers in conference proceedings, and one book chapter. Figure 2 shows a thematic map of the publications and how they are related to each other temporally and methodologically.

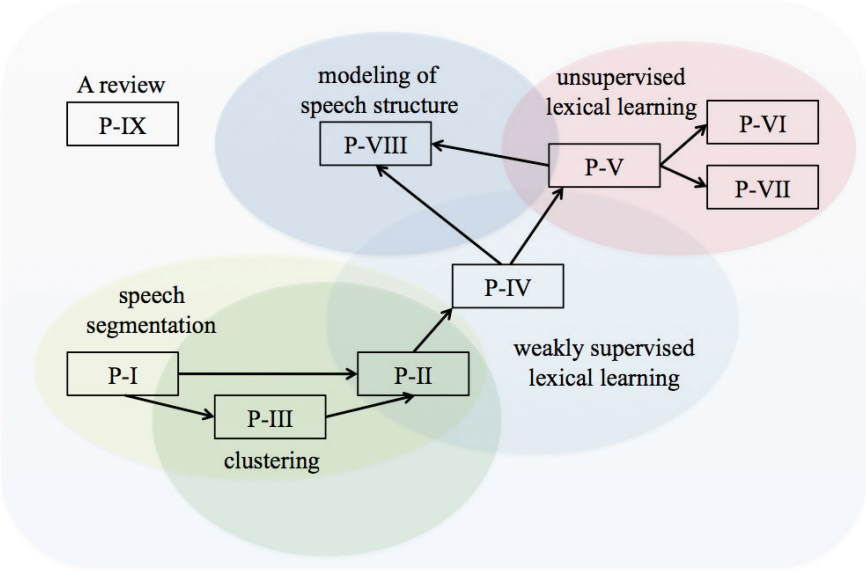


Figure 2: A thematic map of the publications. The arrows illustrate how methodological and theoretical content of the publications connect to earlier work.

1.5 Main contributions of the thesis

The technical contributions of this thesis are the following:

- An algorithm for automatic segmentation of speech into phone-like units (P-I).
- An algorithm for computationally efficient clustering of acoustic feature vectors into categories without requiring a priori manual definition of the number of clusters (P-III).
- An algorithm for discovering patterns from sequential data in the context of weak labeling (P-IV).
- An algorithm for discovering patterns from sequential data in a purely unsupervised manner (P-V).
- A method for deriving spectrotemporal filter-banks optimized for pattern detection in sensory data (P-VIII).

The main scientific contributions to the study of language acquisition are the following:

- A computational model showing that words can be learned from continuous speech when the speech is associated with concurrent visual input related to the speech contents (P-II, P-IV).
- A computational model showing that word-like units can be learned from speech without any a priori linguistic or phonetic knowledge or any concurrent multimodal input by simply analyzing the statistical structure of atomic acoustic events (P-V).
- A result showing that the frequency characteristics of human hearing are matched to the spectrotemporal statistical structure of continuous speech, and that this structure can be derived automatically from speech signals in an incremental manner (P-VIII).
- A simulation result revealing that the statistical learning approach can explain behavioral findings in grammar learning from an artificial language, a finding originally thought to require more abstract rule-based mechanisms (P-VII).
- A preliminary idea and a related simulation suggesting that the visual context can bind together lexical representations that do not originally generalize across acoustically varying realizations of a word (the so called generalization problem; P-VI).

- A review article bringing together existing and partially separate works related to the computational models of phonetic and lexical learning from speech, including the other work presented in this thesis. An interpretation of the proposed models based on the major theoretical frameworks and behavioral findings in language acquisition. Introduction of a taxonomy into word learning models with indirect and direct grounding to the word referents (P-IX).

2. The theoretical background for computational models of language acquisition

The theoretical background of the computational models of LA can be understood from the perspective of two theories of LA that try to integrate the existing findings on early language acquisition: the native language magnet theory expanded (NLM-e; Kuhl et al., 2008) and the PRIMIR framework of LA (Werker & Curtin, 2005). Although the theories do not claim fully explicit sequential ordering of developmental stages, their main connotations are as follows: NLM-e states that language learning starts by learning the statistical properties of native speech sounds, leading to enhanced phonetic perception of native contrasts. Once phonetic perception has achieved a sufficient proficiency level, words can be segmented and learned based on the sequential organization of perceived phonetic units. PRIMIR, on the other hand, states that the organization of the language faculty is driven by the acquisition of word forms directly from the acoustic surface properties of speech signals (or on the "general perceptual plane") that combine both phonetical and indexical features. Later, once sufficiently many lexical tokens have been memorized, the learner is able to discover similar subword patterns across different word tokens, providing a starting point to the subword level organization and perception of language (phonemes). Phoneme representation of spoken language then enables fast accumulation of new vocabulary since the learned tokens automatically generalize to their acoustic variants through phonemic encoding. However, PRIMIR also maintains that the statistical properties of native phonetic units affect the way that spoken words are represented in the general perceptual plane, but this representation does not yield proper generalizations across different contexts and talkers without the help from lexical learning.

The distinction of the learning order of words and subword units between PRIMIR and NLM-e also divides the computational models into two basic categories: those models where the phonetic system is learned before words (cf. NLM-e) and those where proto-lexical items are learned before the

phonetic system (cf. PRIMIR). Despite the intuition that knowledge of subword units such as phones or syllables must precede word learning because they are the basic building blocks of words, the answer to the question of representational learning order is not obvious. As Peter Jusczyk (1993a) wrote,

"*One potential problem with using characterizations of the mature state to guide research about the initial state is that it may lead one to assume that the elementary units that yield the best description for the adult's knowledge function as elementary units during acquisition of the knowledge… … to the extent that a description of the adult state of knowledge of the sound patterns of the language is best captured by assuming phonemic representations, we have to provide an explanation of how these representations develop in the course of language acquisition*."

Researchers in speech sciences have struggled for decades in order to find comprehensive descriptions for the mapping from variable speech sounds of the acoustic domain to the invariant and abstract linguistic units such as phones or even phonemes that can be placed serially to construct larger linguistic units such as words (see Port, 2007, 2010, for reviews). Despite the tremendous amount of work on this issue, the basic problem always seems to be that the variability in the acoustic tokens cannot be captured into segmental models that assume independence of a phone from the preceding and following phones, making accurate categorization of phone-sized units impossible when they are isolated from their context. The standard solution to get away with the difficulties at the segmental level is to extend the units to be context sensitive by making their characteristics dependent on the neighboring phones. Another possibility is to use lexical memory for disambiguation of difficult segments by first retrieving the most likely word, given the sequence of the initial phone hypotheses, and then seeing which phones (or phonemes) correspond to the ambiguous segments (e.g., TRACE, McClelland & Elman, 1986; but see also Norris et al., 2000). Although feasible for segment disambiguation, both of these approaches are not compatible with the idea that speech perception consists of the perception of sequences of independent units that are realized as sequences of phones. Otherwise the surface structure of speech should enable this type of serial segmentation into the discrete building blocks despite the variation introduced by coarticulation. If a phonemic system of sequential discrete elements exists, at least there seems to be no direct access to it from the surface structure of speech.

The second issue from the perspective of language learning is that, even before the categorization of speech sounds into a finite set of categories, the discovery of the phone-like segments themselves is problematic in the absence of a priori knowledge of their structure. While it has been proposed that humans and other primates are capable of primitive segmentation of a continuous acoustic stream into acoustically coherent segments, namely, basic-cuts (Kuhl, 1986; 2004), the correspondence of these units with linguistically motivated phones is not direct. Several diverse computational

methods for blind segmentation of continuous speech into phone-like units have been proposed (Scharenborg et al., 2007; Esposito & Aversano, 2005; Estevan et al., 2007; Aversano et al., 2001; Almpanidis & Kotropoulos, 2008; P-I) and they all systematically fall short of an ideal performance if manually performed phonetic transcription is used as a reference. What is common to these methods is that they analyze changes in spectral content of the speech signal and hypothesize phone boundaries at points of notable discontinuity in the spectrum. While this type of chunking of the speech signal can detect approximately 70–80% of phone boundaries (with ±20 ms accuracy), many of the phone transitions are still detected with very low accuracy or are detected only with an otherwise significant amount of oversegmentation (Räsänen et al., 2009a). For example, the overall quality of the segmentation is too low to be directly utilized as a front-end processing before feature extraction in ASR systems (see also Räsänen & Driesen, 2009). Only when context-sensitive phone models are imposed in a top-down manner and taught to the segmentation algorithm using pre-recorded speech data in supervised training paradigms, the segmentation algorithms reach segmentation performance that starts to converge with the definitions of phone boundaries (e.g., Demuynck et Laureys, 2002; Toledano et al., 2003; Keshet et al., 2005). In this case the segmentation models are essentially built manually upon the criteria that are also used to evaluate their performance.

There is also notable evidence that human listeners do not only pay attention to the sequential evolution of phonetic units, but store detailed supra-segmental and episodic acoustic information regarding speech tokens. Variables such as talker and speaking style characteristics have been shown to affect speech perception performance (e.g., Pisoni, 1997). Young infants' representations of words appear to be holistic and contain information regarding not only phonetic, but also indexical and stress information related to the word forms (Houston & Jusczyk, 2003; Curtin et al., 2001; Curtin et al., 2005). Infants as old as 14 months also fail to discriminate phonetic contrasts in otherwise similar novel words when learning names of external referential objects (Stager & Werker, 1997). Finally, there is some evidence that event-related potentials in the brain, often used to measure categorical auditory perception, are equally sensitive to phonetically relevant and irrelevant acoustic parameters (Aaltonen et al., 1994).

Due to the inability of the phonemic/segmental view of speech perception to explain the acoustic mapping problem and the effects of suprasegmental acoustic details on adult speech perception, contemporary views have emerged that question the entire existence of segmental phonemes as fundamental units of speech perception (Port, 2007; Pisoni, 1997; Warren, 2000). These views are also supported by the detailed analyses of pronunciation errors in young children that point towards suprasegmental or even word-level representations of produced words instead of phonologically motivated encoding of word forms (e.g., Waterson, 1971; see also Markey, 1994, for an overview).

Even if the phonemic representation of language is present in our minds and used to code and decode linguistic messages, the problem is that the mapping from acoustic signals to phonemic representations cannot be easily learned directly from continuous speech by simply analyzing statistical properties of acoustic events without support from some additional source of information. Even if the phone segmentation would succeed with perfect accuracy, the speech sounds from a number of different talkers do not neatly group into clusters of phonetic categories in terms of their acoustic features, but largely overlap in the acoustic space. This is demonstrated in the work of Feldman et al. (2009a), where Bayesian modeling (clustering) with theoretically well-justified mechanisms for learning was used to learn phonetic categories of American English vowels from the formant data of Hillenbrand et al. (1995). Despite the fact that the formant frequencies were estimated from isolated productions of the vowels instead of continuous speech, classification of the segments into correct phone categories was far from perfect. Only when support from the lexical layer was utilized in order to perform context-sensitive classification, the categorization of the segments became successful (Feldman et al., 2009a).

Given all the considerations above, it is not obvious that the infants would learn their native language by *first* acquiring a fully functional phonetic system of the language and only then start learning words as sequences of phones. As for the phonemes, some sort of proto-lexical layer becomes almost necessary as long as phonemes are defined as the smallest units of language that contrast between two words. However, the evidence is not conclusive. First of all, the above discussion does not take into account the fact that human infants are not only equipped with auditory capabilities, but can also use information from other modalities to disambiguate situations that are not separable in the purely auditory domain. Another important factor is that human infants are not only listening, but also experimenting with speech production. Infants are equipped with an articulatory system that gives them access to the constraints and possibilities of speech sound generation, revealing another representation of speech acts that is not linear with respect to the auditory domain.

In general, what kind of sub-lexical and lexical structures can be actually learned from speech with different types of approaches, constraints, and assumptions is not fully known. As will be seen in the following sections, partial success has been achieved with both lexicon-first and subwords-first approaches, but no single model has been so far able to convincingly explain the integral development and interdependence of the two systems.

3. Computational models of phonetic and lexical learning

Based on the distinction of NLM-e and PRIMIR theories of language acquisition, the existing computational models of speech perception development can be roughly divided into two main categories: those attempting to explain acquisition of phonetic categories directly based on the statistical structure of speech input (cf., NLM-e) and the models that start by learning word-like units from speech without explicitly assuming or modeling phonetic or phonemic representations of speech (cf. PRIMIR). In addition, a third group of models, here referred to as integrated models of LA, has recently been emerging and addresses the development of speech production in addition to perceptual development. All three major categories can be further divided into sub-categories according to the type of signals used in the simulations (e.g., a possible visual input in addition to speech) and the type of signal representations used in the statistical learning algorithm. Table 1 shows one way to organize the different methods according to their learning goals and the signal representations they use in their processing.

In the next sub-section, the existing work on models of phonetic learning will be briefly discussed. This is followed by a discussion of what lexical learning is all about, how word semantics are related to word learning, and finally a brief review of work done in computational modeling of lexical learning is given.

Table 1: Classification of computational models of language acquisition from continuous speech into categories according to learning goals and signal representations used in the experiments.

Type of model	***Models of phonetic learning***	***Models of lexical learning***		***Integrated models of perception and production***
How?	Cluster instantaneous spectral features into distinct categories	Discover and model recurring spectrotemporal patterns in continuous speech		Model the learning of speech perception and production simultaneously in an interaction framework
Sub-classes	Audio features only	Indirect lexical grounding (unsupervised learning)	Multivariate pattern matching	Phonetic learning only
			Transitional probability analysis	
	Audiovisual clustering	Direct lexical grounding (weakly supervised learning)	Segment-based representation	Both phonetic and lexical learning
			Fixed-frame representation	

3.1 Models of phonetic learning

One of the basic hypotheses in the NLM-e theory (Kuhl et al., 2008) is that the first stages in LA are dominated by the attunement of the infant to the statistical properties of speech sounds. More specifically, NLM-e states that the exposure to infant-directed speech (IDS) drives statistical learning of native phonetic categories which then form the basis for phonotactic segmentation of words from continuous speech. This theory is supported by multiple behavioral findings. For example, although infants are born with equal sensitivity towards all phonetic contrasts in the world's languages (Eimas et al., 1971; Trehub, 1976), studies indicate that infants show heightened sensitivity to native phonetic contrasts towards the end of their first year (e.g., Kuhl et al., 2006), whereas sensitivity to non-significant non-native contrasts decreases (Werker & Tees, 1984). Moreover, studies show that success in native phonetic category perception predicts later proficiency in the language (e.g., Tsao et al., 2004; Kuhl et al., 2005; see also Kuhl et al., 2008, and references therein for a more comprehensive review on the topic), and that the categorical perception is already reflected in pre-attentive auditory processing (Winkler et al., 1999). In order to understand how the statistical learning of native phonetic categories actually takes place, several computational models have been used to study the acquisition of phonetic categories from speech.

The standard approach in the existing studies has been to apply clustering techniques to formant frequencies or other spectral representations derived from manually segmented vowel sounds and then compare the clustering outcomes to the known category identities of the vowel samples. In some of the studies, the correct number of phonetic categories has been manually provided to the clustering algorithm (de Boer & Kuhl, 2003), whereas more advanced algorithms are able to estimate the correct number of categories automatically (Vallabha et al., 2007; Kouki et al. 2010; Lake et al., 2009; Markey, 1994). The algorithms themselves include different variants of expectation maximization (EM; Dempster et al., 1977) based estimation of Gaussian mixture models (GMMs; see Duda et al., 2001) such as in the OME algorithm of Vallabha et al. (2007) and Lake et al. (2009), feature density histogram estimation in a multidimensional space (TOME; Vallabha et al. 2007), and the use of self-organizing maps, or SOMs (Kohonen, 1990) as in the work of Kouki et al. (2010).

Another possibility is to enhance phone perception by utilizing additional information from other modalities or by using constraints from another level of linguistic representation such as the lexicon. For example, Coen (2006) has shown that the introduction of visual information from the lip movements together with vowel formant data leads to a clustering result that discriminates vowel categories more accurately than using audio or visual information alone. On the other hand, Feldman et al. (2009a) have shown that a Bayesian model of categorical inference is unable to discover proper vowel categories from formant data alone, but when the model is accommodated with constraints from a simultaneously learned lexicon, the category learning succeeds. Finally, a largely unexplored link is the connection between articulatory development and categorical perception of speech sounds. Since the introduction of the motor theory of speech perception (Liberman et al., 1967; Liberman & Mattingly, 1985), a number of studies and models have proposed that the articulatory gestures could play at least some kind of role also in the perception of speech (e.g., Fowler, 1989; Moore, 2007; Skipper et al., 2006; Goldstein et al., 2006; see also Nearey, 1997, and references therein for a discussion), although the exact role of gestural representations varies across theories. However, the basic principle underlying these theories is that the highly variable auditory representation is simplified if the speech can be represented as a series of partially overlapping articulatory gestures with phoneme-specific target positions for the articulators. The original motor theory suggests that speech perception and production are performed using an innate “speech module” and this module is responsible for perceiving speech as a series of invariant intended phonetic gestures (Liberman & Mattingly, 1985). This type of special module receives little support from behavioral or brain imaging studies (e.g., Pulvermüller, 2010, p. 269). Nonetheless, research shows that the motor cortex responsible for the control of articulatory gestures is activated during speech perception (Watkins et al., 2003; Sato et al., 2010) and also has a modulatory effect on speech perception (D’Ausilio et al.,

2009). Also, it has been proposed that the motor representation of speech may be needed to resolve the neural competition between acoustically similar tokens, since spatially segregated motor areas (e.g., areas responsible for motor control of lips and tongue) can inhibit each other, whereas mutual inhibition at the level of the auditory cortex may be difficult due to notable overlap in the acoustic receptive fields (Pulvermuller, 2010). Although it is still early to say whether the motor system is an integral part of the development of successful speech perception, the two processes are evidently connected.

Some computational models have already attempted to utilize the connection between perception and production. For example, the HABLAR model of articulatory and phonological development is based on the hypothesis that "*phonological development emerges from the interaction of auditory perception and hierarchical motor control*" (Markey, 1994). The system first learns to classify incoming (but synthesized) automatically segmented speech sounds into a finite number of acoustic categories and then also learns to imitate these sounds with an articulatory synthesizer equipped with reinforcement learning techniques. However, the model does not change its auditory perception of speech sounds based on articulatory experience, but the speech sound categories are kept fixed after perceptual learning preceding the articulatory development. Still, the preliminary tests reported by Markey indicated notable promise in the approach (Markey, 1994). Unfortunately, the work on the model seems not to have been continued towards more comprehensive experiments, making drawing strong conclusions from the work difficult.

Recently, Howard & Messum (2011) presented an integrative model of phonological development. Their computational learning agent, Elija, learns to produce native speech sounds and words through interaction with a human caregiver. Initially, Elija explores the space of different articulations and receives internal rewards for acoustically salient or motorically diverse productions. After the initial learning, Elija's vocalizations start to draw the caregiver's attention. The caregiver interprets Elija's output in terms of the native phonetic system and provides feedback for successful articulations via the imitative reformulation of Elija's speech output. This then reinforces Elija's native-like articulatory gestures and causes the speech production system to converge towards the set of native speech sound categories. Moreover, mediated by the shared communicative context, Elija is able to associate its own speech to that of the caregiver, allowing Elija to learn the mapping between the acoustics of adult speech and its own articulatory gestures. Later, when the communicative situation is supplemented with referential objects that are being repeatedly named by the caregiver, Elija gradually learns the correspondence between object identifiers ("visual tags") and their respective auditory and motor representations. Although the model concentrates more on the acquisition of articulatory gestures for speech production than modeling the acquisition of categorical perception of phone-like units, it is still an excellent example of how an integrative

framework including the modeling of the learner-caregiver interaction, auditory and motor learning, and the modeling of a shared communicative context can produce human-like learning results.

Other work in the area includes, e.g., the work of Ananthakrishnan & Salvi (2011), Ananthakrishnan (2011), and Rasilo et al. (in preparation; see Räsänen et al. 2012 for a short overview) who have studied the mapping between perceptual categories and speech production. However, similarly to HABLAR and Elija, none of the approaches have utilized caregiver feedback or articulatory learning in the development of the distributional properties of the perceptual categories themselves, and therefore the role of articulatory development in categorical perception of speech is still unstudied in the computational modeling framework.

3.1.1 Conclusions on computational models of phonetic learning

The main finding from the clustering studies is that the distributional structure of signal features corresponding to vowel sounds can be captured with the proposed clustering techniques. In addition, the obtained vowel categories exhibit properties similar to human perceptual biases as described in NLM-e (Kuhl, 2000; Maye et al., 2002), such as the distinctiveness and similarity ratings of tokens belonging to these categories (see Toscano & McMurray, 2010; Feldman et al., 2009b). However, the categorization accuracy is far from perfect, but the average accuracy of classifying any single vowel representation into one of the possible vowel categories is around 70–80 % even for cases where not all vowel categories are considered. This means that assigning a unique and correct phonetic label to each acoustic percept is not possible (cf., Vallabha et al., 2007; Kouki et al., 2010). This is especially pronounced in the generalization across multiple talkers, where distributional clusters of one talker are not compatible to those of another, or where distributions learned from multiple talkers become very ambiguous at the category boundaries. On the other hand, if categorization performance notably below 100% is allowed at the phone perception level (note that human performance is not perfect either for isolated vowel recognition), the question is whether the words can be learned as sequences of such partial and distorted input. Is there a way to learn a robust lexicon directly upon the layer of acquired discrete units that have been learned in an unsupervised manner?

An important limitation in many of the clustering studies is that the speech data do not represent randomly drawn segments from continuous speech, but carefully chosen maximally stable portions of context-limited vowel-segments. The only exception is the work of Kouki et al. (2010), but they obtained only limited success in the clustering of features into vowel categories. In order to increase the ecological plausibility of the other approaches, a mechanism for segmentation of these vowel segments from continuous speech would be needed, or otherwise the methods should be evaluated directly on continuous speech. As already discussed in section 2 and also studied in P-I, the segmentation problem is far from trivial. Related to this, the model of Howard & Messum uses a system for representing

incoming speech as a sequence of categorical units, but their existing work assumes that the perceptual category learning is mainly based on isolated caregiver reformulations of canonical babbling (Howard & Messum, 2011). Although not fully implausible, this behavioral hypothesis is to be confirmed with experimental studies.

Another issue in phone category learning is the generalization across talkers. For example, in the work of Vallabha et al. (2007), the categorization of vowel segments is much more accurate for the experiments in which the same talker is used in the training and in the evaluation of the categorization. When additional talkers are used for evaluation, the performance drops significantly. Although this difficulty is expected due to well-known acoustic variability across different talkers, how human infants may solve this challenge remains unknown. From the computational point of view, statistical learning leads to much more ambiguous category boundaries if the learner receives data from several talkers instead of a single caregiver. On the other hand, modeling each talker separately or using data from only one talker, such as the primary caregiver, leads to much sharper category boundaries, but then these categories are not compatible with those produced by other talkers of the language. The generalization problem is also inherent to the lexicon-first models of LA (see next section), leading to incompatibilities between lexical items learned from the speech of different talkers (P-V).

It should be also noted that, except for the work of Markey (1994), Kouki et al. (2010), and Coen (2006), the inventory of phonetic categories to be learned in the purely acoustic experiments is much smaller than that of the normal number of vowel categories in the world's languages. Currently, how the other proposed methods would scale up to a full vowel repertoire of a language and how they can deal with the temporal ambiguity of vowel boundaries and the coarticulatory effects between subsequent vowels still remains unknown.

3.2 Models of lexical learning

According to the PRIMIR-theory of language acquisition, the development of subword units such as phones is guided by initial learning of a type proto-lexicon – an initial collection of word forms that are coded as holistic acoustic patterns and that may be already grounded to some referential information such as objects or events that the words refer to. Only later, once a sufficient amount of words has been learned, the learner starts to realize that words consist of smaller constituents that overlap between different words, leading to the emergence of phonetic or even phonemic representation of the language (Werker & Curtin, 2005). This type of lexicon-first approach has been also undertaken in many computational models of LA that are briefly reviewed in this section. However, some attention has first to be paid to the nature of lexical representations and the semantic dimension of words before moving to the actual models.

3.2.1 On lexical representations of speech

When literate adults discuss the concept of *words,* they talk about well-defined entities that have both written and spoken form with a finite number of discrete elements (phonemes) in a specific order. The majority of the words either have significant associations that instantly stimulate multimodal perceptions related to the concept denoted by the word, or they play a significant role in the construction of grammatically correct sentences by disambiguating causal and temporal relationships of the actors and events involved in the verbal description at hand. Either way, the words show themselves as meaningful symbols and the symbol is perceived as "incorrect" when it is misspelled or mispronounced, requiring additional cognitive resources to recover from the aberration.

From the viewpoint of a young infant, the situation is very different. This is especially true if the nativist views are completely abandoned and the infant is considered as a tabula rasa cognitive agent with efficient innate learning capabilities and a bias for social behavior. An infant does not know what a lexical item or symbol is. Moreover, it does not even know what speech is about. Much of the learning effort during the first year of the infant's life is about discovering that the objects in the vicinity can be perceived through senses and that they can be also manipulated by motor activity. Through the development of the action-perception loop and maturation of the brain, the infant acquires understanding that the world is a 3-D realm with distinct objects with varying properties, and with actors (living objects) that can have an impact on the state of the other objects in the realm or on the (needs of the) infant itself. Although the development of auditory perception is affected by the exposure to speech associated with social and emotional interaction with the caregiver (cf. NLM-e; Kuhl et al., 2008), the author's claim is that *the first real contact with the language faculty occurs when the infant first realizes that sensory patterns originating from other people's mouths have correspondence to the state of the surrounding world*. This is probably already preceded by the realization that the objects and events in the environment are sometimes associated with distinct non-speech auditory patterns (note that the sensory patterns need not be auditory, but signed language will also do the trick, e.g. Emmorey, 2006; cf. also the "*goes with*" vs. "*stands for*" distinction of words in Golinkoff et al., 1994). The core of the language is in the ability to activate representations in other people's minds about things that are not necessarily available in the present sensory domain, or at least not in the current focus of attention. The learning of these links is necessarily bootstrapped by associating the sensory patterns to internal active representations of the concepts describing the world. For very young infants, these associative links are necessarily tied to the surface form (i.e., acoustic or visual realization) of the patterns, since it is the most directly observable and statistically significant structure that has correspondence to the external world. This type of learning can be accomplished in the absence of any kind of linguistically motivated knowledge in the learner (P-IV; P-V; see also the Ecological Theory of Language Acquisition (ETLA) in Lacerda et al., 2004).

What this all means from the viewpoint of early lexical learning is that the learner does not have an ecological pressure to "find" words from speech nor code these words in any specific format (such as precisely defined sequences of discrete elements like phones or phonemes). Instead, the functional advantage of spoken language emerges from sufficiently detailed but sufficiently general representations of the spectrotemporal acoustic patterns that systematically indicate the co-occurrence of objects and events in the environment, i.e., the useful units of a language are defined by their *semantic content*. These patterns can match individual words, but they can also be part-words, compounds, frequently co-occurring words ("*doyou*"; "*Isee"*), or even entire phrases if they are systematically used in specific situations. As long as there is equally good predictive power (or functional consequences) in "wrong" lexical representations of the language that do not match the adult vocabulary, there is really no need for the learner to refine these representations. As the complexity of the interactions with the environment increases and as the number of proto-lexical representations accumulates, the early representations of spoken language become refined in order to answer to the increasing communicative challenges and to reduce the internal contradictions in the previously acquired lexico-conceptual system (cf., principle of conventionality in Kuhl et al., 2008). For example, the increasing semantic awareness imposes new distinctions to the linguistic representations and gives rise to the concept of lexical synonymy. The increasingly structured parsing of speech and increased size of the lexicon also possibly gives rise to the subword/morphological representation of spoken language as it provides a more efficient means of coding (cf. PRIMIR, Werker & Curtin, 2005).

3.2.2 On the grounding of auditory patterns

In computational models of lexical learning, the word semantics are typically assumed to emerge directly from the established link between an acoustic word form and an internal representation of a word referent, such as a visual or haptic representation of an object or action. Based on this idea, the computational models of lexical learning from continuous speech can be divided into two main categories, based on the principle how words (learned acoustic patterns) are grounded to their referents. These categories will be referred to as models with *indirect* and *direct grounding* of words (Fig. 2; see also the division of methods in Table 1).

Indirect grounding refers to a learning process in which the learning agent first learns speech patterns such as words from continuous speech independently of other modalities. The criterion for the initial word segmentation can be arbitrary, but, assuming the absence of a priori phonetic or linguistic categories, it has to be some type of statistical measure that reacts to the specific organization or recurrence in the acoustic features computed from the auditory signal. Once a pattern is learned and can be recognized from future input, its occurrence can be then studied in the

context of other modalities and internal states2 of the agent in order to find statistical correlations between the pattern and the contextual variables. Once such a correlation is found, it is said that the pattern (or word) is grounded to the contextual variable, providing meaning to the pattern.

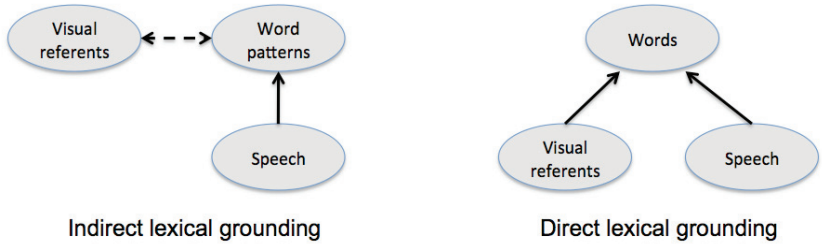


Figure 3: A schematic view illustrating indirect (left) and direct (right) lexical grounding. In indirect grounding, the word representations are first discovered purely on the basis of speech signal statistics, and then later associated to their referents. In direct grounding, the visual context acts as a labeling for the speech input, allowing the use of weakly supervised machine learning techniques in the word discovery.

On the other hand, direct grounding means that the learning agent perceives speech simultaneously with internal active representations of contextual variables. Due to the immediate co-occurrence, the representations of the objects and events in the active internal state become associated with the heard auditory patterns, allowing instantaneous (but originally vague) meaning to emerge for the spoken utterances. In the simplest case, the internal state may simply reflect the visual objects in the immediate surroundings that the learner is attending to, leading to direct cross-modal associations between acoustic patterns and visual objects. For a more complex cognitive system, the internal state may reflect a combination of the task-modulated short-term memory contents (which may consist partly of immediate sensory consequences and partly of items recalled from the long-term memory) and some internal variables of the system, such as the emotional state of the agent. The major difference from indirect grounding is that now the contextual variables such as visual objects can directly affect the patterning of the auditory stream. This provides the learning system with additional statistical constraints that can help in the learning process.

For both indirect and direct grounding of word forms, the main problem is that one exposure to the auditory pattern is not sufficient to obtain meaning of the word since there are typically multiple potential word referents available (Quine, 1960). This is where so-called cross-situational learning mechanism comes into play (Pinker, 1989; Gleitman, 1990), and due to the multiple exposures to situations with several possible referents simultaneously with speech containing the word of interest, the ambiguity in word-referent mappings is gradually resolved (e.g., Smith & Yu, 2008; Smith et al., 2011).

2 Note that while some words such as nouns typically refer to perceivable physical objects, some others such as “sad” or “hungry” are ultimately grounded to the internal needs or emotions of the learner.

The models that discover patterns from continuous speech in the absence of referential information will be discussed first in the next section, and then attention will be turned to the models of direct lexical grounding.

3.2.3 Models of indirect lexical grounding

So far, no computational model of indirect lexical grounding exists that would combine unsupervised acquisition of phonetic categories with lexical learning. Instead, a number of models have been proposed that attempt to learn lexical representations directly from continuous speech without relying on an intermediate phonetic layer.

Table 2: Algorithms for word discovery in indirect lexical grounding (unsupervised learning).

Algorithm	Properties
PERUSE (Oates, 2001, 2001)	Iterative multivariate pattern matching. Prototypical word representations (words are represented as time-variant feature-value distributions). All signals have to be available during computation (batch mode).
P&G (Park & Glass, 2005, 2006)	Multivariate DTW-based discovery of recurring patterns by pair-wise comparison of current and previously heard utterances. Exemplar-based memory (word realizations are stored as multivariate signals that are clustered into word classes using graph clustering). All signals are stored in the memory for processing (batch mode).
Incremental P&G (McInnes & Goldwater, 2011)	Incremental version of the P&G algorithm with higher ecological plausibility. Allows discovery of multiple matching fragments per each pair-wise comparison. Exemplar-based memory similarly to the original P&G. Old inputs are forgotten by the system.
Transition probability-based learning (Räsänen, 2010; publication P-V)	Incremental algorithm based on analyzing transition probabilities (TPs) between subsequent atomic acoustic units (vector-quantized speech features). Words are represented as a collection of transition probabilities characteristic to the word. All signals are immediately forgotten after their TPs have been analyzed.

The computational models of LA from continuous real speech based on indirect lexical grounding include the PERUSE algorithm by Oates (2001, 2002), the P&G algorithm of Park & Glass (2005, 2006) and its incremental variant by McInnes & Goldwater (2011), and the transitional probability-based algorithm of Räsänen (P-V). These algorithms and their main properties are listed in Table 2.

The basic working principle in all these models is to find long recurring patterns from the input that in practice correspond to words or combinations of often co-occurring short words. The algorithms themselves are fully unaware of the concept of word, and since the speech is not accompanied with any other categorical information that would enable grounding of the word forms, the detected structures do not carry any meaning. In each approach, there are several steps in the process of modeling discovery of word segments from raw speech signals. First,

incoming speech has to be transformed into frames of features that describe the temporally local spectral content of the signal in a compact manner. In order to learn recurring structures, or patterns, from speech, the continuous domain feature representations are then analyzed with a statistical model or further transformed into a series of discrete events using vector quantization of the feature vectors before the statistical analysis.

In PERUSE, the learning is based on the assumption that structurally significant patterns occur as sequences of multivariate observations of features, where each temporal spot in the sequence has a unique mean and variance describing the local acoustic properties. The likelihood of a sequence of data for a given pattern model can be directly computed by temporally aligning the model with a pattern using dynamic programming and then summing the log-likelihoods of individual observations across the entire sequence. In order to discover the patterns, PERUSE performs a global and exhaustive search over all available speech data in order to find a set of signal models that each have multiple representative occurrences in the data set. The statistical model itself is updated iteratively by adding new realizations of the pattern class and updating model distributions according to these new observations. Oates has demonstrated the performance of the PERUSE algorithm in a word learning task from English, German and Mandarin speech, where it successfully detected more than 65% of frequent words used by a single talker. Oates has also represented a framework that allows grounding of the detected word forms to contextual sensory data collected by a robot (Oates, 2001). The main drawback of the PERUSE algorithm is that it requires all speech data to be in the memory of the system already at the beginning of the learning. The underlying assumption is that the longest words that have most occurrences in the data have most significance and are therefore learned first. Additionally, the algorithm is computationally complex, as it has to search and evaluate the data set iteratively numerous times in order to converge to the final set of words. In addition, each word has to occur several times in the data before a representation can emerge for it. This makes the approach implausible for a biological system that needs to deal with the continuous flow of sensory information here and now without access to globally determined statistical significances between different choices of signal patterning. Oates has also acknowledged this limitation, noting that iterative batch processing is an unreasonable requirement for a computational agent that should support continuous long-term LA (Oates, 2001).

The word discovery algorithm by Park and Glass (2005, 2006), hence the P&G algorithm, is based on a modified dynamic time warping (DTW) of feature representations of auditory patterns. In the DTW, speech signals are represented as multivariate spectral time series and the aim of the DTW algorithm is to discover the cheapest path across a distance matrix whose elements describe the distances between the spectral frames of two signals. As an outcome, the obtained shortest path describes the temporal correspondence between spectrotemporal patterns in both sequences, and

DTW is therefore especially suitable for temporal alignment of signals that are known to contain identical utterances but spoken at a different tempo. Since individual words in speech represent only a small part of long utterances, the alignment is performed in limited temporal slices in the P&G algorithm. The alignment process is repeated for all pairs of utterances in the data-set, leading to a collection of pairs of aligned specrotemporal signals. Then the signals are clustered using an agglomerative graph clustering method (Newman, 2004) in order to find categories of similar patterns that in case of speech correspond to often occurring words. Note that the original P&G algorithm was not intended to model infant language acquisition, but was designed as an engineering tool for unsupervised pattern discovery from speech signals. Lately, McInnes and Goldwater (2011) have modified the P&G algorithm in order to achieve a higher ecological plausibility for LA simulations. Instead of performing word discovery as a batch process as in the original P&G algorithm, their system works incrementally by comparing the current input only to the previously discovered word fragments and to a finite number of previously perceived utterances. The word fragment extraction of the algorithm was also modified in order to allow the discovery of multiple separate word fragments from each pair-wise alignment between signals. They showed that their algorithm is able to discover recurring words from audio recordings of mother-infant interaction, especially in cases where the speech contains an infant-directed speaking style with multiple repetitions of salient words occurring close in time (McInnes & Goldwater, 2011).

Finally, publication P-V presents a computational model for unsupervised word discovery from speech that is based on transitional probabilities (TPs) between atomic acoustic events. The model is inspired by the finding that eight-month-old infants can already segment recurring words from speech by analyzing TPs of subsequent syllables (Saffran et al., 1996a; 1996b) and may treat the detected segments as lexical items when presented in a proper linguistic context (Saffran, 2001; see also the introduction to statistical learning in section 1). However, the model of Räsänen does not assume that the learner can recognize phonetic or linguistic units such as phones or syllables from continuous speech, but simply represents the acoustic speech signal as a sequence of discrete elements obtained by unsupervised vector quantization of spectral vectors. Recurring speech patterns are modeled by analyzing the TPs between the discrete acoustic events, i.e. each unique pattern model is characterized by a specific set of TPs between the discrete elements in the signal. However, the modeling is not performed only for transitions between the two subsequent elements, but in parallel for a number of different temporal distances in order to capture long-range statistical dependencies and to enhance robustness of the model against noise variability in the signals.

When tested on an English corpus containing child-directed speech (CAREGIVER; Altosaar et al., 2010), the results showed that the TP-based algorithm successfully learned a number of ungrounded word models that

were selective towards specific words in the material when compared against the word-level annotation (many of the models responded only to one word above 80% of the time; see also Räsänen, 2010, for results in the Finnish language). The word segmentation accuracy was also notably above chance level. It was also observed that the learning performance was much higher when the training and recognition was performed with data from the same talker. When data from multiple talkers were used, the generalization across talkers was relatively poor in terms of word recognition accuracy, although the segmentation accuracy generalized better for models learned from one talker to speech from another talker. This replicated the common finding from ASR research that acoustic models learned for one speaker are not easily generalized to other speakers, especially to speakers of different gender. The results also showed that the typical errors in word segmentation and model selectivity were either related to a situation where a model had considered that two frequently co-occurring short words are one word (such as "*doyousee*", "*doyoulikethe*"), or when the words were acoustically similar ("*small*" and "*ball*", or "*cow*" and "*cat*").

In later work (P-VI), it was shown that when the learned word models are grounded to visual referents through cross-situational learning, the acquired word-referent mappings also allow linking of synonymous but acoustically distinct word models together. This is one possible explanation of how initially acquired speaker-dependent word representations can be later generalized across a larger talker population based on the environmental contexts in which these words are used. Also, it was shown in P-VII that the TP-based model of P-V is able to explain behavioral findings in a rule-learning task from an artificial language, originally thought to require more abstract rule-based computational mechanisms (P-VII; Laakso & Calvo, 2011; Endress & Bonatti, 2007; Peña et al., 2002).

All in all, the above models demonstrate that learning of words (or "proto-words") from continuous speech is possible in the absence of contextual support or external feedback and without any a priori linguistic or phonetic knowledge. However, these word representations are not always perfectly aligned with the words defined by a proficient language user, but more likely represent statistically significant continuous spectrotemporal structures that systematically recur in the speech data. Notably, the existing studies use three totally different methodological approaches. The P&G approach and its incremental version (McInnes & Goldwater, 2011) basically perform exemplar-based learning by extracting recurring fragments of speech and then comparing these fragments to novel utterances. The approach in P-V does not store word exemplars per se, but represents each word as a construct that defines probabilities at which specific acoustic events follow each other in the word. Finally, the PERUSE algorithm (Oates, 2001, 2002) treats each word as a probabilistic construct, but solves the problem in purely continuous time and feature domains. Despite their differences, all the algorithms show a similar pattern of results, suggesting that the

recurring word structure in speech can be captured using a variety of pattern representations.

3.2.4 Models of direct lexical grounding

Table 3 shows a number of algorithms that have been used to study word learning under direct lexical grounding conditions (see Fig. 2). In a typical direct lexical grounding simulation, the acoustic speech utterances are presented together with categorical label information denoting what type of objects or entities are related to the linguistic contents of the utterance. However, the labels are not ordered or aligned in time as in typical supervised learning (cf. standard ASR), but the alignment only takes place at the utterance level. In some cases, real visual input is used in parallel to the audio and processed into clusters of visual features (e.g., Hörnstein et al., 2009). The task of the algorithm is then to discover what parts of speech correspond to the categorical labels in the visual domain. This type of experimental setup aims to simulate a situation where the language learner listens to the speech of a caregiver and simultaneously attends to the objects and events related to the interaction situation. It is suggested in Räsänen (2012) that this is not an unreasonable simplification since infants at the age of early word learning (around 6–12 months) are already able to perceive the world as a set of discrete entities with separate identities and follow the attention of a caregiver during the infant-caregiver interaction.

Table 3: Algorithms for word discovery in direct lexical grounding (weakly supervised learning).

Algorithm	Properties
Statistical word discovery (SWD) (ten Bosch & Cranen, 2007)	Segments speech into phone-like units and clusters them into 25 categories. Each context label is linked to audio by collecting a list of segment sequences that occur many times during the presence of the label. Most frequent sequences are treated as best exemplars for each label.
Non-negative matrix factorization (NMF). (ten Bosch et al., 2009a; Van hamme, 2008)	Represents speech signals as histograms of acoustic co-occurrences (HACs) of atomic acoustic units. These histograms are then combined with visual information and decomposed with matrix factorization into basis vectors (words) and weights that denote the presence of these words in each signal.
Weakly-supervised transitional probability analysis (publications P-II, P-IV)	Analyzes transition probabilities of atomic acoustic events in the context of visual information. First version (P-II) operates on automatically derived phone-like segments, but the generalized algorithm (P-IV) was studied using a fixed-frame signal representation.
DP-ngrams (Aimetti, 2009)	Dynamic-programming-based pattern matching. Words stored as exemplars that are linked to the contextual variables such as visual percepts.
Mutual information based audiovisual clustering (Hörnstein et al., 2009).	Searches for temporally local and repeated target words using prosodic features and analyzes attended visual objects into features. Agglomerative clustering is performed in both modalities and audiovisual associations are defined as the pairs of cluster nodes having the highest mutual information.

As one of the first attempts of weakly supervised learning of words from continuous speech, the statistical word discovery (SWD) algorithm described by ten Bosch and Cranen (2007) utilizes segmental representation of speech by blindly segmenting the input signals into phone-like units based on spectral changes in the signal. These segments are then aligned with DTW and clustered with the k-means algorithm (MacQueen, 1967) so that each segment becomes represented by a categorical integer index. In other words, the utterances are converted into discrete sequences, one element spanning approximately one phone-sized unit. During the word learning process, the segmental representation of each utterance is represented in association with a bag of abstract tags that describes which words are present in the utterance but do not reveal the temporal locations or ordering of the words. Each utterance is then compared to all previously perceived utterances and the best matching subsequence of each pair is extracted. If the current utterance shares the same abstract tag with the one in the memory that it is being compared to, the best matching subsequence is appended to a B_{match} list. Otherwise it is added to $B_{no\text{-}match}$. When the learning process is repeated across several utterances, the match and no-match lists grow in number. The lists are sorted so that the most frequently occurring sequences are placed at the top of the lists, revealing the most typical sequential representations of each word. The sequences in the $B_{no\text{-}match}$ are considered as negative examples of a word, and therefore the equivalent sequences in the B_{match} list are eliminated in order to facilitate the contrast in the cross-situational learning (ten Bosch & Cranen, 2007).

Ten Bosch and Cranen evaluated the performance of the algorithm using the Aurora 2.0 database (Hirsch & Pearce, 2000) that contains continuously spoken English digit sequences with 1–7 digits per utterance and speech from a large number of talkers. The results showed that their algorithm achieved approximately a 90% word recognition rate after perceiving 1000 tokens per digit when the hypothesized word tags were compared to the ground truth. The authors also noted that the number of false alarms (words being hypothesized to points in time where there are no corresponding words) was relatively high (above 10%). They hypothesized that it may indicate that the learned word representations were somewhat shorter than the true lengths of the words, since the correct recognitions did not cover the entire timeline of the utterances.

In general, the SWD algorithm is interesting because it is one of the rare attempts to segment speech into phone-like units before further processing (cf. Kuhl's basic cuts; Kuhl, 1986, 2004). While the idea of making an exhaustive comparison of the current speech token against all previously heard utterances with a shared context seems drastic, it is not totally unreasonable from the perspective of exemplar-based theories of human memory. Still, the work of ten Bosch & Cranen focuses mainly on the of question whether statistical regularities in the automatically learned phone-like segments can form a basis for a lexicon. Analogues to human-like processing are not given by the authors.

Blind segmentation of speech into phone-like units was also used in the work of Räsänen, Laine and Altosaar (P-II). Similarly to ten Bosch and Cranen (2007), the segments were vector quantized with an early version of the method described in P-III in order to obtain discrete sequences of phone-like units representing the speech signals. Also, the utterances were paired with abstract tags denoting the "visual objects" perceived simultaneously with the utterance. However, the learning procedure was now based on TPs between subsequent phone-like units (cf. Saffran et al., 1996a) in the context of each tag, i.e. the probability of a transition from discrete phone-like segment S_t to segment S_{t+1} was measured separately for the presence of each contextual tag. During recognition, the probability of each tag was computed by following the transitions through the sequential representation of the utterance and retrieving the corresponding tag-specific TPs from the memory. When evaluated with the CAREGIVER Y1 FIN corpus (Altosaar et al., 2010) with a total of ten unique keywords (the visual tags), one keyword embedded in each utterance in addition to the surrounding carrier sentences, the algorithm obtained a keyword recognition rate of 74.5% for speech from one talker.

The experiments of P-II showed that there is some feasibility in the transitional probability approach in word recognition, but the overall word recognition rate was relatively low considering the simplicity of the task. Räsänen & Driesen (2009) found that the low performance was mainly due to segmental representation of speech that did not capture spectrotemporal details with sufficient accuracy in order to obtain efficient models for words. Insertions and deletions of segments were also a concern. Finally, the analysis of only subsequent segments (bigrams) of P-II did not yield sufficiently strong statistical models for speech. This led to the discarding of the segmentation-based approach, and a further developed mathematical framework for TP analysis is presented in P-IV (see also Räsänen et al., 2009b). The approach makes use of the normal fixed-frame windowing with 10-ms frame shifts and vector quantization of speech signals. In addition, similarly as in P-V, the TP analysis is performed at a variety of different lags, increasing notably the robustness of the learned models. When evaluated with the CAREGIVER Y2 UK corpus with 50 unique keywords, 1–4 keywords occurring in each utterance, a word recognition rate of above 92% was obtained for data from four different talkers (two male, two female; P-IV).

Also based on vector quantized short-term acoustic features, ten Bosch et al. (2009a) and van Hamme (2008) have presented a non-negative matrix factorization (NMF) based approach to word learning. In NMF, spoken utterances are represented as histograms of acoustic co-occurrences (HACs) of atomic acoustic units (vector quantization indices) similarly a in P-IV and P-V. These histograms are then combined with related visual information into a large matrix $\mathbf{V}$ and decomposed with matrix factorization into basis vectors $\mathbf{W}$ and weights $\mathbf{H}$ so that $\mathbf{V} \approx \mathbf{WH}$. When factorized, $\mathbf{V} \rightarrow \mathbf{WH}$, matrix $\mathbf{W}$ represents the typical audiovisual patterns that recur in the data,

whereas **H** describes the activation level of these patterns in each utterance. During word recognition, the activation matrix **H** is computed from the HAC formed from the utterance under consideration (without the visual part). In order to obtain the activation values for the visual tags, the obtained **H** matrix is then multiplied by the submatrix $\mathbf{W}_g$ of the **W** that contains only the vectors related to the visual grounding information. When evaluated with the CAREGIVER corpus (Altosaar et al., 2010), the results showed that the NMF-based system can learn the ten unique keywords with high accuracy when grounded directly with the related visual tags during learning. For further ecological plausibility, the NMF-based LA model was modified in Driesen et al. (2009) so that the processing is no longer performed in a batch mode, but allows incremental learning. This makes the approach intriguing for LA studies since the NMF, and especially the HAC representation of the sensory input, share many properties with human-like information processing, including the representation of information in a distributed form and the ability to include multiple sources of information at different granularities into the same computational framework.

Aimetti (2009) has proposed a dynamic-programming-based system for word learning from continuous speech. However, unlike the P&G algorithm (Park & Glass, 2005; 2006), the system also utilizes direct grounding of the lexical items to the co-occurring visual objects (visual objects are simulated with abstract and discrete semantic tags). The learning proceeds by first comparing the visual tags of the current utterance to the tags of previously perceived utterances in the short-term memory (STM) of the system. For a tag never perceived before, the entire utterance is stored as an acoustic entry for the new tag, and the system proceeds to the next utterance. In case of matching tags, the utterances are aligned with a method called DP-ngrams, which is a modification of the standard DTW that allows efficient extraction of best matching temporally contiguous aligned sequences. The part of the novel utterance containing the best matching alignment with existing memory entries is then extracted and appended to the list of exemplars representing the corresponding tag. When the process is repeated over the entire training data, each tag becomes associated with a list of exemplar occurrences of the corresponding word. These exemplars can then be matched with a novel utterance in order to determine which tag is the most likely in the speech signal. Also, a clustering process can be applied to the list of exemplars in order to obtain a prototypical representation of each word (Aimetti, 2009).

Evaluation of Aimetti's algorithm was carried out with the Y1 UK version of the CAREGIVER corpus (Altosaar et al., 2010). When evaluated in terms of word recognition accuracy (correspondence between true and hypothesized visual tag of a novel utterance), convergence to a word recognition rate of approximately 90% was observed with exemplar-based recognition after perceiving approximately 140 utterances. For prototype-based recognition, the accuracy was notably worse (around 70%), suggesting that a word even from a single talker is not very well represented by an

"acoustic mean" of its realizations. For experiments with four talkers instead of one, a recognition rate of slightly below 50% was obtained after observing 200 utterances, again giving a clear indication of the notable acoustic mismatch between different talkers even on material with very limited vocabulary.

Finally, Hörnstein et al. (2009) have proposed a word learning model that can be actually considered as a hybrid of indirect and direct grounding. The first stage of their model is based on discovery of repeated word forms based on their approximate acoustic similarity within a finite (10-20 s) time window. The discovered repetitions of a word are then moved to a long-term memory where hierarchical agglomerative clustering is performed the word tokens based on their acoustic similarity. Simultaneously, the objects located in the center of the attended visual scene (captured by a camera) are analyzed into shape features and these shape representations are also clustered using the same agglomerative clustering algorithm but using a separate cluster space. The general problem in agglomerative clustering is to find the correct cut-off level in the hierarchical cluster tree at which the clusters are general enough to capture all the variability in the individual realizations of words and objects of the same category, but specific enough so that tokens from different categories do not become members of the same cluster. Hörnstein et al. (2009) solve this problem by finding the nodes of the cluster trees at which the mutual information between the auditory word clusters and the visual object clusters is maximized. In other words, the acoustic boundaries of the final word categories become determined by their statistical dependency with the most salient objects present in the concurrent visual context.

Hörnstein et al. demonstrate that their algorithm is capable to learn at least a small number of keywords related to salient visual objects such as dolls and balls that are discussed in natural interaction settings between the learner and a caregiver. Moreover, they show that the inclusion of cues typical to infant directed speech (IDS), namely F_0 accent and word-final keyword position information, improve the word learning performance of the system. This provides some of the first simulation results indicating that paying attention to the characteristics of IDS can aid in the word learning process, as already suggested by many studies with real infants (see Hörnstein et al., 2009, Werker & Curtin, 2005, and Kuhl et at al., 2008, for numerous references; see also de Boer & Kuhl, 2003 for the use of IDS in phonetic learning).

3.2.5 Other experiments with direct lexical grounding

In addition to the methods described above, a number of additional experiments of LA have been reported using the described algorithms or their further modifications. In ten Bosch et al. (2009b), the TP-based learning algorithm presented in P-IV was studied in different caregiver-learning agent interactions. Different interaction strategies were simulated by varying the reliability of the visual labels associated with the spoken utterances, revealing the somewhat expected result that the more there is

ambiguity in the referential context of spoken words, the slower the learning rate will be.

In ten Bosch et al. (2009c), the effect of the number of different caregivers was studied using the TP-based algorithm of P-IV, the DP-ngram algorithm of Aimetti (2009), and the NMF-based system presented in ten Bosch et al. (2009a) and Van hamme (2008). In the learning phase, the learner perceived speech from either one or four different caregivers. The word recognition accuracy was then probed for the four main caregivers and for six additional talkers available in the CAREGIVER Y2 corpus. The results from all three learning algorithms revealed that the generalization from one caregiver to the novel talkers was much worse than the word recognition performance evaluated for novel speech from the caregiver used in the training. When several different caregivers were used in the training stage (two male, two female), the generalization to six additional talkers was slightly better but still far from the matched talker condition. The NMF was also noted to have obtained the best generalization to unseen talkers from the three algorithms (ten Bosch et al., 2009c).

Lately, Versteegh et al. (2010) have studied how the word learning performance is affected when the learning agent can actively decide whether the visual information (tags) is sufficiently reliable to be used in grounding of the co-occurring speech input. They devised a confidence measure that indicated the reliability of the utterance-tag pair. Based on that confidence measure and a user-set threshold, the agent was able to decide whether the content of an utterance was related to the concurrent visual tag. If the confidence was too low, no learning occurred. Otherwise the model contents were updated according to the standard NMF learning procedure (e.g., Van hamme, 2008). The results of the experiments revealed that the learner was partially able to overcome the uncertainty in the visual domain when actively questioning the reliability of the correspondence between visual and audio domains. If active learning was disabled, the word learning performance was notably hindered (Versteegh et al., 2010).

3.2.5 Conclusions from models of lexical learning

The computational studies reviewed in the previous subsection have successfully demonstrated that the word learning from continuous speech is indeed feasible without explicit top-down information using a variety of techniques. The models with indirect lexical grounding show that, in principle, proto-lexical representations of recurring word forms can be discovered based on the acoustic similarity of word tokens and in the absence of any linguistically motivated expert knowledge in the task. This type of discovery can be based on global matching of similar subsequences (exemplar-based view; Oates, 2002; Park & Glass, 2005; 2006) or on incremental analysis of TPs between automatically discovered speech sounds (P-V; see also Miller & Stoytchev, 2009).

When contextual support in the form of visual abstract tags or visual features representing the objects in the concurrent visual scene are available, the discovery of auditory patterns that are relevant for each tag

can be made efficiently using a variety of approaches (ten Bosch & Cranen, 2007; ten Bosch et al., 2009a; Van hamme, 2008; P-II; P-IV; Aimetti, 2009; Hörnstein et al., 2009). This simulates a learning situation where the learning agent simultaneously hears the speech of the caregiver and shares attention with the caregiver towards some specific objects in the environment.

In theory, direct grounding enables more efficient pattern models due to additional statistical constraints. If the learning is based on the assumption that the visual objects are always present when they are being discussed, the learning algorithm can assume that the auditory patterns occurring in the absence of the visual object are not related to it. This makes it possible to contrast the statistical models so that those aspects of the models that are relevant only for the given visual tag are given higher priority (cf. ten Bosch & Cranen, 2007; P-IV). However, this also means that auditory patterns that are not systematically represented in the visual domain as possible referents do not obtain their own representations.

Indirect grounding, on the other hand, enables acquisition of word forms independently of the surrounding context, making accumulation of vocabulary faster since any words can be learned without requiring evident referents in the external world. However, the obtained word models are initially weaker since the relevance of the patterns must be inferred solely from the statistical properties of the auditory stream. Because the models are learned in isolation from other domains, such as the visual world, there is no guarantee that the learned patterns have optimal correspondence to objects and events in the world (e.g., the agent may learn "*redball*" instead of "*red*" and "*ball*" if the combination occurs sufficiently often, although they are clearly dissociative entities in the world of visuomotoric experience). It is also possible that no distinct model emerges for a word at all, even though the word constantly has a visual referent in a normal learning situation. For example, the word models obtained from the unsupervised TP-based algorithm of P-V can be afterwards grounded to co-occurring visual tags in the 50 keyword recognition task of the CAREGIVER Y2 corpus by simply analyzing the co-occurrence frequencies of the words and tags. If these associative links are then used to recognize the most likely visual objects associated with novel utterances, a word recognition rate of approximately 67% can be obtained (P-VI). This is notably worse than the result obtained with an algorithm applying the direct grounding approach for which above 92% word recognition rate has been achieved using the same material (P-IV).

In general, indirect and direct grounding, when used in isolation, do not seem to directly correspond to the learning challenge faced by infants in early word learning. Requiring the learner always to have a concrete and attended referent for the spoken language it hears in order to learn something seems an unreasonable limitation. On the other hand, as language serves the purpose of lighting up associations in the minds of the receivers, learning a language in total isolation from the environment does

not make sense either. The need for some kind of contextual support is already indicated by the fact that detection and the precise modeling of the word patterns is not an easy task due to immense acoustic and temporal variability in speech. However, one should note that the models of lexical acquisition presented in this work aim to explain the very first steps of the LA, i.e. they show possible ways to bootstrap the learning system. After being able to segment a novel utterance into word-like units and unfamiliar segments, the demands of the acquisition process change and additional hybrid mechanisms of unsupervised pattern discovery and cross-situational grounding may become feasible. Also, since the human memory is highly based on associative links between perceived patterns and events, and since the processing at early sensory cortices is modulated by the multimodal context (Brosch & Scheich, 2005), it may well be that there never is truly unsupervised unimodal learning, but all sensory processing takes place in the context of other modalities and internal states of the learner, even in the absence of "correct" referents. These contextual cues can be then used to store and retrieve patterns from the memory and to categorize them according to the similarity of the contexts in which they occur (cf., P-VI). Over time, the relationships between the patterns and their referential contexts simply become more distinct, enhancing the predictive value of spoken messages.

As for the ecological plausibility of the proposed algorithms, PERUSE is the only one that directly runs into trouble with its batch training. The ecological plausibility of the DTW-based approaches (the P&G algorithm and the work of McInnes and Goldwater, 2007) is greatly enhanced by the work of Unal & Tepedelenlioglu (1992), who showed that the DTW computations can be accomplished with artificial neural networks. In a similar manner, the TP-based modeling of the signal structure in terms of short-term acoustic events is possible with recurrent neural networks with sufficient temporal memory (see, e.g., Elman, 1990), and the average temporal dependency structure explicitly modeled by the TPs has even been shown to have a close correspondence to the integration times measured in the human auditory system (Räsänen & Laine, submitted for publication).

It should be also noted that despite the absence of an explicit phonetic layer, the statistical properties of speech sounds are implicitly taken into account in the approaches utilizing vector quantization of speech frames (e.g., P-II; P-IV; P-V; P-VI; ten Bosch & Cranen, 2007; ten Bosch et al., 2009a; Van hamme, 2008). However, these clusters and the corresponding sequence elements are by no means comparable to phones, not least because they are not defined in duration, but have a fixed and short (typically 10 ms) length, they are not fully selective to speech sounds from only one phonetic category independently of the talker, and because the typical number of elements is much higher than the number of phones in any language. The basic reason for the conversion is not the belief that human infants would perceive speech as sequences of symbolic elements, but because the computational pattern discovery problem is simplified notably. Since the

discretization can be considered as lossy compression, the word modeling results obtained with discrete representations provide a lower bound to the learnability from the data from which semi-continuous or fully continuous methods designed to do the same task should be able to improve, if properly formulated.

4. Summary of publications

Publication I: "Blind segmentation of speech using non-linear filtering methods"

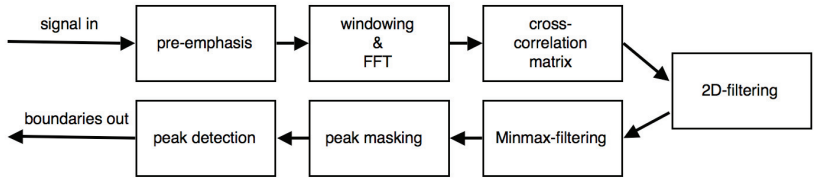


Figure 4: A schematic overview of the blind segmentation algorithm of P-I.

The first paper in the thesis introduces a novel algorithm for unsupervised segmentation of continuous speech into phone-like units. The algorithm is based on the hypothesis that phone segment boundaries can be detected by finding the time instances of sudden spectral changes in the speech signal. The study shows that approximately 75% of phone boundaries can be automatically discovered from speech when manually performed phonetic annotation is used as the reference in the evaluation. These results are in line with previous, methodologically different, approaches to blind segmentation of speech. Therefore the results strengthen the proposition that there is an inevitable upper limit in performance in purely bottom-up approaches to segmentation, making accurate segmentation of speech into phones impossible when only local spectral features are used.

Publication II: "Computational language acquisition by statistical bottom-up processing"

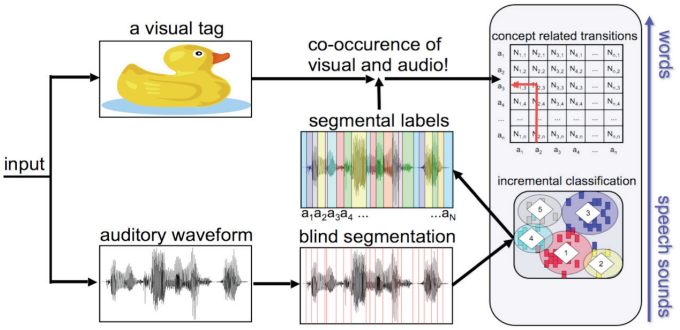


Figure 5: A schematic view of the cross-modal associative learning system in P-II.

It has been proposed in the literature that human infants might track transitional probabilities between phones or syllables in order to segment continuous speech into words. Publication II demonstrates a computational system that is able to learn associations between words in continuous speech and systematically co-occurring variables that represent simultaneously perceived objects in a visual context. More specifically, the system is based on blind segmentation of speech into phone-like units and the tracking of transitional probabilities of these segments in the context of visual objects. The study shows that the transitional probabilities between acoustic segments can lead to a word recognition rate that is notably better than chance when supported by contextual information from the visual stream during the learning stage.

Publication III: "Self-learning vector quantization for pattern discovery from speech"

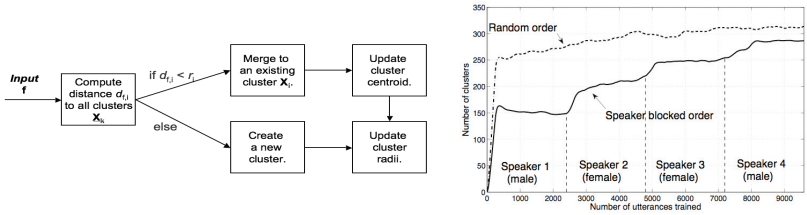


Figure 6: A schematic view of the clustering method in P-III is shown on the left. The right panel shows a result from simulations where the number of acoustic categories (clusters) increases when speech from new talkers is introduced.

Vector quantization of speech signals into sequences of discrete elements is an effective way to simplify the task of pattern discovery from continuous speech. However, the majority of the standard clustering algorithms either require that the number of clusters is specified in advance or they are not suitable for incremental clustering of data, making them implausible approaches for an on-line language learning system. In Publication III, a novel and computationally flexible method called self-learning vector

quantization (SLVQ) for incremental clustering of data is presented. The experiments described in the paper show that the SLVQ algorithm can learn a non-specified number of data clusters from speech features, and that the obtained clusters show comparable quality to a computationally more expensive and non-incremental k-means algorithm. It is also demonstrated that the learning process is relatively stable across a wide range of parameter settings and adapts correctly to the introduction of new acoustically deviant talkers.

Publication IV: "A method for noise-robust context-aware pattern discovery from categorical sequences"

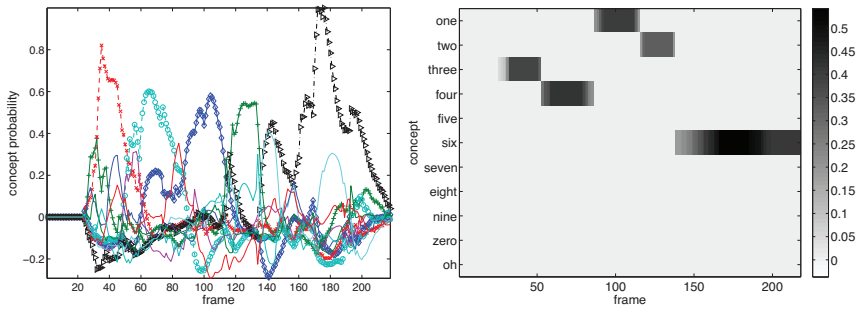


Figure 7: An example of the recognition output from the Concept Matrix (CM) algorithm in a digit recognition task (P-IV). The left panel shows the activity curves for all learned word models and the right panel shows the result after inhibiting the non-winning models.

Publication IV extends the analysis of transitional probabilities between sequential elements by introducing a novel, incremental, and fast algorithm for weakly supervised pattern discovery and recognition from sequential data. Unlike with previously existing techniques such as Markov chains and hidden-Markov models, the present algorithm avoids the problematic Markov assumption by modelling the temporal dependencies of signals at a number of different distances, making it robust against variability and noise in the signals. It is shown in the study that the algorithm performs well in weakly supervised word learning tasks where the precise forms and locations of target words in the training signals are unknown. Also, the algorithm compares favourably against supervised speech recognition approaches in word recognition in noise.

Publication V: "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events"

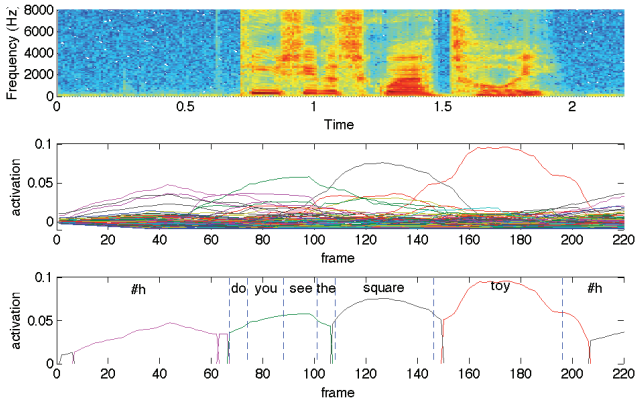


Figure 8: An example of word patterns automatically discovered by the algorithm presented in P-V. The top panel shows a spectrogram of the original speech signal, the middle panel shows activation curves for each learned word, and the bottom-panel shows the final recognition output. True spoken words are also shown in the bottom panel.

Publication V presents a computational system that is able to perform fully unsupervised segmentation and recognition of word-like units from continuous speech without a priori linguistic or phonetic assumptions. Unlike in Publication IV, the present model does not incorporate any type of contextual information such as visual labels in its processing, but simply analyzes transitional probabilities between small timescale acoustic events that are obtained automatically by feature extraction and bottom-up clustering. The experimental results reported in the publication support the theory that, instead of first mastering the phonetic unit of their native language, infants may first acquire proto-lexical representations of speech based on recurring acoustic patterns and only later discover the subword units that are shared between different lexical items. The results also lend support to the statistical learning hypothesis, showing that lexical learning is possible in the absence of innate linguistic knowledge.

Publication VI: "Context induced merging of synonymous word models in computational modeling of early language acquisition"

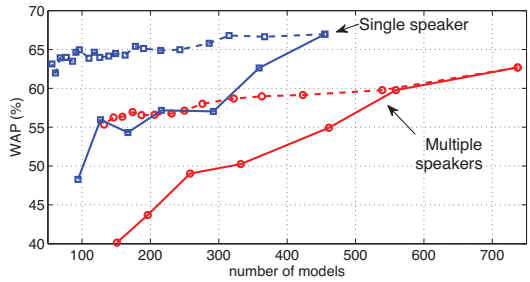


Figure 9: Word-referent association performance in the word learning task of P-VI for multiple talkers and a single talker as a function of the number of word patterns learned from speech. The red solid line shows the performance for normal bottom-up acoustic learning as in P-V, and the blue dashed line shows how the number of parallel word representations can be decreased with the help of word semantics.

Publication VI extends the unsupervised model of word learning of Publication V to a system that first discovers word-like units from speech and then grounds the learned words to contextual variables such as visual objects and events through cross-situational learning. Moreover, the system is able to utilize cross-situational learning in order to discover synonymy of learned word models, allowing merging of models that represent differing acoustic variants of the same word. The experiments performed in the work show that 1) the grounding of word forms to external contextual variables is successful, although the learned models are not as selective as in the case of weakly supervised learning where contextual variables provide additional constraints to the word learning problem, and 2) the merging of synonymous word models allows to reduce the overall number of parallel representations for each word without significant loss in word recognition accuracy.

Publication VII: "Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language"

Research on artificial language acquisition has shown that insertion of short subliminal gaps into a continuous stream of speech has a notable effect on how human listeners interpret speech tokens constructed from syllabic constituents of the language. Based on this finding, it has been argued that the observed results in artificial language acquisition cannot be explained by a single statistical learning mechanism. Publication VII shows that a system performing unsupervised learning of transition probabilities between short-term acoustic events can replicate the main findings of the related behavioral studies. However, success of this learning calls for a specific constraint on the temporal processing of dependencies at the acoustic level, raising the question whether the human auditory system is also limited to the learning of relatively short-term acoustic dependencies.

Publication VIII: "Average spectrotemporal structure of continuous speech matches with the frequency resolution of human hearing"

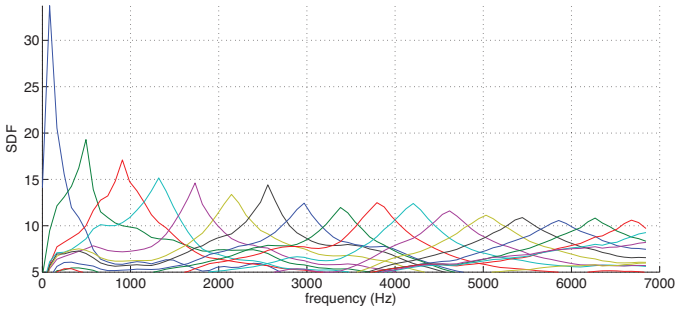


Figure 10: Auditory filterbanks automatically derived from speech signals in P-VIII.

Publication VIII describes how the average spectrotemporal structure of continuous speech can be measured using the so-called spectrotemporal dependency function. The obtained dependency measure can be interpreted as a matched filter for performing signal detection from continuous speech, providing a qualitative formulation for the integration characteristics that would be expected from a hearing system optimized to recognize patterns from speech. Simulations presented in the work show how the average spectrotemporal structure of continuous speech matches with the frequency resolution of human hearing.

Publication IX: "Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions"

Publication IX is a review article on the existing computational models of phonetic and lexical learning. The goal of the publication is to integrate approaches and findings from different studies together in order to see what has been learned from the existing work, how these findings fit to the other studies of language acquisition, and what are the central issues that should be addressed in future studies. The article also represents the main contributions of the current thesis in the context of other studies in the field.

5. Conclusions

It is evident from cognitive science studies that language exposure shapes the perception of speech sounds and that word segmentation is affected by the statistical structure of speech signals. This thesis set out to explore the question of what kind of patterns can be automatically extracted from speech without assuming innate linguistic knowledge. A number of computational algorithms were developed and simulations carried out in order to segment speech into phone-like units, to cluster segments or spectral features into discrete acoustic atomic events, and to learn words from continuous speech using the derived signal representations in weakly supervised and unsupervised language learning conditions. The main findings of this thesis can be summarized as follows:

- Continuous speech contains statistical regularities that can be extracted with unsupervised and weakly supervised pattern discovery algorithms without ASR-like precise annotation and without a priori linguistic knowledge.
- Bottom-up segmentation of speech into phone-like units without a priori language knowledge is possible to some degree. However, the current computational algorithms are not sufficiently accurate for systematic detection of all phone transitions in speech. Also, there is no known way to classify the obtained context-independent segments into discrete and speaker independent phone (or phoneme) categories.
- A small vocabulary of words can be learned automatically without an intermediate phonetic or phonemic representation of speech. Moreover, learning of this proto-lexicon may be essential in the learning of efficient subword coding of language.
- Statistical word learning is highly facilitated by contextual constraints from shared visual attention between the learner and the caregiver.

- Transition probability based learning does not necessarily imply tracking of phone or syllable probabilities, but that the statistical regularities at the level of linguistic units are also necessarily reflected in the statistical regularities of the acoustic speech signals.
- Time-frequency structure of continuous speech is matched to the frequency resolution of the human auditory system when the structure is measured in terms of statistical dependency of speech energy at different frequencies. This match is motivated by the optimality principles of pattern detection.

In general, together with the results from simultaneous work carried out in other research groups[3], the results reveal that word learning is possible purely on the basis of the statistical structure available in the speech signals and that this learning does not require an intermediate layer of phonetic or syllabic representations of language. In contrast, the unsupervised acquisition of robust speaker invariant subword units for speech perception is a complex task that is not satisfactorily solved by the existing computational methods and may require additional sources of information in addition to the distributional characteristics of low-level acoustic features. The current evidence, however, is not conclusive, proving only what is possible instead of proving what is impossible. For example, none of the existing models have been able to address the hypothesis of PRIMIR that the phonemic representation of language emerges later in the development through accumulation of lexicon and discovery of similarities across different lexical items.

If a learning agent could somehow obtain sufficiently a systematic and invariant sequential representation of speech sounds comparable to the phonetic or syllabic transcriptions made by expert phoneticians, the various word segmentation methods described in the literature show how word learning can be accomplished using this type of representation (e.g., de Marcken, 1995; Christiansen et al., 1998; Brent & Cartwright, 1996; Brent, 1999; Venkataraman, 2001; Swingley, 2005; Blanchard et al., 2010), and how syntactic categories could be also inferred from phonological representations using distributional information (Christiansen et al., 2009). Before that, more research is required in order to understand how the interplay between lexical and sub-lexical representations drives the development of language proficiency and communicative capability in early development.

In addition to demonstrating different strategies for statistical word learning, the thesis demonstrates that statistical learning is not necessarily

3 Much of the research was carried out in collaboration with L. Boves and L. ten Bosch from Radboud University Nijmegen, Netherlands; R. K. Moore and G. Aimetti from University of Sheffield, England; H. Van hamme, J. Driesen and K. Demuynck from University of Leuven, Belgium; and B. Kleijn, G. Henter and C. Koniaris from KTH, Sweden, during the Acquisition of Communication and Recognition Skills (ACORNS)-project funded by EU FP6 FET.

limited to linguistically motivated representations such as phones or words, but, in theory, the frequency resolution characteristics of the human auditory system can also be derived from speech signals with a neurally plausible learning rule.

5.1 Open issues and future work

The current computational models face a number of issues that are not fully addressed by any of the existing models. One of the biggest problems is the generalization across tokens with variable acoustic properties, such as different talkers. Distributional categories of speech sounds learned from purely acoustic signals are talker specific, and speaker-independent representations overlap so largely that discrimination of context-independent vowels is far from perfect. The same is true for learned lexical items, where generalization towards new talkers is poor (see, e.g., P-V; ten Bosch et al., 2009c). How infants overcome the generalization problem is currently not understood, although first indications of the important role of the communicative context were provided in P-VI. Also, the manner that articulatory development, lexical learning, or, e.g., speaker normalization in the acoustic domain affect speech sound perception are not yet completely understood.

Another largely unexplored area is the role of prosody in early LA. Although behavioral studies indicate that infants are sensitive to prosodial aspects such as intonation and stress (Thiessen & Saffran, 2003; Thiessen & Saffran, 2004; Cutler, 1994; Jusczyk, 1993b), no computational model dealing with continuous speech has so far been able to utilize these features efficiently in its functionality. One should however note that the prosodial features are inherently included in all standard features that encode the wide-band spectrum of the speech signal such as the FFT and Mel frequency cepstral coefficients (MFCCs). The question then remains whether young infants treat prosody or other suprasegmental cues as a separate source of information and process them in isolation from the systems dealing with phonetic and lexical identity. If so, inclusion of a separate mechanism for prosodial processing should show some value in computational simulations. If not, a mechanism explaining the development of the ability to separate linguistic and paralinguistic information from the one and same signal is required.

Temporal representation of speech signals is also an open question. Current computational models typically describe speech as a sequence of feature frames extracted at fixed intervals. While the segmentation of the signal into phone-like units before lexical access has been studied (P-I and P-II; Markey, 1994; ten Bosch & Cranen, 2007), bottom-up discovery of phone-like units is evidently difficult (see section 3). Interestingly, no computational approach has truly utilized syllabification of speech signals, although, e.g., the WRAPSA model of LA (Jusczyk, 1993a) directly states that the syllables serve as the basic temporal slices of speech input in perception. This is strengthened by the argument that seeing any other units

than syllables that would enable automatic temporal normalization of speech is difficult (Mehler et al., 1990). Also, Werker and Curtin (2005) argue in their PRIMIR theory that there is an innate preference for syllabification of speech input. The unsupervised segmentation of speech signals into syllabic units is known to be much more systematic and accurate with computational algorithms than the blind segmentation into phone-like units (Villing et al., 2006). Finally, EEG studies suggest that human speech comprehension performance correlates with the synchronization of the auditory cortex to the energy envelope of speech, and this synchronization shows deficits in dyslexic patients (Ahissar & Ahissar, 2005). Therefore the role of syllables in early LA should be further investigated with computational models.

The question of the role of grounding in the learning of internal representations for acoustic patterns corresponding to words also needs more attention in the future. There is clear evidence that highly accurate models for words can be learned in cross-situational learning simulations, where the learning mechanism receives utterances paired with a set of possible word referents and that this scales up to vocabularies of at least 50 words and multiple target words in each utterance (P-II; P-IV; P-VI; Van hamme, 2008; ten Bosch et al. 2009a; Aimetti, 2009). However, the statistical linkage between the acoustics and the referents is direct, and the algorithms cannot learn models for word patterns without clear referential information. Moreover, no lexical learning occurs at all if no referents are available. This also means that the models cannot learn words that do not have distinctive contextual referents. Also, the simulations have very strong assumptions (but not necessarily unreasonable; see Räsänen, 2012) regarding the coherence between the contents of the spoken utterances and the attention of the learner. On the other hand, unsupervised acquisition of words in the absence of referential information is demonstrated in the works of Oates (2001, 2002), Park and Glass (2005, 2006) and in P-V, but the generalization performance of these word models is worse due to their inability to overcome significant acoustic differences between word tokens in the absence of any contextual constraints, possibly leading to multiple parallel models for each word spoken in different acoustic conditions.

The general problem is that none of the existing models provide a systematic strategy to exploit both bottom-up statistical cues and cross-situational cues in concert in order to find the best possible representations for the incoming speech. In language learning literature, the word forms are often assumed to be first segmented from the speech stream before meaning can be attached to them (e.g., Werker & Curtin, 2005), being in line with the idea of indirect lexical grounding. On the other hand, the contents of the early receptive vocabulary of infants mainly consists of nouns with very distinctive external referents (e.g., Gentner, 1983; see also MacArthur–Bates communicative development inventories, Fenson et al., 2003), suggesting that the bootstrapping of early lexicon could be also explained by acoustic patterning based on direct cross-situational learning. Given the current

evidence, it is difficult to say whether word forms come first based solely on their acoustic properties or whether contextual constraints play a role in the lexical learning all the way from the beginning.

Joint attention and intentionality in language learning are also topics of future research. The social-pragmatic theory of word learning states that a linguistic symbol itself is a tool to share attention between many persons and that learning children already know that the function of language is to direct the attention of others (Tomasello, 2000). The theory also assumes that the linguistic competence arises from the learner's ability to infer the intentions of the speaker and then relate these to the spoken messages instead of just associating superficially perceived objects and actions to concurring words. Intentionality is also claimed to drive attention so that the task of the perceptual system is to search for goal-relevant features in the environment (Tomasello, 1995). In other words, the entire language use is about modulating attention toward internal and external concepts. However, the social-pragmatic theory does not state how the intentions are extracted or represented by the learner. This poses a difficult task for computational models of LA, since modeling of intentionality calls for agents with a highly developed ability to understand and reason the state of matters in the surrounding world, not only for their own percepts and actions, but also for actions and states of other agents. The question how intentionality could be modeled in simulated environments in a plausible manner is way beyond the scope of this discussion, but the reader is recommended to see Kaplan & Hafner (2006) for a related review.

Finally, there are numerous other important phenomena that are barely touched in the existing research on computational models of LA. For example, syntax has been largely ignored in the existing work, assuming that syntactic learning follows from lexical knowledge. However, syntax may actually provide additional cues to the word learning problem: instead of assuming that the spoken words are purely independent of each other, the statistical dependencies across hypothesized word-like patterns may be used as an additional criterion in the learning process. In a similar vein, articulatory learning and the role of feedback in perceptual learning are not currently understood, although the hypothetical role of the articulatory domain in speech perception has been disputed since the introduction of the motor theory of speech perception (Liberman & Mattingly, 1985). There is also a plethora of knowledge from experimental studies related to bilingualism, language-related developmental disorders, and, e.g., statistical learning in non-speech domains that could be used to inspire and to evaluate future models of language acquisition.

So far, we have barely started the journey towards understanding how language is represented in the human mind, how it is be learned by infants, and how it could be learned by machines. On this journey, computational modeling will serve as an important tool for testing new theories and integrating different pieces of language-related knowledge into coherent functional models. The development of new techniques for autonomous

machine learning also provides new solutions to different research and industrial applications dealing with language or other patterned data. However, the ultimate success in understanding language learning and processing will depend on the advances and cross-fertilization across the different disciplines seeking to understand aspects of human communication and cognitive processes behind the use of language.

References

Aaltonen, O., Eerola, O., Lang, H., Uusipaikka, E., & Tuomainen, J., 1994. Automatic discrimination of phonetically relevant and irrelevant vowel parameters as reflected by mismatch negativity. *Journal of the Acoustical Society of America*, 96, 1489–1493.

Abler, W., 1989. On the particulate principle of self-diversifying systems. *Journal of Social and Biological Structures*, 12, 1–13.

Ahissar, E. & Ahissar, M., 2005. Processing of the temporal envelope of speech. In R. König, P. Heil, E. Budinger, & H. Scheich (Eds.), *The auditory cortex: A synthesis of human and animal research* (pp. 295–314). New Jersey: Lawrence Erlbaum Associates.

Aimetti, G., 2009. Modelling early language acquisition skills: Towards a general statistical learning mechanism. *Proc. EACL-2009-SRWS*, Athens, Greece, pp. 1–9.

Almpanidis, G. & Kotropoulos, C., 2008. Phonemic segmentation using the generalized Gamma distribution and small sample Bayesian information criterion. *Speech Communication*, 50, 38–55.

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H., 2010. A Speech Corpus for Modeling Language Acquisition: CAREGIVER. *Proc. International Conference on Language Resources and Evaluation (LREC)*, Malta, pp. 1062–1068.

Ananthakrishnan, G., 2011. *From Acoustics to Articulation*. Doctoral Thesis, KTH, Stockholm, Sweden.

Ananthakrishnan, G. & Salvi, G., 2011. Using Imitation to Learn Infant-Adult Acoustic Mappings. *Proc. Interspeech'2011*, Florence, Italy, pp.765–768.

Aslin, R. N. & Newport, E. L., 2012. Statistical Learning: From Acquiring Specific Items to Forming General Rules. *Current Directions in Psychological Science*, 21, 170–175.

Arciuli, J. & Simpson, I. C., 2012. Statistical Learning is Related to Reading Ability in Children and Adults. *Cognitive Science*, 26, 286–304.

Aversano, G., Esposito, A., Esposito, A. & Marinaro, M., 2001. A New Text-Independent Method for Phoneme Segmentation. *Proc. IEEE international Workshop on Circuits and Systems*, Dayton, Ohio, USA, pp. 516–519.

Blakemore, C., & Cooper, G., 1970. Development of the brain depends on the visual environment. *Nature*, 228, 477–478.

Blanchard, D., Heinz, J., & Golinkoff, R., 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, 37, 487–511.

de Boer, B., & Kuhl, P., 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4, 129–134.

ten Bosch, L., Van hamme, H., Boves, L. & Moore, R.K., 2009a. A computational model of language acquisition: the emergence of words. *Fundamenta Informaticae*, 90, 229–249.

ten Bosch L., Boves L. & Räsänen O., 2009b. Learning meaningful units from multimodal input - the effect of interaction strategies. *Proc. Workshop on Child, Computer and Interaction 2009 (WOCCI)*, Boston, MA, United States.

ten Bosch L., Räsänen O., Driesen J., Aimetti G., Altosaar T. & Boves, L., 2009c. Do Multiple Caregivers Speed up Language Acquisition? *Proc. Interspeech'09,* Brighton, England, pp. 704–707.

Brent, M. R., & Cartwright, T. A., 1996. Distributional regularity and phonotactics are useful for segmentation. *Cognition*, 61, 93–125.

Brent, M. R., 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34, 71–105.

Brosch, M., & Scheich, H., 2005. Non-acoustic influence on neural activity in auditory cortex. In R. König, P. Heil, E. Budinger, & H. Scheich (Eds.), *The auditory cortex: A synthesis of human and animal research* (pp. 127–144). New Jersey: Lawrence Erlbaum Associates.

Bulf, H., Johnson, S. P. & Valenza, E., 2011. Visual statistical learning in the newborn infant. *Cognition*, 121, 127–132.

Buttery, P., 2006. *Computational Models for First Language Acquisition*. Technical Report No. 675, University of Cambridge, Computer Laboratory, UK.

Chomsky, N., 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, N., 1975. *Reflections on Language*. Pantheon Books, New York.

Chomsky, N., 1980. *Rules and Representations*. Columbia University Press.

Christiansen, M. H., Allen, J. A., & Seidenberg, M. S., 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221–268.

Christiansen, M. H., Onnis, L., & Hockema, S. A., 2009. The secret is in the sound: from unsegmented speech to lexical categories. *Developmental Science*, 12, 388–395.

Coen, M. H., 2006. Self-supervised acquisition of vowels in American English. *Proc. 21st national conference on Artificial intelligence*, Boston, USA, 2, pp. 1451–1456.

Curtin, S., Mintz, T. H. & Byrd, D., 2001. Coarticulatory cues enhance infants' recognition of syllable sequences in speech. *Proc. 25th Annual Boston University Conference on Language Development*, Somerville, MA: Cascadilla, pp. 190–201.

Curtin, S., Mintz, T. H. & Christiansen, M. H., 2005. Stress changes representational landscape: Evidence from word segmentation. *Cognition*, 96, 233–262.

Cutler, A., 1994. Segmentation problems, rhythmic solutions. *Lingua*, 92, 81–104.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C. & Fadiga, L., 2009. The motor somatotopy of speech perception. *Current Biology*, 19, 381–385.

Daland, R., & Pierrehumbert, J., 2011. Learning Diphone-Based Segmentation. *Cognitive Science*, 35, 119–155.

Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society - Series B (Methodological)*, 39, 1–38.

Demuynck, K., & Laureys, T., 2002. A Comparison of Different Approaches to Automatic Speech Segmentation. *Proc. 5th International Conference on Text, Speech and Dialogue*, pp. 277–284.

Driesen, J., ten Bosch, L. & Van hamme, H., 2009. Adaptive Non-negative Matrix Factorization in a Computational Model of Language Acquisition. *Proc. Interspeech'09*, Brighton, England, pp. 1731–1734.

Duda, R., Hart, P., & Stork, D., 2001. *Pattern Classification (2nd Edition)*. Wiley-Interscience, New York.

Edeline, J-M., 1999. Learning-induced physiological plasticity in the thalamo-cortical sensory systems: a critical evaluation of receptive field plasticity, map changes and their potential mechanisms. *Progress in Neurobiology*, 57, 165–224.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. & Vigorito, J., 1971. Speech perception in infants. *Science*, 171, 303–306.

Elman, J., 1990. Finding Structure in Time. *Cognitive Science*, 14, 179–211.

Emmorey, K., 2006. The signer as an embodied mirror neuron system: neural mechanisms underlying sign language and action. In M. Arbib (Ed.), *From Action to Language via the Mirror Neuron System*. Cambridge University Press, New York, pp. 110–135.

Endress, A. D. & Bonatti, L. L., 2007. Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, 105(2), 247–299.

Esposito, A. & Aversano, G., 2005. Text Independent Methods for Speech Segmentation. In G. Chollet et al. (Eds.), *Lecture Notes in Computer Science: Nonlinear Speech Modeling*. Springer Verlag, Berlin Heidelberg, pp. 261–290.

Estevan, Y.P., Wan, V. & Scharenborg, O., 2007. Finding Maximum Margin Segments in Speech. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, Honolulu, Hawaii, USA, pp. IV–937–940.

Feldman, N., Griffiths, T. & Morgan, J., 2009a. Learning phonetic categories by learning a lexicon. *Proc. 31st Annual Conference of the Cognitive Science Society*, Austin, Texas, pp. 2208–2213.

Feldman, N., Griffiths, T. L., & Morgan, J. L., 2009b. The Influence of Categories on Perception: Explaining Perceptual Magnet Effect as Optimal Statistical Inference. *Psychological Review*, 116, 752–782.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S. & Bates, E., 2003. *MacArthur–Bates communicative development inventories (CDIs) (2nd ed.)*. Baltimore, MD: Brooks Publishing.

Frank, S. L., Bod, R. & Christiansen, M. H., 2012. How hierarchical is language use? *Proceedings of the Royal Society B*, published online before print.

Fowler, C., 1989. Real objects of speech perception: a commentary on Diehl and Kluender. *Ecological Psychology*, 1, 145–169.

Gentner, D., 1983. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language Development, Vol. 2, Language, cognition and culture*. Hillsdale, NJ: Erlbaum.

Gleitman, L. R., 1990. The structural sources of verb meanings. *Language Acquisition*, 1, 3–55.

Goldstein, L., Byrd, D., & Saltzman, E., 2006. The role of vocal tract gestural action units in understanding the evolution of phonology. In M. Arbib (Ed.), *From Action to Language via the Mirror Neuron System* (pp. 215–249). Cambridge University Press, New York.

Golinkoff, R. M., Mervis, C. B. & Hirsh-Pasek, K., 1994. Early object labels: the case for a developmental principles framework. *Journal of Child Language*, 21,125–155.

Hauser, M. D., Newport, E. L. & Aslin, R. N., 2001. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, 78, B53–B64.

Hawkins, J., & Blakeslee, S., 2005. *On Intelligence*. Times Books, Henry Holt and Co.

Hebb, D. O., 1949. *The organization of behavior*. New York: Wiley & Sons.

Hillenbrand, J., Getty, L. A., Clark, M. J. & Wheeler, K., 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099–3111.

Hirsch, H. G., & Pearce, D., 2000. The AURORA Experimental framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions. *Proc. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, pp. 29–32.

Hockett, C., 1960. The Origin of Speech. *Scientific American*, 203, 89–97.

Houston, D. M. & Jusczyk, P. W., 2003. Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 1143–1154.

Howard, I. S, & Messum, P., 2011. Modeling the Development of Pronunciation in Infant Speech Acquisition. *Motor Control*, 15, 85–117.

Hörnstein, J., Gustavsson, L., Lacerda, F. & Santos-Victor, J., 2009. Multimodal Word Learning from Infant Directed Speech. *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'09)*. St. Louis, USA, pp. 2748–2754.

Jusczyk, P. W., 1993a. From general to language-specific capacities: the WRAPSA model of how speech perception develops. *Journal of Phonetics*, 21, 3–28.

Jusczyk, P. W., 1993b. Discovering sound patterns in the native language. *Proc. 15th Annual Meeting of the Cognitive Science Society*, Colorado, Boulder, pp. 49–60.

Keshet, J., Shalev-Shwartz, S., Singer, Y., & Chazan D., 2005. Phoneme Alignment Based on Discriminative Learning. *Proc. Interspeech'05*, pp. 2961– 2964.

Kirkham, N. Z., Slemmer, J. A. & Johnson, S. P., 2002. Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition*, 83, B35–B42.

Kohonen, T., 1990. The Self-organizing Map. *Proceedings of the IEEE*, 78, 1464–1480.

Kouki, M., Kikuchi, H., & Mazuka, R., 2010. Unsupervised Learning of Vowels from Continuous Speech Based on Self-Organized Phoneme Acquisition Model. *Proc. Interspeech'2010*, pp. 2914–2917.

Krunic, V., Salvi, G., Bernardino, A., Montesano, L. & Santos-Victor, J., 2009. Affordance based word-to-meaning association. *Proc. IEEE international conference on Robotics and Automation (ICRA'09)*, Kobe, Japan, pp. 806–811.

Kuhl, P. K., 1986. Theoretical contributions of tests on animals to the special mechanisms debate in speech. *Experimental Biology*, 45, 233–265.

Kuhl, P. K., 2000. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97, 11850–11857.

Kuhl, P. K., 2004. Early Language Acquisition: Cracking the Speech Code. *Nature Reviews Neuroscience*, 5, 831–843.

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T. & Pruitt, J., 2005. Early speech perception and later language development: implications for the "critical period". *Language Learning and Development*, 1, 237–264.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. & Iverson, P., 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9, F13–F21.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M. & Nelson T., 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded, (NLM-e). *Phil. Trans. Royal Society B*, 363, 979–1000.

Laakso, A., & Calvo, P., 2011. How Many Mechanisms Are Needed to Analyze Speech? A Connectionist Simulation of Structural Rule Learning in Artificial Language Acquisition. *Cognitive Science*, 35, 1243–1281.

Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, L., Marklund, E. & Sundberg, U., 2004. Ecological Theory of Language Acquisition. *Proc. Fourth International Workshop on Epigenetic Robotics (Epirob 2004)*.

Lake, B., Vallabha, G. & McClelland, J., 2009. Modeling Unsupervised Perceptual Category Learning. *IEEE Transactions on Autonomous Mental Development*, 1, 35–43.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M., 1967. Perception of the Speech Code. *Psychological Review*, 74, 431–461.

Liberman, A., & Mattingly I., 1985. The motor theory of speech perception revised, *Cognition*, 21, 1–36.

Lieberman, P., 2007. The Evolution of Human Speech. *Current Anthropology*, 48, 39–66.

Lippmann, R., 1997. Speech recognition by machines and humans. *Speech Communication*, 22, 1–15.

MacQueen, J. B., 1967. Some Methods for Classification and Analysis of Multivariate Observations. *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 281–297.

de Marcken, C., 1995. *The unsupervised acquisition of a lexicon from continuous speech*. AI Memo No. 1558, Massachusetts Institute of Technology, MA.

Marr, D., 1982. *Vision: A Computational Approach*. San Francisco, Freeman & Co.

Mehler, J., Dupoux, E. & Segui, J., 1990. Constraining models of lexical access: the onset of word recognition. In G. T. Altmann (Ed.), *Cognitive models of speech processing*. Hillsdale, NJ:Erlbaum.

Nearey, T. M., 1997. Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, 101, 3241–3254.

Newport, E. L. & Aslin, R. N., 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127–162.

Newport, E. L., Hauser, M. D., Spaepen, G. & Aslin, R. N., 2004. Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, 49, 85–117.

Markey, K. L., 1994. *The Sensorimotor Foundations of Phonology: A Computational Model of Early Childhood Articulatory and Phonetic Development*. Doctoral Thesis, University of Colorado, Dept. Computer Science, Colorado, USA.

Maye, J., Werker, J. F. & Gerken, L., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101–111.

Maye, J., Weiss, D. J. & Aslin, R. N., 2008. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, 11, 122–134.

McClelland, J., & Elman, J., 1986. The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.

McInnes, F. & Goldwater, S., 2011. Unsupervised extraction of recurring words from infant-directed speech. *Proc. 33rd Annual Meeting of the Cognitive Science Society*, Boston, MA, pp. 2006–2012.

McNeilage, P. F., & Davis, B. L., 2005. The Evolution of Language. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology*. John & Wiley & Sons, Inc., Hoboken, New Jersey.

Miller, M. & Stoytchev, A., 2009. An Unsupervised Model of Infant Acoustic Speech Segmentation. *Proc. 9th International Conference on Epigenetic Robotics*, Venice, Italy, 12–14.

Meltzoff, A. N., Kuhl, P. K., Movellan, J. & Sejnowski, T. J., 2009. Foundations for a New Science of Learning. *Science*, 325, 284–288.

Moore, R. K., 2007. Spoken language processing: Piecing together the puzzle. *Speech Communication*, 49, 418–435.

Mountcastle, V., 1978. An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System. In Edelman, G. M. & Mountcastle, V. (Eds.): *The Mindful Brain*. MIT Press, Cambridge, MA.

Newman, M. E. J., 2004. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 0066133–1–5.

Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52, 189–234.

Norris, D., McQueen, J., & Cutler, A., 2000. Merging information in speech: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 299–370.

Oates, T., 2002. PERUSE: An unsupervised algorithm for finding recurrent patterns in time-series. *Proc. IEEE International Conference on Data Mining (ICDM)*, Maebashi City, Japan, pp. 330–337.

Oates, T., 2001. *Grounding Knowledge in Sensors: Unsupervised Learning for Language and Planning*. Doctoral Thesis, University of Massachusetts Amherst, MA, USA.

Park, A. & Glass, J. R., 2005. Towards Unsupervised Pattern Discovery in Speech. *Proc. 2005 IEEE Workshop Automatic Speech Recognition and Understanding (ASRU'05)*, Cancún, Mexico, pp. 53–58.

Park, A. & Glass, J. R., 2006. Unsupervised word acquisition from speech using pattern discovery. *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, pp. 409–412.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J., 2002. Signal-driven computations in speech processing. *Science*, 298, 604–607.

Pinker, S., 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pisoni, D. B., 1997. Some thoughts on "Normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing*. San Diego: Academic Press, pp. 9–32.

Port, R., 2007. How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, 25, 143–170.

Port, R., 2010. Language as a Social Institution: Why Phonemes and Words Do Not Live in the Brain. *Ecological Psychology*, 22, 304–326.

Ptito, M., Moesgaard, S. M., Gjedde, A., & Kupers, R., 2005. Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. *Brain*, 128, 606–614.

Pulvermüller, F., 2010. Brain-Language Research: Where is the Progress? *Biolinguistics*, 4, 255–288.

Quine, W. V. O., 1960. *Word and object*. Cambridge, MA: MIT Press.

Rasilo, H., Räsänen, O., & Laine, U. K., submitted for publication. Feedback and imitation by caregiver guides a virtual child to learn native phonemes and the skill of speech inversion.

Roy, D., 2005. Grounding words in perception and action: computational insights. *TRENDS in Cognitive Sciences*, 9, 389–396.

Räsänen, O., 2010. Fully Unsupervised Word Learning from Continuous Speech Using Transitional Probabilities of Atomic Acoustic Events. *Proc. Interspeech'10*, Chiba, Japan, pp. 2922–2925.

Räsänen, O. & Driesen, J., 2009. A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition. *Proc. 17th Nordic Conference on Computational Linguistics*, Odense, Denmark, pp. 255–262.

Räsänen, O., Laine U.K. & Altosaar T., 2009a. An Improved Speech Segmentation Quality Measure: the R-value. *Proc. Interspeech'09*, Brighton, England, pp. 1851–1854.

Räsänen, O., Laine, U. K. & Altosaar, T., 2009b. A noise robust method for pattern discovery in quantized time series: the concept matrix approach. *Proc. Interspeech'09*, Brighton, England, pp. 3035–3038.

Räsänen, O., Rasilo, H. & Laine U. K., 2012. Modeling spoken language acquisition with a generic cognitive architecture for associative learning. *Proc. Interspeech'2012*, Portland, Oregon.

Räsänen, O., 2012. Non-auditory cognitive capabilities in computational modeling of early language acquisition. *Proc. Interspeech'2012*, Portland, Oregon.

Räsänen, O. & Laine, U., submitted for publication. Time-frequency integration characteristics of hearing are optimized for perception of speech-like acoustic patterns.

Saffran, J. R., Aslin, R. N. & Newport, E. L., 1996a. Statistical Learning by 8-Month-Old Infants. *Science*, 274, 1926–1928.

Saffran, J. R., Newport, E. L. & Aslin R. N., 1996b. Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, 35, 606–621.

Saffran, J. R., Johnson, E. K., Aslin, R. N. & Newport, E. L., 1999. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52.

Saffran, J. R., 2001. Words in a sea of sounds: The output of statistical learning. *Cognition*, 81, 149–169.

Sato, M., Buccino, G., Gentilucci, M., & Cattaneo, L., 2010. On the tip of the tongue: Modulation of the primary motor cortex during audiovisual speech perception. *Speech Communication*, 52, 533–541.

Saussure, F. D., 1916. *Course in general linguistics*. (W. Baskin, Trans.), New York, NY: Philosophical Library.

Scharenborg, O., 2007. Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49, 336–347.

Scharenborg, O., Ernestus, M., & Wan, V., 2007. Segmentation of speech: Child's play? *Proc. Interspeech'07*, Antwerp, Belgium, pp. 1953–1956.

Scharenborg, O. & Boves, L., 2010. Computational modelling of spoken-word recognition processes. *Pragmatics & Cognition*, 18, 136–164.

Skipper, J. I., Nusbaum, H. C. & Small, S. L., 2006. Lending a helping hand to hearing: another theory of speech perception. In M. Arbib (Ed.), *From Action to Language via the Mirror Neuron System* (pp. 250–286). Cambridge University Press, New York.

Smith, K., Smith, A. D., Blythe, R. A. & Vogt, P., 2006. Cross-situational learning: a mathematical approach. *Proc. Third International Workshop on the Emergence and Evolution of Linguistic Communication*, Rome, Italy, pp. 31–44.

Smith, K., Smith, A. D., & Blythe, R. A., 2011. Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms. *Cognitive Science*, 35, 480–498.

Smith, L. B., & Yu, C., 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558–1568.

Stager, C. L., & Werker, J. F., 1997. Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388, 381–382.

Studdert-Kennedy, M. & Goldstein, L., 2003. Launching language: The gestural origin of discrete infinity. In. M. H. Christiansen & S. Kirby: *Language Evolution: The States of the Art*. Oxford University Press

Sur, M., Garraghty, P. E., & Roe, A. W., 1988. Experimentally induced visual projections into auditory thalamus and cortex. *Science*, 242, 1437–1441.

Swingley, D., 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86–132.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P. & Huotilainen, M., 2009. Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neurosicence*, 10:21.

Thiessen, E., & Saffran, J. R., 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39, 706–716.

Thiessen, E., & Saffran, J. R., 2004. Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception and Psychophysics*, 65, 779–791.

Toledano, D., Hernández Gómez, L. & Villarubia Grande, L., 2003. Automatic Phonetic Segmentation. *IEEE Transactions in Speech and Audio Processing*, 11, 617–625.

Tomasello, M., 2000. The Social-Pragmatic Theory of Word Learning. *Pragmatics*, 10, 401–413.

Tomasello, M., 1995. Joint attention as social cognition. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Hillsdale, NJ: Erlbaum.

Toscano, J. C. & McMurray, B., 2010. Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science*, 34, 434–464.

Trehub, S. E., 1976. The discrimination of foreign speech contrasts by infants and adults. *Child Devevelopment*, 47, 466–472.

Tsao, F.-M., Liu, H.-M. & Kuhl, P. K., 2004. Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Development*, 75, 1067–1084.

Unal, F. A., & Tepedelenlioglu, N., 1992. Dynamic time warping using an artificial neural network. *Proc. International Joint Conference on Neural Networks (IJCNN)*, Baltimore, Maryland, pp. 715–721.

Vallabha, G. K., McLelland, J. L., Pons, F., Werker, J. F. & Amano, S., 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of National Academy of Sciences*, 104, 13273–13278.

Van hamme, H., 2008. HAC-models: a Novel Approach to Continuous Speech Recognition. *Proc. Interspeech'08*, Brisbane, Australia, pp. 2554–2557.

Venkataraman, A., 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27, 351–372.

Versteegh, M., ten Bosch, L. & Boves, L., 2010. Active word learning under uncertain input conditions. *Proc. Interspeech'10*, Chiba, Japan, pp. 2930–2933.

Villing, R., Ward, T. & Timoney, J., 2006. Performance Limits for Envelope based Automatic Syllable Segmentation. *Proc. Irish Signals and Systems Conference (ISSC2006)*, Dublin, Ireland, pp. 521–526.

Vouloumanos, A., 2008. Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107, 729–742.

Vygotsky, L. S., 2012. *Thought and Language – Revised and expanded edition. Translation and foreword by A. Kozulin*. MIT Press.

Warren, R. M., 2000. Phonemic organization does not occur: Hence no feedback. Commentary to Norris, D., McQueen, J. M., & Cutler, A., 2000. Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 350–351.

Waterson, N., 1971. Child phonology: a prosodic view. *Journal of Linguistics*, 7, 179–211.

Watkins, K. E., Strafella, A. P., & Paus, T., 2003. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41, 989–994.

Werker, J. F. & Tees, R. C., 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavioral Development*, 7, 49–63.

Werker, J. F., & Lalonde, C. E., 1988. Cross-Language Speech Perception: Initial Capabilities and Developmental Change. *Developmental Psychology*, 24, 672–683.

Werker, J. F., & Curtin, S., 2005. PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, 1, 197–234.

Wiesel, T. N., & Hubel, D. H., 1963. Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Neurophysiology*, 26, 1003–1017.

Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csépe, V., Aaltonen, O., Raimo, I., Alho, K., Lang, H., Iivonen, A., & Näätänen, R., 1999. Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. *Cognitive Brain Research*, 7, 357–369.

Witner, S., 2010. Computational Models of Language Acquisition. *Proc. 11th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'2010)*, Iași, Romania, pp. 86–99.

Yang, C. D., 2004. Universal Grammar, statistics or both? *TRENDS in Cognitive Sciences*, 8, 451–456.

Errata

Publication P-VII

Equation (6) should read $A_{\text{tot}} = \max_{c \in C, t \in [0,T]} (A_c(t) \mid X)$, where T is the duration of the stimulus.

Human infants acquire their native language almost effortlessly and without explicit training. However, answers to questions such as how the learning takes place, what aspects of language can be learned purely from sensory experience, and, e.g., how much of language learning requires language-specific innate constraints are yet largely unknown. This thesis attempts to address these questions by studying how the language acquisition process could be modeled using domain general statistical learning mechanisms. The focus is on understanding what type of linguistic and statistical structures are learnable from acoustic speech signals, and what type of computational algorithms are required to perform these learning tasks.

Aalto-DD 55/2013

ISBN 978-952-60-5096-6
ISBN 978-952-60-5097-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS