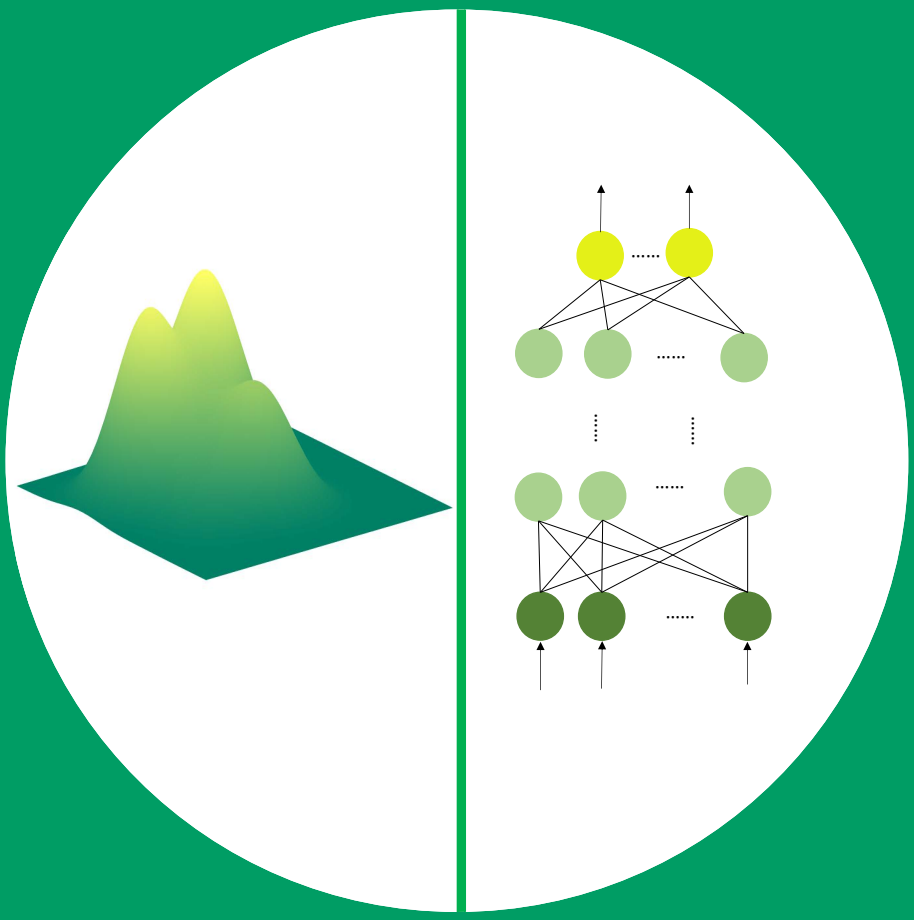
Department of Signal Processing and Acoustics

Machine learning methods for suprasegmental analysis and conversion in speech

Shreyas Seshadri



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 201/2020

Machine learning methods for suprasegmental analysis and conversion in speech

Shreyas Seshadri

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, Remote connection link (e.g. Zoom), on 18 December 2020 at 12:00.

Zoom link-
https://aalto.zoom.us/j/62070164113?pwd=bTV4cFZXRmcyUExESi81bGk3Nm45Zz09

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics

Supervising professors
Professor Paavo Alku, Aalto University, Finland

Thesis advisor
Assistant Professor Okko Räsänen, Tampere University, Finland

Preliminary examiners
Associate Professor Hung-yi Lee, National Taiwan University, Taiwan
Assistant Professor Yan Tang, University of Illinois, USA

Opponent
D.Sc. Ville Hautamäki, University of Eastern Finland, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 201/2020

ISBN 978-952-64-0166-9 (printed)
ISBN 978-952-64-0167-6 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0167-6

Unigrafia Oy
Helsinki 2020

Finland

Printed matter
4041-0619

Aalto University

Abstract

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

Author
Shreyas Seshadri

Name of the doctoral dissertation
Machine learning methods for suprasegmental analysis and conversion in speech

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 201/2020

Field of research Speech and Language Technology

Manuscript submitted 22 May 2020 **Date of the defence** 18 December 2020

Permission for public defence granted (date) 22 October 2020 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Speech technology is a field of technological research focusing on methods to process spoken language. Work in the area has largely relied on a combination of domain-specific knowledge and digital signal processing (DSP) algorithms, often combined with statistical (parametric) models. In this context, machine learning (ML) has played a central role in estimating the parameters of such models. Recently, better access to large quantities of data has opened the door to advanced ML models that are less constrained by the assumptions necessary for the DSP models and are potentially capable of achieving higher performance.

The goal of this thesis is to investigate the applicability of recent state-of-the-art (SoA) developments in ML to the modelling and processing of speech at the so-called suprasegmental level to tackle the following topical problems in speech research: 1) zero-resource speech processing (ZS), which aims to learn language patterns from speech without access to annotated datasets, 2) automatic word (WCE) and syllable (SCE) count estimation which focus on quantifying the amount of linguistic content in audio recordings, and 3) speaking style conversion (SSC), which deals with the conversion of the speaking style of an utterance while retaining the linguistic content, speaking identity and quality. In contrast to the segmental level which consists of elementary speech units known as phone(me)s, the suprasegmental level encodes more slowly varying characteristics of speech such as the speaker identity, speaking style, prosody and emotion. The ML-approaches used in the thesis are non-parametric Bayesian (NPB) models, which have a strong mathematical foundation based on Bayesian statistics, and artificial neural networks (NNs), which are universal function approximators capable of leveraging large quantities of training data. The NN variants used include 1) end-to-end models that are capable of learning complicated mapping functions without the need to explicitly model the intermediate steps, and 2) generative adversarial networks (GANs), which are based on training a minimax game between two competing NNs.

In ZS, NPB clustering methods were investigated for the discovery of syllabic clusters from speech and were shown to eliminate the need for model selection. In the WCE/SCE task, a novel end-to-end model was developed for automatic and language-independent syllable counting from speech. The method improved the syllable counting accuracy by approximately 10 percentage points from the previously published SoA method while relaxing the requirements of the data annotation used for the model training. As for SSC, a new parametric approach was introduced for the task. Bayesian models were first studied with parallel data, followed by GAN-based solutions for non-parallel data. GAN-based models were shown to achieve SoA performance in terms of both subjective and objective measures and without access to parallel data. Augmented CycleGANs also enable manual control of the degree of style conversion achieved in the SSC task.

Keywords suprasegmental speech processing, Bayesian learning, deep learning, zero-resource speech processing, word and syllable count estimation, speaking style conversion.

ISBN (printed) 978-952-64-0166-9 **ISBN (pdf)** 978-952-64-0167-6

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2020

Pages 175 **urn** http://urn.fi/URN:ISBN:978-952-64-0167-6

Preface

I still distinctly remember my first day at Aalto University back in the autumn of 2014. My initial impression was that this was a calm, serene and very creative space. I knew right away that I had made the right decision in taking up this PhD program. That impression has only been augmented during the last six years I have spent here. I came here as a timid but ambitious student. I feel that I have grown into a confident researcher willing to take up challenging projects. The field of machine learning solutions to problems in speech processing has, of late, exploded with new possibilities. I was fortunate to be part of a very competent research team at SPA who helped me to address the new developments.

First and foremost, I would like to thank my thesis advisor Asst. Prof. Okko Räsänen for being a patient and inspirational guide throughout the duration of my thesis research. His contribution towards my growth has been immense. He has been a mentor to me, in the true sense of the word. I would like to thank my thesis supervisor Prof. Paavo Alku for accepting me as part of his group. He allowed me the perfect balance between freedom and guidance. I would like to thank him as well for giving me the opportunity to work on the speaking style conversion project. Thanks to Prof. Unto Laine for the interesting research avenues we explored in the early days of my PhD program.

I would like to acknowledge the contributions of the co-authors of the publications attached in this thesis. Special thanks to Dr. Ulpu Remes, who gave me the confidence to tackle even the most complicated mathematical equations. I enjoyed our work together on Bayesian modelling. Thanks, as well to Dr. Lauri Juvela who helped kick-start my hands-on work with deep learning models and generative adversarial networks. Thanks to Prof. Junichi Yamagishi for our collaboration. Thanks to the entire ACLEW team, Dr. Marisa Casillas, Dr. Alejandrina Cristia, Associate Prof. Florian Metze, Associate Prof. Melanie Soderstrom, Asst. Prof. Elika Bergelson, Prof. Celia Renata Rosemberg, Prof. Björn Schuller and Prof. Emmanuel Dupoux, for involving me in the research work. I will cherish the memories of our group meetings in Paris and Buenos Aires. Thanks to Ana Ramíres

López for our work together. Thanks to the organisers of the zero-resource speech processing events at Interspeech, the ACLEW team and to Dr. Emma Jokinen for the data used in the experiments in this thesis.

Special thanks to Ilkka Huhtakallio, Dr. Sofoklis Kakouros and Dr. Sudarsana Reddy Kadiri for being great office roommates. Thanks as well to the rest of my current and past colleagues at SPA including Prof. Mikko Kurimo, Associate Prof. Tom Bäckström, Dr. Bajibabu Bollepalli, Dr. Narendra N P, Dr. Manu Airaksinen, Dr. Jouni Pohjalainen, Dr. Dhananjaya N. Gowda, Dr. Rahim Saeidi, Sneha Das, Pablo Perez, Farhad Javanmardi and Hilla Pohjalainen.

I would also like to thank Dr. Ramya Rasipuram, Dr. Ladan Golipour, David Winarski, and the rest of the team at Apple for giving me the opportunity to be an intern and to start my career in industry with them.

I extend my gratitude to the pre-examiners of the thesis Associate Prof. Hung-yi Lee and Asst. Prof. Yan Tang for their invaluable comments which greatly helped in elevating the quality of this thesis. Thanks to Dr. Ville Hautamäki for accepting to be the opponent of this thesis.

I would like to acknowledge my family without whom I would not be here. First and foremost, I would like to thank my father K S Narayanaswamy who has been a pillar of support and integrity. Throughout my education, he has been intensely involved with every aspect of my work. I would certainly not be where I am today without his presence. Thanks to my mother Gowri Narayanaswamy for being the ever-present joyful light in my life. Thanks to my sisters Dr. Janani Akhilandeswari and Dr. Priyanka Parvathi for inspiring me to continue with the sibling tradition of PhD in the family. Their guidance through every step of my career has been invaluable. Special thanks to my brother-in-law Dr. Balasubramanian Ramani who was responsible for exposing me to the research environment in Europe back in 2011. Thanks to my uncle K. S. Ramasubramanian, my aunt Vatsala and my brother-in-law Santosh for always being there. Thanks also to my in-laws Vijayakrishnan Narayanaswamy, Dhanasree and Lavanya for their support. And most importantly, I would like to thank my beloved wife Pavithra Vijayakrishnan for infusing in me the necessary focus and drive to complete this thesis work. The ambience of joy, love and laughter that she created around me was essential in helping me through the most difficult phases towards the end of the PhD research. I would not have been able to complete this thesis work without her support.

Helsinki, November 14, 2020,

Shreyas Seshadri

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Seshadri, S., Remes, U. & Räsänen, O.. Comparison of non-parametric Bayesian mixture models for syllable clustering and zero-resource speech processing. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, pp. 2744–2748, August 2017.

II Seshadri, S., Remes, U. & Räsänen, O.. Dirichlet process mixture models for clustering i-vector data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, pp. 5740–5744, March 2017.

III Räsänen, O., Seshadri, S., Karadayi, J., Riebling, E., Bunce, J., Cristia, A., Metze, F., Casillas, M., Rosemberg, C., Bergelson, E. & Soderstrome, M.. Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech. *Speech Communication*, vol. 113, pp. 63–80, October 2019.

IV Seshadri, S. & Räsänen, O.. SylNet: An adaptable end-to-end syllable count estimator for speech. *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1359–1363, July 2019.

V Seshadri, S., Juvela, L., Räsänen, O. & Alku, P.. Vocal effort based speaking style conversion using vocoder features and parallel learning. *IEEE Access*, vol. 7, pp. 17230–17246, January 2019.

VI Seshadri, S., Juvela, L., Yamagishi, J., Räsänen, O. & Alku, P.. Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 6835–6839, May 2019.

VII Seshadri, S., Juvela, L., Alku, P. & Räsänen, O.. Augmented CycleGANs for continuous scale normal-to-Lombard speaking style conversion. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, pp. 2838–2842, September 2019.

Author's Contribution

Publication I: "Comparison of non-parametric Bayesian mixture models for syllable clustering and zero-resource speech processing"

The author was involved closely in planning the use of different non-paramteric Bayesian models for zero-resource speech processing. The author implemented the Bayesian Gaussian mixture model (BGMM) as well as the different weight distributions: the Dirichlet distribution (DD), the Dirichlet Process (DP) and the Pitmann-Yor process (PYP). The author ran all the experiments and was the primary writer of the manuscript.

Publication II: "Dirichlet process mixture models for clustering i-vector data"

The author was involved closely in planning the use of different Dirichlet process (DP) models for speaker clustering using i-vector features. The author implemented the Dirichlet process Gaussian mixture model (DPGMM), ran all the experiments, and was the primary writer of the manuscript.

Publication III: "Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech"

The author had a major role in developing the syllable-based feature extractor for the paper. The author also participated to writing of Sections 2 and 4 of the paper.

Publication IV: "SylNet: An adaptable end-to-end syllable count estimator for speech"

The author designed and implemented the SylNet algorithm introduced in the paper, ran the experiments, and was the primary writer of the manuscript.

Publication V: "Vocal effort based speaking style conversion using vocoder features and parallel learning"

The author implemented the machine learning models used for the speaking style conversion (SSC) system and performed all the experiments. The author also played a major role in the writing of the whole manuscript, except for Section 3 (Vocoders) which was primarily written by the second author.

Publication VI: "Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion"

The author implemented the CycleGAN model and performed all the experiments. The author also played a major role in the writing of the whole manuscript, except for Section 3.1 (PML Vocoder) which was primarily written by the second author.

Publication VII: "Augmented CycleGANs for continuous scale normal-to-Lombard speaking style conversion"

The author implemented the Augmented CycleGAN model and performed all the experiments. The author also played a major role in the writing of the whole manuscript, except for Section 3.1 (PML Vocoder) which was primarily written by the second author.

Language check

The language of my dissertation has been checked by Lingsoft Language Services Oy. I have personally examined and accepted/rejected the results of the language check one by one. This has not affected the scientific content of my dissertation.

List of Figures

Abbreviations

1D	one-dimensional
2D	two-dimensional
ASR	automatic speech recognition
BGMM	Bayesian Gaussian mixture model
BLSTM	bi-directional long-short term memory network
BM	basilar membrane
CNN	convolutional neural network
CycleGAN	cycle-consistent generative adversarial network
DD	Dirichlet distribution
DDGMM	Dirichlet distribution Gaussian mixture model
DFT	discrete Fourier transform
DNN	deep neural network
DP	Dirichlet process
DPGMM	Dirichlet process Gaussian mixture model
DSP	digital signal processing
DTW	dynamic time warping
E-step	expectation step
ELBO	evidence lower bound
EM	expectation maximisation
ERB	equivalent rectangular bandwidth

GAN	generative adversarial network
GMM	Gaussian mixture model
GRU	gated recurrent unit
HMM	hidden Markov model
INCA	iterative combination of a nearest neighbour search step and a conversion step alignment method
JSD	Jensen-Shannon divergence
KLD	Kullback–Leibler divergence
LDA	linear discriminant analysis
LSGAN	least squares generative adversarial network
LSTM	long-short term memory
M-step	maximization step
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MFCC	mel-frequency cepstral coefficients
MGC	mel-generalised cepstrum
MH	Metropolis Hastings
ML	machine learning
MLE	maximum likelihood
MLM	machine learning model
MLP	multi-layer perceptron
NIW	normal-inverse Wishart distribution
NLG	natural language generation
NLU	natural language understanding
NN	neural network
NPB	non-parametric Bayesian
PML	pulse model in log-domain
PYP	Pitmann-Yor process

PYPVMM	Pitmann-Yor process von Mises-Fisher mixture model
PDF	probability density function
RNN	recurrent neural network
SAD	speech activity detector
SCE	syllable count estimation
SIIB	speech intelligibility in bits
SGD	stochastic gradient descent
SGMM	standard Gaussian mixture model
SoA	state-of-the-art
SPSS	statistical parametric speech synthesis
SSC	speaking style conversion
SVM	support vector machine
TTS	text-to-speech
USM	unsupervised sub-word modelling
UTD	unsupervised spoken term discovery
VC	voice/speaker identity conversion
VMF	von Mises-Fisher distribution
WCE	word count estimation
WGAN	Wasserstein generative adversarial network
ZS	zero-resource speech processing

Symbols

List of Latin symbols

$\mathbf{a}$	notation for vector
$\mathbf{A}$	notation for matrix
$\mathbf{b}$	bias of an NN
Beta(v_k; $1,\gamma$)	Beta distribution modelling the stick-breaking process with v_k representing the proportion of the stick broken off at time k and parameterised by γ
d	dimensionality of data
d_l	dimensionality the lth layer of an NN
D	discriminator of a GAN
D^*	optimal discriminator of a GAN
DD($\boldsymbol{\zeta};\boldsymbol{\alpha}$)	PDF of Dirichlet distributed data $\boldsymbol{\zeta}$ parameterised by α
DP(γ,H)	a DP parameterised by concentration parameter γ and a distribution over model parameters H
$f()$	a certain function over the analytically intractable posterior $p(\boldsymbol{\theta}\|\mathbf{X})$ that Bayesian inference methods aim to approximate
f_p	frequency (Hz) of an acoustic signal
$g_l()$	activation function of the lth layer of an NN
g'_l	derivative of activation function of the lth layer of an NN
G	generator of a GAN

G^*	optimal generator of a GAN
$\mathbf{h}_t$	hidden state of an RNN at time t
H	distribution over model parameters H that is one of the parameters of a DP
$\mathbf{j}$	a sequence of words that are the output of an ASR system
$\mathbf{j}^*$	optimal sequence of words that are the output of an ASR system
KLD$(q\|\|p)$	KLD between distributions $p()$ and $q()$
K	number of Gaussian components in a GMM
$\mathbf{l}$	sequence of sub-word units in an ASR system
L	number of layers of an NN
$\mathcal{L}$	cost function of an NN
m	perceived pitch (mel)
$\mathbf{m}$	mean parameter of an NIW
M	the number of independent samples drawn from the posterior distribution $p(\boldsymbol{\theta}\|\mathbf{X})$
N_k	the amount of data that the kth Gaussian is responsible for explaining in a GMM
$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$	PDF of frequentist Gaussian distributed data $\mathbf{x}$ parameterised by mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma})$	PDF of Bayesian Gaussian distributed data $\mathbf{x}$ given the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
NIW$(\boldsymbol{\mu},\boldsymbol{\Sigma};$ $\mathbf{m},\beta,\mathbf{V},\nu)$	PDF of NIW distributed data $\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ parameterised by the mean $\mathbf{m}$, number of prior measurements β, scale matrix $\mathbf{V}$ and degrees of freedom ν
$p()$	a PDF
$\mathcal{P}_{\text{Data}}$	distribution of the dataset
$\mathcal{P}_{\text{MLM}}$	distribution of the output of an MLM
$q_\lambda(\boldsymbol{\theta})$	variational distribution with hyperparameters λ used to approximate the analytically intractable posterior $p(\boldsymbol{\theta}\|\mathbf{X})$
r_m	mth sample, of M independent samples, drawn from the posterior distribution $p(\boldsymbol{\theta}\|\mathbf{X})$

v_k	proportion of the stick broken off at time k in a stick breaking process
$\mathbf{V}$	scale matrix of an inverse-Wishart distribution
$\mathbf{w}$	d dimensional weights of a single neuron of an NN of the form $\mathbf{w} = \{w_0, w_1, ...w_d\}$
$\mathbf{W}_l$	$\{d_{l-1} \prod d_l\}$ dimensional weights of the lth layer of an NN
$\mathbf{W}_h$	one of the parameters of an RNN
$\mathbf{W}_x$	one of the parameters of an RNN
$\mathbf{W}_o$	one of the parameters of an RNN
$\mathbf{W}^*$	optimal set of weights of an NN
$\mathcal{W}^{-1}(\boldsymbol{\Sigma}; \mathbf{V}, \nu)$	PDF of an inverse-Wishart distribution used to model the covariance of a Gaussian $\boldsymbol{\Sigma}$ parameterised by the scale matrix $\mathbf{V}$ and degrees of freedom ν
$x(n)$	time-varying discrete signal at time instant n
$\mathbf{X}$	data matrix of N data points as $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_N}\}$
$\mathbf{x}$	data point from a dataset
$\hat{\mathbf{x}}$	fake data point as the output of the generator of a GAN
$\mathbf{x}_t$	data point from a dataset at time t
$X(k)$	discrete Fourier transform of $x(n)$
$\mathbf{y}$	output of an NN for the input data point $\mathbf{x}$
$\mathbf{y}_t$	output of an RNN at time t
z_{nk}	a binary latent variable that encodes the identity of the kth Gaussian that each of the data $\mathbf{x}_n$ is associated with in a GMM
$\mathbf{z}$	a 1-of-K binary latent variable that encodes the identity of the Gaussian that each of the data $\mathbf{x}$ is associated with in a GMM

List of Greek symbols

$\boldsymbol{\alpha}$	parameter of a DD
β	number of prior measurements parameter of an NIW
γ	concentration parameter of a DP

$\Gamma()$	Gamma function
$\boldsymbol{\zeta}$	weights on each Gaussian of a GMM
$\boldsymbol{\theta}$	parameters of a Bayesian model
$\boldsymbol{\theta}^*_{\text{MLE}}$	optimal set of parameters for a Bayesian model in the sense of MLE
$\boldsymbol{\theta}^*_{\text{MAP}}$	optimal set of parameters for a Bayesian model in the sense of MAP
$\kappa(z_{nk})$	responsibilities representing the contribution of each of the kth Gaussian in 'explaining' the data point $\mathbf{x}_n$ in a GMM
λ	hyperparameters of variational distribution $q_\lambda(\boldsymbol{\theta})$
λ^*	optimal hyperparameters of variational distribution $q_\lambda(\boldsymbol{\theta})$
$\boldsymbol{\mu}$	mean of a Gaussian distribution
ν	degrees of freedom of an inverse-Wishart distribution
$\boldsymbol{\xi}_l$	outputs of the lth layer of an NN
$\boldsymbol{\Sigma}$	covariance matrix of a Gaussian distribution
ϕ	learning rate of the gradient descent algorithm used to train an NN
$\boldsymbol{\omega}$	samples from a known distribution that are fed to a generator in a GAN

1. Introduction

Communication through acoustic signals is widely observable among animals. Over the millennia, this acoustic communication has evolved into a complex phenomenon in humans called speech. Spoken language has taken on a variety of different forms in various human civilisations and cultures (Lieberman and McCarthy, 2015). In order to communicate over distances in space and time, several human societies have also developed orthographic representations of the spoken language that come in many forms and shapes, such as the Latin script or the Chinese writing system (Poole, 1999). Speech, together with its orthographic representation, can be considered as a central component to human intelligence and cultural evolution (Lieberman and McCarthy, 2015; Smith and Kirby, 2008). The speech signal is transmitted as vibrations in air, and is produced and perceived in humans by highly complex physiological and neurological systems. Speech encompasses several levels of linguistic structure, which are not all transparent in the textual orthographic representation. At the short time-scale, the rapidly varying characteristics of the signal are largely related to phonetic categories, also known as phonetic segments. These phonetic segments often have a direct correspondence to their symbolic representation in written text. The suprasegmental level of speech, on the other hand, encodes slower varying characteristics such as speaker identity, speaking style, prosody, dialect, or other speaker states and traits such as emotions. Paralinguistic aspects of this information are lost when spoken language is represented in terms of symbols aimed at conveying the linguistic content of the message, as the written symbols focus purely on the lexical and syntactical contents. In other words, the speech signal is packed with various levels of information that are partially absent from written language.

The last several decades of research in the field of speech processing have dealt with the problem of enhancing or computationally emulating different aspects of the production or perception chains of speech. Speech processing has several areas of research, including *analysis*, *synthesis* and *conversion*. Analysis deals with the extraction of some of the modalities of information

present in the speech signal, for example, automatic speech recognition (ASR), speaker identity recognition, emotion classification etc. Synthesis deals with the creation of a speech signal that contains some specified information such as text-to-speech (TTS). Finally, conversion deals with the change of some aspect of information in a given speech utterance while the quality and the rest of the modalities of information remain unmodified, for example, speaker identity conversion, style conversion etc.

Nearly all speech processing problems involve some type of processing at both segmental and suprasegmental levels, be it implicit or explicit. However, while applications such as ASR focus primarily on the linguistic content of speech, others have a stronger emphasis on either analysing, converting, or otherwise utilising the structure of speech beyond the linguistic content. For instance, some studies and algorithms deal with the analysis of prosodic cues in speech (e.g., Kalinli and Narayanan, 2009; Kakouros and Räsänen, 2016). Others deal with the analysis of paralinguistic features such as speaker identity (e.g., Atal, 1976; Rosenberg, 1976; Campbell, 1997), emotion (e.g., Schuller, 2018; Akçay and Oğuz, 2020) and speech disorders (e.g., Narendra and Alku, 2019). These aspects of suprasegmental information are also useful in the context of speech conversion and synthesis. For example, the task could be one where a speech utterance has to be generated not only with certain linguistic information but also with a specific prosody or using a specific voice. While speech synthesis can be controlled to generate speech with given characteristics (Wang et al., 2018; Hsu et al., 2019), voice conversion (Stylianou, 2009; Lorenzo-Trueba et al., 2018; Toda et al., 2016) and speaking style conversion can be applied to modify one of these modalities of suprasegmental information when the speech signal is already given.

As for the technology underlying the speech processing applications, data-driven approaches called machine learning (ML) have played an important role in the creation of statistical models aimed at solving speech processing problems, including the analysis (e.g., Povey et al., 2011; Chiu et al., 2018; Rouhe et al., 2020; Schmidt et al., 2014), synthesis (e.g., Taylor, 2009; Airaksinen et al., 2016; van den Oord et al., 2018) and conversion (e.g., Stylianou et al., 1998; Mohammadi and Kain, 2017; Inanoglu and Young, 2009) listed above. The complexity of the ML models that can be reliably trained is proportional to the amount of data available. Since speech is such a complex phenomenon, the amount of data required to model several of the problems listed above is quite large. Hence several speech processing systems have relied on a combination of signal processing (DSP) techniques and domain-specific knowledge in conjunction with statistical ML models. Such systems are based on certain approximations that simplify the real world phenomena, which ultimately sets an upper limit to their performance. Recently, increasing access to large quantities of speech data together with steady increases in computational power have made

it possible to create advanced ML models that rely on fewer assumptions. Such models are potentially capable of achieving higher performance than their DSP-based counterparts.

The research topics studied in our department, when I joined in late 2014, included topics related to human language acquisition and speaking style control. I was interested in the application of state-of-the-art ML to suprasegmental solutions in these broad areas, which led to the main topics of focus for the current thesis. This thesis therefore aims to apply state-of-the-art ML to the modelling and processing of speech at the suprasegmental level to tackle the following three topical problems in speech research:

- *Zero-resource speech processing (ZS):* At a high level, ZS systems aim to emulate several speech processing models without access to annotated data. This field of research is primarily motivated by the human learning of speech and language, which takes place without access to annotated data. In this context, this thesis investigates the use of syllabic patterns in speech as a gateway for unsupervised vocabulary acquisition from speech data.

- *Automatic word and syllable count estimation (WCE/SCE):* These systems aim to quantify the amount of linguistic content in real-world audio recordings. For example, this could be used to quantify the amount of speech in different social scenarios, as captured by wearable microphones, or to quantify the amount of speech that a child is exposed to in different socio-economic scenarios, thereby enabling the study of child-language acquisition using large-scale naturalistic data. This thesis attempts to solve the WCE/SCE problem by focusing on the analysis of units that occur at the rhythmic level, namely, syllables.

- *Speaking style conversion (SSC):* SSC deals with the technology of converting the speaking style of utterances from one style to another while retaining the quality, speaker identity, and linguistic information of the original utterance. This thesis focuses on vocal effort-based SSC—an area where a large amount of training data is typically not available—and uses the conversion between normal and Lombard styles as a case study of the topic. This thesis proposes a parametric framework for the SSC task combining parametric vocoders and ML models, to investigate the extent that the problem can be solved using both parallel and non-parallel training data.

In order to tackle these intricate problems, this thesis explores the use of several ML methods. The major ML-approaches investigated are Bayesian models and artificial neural networks (NNs). Bayesian models are probabilistic models that are based on Bayesian statistics. The parameters in

these models are random variables with their own distributions. NNs are loosely based on the neurons in the brain. These are universal function approximators that typically include several layers of non-linear transformations. Specifically, extensions of the NN that are used are 1) end-to-end models that are capable of learning complicated mapping functions without the need to explicitly model intermediate steps and representations within the function, and 2) generative adversarial networks (GANs), that are powerful generative models based on training a minimax game between two competing NNs.

This thesis is organised in two parts: 1) the introduction and 2) a collection of peer-reviewed articles published in speech technology journals and conferences. The first part provides an introductory overview of the topics covered in the thesis, comprising eight chapters. Specifically, Chapter 2 introduces the theoretical background of the production, perception, and some basics of digital speech processing. Chapter 3 outlines the framework for the Bayesian and NN-based ML models used in this thesis. Chapters 4, 5 and 6 describe the motivation and prior art in the speech processing topics addressed in the thesis. Chapter 7 gives a short summary of the publications in this thesis. Finally, Chapter 8 discusses the conclusions and implications of the research carried out in the thesis.

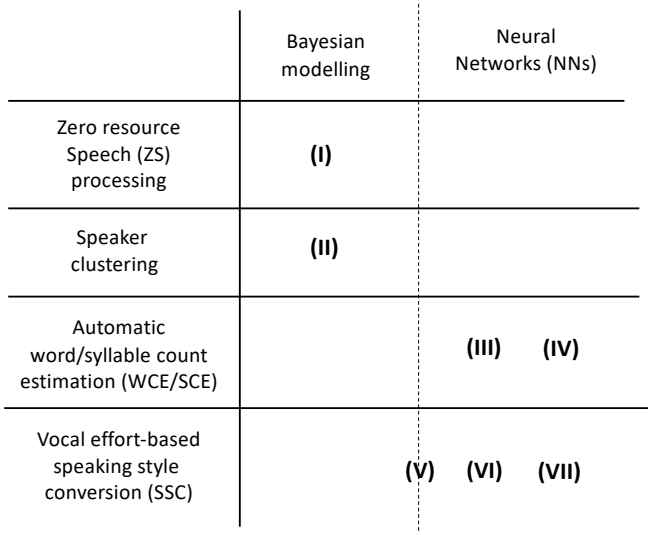


Figure 1.1. Overview of the topics of the papers included in this thesis and the primary machine learning approaches investigated in each of them. The Roman numerals refer to the publication numbers, as listed in Chapter 7.

The second part consists of seven peer-reviewed publications that are organised as shown in Figure 1.1. Publication I deals with the task of linguistic pattern learning in the ZS setting. Non-parametric Bayesian (NPB) methods were explored in the clustering of syllabic (rhythmic) units.

Publication II deals with the task of speaker clustering using NPB methods. Publications III and IV deal with WCE/SCE using recurrent NNs and end-to-end NNs, respectively, in order to achieve state-of-the art performance. Publication V outlines a parametric system for vocal effort-based SSC. It compares three different vocoders and three ML approaches for SSC with parallel data. Publications VI and VII extend the work on the parametric system developed in Publication V to non-parallel data with GAN-based NN models. These models are able to solve the SSC problem without access to speech data with the same linguistic content in the source and target speaking styles while achieving manual control of the strength of the style change during the conversion.

2. Structure, production, and perception of speech signals

This thesis deals with speech technology algorithms focusing on the processing of speech signals at the suprasegmental level. This section will briefly discuss the different aspects of the speech signal in general, its structure and the principles of its production and perception. It also briefly highlights speech processing models based on the mathematical approximations of these complex physical systems. Section 2.1 starts with a basic description of the structure of speech signals and briefly discusses their properties from the time-frequency perspective. Section 2.2 outlines the principles of the human speech production system and the *source-filter* model that is a commonly used approximation. Section 2.3 discusses the principles of the human auditory perception system and some scales of measurement that approximate its frequency sensitivity. Section 2.3 also highlights the *mel-frequency filter bank*, which is a commonly used method to approximate human auditory perception in technological applications. The same section also examines *vocoders*, which are modules capable of compressing a speech signal to a set of parameters and synthesising a speech signal back from the same set of parameters that can be potentially modified between the analysis and synthesis.

2.1 Basic structure of speech

Speech constitutes one of the major modalities of human communication. It is an auditory signal produced by the *human speech production system* that carries information through vibrations in the air. These vibrations are captured and decoded into meaningful information by the *human auditory perception system*. From the perspective of linguistics, the speech signal can be thought of as a collection of units called *phones* that are defined in terms of their unique acoustic properties. A collection of one or more phones make up rhythmic patterns known as *syllables*, and one or more syllables make up units known as *words*. One or more words make up *utterances*, which are the basic units of communicative acts in spoken language and are

often separated by pauses. While words are the smallest meaning bearing entities in a language as stand-alone units, another central concept is the *phoneme*: an abstraction of a phone, defined in terms of its capability to contrast meanings. In short, phonemes are the smallest unit of language that, when changed from one to another, can also change the meaning of the word they belong to. The structure of speech can also be viewed in terms of the time-scale and phenomena of interest. The *segmental* properties of a speech signal are those that change at the phonetic level, that is, at the level of phone segments, and are therefore concerned with the linguistic content of the spoken message. In contrast, *suprasegmental* properties are the slower varying characteristics of the speech signal that span beyond individual phone segments by modifying the intonation, stress patterns and durational properties of the speech signal. Suprasegmentals play a grammatical role in structuring the speech into constituent units at various temporal scales, and also indirectly contribute to the meaning of the spoken message by providing countless ways to express the same phonemic content. The suprasegmental level is also largely affected by paralinguistic factors (see below) such as the general speaking style or the speaker's emotional state. As examples of different time-scales of suprasegmental phenomena, syllabic information varies at the syllable-rhythm level, while *style*, *emotion* and *prosodic* information typically varies at the level of one or more syllables or words. However, much slower changes also take place in the characteristics of speech signals due to factors such as the age and health state of the speaker that may change at the time-scale of several days or even years.

The information present in the speech signal can also be categorised as *linguistic* and *paralinguistic*. Linguistic information refers to the lexical and syntactical content of speech, and can be represented in terms of symbolic entities in phonetic or phonemic transcriptions or using letters of written text. Paralinguistic information in speech includes all the other types of information in a speech signal that cannot be captured via its textual representation, such as prosody, speaker identity, style, or emotion. Hence, paralinguistic information is largely conveyed at the suprasegmental level of speech.

In terms of digital signal processing, a speech signal can be represented as a time-varying discrete signal $x(n)$ at a certain *sampling frequency* representing the air pressure variation as a function of time. Another very useful domain of representation for a speech signal is the frequency domain. A given time domain speech signal of length N can be represented in the frequency domain using the discrete Fourier transform (DFT) as

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-\frac{2\pi i n k}{N}} \tag{2.1}$$

where X is the N-length frequency domain representation of the speech

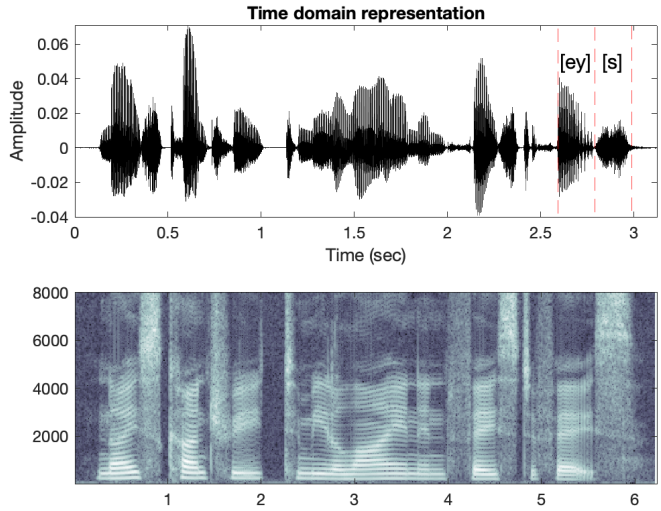


Figure 2.1. From top to bottom showing a) the time domain and b) log-magnitude spectrogram representations for the English utterance *'Nice country to meet a lion in face to face.'* spoken by a female speaker sampled at 16kHz (taken from the TIMIT corpus (Garofolo et al., 1993)). Positions of the phones [ey] and [s] are highlighted.

signal. Since the spectrum is a complex valued vector, it can be represented in terms of its magnitude and phase. Since speech is not stationary but changes over time, speech signals are often analysed using a combination of time and frequency domains. In this case, the DFT is calculated over a series of fixed length windows applied on the speech signal, resulting in a *spectrogram* representation. As an example, Figure 2.1 shows the time-domain signal and log-magnitude spectrogram (calculated as $20\ \log_{10}(|X(k)|)$ for each window) for an English utterance. The lighter colours on the spectrogram indicate a higher energy at that time-frequency position.

2.2 Principles of human speech production and its modelling

A schematic illustration of the human speech production system is shown in Figure 2.2. Speech is produced by the movement of air from the lungs that is first converted into a periodic pulse train in the vocal folds. This pulse train is further modulated by the positioning of articulators in the human speech production system and finally radiated through the lips (and nose) into the surrounding air. In more detail, the movement of the diaphragm constricts the lungs, pushing air upwards through the trachea and larynx to the vocal folds. The vocal folds are composed of twin infoldings of mucous membrane that control the opening between

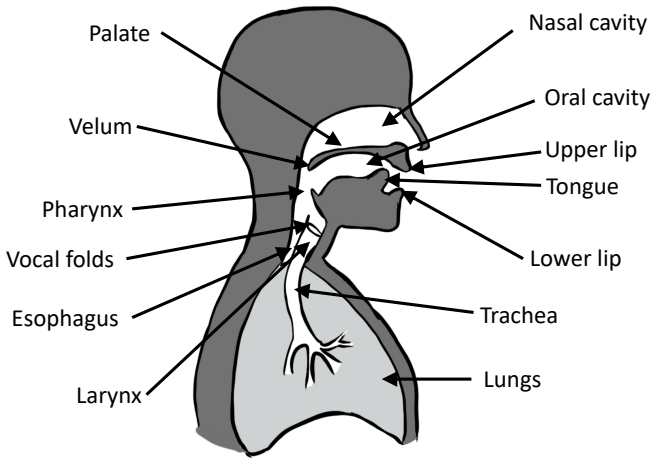


Figure 2.2. Illustration of the human speech production system. Figure adapted from Flanagan, 1972.

the larynx and pharynx. During voiced speech the vocal folds vibrate at a certain frequency which is the *fundamental frequency* of speech. The vocal tract modulates the frequency characteristics of the air flow by creating resonances and antiresonances called *formants*. This is done by changing the shape of the vocal tract and by potentially obstructing the flow of air at certain points of the vocal tract. The air flow coming up the larynx can also be directed to the nasal cavities with the help of a soft tissue called the velum. The shape of the vocal tract can be voluntarily changed by the movement and positioning of the pharynx, tongue, lips and jaw. Full obstructions to the flow of air are mainly created by the upper and lower lips and the contact of different parts of the tongue with the back of the teeth, alveolar ridge, or with the top of the palate. The speech is finally radiated through the lips and nostrils into the surrounding air.

Phones can be classified based on whether or not there is vibration of the vocal folds as *voiced* sounds or *unvoiced* sounds, respectively. Phones can also be classified based on the amount of obstruction in the vocal tract. Vowels are high-energy speech sounds that are produced when there is no obstruction to the flow of air through the vocal tract. The properties of vowel sounds are primarily adjusted by the positioning of the tongue in the mouth and potential rounding of the lips. As a result, each vowel is characterised by a unique pattern of formants, and the first two formants are the most central for the identification of the vowel. *Consonants* are phones produced when there is some obstruction to the flow of air, and come in many varieties. Speech is generally an alternating pattern of vowels and consonants. This creates the rhythmic pattern in the energy of the speech signal responsible for syllabic units.

Several of these structures can be observed in the log-magnitude spectro-

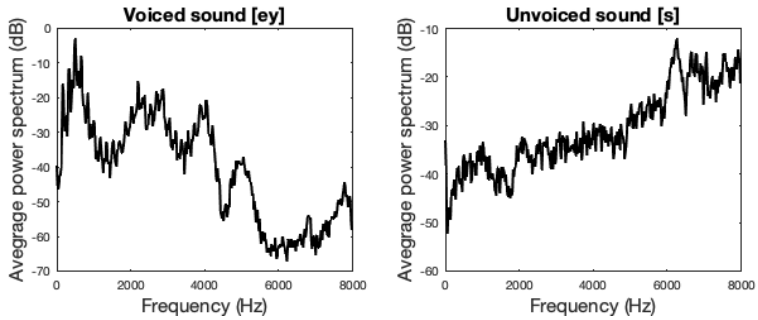


Figure 2.3. From left to right the average power spectra in dB for a) the voiced sound [ey] and b) unvoiced sound [s].

gram as shown in Figure 2.1. The fundamental frequency (~280Hz for this speaker) and its harmonics (at multiples of the fundamental frequency) are clearly visible. The formants can be seen as bands of high energy. For example the vowel [ey] (separately marked in the waveform) has the first and second formants at ~520 Hz and ~2240 Hz, respectively. The positions of the formants largely determine the phoneme the vowel is associated with. The unvoiced consonant [s] is also highlighted, where neither the harmonic structure nor the formant patterns are visible due to lack of voicing. Instead, high-frequency friction noise which is characteristic to the sound can be seen at higher frequencies. The frequency characteristics of the two sounds can be more clearly observed from the *average spectra* calculated as a mean over time of the spectrogram for the sounds [ey] and [s], as shown in Figure 2.3.

Speech production also differs in terms of different speaking styles. For instance, a number of speaking styles can be placed on a continuum of vocal effort needed to produce the speech, ranging from *whispered* speech through *normal* and *Lombard* speech (Lombard, 1911; Lane and Tranel, 1971) to *shouted* speech. Whispered speech is produced when the vocal folds do not achieve a complete closure, and thus the produced speech is lower in energy and has an aperiodic component instead of regular voicing. Normal speech is produced when the vocal folds oscillate in an efficient manner. Lombard speech is the style of speech produced spontaneously by humans under noisy acoustic conditions as a result of a modification of vocal effort. Increased vocal effort in the production of loud speech signals, such as Lombard speech, is brought about by the speaker using increased lung pressure, the raising of the larynx and increased tension in the vocal folds (Isshiki, 1964; Ladefoged and McKinney, 1963; Hertegård et al., 1995; Alku et al., 2006). These phenomena increase the fundamental frequency (Lu and Cooke, 2009a; Summers et al., 1988) and the sound intensity (Dreher and O'Neill, 1957; Summers et al., 1988) of the produced speech signal. The length of the vocal tract is also shortened, which raises the formants and further increases the concentration of energy at higher frequencies

and flattens the spectral tilt compared to normal speech (Hansen and Varadarajan, 2009; Lu and Cooke, 2009a; Summers et al., 1988; Tartter et al., 1993). In addition, the duration of the phonetic segments (Dreher and O'Neill, 1957; Hansen and Varadarajan, 2009; Summers et al., 1988; Tartter et al., 1993) in Lombard speech are also varied. The increase in intelligibility in noisy surroundings obtained by Lombard speech is due to a combination of several of these factors, especially the increase in vocal intensity and the decrease in spectral tilt (Cooke and Lu, 2010; Cooke et al., 2014; Lu and Cooke, 2009a). However, other effects of Lombard speech, such as the increase in F0 and the lengthening of phone durations have been shown not to increase speech intelligibility (Lu and Cooke, 2009a,b; Cooke et al., 2014). The increase of F0 in the production of loud speech (such as Lombard speech), for example, has been interpreted to be a secondary effect caused by increasing lung pressure to raise the vocal intensity, therefore the raising of F0 per se does not improve speech intelligibility in noise (Alku et al., 2002; Gramming et al., 1988; Lu and Cooke, 2009a,b). Shouted speech can be considered the ultimate endpoint of the vocal effort continuum and involves a much larger increase in the overall energy of the speech signal (Rostolland, 1985; Raitio et al., 2013). Shouted speech has decreased intelligibility (Pickett, 1956; Rostolland, 1985). Speaking styles may also be placed on a continuum based on the degree of articulation (Lindblom, 1990) as *clear* speech (Bradlow and Bent, 2002; Baker and Hazan, 2009; Grynpas et al., 2011; Granlund et al., 2012; Hazan et al., 2018) and *casual* speech (Manuel, 1992; Dilley et al., 2013). Clear speech refers to a speaking style which is characterised by a slower speaking rate and it is used in order to increase intelligibility, for example, when communicating with hearing-impaired listeners (Bradlow and Bent, 2002). Casual speech refers to a speaking style that is used in normal conversations and is similar to normal speech from the point of view of vocal effort.

2.2.1 The source-filter model

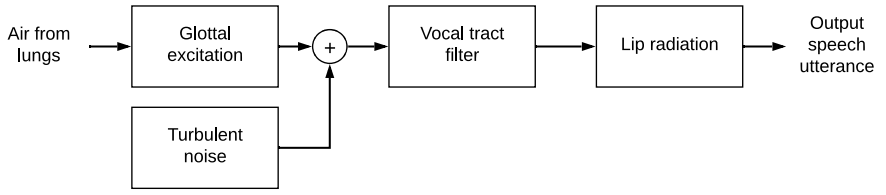


Figure 2.4. Schematic of the source-filter model of human speech production.

The human speech production system can be mathematically approxi-

mated by a simple model called the source-filter model that consists of three independent components. This linear model can be simply represented in the z-transform domain as

$$P(z) = U(z)T(z)R(z) \tag{2.2}$$

where $P(z)$, $U(z)$, $T(z)$ and $R(z)$ are the z-transforms of the radiated speech pressure waveform, the glottal flow excitation generated by the vocal folds, the transfer function of the vocal tract filter, and the radiation characteristics at the lips, respectively (Fant, 1970). A basic schematic of the source-filter model is shown in Figure 2.4. The excitation block is a combination of the glottal excitation, which models excitation of the voiced segments, and a turbulent noise block that models the unvoiced segments. The glottal excitation is a quasi-periodic excitation that is a sequence of air volume velocity pulses in time. At the most basic level, it can be modelled by a pulse train at the fundamental frequency. The turbulent noise block is responsible for the aperiodic noise excitation characteristic of the unvoiced segments and is modelled either as stationary turbulent noise or as sudden noise bursts depending on the type of unvoiced segment they model. The vocal fold excitation block usually has a spectral tilt of ~−12 dB/octave. The vocal tract filter frequency response is dependent on the shape and obstructions present in the vocal tract, giving rise to formant resonances at certain frequencies. The lip radiation acts as a high-pass filter with the spectral tilt of ~6 dB/octave. Hence, speech on average has a spectral tilt of ~−6 dB/octave. In reality, the glottal excitation and the vocal tract are not independent, nor is the vocal tract lossless. The source-filter model, however, works very well for many applications such as speech synthesis, speech coding and speech analysis, and provides a useful starting point for more advanced models. For example it is fundamental to several vocoders (see Section 2.3.2).

2.3 Principles of human auditory perception and its modelling

The human auditory perception system is responsible for decoding the different aspects of information present in the sounds in the environment of the listener. The peripheral human auditory system is shown in Figure 2.5. It consists of the outer, middle and inner-ear, and of the auditory nervous system that connects the input from the two ears together and with other parts of the central nervous system.

Air pressure waves carrying the auditory information are first collected and funnelled into the ear canal with the help of the pinna. This air pressure wave excites the eardrum (also referred to as the tympanic membrane). The eardrum transfers this excitation to the cochlea via three bones, the malleus, incus and stapes. The tympanic cavity is a small cavity

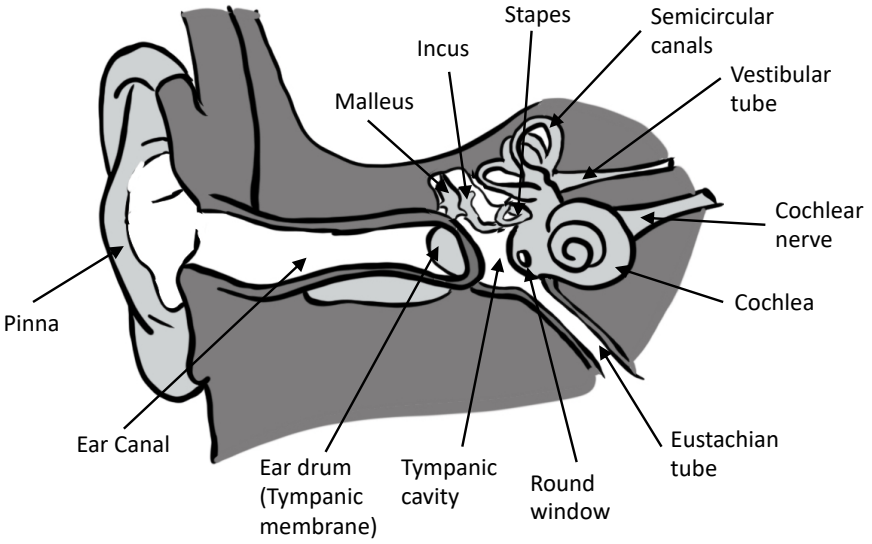


Figure 2.5. Illustration of the peripheral human auditory perception system. Adapted from Chittka and Brockmann, 2005.

that houses these three bones. The eustachian tube connects the tympanic cavity to the upper throat and the back of the nasal cavity equalising the air pressure within the middle ear and outside the body. The cochlea is a coiled tube containing a liquid membrane that is housed in a hard outer shell (Schnupp et al., 2011). The basilar membrane (BM), which is located inside the cochlea, is of gradually decreasing stiffness going from the rounded coils to the far end of the cochlea. This allows for different locations of the BM to vibrate in response to different frequencies of mechanical excitation from the stapes. Hair cells positioned atop the BM transfer the vibrations from different locations on the BM to mechanosensing organelles called the stereocilia. The stereocilia transform the mechanical vibrations into electrical signals. These electrical signals pass through the auditory nervous system where various properties of the acoustic signal are extracted before they are processed by the cerebral cortex in the brain (Moore, 2012).

Humans can perceive sounds that are 20 Hz–20 kHz in frequency. Human speech is approximately in the range of 85 Hz to 8 kHz. Because of the structure of the BM and the cochlea that it is housed in, the sensitivity of the human auditory perception system to different speech sounds is non-linearly dependent on its frequency. The perceived "height" of a tone—a sound consisting of only one sinusoidal component at a specific frequency—is referred to as *pitch*. The *mel* scale (Stevens et al., 1937) is one measure of pitch on which equidistant tones are also perceptually at an equal distance from each other. Mathematically, the relationship between the frequency (Hz) f_p of an acoustic wave and the perceived pitch (mel) m can

be approximated as

$$m = 2595 \log_{10}(1 + \frac{f_p}{700}) \tag{2.3}$$

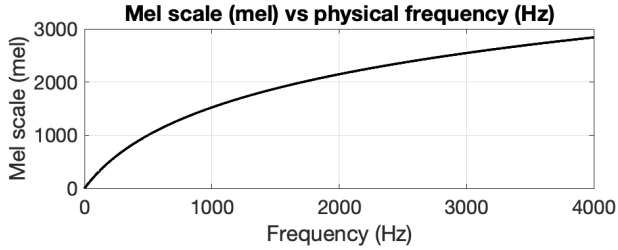


Figure 2.6. Plot comparing the physical frequency of an audio signal (Hz) and the perceptual mel scale (mel) as defined by Equation 2.3 (O'Shaughnessy, 1987).

An illustration of the mel-scale vs the physical frequency is shown in Figure 2.6. In addition to the logarithmic perception of tone heights, the hearing system's frequency selectivity—the capability to distinguish close-frequency sounds from each other—differs as a function of the absolute frequency of the sounds. Measures attempting to capture this phenomenon include the *Bark scale* (Rossing et al., 2002) which is based on dividing the frequency scale into 24 critical bands (Zwicker, 1961) with approximately logarithmically increasing bandwidth as a function of the centre frequency at above 500 Hz. Within each critical band, concurrent sounds (e.g., tones) either merge into a unified percept of the sounds or louder sounds may mask the less energetic sounds. Another similar perceptive scale is the *equivalent rectangular bandwidth* (ERB, Moore and Glasberg, 1983; Glasberg and Moore, 1990) scale, which is also based on the concept of critical bands, but uses another measurement principle to derive the bandwidths of the critical bands.

2.3.1 Mel-frequency spectrum

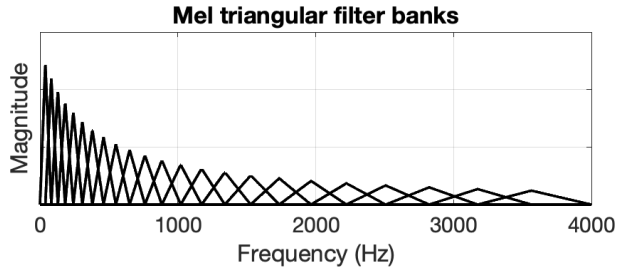


Figure 2.7. Plot showing the 24 triangular filters of the mel filter bank.

The perceptual sensitivity of the human auditory system varies non-linearly with frequency. It is hence useful to define a separate time-

frequency scale where the frequency representation varies more closely with the frequency selectivity of the human auditory system. A commonly used, perceptually motivated, time-frequency feature space based on the ERB scale is the mel-spectrum. The mel-spectrum is obtained by calculating the amount of energy present in a series of sub-bands called the mel filter bank. The mel filter bank is typically represented in the frequency domain using a series of overlapping, triangular filters as shown in Figure 2.7. The mel-spectrum (or its logarithm) has several commonly used extensions based on different methods of compressing the filter outputs, such as the mel-frequency cepstral coefficients (MFCC, Davis and Mermelstein, 1980) or the mel-generalised cepstrum (MGC, Tokuda et al., 1994).

The mel-spectrum and its extensions are used in Publications I and III–VII to represent the time-frequency structure of speech before further processing. It is used in Publication I as a starting point for discovering fixed-dimensional time-frequency representations of syllable-rhythmic units from running speech. In Publications III and IV, the mel frequency spectra are used as the features the machine learning models (MLMs, see Section 3) operate on for automatic syllabification. In Publications V–VII, MGCs are used to represent the spectral envelope of speech in different speaking styles as a part of several vocoder architectures (see Section 2.3.2). These representations, among a number of other parameters describing the speech, are then subjected to a mapping between speaking styles.

2.3.2 Vocoders

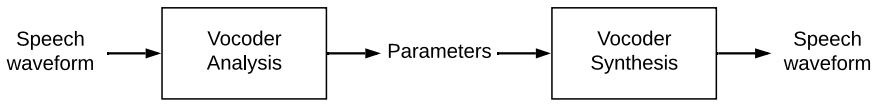


Figure 2.8. Schematic of the analysis and synthesis blocks of a vocoder.

A vocoder is a key component in several areas of speech technology such as text-to-speech (Zen et al., 2009), voice conversion (Stylianou, 2009) and style conversion (see also Section 6). The main task of the vocoder is to represent speech as a set of parameters in the analysis stage, and to synthesise the speech waveform back from the parameter set in the synthesis stage. Vocoders generate a more compact representation of the speech signal compared to the original time-frequency content of the speech signal (c.f., e.g., complex spectrogram). In addition, vocoders typically use speech signal representations that are easy for humans to interpret and also enable the manipulation of desired properties of the speech signal without affecting the others. The basic schematic of a vocoder is shown in Figure 2.8.

Several vocoders use signal processing (DSP) techniques for the design of

the analysis and synthesis stages. Most classical vocoders are based on the source-filter model (see Section 2.2.1). Examples of these vocoders include glottal vocoders, which use waveforms computed from natural speech as the excitation (e.g., GlottHMM (Raitio et al., 2010) and GlottDNN (Airaksinen et al., 2016)). Mixed/impulse excited vocoders, such as STRAIGHT (Kawahara et al., 1999) and WORLD (Morise et al., 2016), use simple impulse trains or mixtures of impulse trains and noise as the excitation. In addition to the differences in the generation of the excitation, source-filter vocoders might also differ in terms of how the filter of the source-filter model is represented (e.g., line spectral frequencies in GlottHMM vs. mel-generalised cepstrum in STRAIGHT). On the other hand, sinusoidal vocoders (e.g., quasiharmonic model (Erro et al., 2013) and dynamic sinusoidal model (Hu et al., 2014)) represent speech as a sum of sinusoidal functions evolving over time. Some vocoders use the source-filter model but use sinusoidal signal analysis as a measure of harmonicity (e.g., pulse model in log domain, PML (Degottex et al., 2016)).

Particularly in text-to-speech synthesis, there is increasing interest in vocoders called neural vocoders. These vocoders do not use a DSP-oriented computation of the parameters of a simple speech signal model, but instead take advantage of MLMs (typically NNs) to train a network to represent the signal. These models typically produce better quality speech synthesis, as they are not restricted by the assumptions (such as the source-filter model) of the classical, DSP-motivated vocoders described in the previous paragraph. Neural vocoders, however, have a much higher data requirement in training. The first neural vocoders were based on auto-regressive MLMs, whose synthesis modules generate the speech signal sample by sample (e.g., WaveNet (van den Oord et al., 2016) and WaveRNN (Kalchbrenner et al., 2018)). Auto-regressive neural vocoders suffer from slow signal generation. More recent research on neural vocoders has focused on speeding up the generation time, resulting in new neural vocoders such as parallel WaveNet (van den Oord et al., 2018), WaveGlow (Prenger et al., 2019) and WaveGrad (Chen et al., 2020). Finally, hybrid vocoders use both DSP-based assumptions as well as MLMs to get the best of both approaches (Valin and Skoglund, 2019; Wang et al., 2019).

The GlottDNN, STRAIGHT and the PML vocoders are compared in Publication V of this thesis for the task of speaking style conversion (SSC) between normal and Lombard speaking styles. Given the findings of Publication VI, the PML vocoder was chosen for use in Publications VI and VII that focus on further developments of speaking style conversion.

3. Machine learning methods

Machine learning is a field of technology that deals with finding recurring patterns in data that are generalisable to new unseen data. These recurring patterns are represented by mathematical models called machine learning models (MLMs) that contain a set of parameters that can be tuned. *Training* is the process by which an MLM *learns* to find these recurring patterns by adjusting the parameters based on the input data. Training is usually done in such a way that the model parameters minimise some appropriately chosen *cost function* using a chosen training algorithm. The parameters of the MLM that cannot be optimised by the training algorithm are called the *hyper-parameters* of the MLM. These hyper-parameters have to be tuned separately by using *model selection* (Bishop, 2006). This can be done, for example, by checking the fit of the model on a separate held-out set not used by the training algorithm, referred to as the *validation set* (Bishop, 2006), or by using some other metric such as the Bayesian information criterion (Schwarz et al., 1978; Bishop, 2006). Once an MLM has been trained, it can be used to make predictions on new unseen data by *inference*. *Probabilistic MLMs* are MLMs whose predictions have a probabilistic interpretation.

Listed below are some of the categories of MLMs that are referenced throughout this thesis (this is by no means exhaustive of the topics covered by ML as a whole.)

- *Supervised MLMs* - Supervised MLMs are predictive models that learn a mapping function from inputs to certain outputs, both of which are explicitly available in the training data (examples of such methods include, support vector machines (SVMs, Cortes and Vapnik, 1995) and feed-forward neural networks (NNs)).

- *Unsupervised MLMs* - Unsupervised MLMs, on the other hand, aim to model the inherent structure or distribution of the training data. They are used in situations where the training data does not have explicit output labels or targets, but where it would be desirable to understand the structure of the data, such as in cluster analysis,

data visualisation or where a model is pre-trained for later use in supervised training with fewer labelled samples (examples include K-means clustering (Lloyd, 1982; Bishop, 2006) and Gaussian mixture models (GMMs, McLachlan and Basford, 1988; Bishop, 2006)).

- *Semi-supervised MLMs* - Annotating training data with appropriate output labels can often be expensive or cumbersome. However, large pools of unannotated data are often readily available. Semi-supervised MLMs draw from concepts of both supervised and unsupervised MLMs and are designed to handle scenarios where there is access to a pool of data, where only a small subset of the data is annotated. A typical semi-supervised method attempts to learn an initial model from the labelled data and then uses its own label inferences on the unlabelled data to further train the model.

This thesis aims to solve a variety of supervised and unsupervised problems in speech technology. This section provides the theoretic background for the MLMs used in the thesis. It is organised as follows: Section 3.1 discusses and compares the modelling options and inference methodologies of probabilistic MLMs with Bayesian and frequentist models. The GMM which is used in Publications I, II and VI, as an example illustrates these differences between frequentist and Bayesian modelling, as well as the different Bayesian modelling choices. Section 3.2 gives a brief overview of artificial neural networks (NNs), specifically detailing the feed-forward NN (used in Publication VI), recurrent neural networks (RNNs; used in Publications III, IV and V), convolutional neural networks (CNNs; used in Publications IV and V), end-to-end learning (used in Publications IV and V) and generative adversarial networks (GANs; used in Publications VI and VII).

3.1 Bayesian modelling

Probabilistic MLMs can be broadly categorised based on two major underlying philosophies - the *Frequentist* approach and the *Bayesian* approach. These philosophies primarily differ in their definitions of probabilities. Frequentist modelling assumes that probabilities make sense only in the context of repeatable events whose frequencies can be recorded, for example, the tossing of a coin or the roll of a die. In the Bayesian interpretation, probabilities can more loosely be associated with events that are not necessarily repeatable, for example, the probability of a certain politician becoming the president or the probability of a certain team winning the World Cup.

More formally, let us assume a given dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ such that $\{\mathbf{x}_n\}_N^{n=1} \sim \mathcal{P}_{\text{Data}}$ with dimensionality d is modelled with a frequentist model

parameterised by $\boldsymbol{\theta}$, which are assumed to be fixed-valued unknowns. The goal of frequentist modelling is to find the best estimate of these parameter values $\boldsymbol{\theta}$ so that the likelihood of the data is maximised. This is called the maximum likelihood (MLE) estimate of $\boldsymbol{\theta}$, and can be formulated as

$$\boldsymbol{\theta}^*_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X};\boldsymbol{\theta}) \tag{3.1}$$

Let us now consider that the dataset $\mathbf{X}$ is modelled by a Bayesian model. In this case, the model parameters $\boldsymbol{\theta}$ are considered to be random variables, and their distributions are governed by Bayes' rule (Gelman et al., 2013; Bishop, 2006):

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} \tag{3.2}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the parameters, representing our belief of the potential values of the parameters before any data is observed and $p(X|\theta)$ is once again the likelihood of the data[1]. $p(X)$ is the evidence of the data that can be represented as the marginal over the different values of $\boldsymbol{\theta}$ as $p(\mathbf{X}) = \int p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. $p(\boldsymbol{\theta}|\mathbf{X})$ is the posterior distribution of the parameters representing our updated belief of the distribution of the parameters after the data $\mathbf{X}$ has been observed under the prior model. Learning in the Bayesian setting corresponds to finding the set of parameter values that have the highest probability value in the posterior distribution, called the maximum a posteriori (MAP) estimate, defined as

$$\begin{aligned}\boldsymbol{\theta}^*_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) \\ &= \arg\max_{\boldsymbol{\theta}} \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}\end{aligned} \tag{3.3}$$

Let us consider the Gaussian mixture model (GMM) as an example to better understand the frequentist and Bayesian modelling perspectives in practice. A GMM is a generative model consisting of a collection of K different Gaussians that can be used to model arbitrary-shaped uni- or multivariate data distributions. The kth Gaussian ($k \in [1,K]$) has two types of parameters: the mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$, that define the position and shape of the Gaussian in the input space. In addition, each Gaussian has a weight ζ_k, such that $\sum_{k=1}^{K} \zeta_k = 1$, which describes the relative contribution of that Gaussian to the overall mixture of distributions. In the frequentist interpretation of the GMM (referred to here as the standard GMM or SGMM), the parameters are considered to be fixed values, and

[1]Note that the known variables in the probability distribution are represented here to the right of the '|' symbol, representing that they are also random variables with a particular value, whereas for the previous case of frequentist modelling in Equation 3.1 they were represented by a ';' signifying that they are constant.

the likelihood can be written as

$$p(\mathbf{x};\boldsymbol{\zeta},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{k=1}^{K} \zeta_k \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$
$$\text{where, } \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{3.4}$$

In the Bayesian interpretation, the parameters are considered as random variables and hence have prior distributions. The Bayesian extension of the GMM is referred to as the Bayesian GMM (BGMM) in this thesis. We choose the prior distributions over the parameters of the GMMs to be *conjugate priors* (Bishop, 2006; Murphy, 2012). This will ensure that the posterior distributions will be from the same family of distributions as the priors, and hence make analytical inference easier (see Section 3.1.2 for further details). In practice, of course, any prior distribution can be chosen. In our example of a BGMM, we could set the prior of the weights ζ_k to a Dirichlet distribution (DD) of dimensionality K as $\boldsymbol{\zeta} \sim \text{DD}(\boldsymbol{\alpha}_0)$. This will ensure that the posteriors are also from a DD. The prior distribution over the weights can then be written as

$$p(\boldsymbol{\zeta};\boldsymbol{\alpha}_0) = \text{DD}(\boldsymbol{\zeta};\boldsymbol{\alpha}_0)$$
$$\text{where, DD}(\zeta;\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2).....\Gamma(\alpha_K)} \prod_{k=1}^{K} \zeta_k^{\alpha_k - 1} \tag{3.5}$$

where $\boldsymbol{\alpha}_0 = [\alpha_1, \alpha_2, ..., \alpha_K]$ is the hyper-parameter of the DD. Intuitively, the higher the value of α_k, the greater is the probability that the weight over the kth Gaussian ζ_k is higher. Similarly the priors of the means, $\boldsymbol{\mu}_k$, and covariances, $\boldsymbol{\Sigma}_k$ of Gaussians are chosen to be conjugate priors of a Gaussian distribution, which is the Normal-inverse Wishart (NIW) distribution, denoted as $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\} \sim \text{NIW}(\mathbf{m}_0, \beta_0, \mathbf{V}_0, \nu_0)$.

$$p(\boldsymbol{\Sigma};\mathbf{W}_0,\nu_0) = \mathcal{W}^{-1}(\boldsymbol{\Sigma};\mathbf{V}_0,\nu_0)$$
$$p(\boldsymbol{\mu}|\boldsymbol{\Sigma};\mathbf{m}_0,\beta_0) = \mathcal{N}(\boldsymbol{\mu};\mathbf{m}_0,\frac{\boldsymbol{\Sigma}}{\beta_0}) \tag{3.6}$$
$$\text{where,} \mathcal{W}^{-1}(\boldsymbol{\Sigma};\mathbf{V}_0,\nu_0) = \frac{\mathbf{V}_0^{\frac{\nu_0}{2}}}{2^{\nu_0\frac{d}{2}}\Gamma_d(\frac{\nu_0}{2})}|\boldsymbol{\Sigma}|^{\frac{-(\nu_0+d+1)}{2}} e^{-\frac{1}{2}\mathbf{tr}(\mathbf{V}_0)\boldsymbol{\Sigma}^{-1}}$$

where $\mathbf{m}_0$, β_0, $\mathbf{V}_0$ and ν_0 are the hyper-parameters of the NIW distribution (see Bishop, 2006; Murphy, 2012, 2007 for further details). Intuitively $\mathbf{m}_0$ represents our belief of the expected value of the means $\boldsymbol{\mu}_k$ of each of the k Gaussians, β_0 represents how strongly we believe the prior $\mathbf{m}_0$, $\mathbf{V}_0$ is proportional to the expected value of the covariances of each of the k Gaussians, $\boldsymbol{\Sigma}_k$, and ν_0 controls how strongly we believe the prior $\mathbf{V}_0$ (Murphy, 2012).

We can now write the likelihood very similarly to Equation 3.4, with the only difference being that the parameters are represented as random variables.

$$p(\mathbf{x}|\boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \zeta_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \tag{3.7}$$

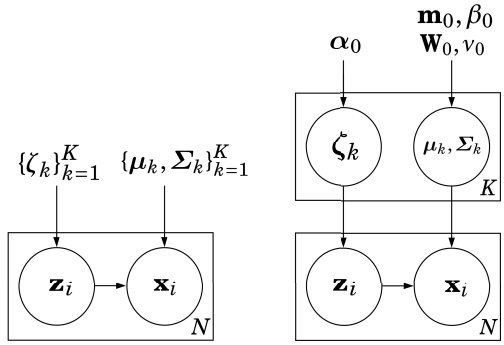


Figure 3.1. Probabilistic graphs showing from left to right, the generative process for (a) the SGMM and (b) the BGMM respectively. The random variables are shown in circular blocks and plates represent repeated variables.

In order to simplify the mathematical treatment of the generative process of both the SGMM and BGMM, we introduce a 1-of-K binary latent variable $\mathbf{z}$ that encodes the identity of the Gaussian that each of the data $\mathbf{x}$ is associated with, as shown in Figure 3.1. See Section 3.1.2 to see how $\mathbf{z}$ also helps with the inference process by defining the *responsibilities*, $p(z_k = 1|\mathbf{x}_n)$, of each Gaussian component in explaining each data point $\mathbf{x}$. We can write

$$\begin{aligned} p(\mathbf{z}) &= \prod_{k=1}^{K} \zeta_k^{z_k} \\ p(\mathbf{x}|\mathbf{z}) &= \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \end{aligned} \tag{3.8}$$

The generative process from both the SGMM and BGMM can now be represented as shown in Figure 3.1. Section 3.1.1 details the parametric and non-parametric modelling options for Bayesian models. Section 3.1.2 details the inference methodologies used in Bayesian modelling and contrasts them to the frequentist approach continuing with the GMM as an example.

3.1.1 Parametric and non-parametric modelling

Bayesian models can be categorised as parametric and non-parametric. *Parametric Bayesian models* are those whose parameter space is bounded to be finite dimensional. The example mentioned earlier of a BGMM

with a DD prior (DDGMM) on the weights is an example of a parametric Bayesian model as the number of parameters with regards to the weights is fixed to K, the number of Gaussians in the model. Typically, model selection (as mentioned at the beginning of Section 3), involves comparing different variants of the model while varying the hyper-parameters. In the training of a parametric Bayes model, an important characteristic of the MLM that has to be varied during model selection is its complexity, that is, the number of parameters. For instance, in the case of the DDGMM, model selection needs to be performed not just on the hyper-parameters of the model, such as $\boldsymbol{\alpha}_0$, $\mathbf{m}_0$, β_0, $\mathbf{V}_0$ and ν_0 for the DDGMM, but also for the number of Gaussian components, K. This often makes the training of parametric Bayesian models cumbersome. Moreover, it also makes parametric Bayesian models difficult to adapt to new data after their initial optimisation, as one would have to re-estimate the ideal model complexity. There are also many scenarios, as in the case of training unsupervised MLMs, where it may be difficult to come up with appropriate criteria for model selection.

Non-parametric Bayesian (NPB) models are a class of Bayesian models (Hjort et al., 2010) that are able to vary their complexity depending on that of the dataset. These are models that have an infinite number of parameters, that is, the prior distribution of the parameters lies on an infinite dimensional space. The main advantage of these models is that they theoretically eliminate the need for model selection for the hyper-parameters controlling the model complexity. One way of extending our example of the BGMM model to the non-parametric case is by using a special class of priors referred to as the Dirichlet Process (DP, Antoniak, 1974; Gershman and Blei, 2012; Ferguson, 1973), which is a non-parametric extension of the DD. A DP is uniquely defined by a distribution H on the model parameter values $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and a positive scalar called the concentration hyperparameter γ, the distribution being denoted as $G \sim \mathrm{DP}(\gamma, H)$. The weights $\{\zeta_k\}_{k=1}^{\infty}$ on the Gaussians decrease exponentially and their generation can be defined by the stick-breaking process (Sethuraman, 1994) (see also Chinese restaurant process (Aldous, 1985; Pitman, 1995; Pitman et al., 2002) or Pólya urn scheme (Hoppe, 1984)). In the stick-breaking process, we assume to have a unit-length stick from which we break off pieces one at a time. The length of the stick broken off at time k represents the weight on the kth Gaussian component, ζ_k, and hence the weights of the new Gaussians follow exponential decay. The stick-breaking process-based construction of

the DP weights can be formulated as shown below

$$\begin{aligned} \zeta_k &= v_k \prod_{i=1}^{k-1}(1-v_i) \\ \text{where, } v_k &\sim \text{Beta}(1,\gamma) \\ \text{with, } \text{Beta}(v_k;1,\gamma) &= (1-v_k)^{\gamma-1}\frac{\Gamma(1+\gamma)}{\Gamma(1)+\Gamma(\gamma)} \end{aligned} \tag{3.9}$$

where v_k is the proportion of the stick broken off at time k and $\prod_{i=1}^{k-1}(1-v_i)$ represent the length of the stick left after the $(k-1)$th break. This process is continued until $k \to \infty$ for the infinite Gaussian components of the DPGMM.

Publication I compares both parametric (DDGMM) and non-parametric (DPGMM and PYPGMM) Bayesian models for the zero resource speech processing (ZS) task. Publication II uses NPB models, that is, the DPGMM and DPVMM (the DP mixture model using the Von-Mises Fischer distribution) for the task of clustering speech data by speaker identity when the number of speakers is unknown. Finally, Publication V uses the parametric Bayesian DDGMM for the task of modelling the mapping functions for the speaking style conversion (SSC) system.

3.1.2 Model inference

In the context of Bayesian learning, model inference is the process of obtaining the best possible set of parameters for a given model structure with a given input dataset. As an example of model inference with frequentist modelling, let us first consider the SGMM. These models can be solved with maximum likelihood as shown in Equation 3.1 as

$$\begin{aligned} \{\boldsymbol{\zeta}^*_{\text{MLE}}, \boldsymbol{\mu}^*_{\text{MLE}}, \boldsymbol{\Sigma}^*_{\text{MLE}}\} &= \arg\max_{\boldsymbol{\theta}} p(\mathbf{X};\boldsymbol{\zeta},\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N}\sum_{k=1}^{K} \zeta_k \mathcal{N}(\mathbf{x}_n;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \end{aligned} \tag{3.10}$$

This leads to the iterative expectation maximisation (EM) algorithm as shown below. The EM-algorithm alternates between the expectation-step (E-step) and the maximisation-step (M-step). The E-step involves the calculation of the *responsibilities* $\kappa(z_{nk})$ which represent the contribution of each of the kth Gaussian in 'explaining' the data point $\mathbf{x}_n$ for the current set of models (Bishop, 2006).

$$\kappa(z_{nk}) = p(z_k = 1|\mathbf{x}_n) = \frac{\zeta_k \mathcal{N}(\mathbf{x}_n;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\zeta_j \mathcal{N}(\mathbf{x}_n;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} \tag{3.11}$$

The M-step uses the current set of responsibilities, $\kappa(z_{nk})$, to update the model parameters, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and ζ_k in order to maximise the likelihood as shown in Equation 3.10. The optimal model parameters can be calculated as

$$\begin{aligned}
\boldsymbol{\mu}^*_{k,\text{MLE}} &= \frac{1}{N_k}\sum_{n=1}^{K}\kappa(z_{nk})\mathbf{x}_n \\
\boldsymbol{\Sigma}^*_{k,\text{MLE}} &= \frac{1}{N_k}\sum_{n=1}^{K}\kappa(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^T \\
\zeta^*_{k,\text{MLE}} &= \frac{N_k}{N}
\end{aligned} \tag{3.12}$$

where N_k is defined as the amount of data that the kth Gaussian is responsible for 'explaining', defined as $N_k = \sum_{n=1}^{N}\kappa(z_{nk})$. Note that this value does not have to be an integer as the responsibilities, $\kappa(z_{nk})$, are probabilistic associations of each data point $\boldsymbol{x}_n$ to the kth Gaussian.

Model inference in the case of Bayesian modelling involves the estimation of the posterior distribution over the parameters of the model given by Equation 3.2. The problem, however, is that the evidence term $p(X)$ in the denominator is often intractable, and hence in most cases the posterior does not have an analytic solution. Sampling-based inference and variational approximation provide two avenues where the posterior can be estimated.

Sampling methods

Generally, the aim of inference is not to obtain the posterior distribution over the parameter, but rather to estimate some value based on the posterior distribution. For example, in Equation 3.3, the goal is to estimate the parameters that maximise the posterior distribution. Sampling-based inference methods aim to approximate a certain function $f()$ over the analytically intractable posterior, $p(\boldsymbol{\theta}|\mathbf{X})$, by drawing a set of independent samples, $\{r_m\}_{m=1}^{M}$, from it as $r_m \sim p(\boldsymbol{\theta}|\mathbf{X})$. Now expectations with the function $f()$ over the posterior may be calculated as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}|\mathbf{X}}[f(\boldsymbol{\theta})] &= \int_{-\infty}^{\infty}[f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta} \\
&\approx \frac{1}{M}\sum_{m=1}^{M}f(r_m)
\end{aligned} \tag{3.13}$$

Sampling-based approaches rely on the fact that we can very easily estimate the probability of a certain data point from the posterior distribution up to a certain normalising constant, even if we do not know its exact analytical form. The simplest sampling-based approaches such as rejection sampling and importance sampling (Gelman et al., 2013; Bishop, 2006) are based on using another distribution called the *proposal distribution* from which samples can be easily drawn. A choice is then made whether to accept or reject that sample based on some acceptance criterion.

An extension of this simple approach is to sample from the proposal distribution in a Markovian process where the current sample depends on the previous sample. This gives rise to Markov chain Monte Carlo (MCMC, Gelman et al., 2013; Bishop, 2006) methods. The samples drawn using MCMC will of course be highly correlated, and hence much of these samples have to be rejected before calculating the approximated expectations as in Equation 3.13. MCMC-based methods, such as the Metropolis Hastings (MH) and Gibbs sampling (which is an extension of MH (Gelman et al., 2013; Bishop, 2006)), have the advantage of working better at drawing samples in high-dimensional distributions than the simpler sampling methods.

The advantage of sampling-based approaches is that we are able to get samples from the true posterior distribution, while the disadvantage is that they are often slow to converge and to draw enough samples M. Moreover, several of the accepted samples have to be rejected if a method like MCMC is used in order to ensure that the samples drawn are independent.

Variational methods

As opposed to sampling-based solutions, variational approximation-based models approximate the analytically intractable posterior $p(\boldsymbol{\theta}|\mathbf{X})$ with a tractable distribution $q_\lambda(\boldsymbol{\theta})$, with hyperparameters λ, so that the Kullback–Leibler divergence (KLD, Kullback and Leibler, 1951) between the true posterior and the approximate distribution is minimised, formally denoted as

$$\begin{aligned} \lambda^*_{\text{Variational}} &= \arg\min_{\boldsymbol{\lambda}} \text{KLD}(q_\lambda(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{X})) \\ \text{where, KLD}(q||p) &= \int_{\boldsymbol{\theta}} q \ \log(\frac{q}{p}) d(\boldsymbol{\theta}) \end{aligned} \tag{3.14}$$

By using Equation 3.2, and the fact that the evidence $p(\mathbf{X})$ is constant for a given dataset $\mathbf{X}$, Equation 3.14 can be shown to be equivalent to

$$\lambda^*_{\text{Variational}} = \arg\max_{\boldsymbol{\lambda}} \mathbb{E}_{q_\lambda}[\log \ p(\mathbf{X}|\boldsymbol{\theta})] + KLD(q_\lambda(\boldsymbol{\theta})||p(\boldsymbol{\theta})) \tag{3.15}$$

where the term on the right to be maximised is the evidence lower bound (ELBO), $\text{ELBO} = \mathbb{E}_{q_\lambda}[\log \ p(\mathbf{X}|\boldsymbol{\theta})] + KLD(q_\lambda(\boldsymbol{\theta})||p(\boldsymbol{\theta}))$, as it represents a lower bound on the evidence, $p(\mathbf{X})$. The first term of the ELBO represents the likelihood of the data over the variational distribution. The second term is a measure of the distance between the variational distribution and the prior, which can be intuitively viewed as a regularisation term to the standard likelihood cost function (Bishop, 2006).

Let us consider the variational inference-based posterior for a BGMM. The posterior $p(\mathbf{z}, \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x})$ is approximated by the distribution $q(\mathbf{z}, \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ which we assume to be factorisable as

$$q(\mathbf{z}, \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = q(\mathbf{z})q(\boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.16}$$

This simple assumption (Bishop, 2006) along with our choice of conjugate priors (see Equation 3.5) will enable us to write the variational posteriors by maximising the ELBO as

$$\begin{aligned} q(\boldsymbol{\zeta}) &= \mathrm{DD}(\boldsymbol{\zeta}|\boldsymbol{\alpha}) \\ q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \mathrm{NIW}(\mathbf{m}_k, \beta_k, \mathbf{V}_k, \nu_k)) \text{ where, } k \in [1, K] \end{aligned} \tag{3.17}$$

where the hyperparameters $\boldsymbol{\alpha}, \{\mathbf{m}_k, \beta_k, \mathbf{V}_k, \nu_k\}_{k=1}^{K}$ can be analytically calculated similar to the EM algorithm in the SGMM inference with an E-step that calculates the responsibilities of each component in explaining the data and an M-step that updates the model parameters based on the responsibilities (see Bishop, 2006 for details). If a non-parametric prior distribution is used, the theoretically infinite number of components K is truncated to a truncation limit T for practical computations (Blei and Jordan, 2006).

The advantage of variational approximation-based methods is that they are generally much faster to compute than sampling-based approaches. The disadvantage is that, since they are based on an approximation of the true posterior, there is a theoretical upper limit to their performance. Bayesian models were used in Publications I, II and V of this thesis, where they are trained using variational inference.

3.2 Artificial neural networks

Artificial neural networks are a class of MLMs that are loosely based on the functioning of neurons in the human brain. In an NN, a neuron is a simple mapping function from a D dimensional input, $\mathbf{x} = [x_1, x_2, ..., x_d]$ to a 1-D output, y.

$$\begin{aligned} y &= \sum_{d=1}^{D} g(x_d w_d + b_d) \\ &= g(\mathbf{x}\mathbf{w} + \mathbf{b}) \end{aligned} \tag{3.18}$$

where $\mathbf{w} = [w_1, w_2, ..., w_d]$ is a d-dimensional weight parameter, $\mathbf{b} = [b_1, b_2, ..., b_d]$ is a bias term, and $g()$ is the activation function. Activation functions are typically non-linear functions, such as the sigmoid, hyperbolic tangent, or rectified linear units (Goodfellow et al., 2016). The schematic of a single neuron of an NN is shown in Figure 3.2.

NNs are structured in general as graphs with layers of neurons through which information propagates from input to output. An NN with just a single hidden layer with a sufficiently large finite number of neurons can be proven to be a universal approximator of any function (Cybenko, 1989; Hornik, 1991; Hornik et al., 1989; Lu et al., 2017). However the number of neurons required could be exponentially larger than the dimensionality of

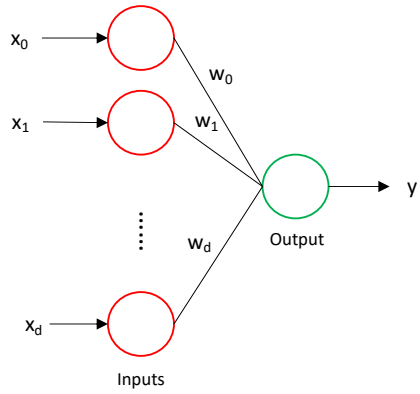


Figure 3.2. Basic structure of a neuron in an NN. The inputs are shown in red and the output in green. The bias term **b** is omitted for clarity.

the input. In addition, the training of a real network might not converge on the correct solution. Hence it is often more efficient to train a network with several layers (with fewer neurons each). Deep neural networks, or DNNs (Goodfellow et al., 2016), are such NN architectures with many layers.

NNs are trained so that their parameters $\mathbf{W}$ are tuned to minimise the *cost function* $\mathcal{L}$:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{t}, \mathbf{y}) \tag{3.19}$$

The cost function is some function of the outputs $\mathbf{y}$ of the NN and its expected outputs $\mathbf{t}$, as defined by the task being solved. Examples of the most commonly used cost function include the L1 and L2 distances (for regression problems, where the targets, $\mathbf{t}$, are continuous variables) and cross entropy (for classification problems, where the targets $\mathbf{t}$ are categorical variables) (Goodfellow et al., 2016).

Another important aspect to training NNs is regularisation, which aims to reduce the effects of overfitting. This is usually an additional term added to the cost function $\mathcal{L}$, such as the L1 or L2 norm of the parameters $\mathbf{W}$. Other methods include modifications of the training algorithm. An example of such a modification is dropout, where random neurons of a given layer are disabled during each forward and backward cycle of the training procedure, forcing the network to find solutions that are not dependent on individual neurons (Hinton et al., 2012; Goodfellow et al., 2016). The primary method of training an NN is *backpropagation* and is explained in detail for the example of feed-forward NNs in the next sub-section. A few of the commonly used NN architectures are also described below.

Feed-forward neural networks

Feed-forward neural networks, also called multi-layer perceptrons (MLPs), are the simplest variant of artificial neural networks. Let us consider an L

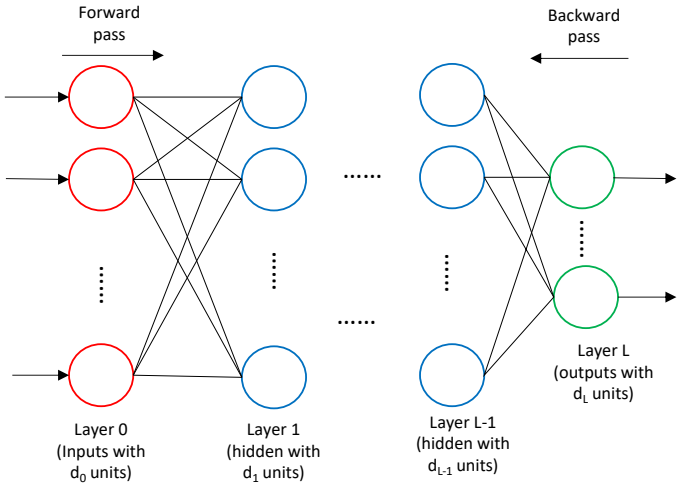


Figure 3.3. Basic structure of an MLP with $L-1$ hidden layers. The input, output and hidden layers are shown in red, green and blue respectively.

layer MLP as shown in Figure 3.3 with d_0 dimensional inputs $\{\mathbf{x}_n\}_{n=1}^N \sim \mathbf{X}$ with corresponding d_L dimensional outputs $\{\mathbf{y}_n\}_{n=1}^N$. Each of the layers has weights $\{\mathbf{W}_l\}_{l=1}^L$ that are $\{d_{l-1} \times d_l\}_{l=1}^L$ dimensional with activation functions $\{g_l\}_{l=1}^{L-1}$. The flow of information from the inputs to the calculation of the output is called the forward pass. During the forward pass, let the outputs of layer $l-1$ be represented as $\boldsymbol{\xi}_{l-1}$. The forward pass from layer $l-1$ to layer l can then be written as follows[2]

$$\begin{aligned} \boldsymbol{\xi}_l &= g_l(\boldsymbol{\xi}_{l-1}\mathbf{W}_l) \\ \text{where, } \boldsymbol{\xi}_0 &= \mathbf{x} \text{ and } \boldsymbol{\xi}_L = \mathbf{y} \end{aligned} \tag{3.20}$$

Training is done using an iterative method called gradient descent (Cauchy, 1847) (or its variations such as stochastic gradient descent (SGD) or mini batch gradient descent (Robbins and Monro, 1951; Kiefer et al., 1952; Goodfellow et al., 2016)) where the weights $\mathbf{W}_l$ are modified at iteration t as follows

$$\mathbf{W}_l^t = \mathbf{W}_l^{t-1} - \phi \frac{\partial \mathcal{L}}{\partial \mathbf{W}_l} \tag{3.21}$$

where ϕ is the learning rate and $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_l}$ is the gradient of the cost function $\mathcal{L}$ with respect to the weights $\mathbf{W}_l$ at layer l. The training is usually continued until there is no longer any significant change in the cost function $\mathcal{L}$. The gradients, $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_l}$, can be calculated efficiently by keeping track of the derivatives with respect to the outputs at that layer $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}_l}$ from Equation 3.20

[2]The bias term has been omitted in the following equations for notational simplicity. In practice, the bias term can be incorporated to the given notation if a constant of 1 is always concatenated to each $\boldsymbol{\xi}_l$.

as

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{W}_l} &= \frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}_l}\frac{\partial \boldsymbol{\xi}_l}{\partial \mathbf{W}_l} \\ &= \frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}_l} g'_l \; \boldsymbol{\xi}_{l-1}\end{aligned} \tag{3.22}$$

where g'_l is the derivative of the activation function. The mechanism of training where the gradient is propagated backwards through the layers of the NN from the outputs to the inputs is called *backpropagation*. Feed-forward NNs were used in Publication V as one of the several mapping function candidates for the speaking style conversion (SSC) system.

Convolutional neural networks (CNN)

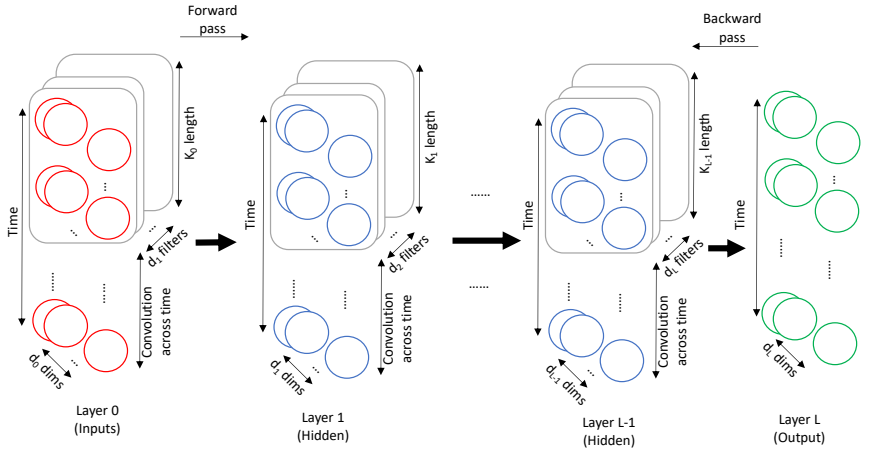


Figure 3.4. Basic structure of a CNN. A convolution operation is applied at each layer with a set of filters. The time axis is shown here vertically. The connections between neurons are not shown for simplicity.

When the inputs to an NN are signals such as speech or images where the same recurring patterns can be expected at different positions of the input, an MLP is not the most efficient NN architecture for modelling. This is because an MLP will have to separately learn to detect the same patterns at each possible position. Creating an MLP that has such a high modelling capacity also has a high risk of overfitting. An ideal architecture would be one where the detection of patterns is translation ("position") invariant. One way to achieve this is to replace the hidden layer of an MLP with a bank of convolution operations as shown in Figure 3.4. The d_0 dimensional input of an MLP is replaced by a time varying $T \times d_0$ dimensional input (note that T could also correspond to the spatial dimensions of an image). The convolution operation is performed across time at layer l with d_{l+1} filters of length K_l. Such an NN is referred to as a convolutional neural network (CNN, LeCun et al., 1998, 1990; Goodfellow et al., 2016).

The *receptive field* of a CNN is the length of input across the temporal or spatial dimension that directly influences the value of any given output frame. This value signifies how much temporal context the NN is able to keep track of in order to make a prediction for a particular frame of the output.

CNNs are one of the major NN architectural units in the novel SylNet architecture developed for the syllable count estimation task in Publication IV. CNNs are also used for the mapping function of the SSC system described in Publications VI and VII.

Recurrent neural networks (RNN)

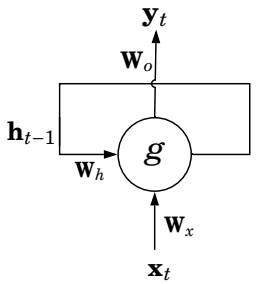


Figure 3.5. Basic structure of an RNN. At time t an RNN takes as input the current frame from the input, $\mathbf{x}_t$, and the previous hidden state value, $\mathbf{h}_{t-1}$.

In applications such as natural language processing and speech processing, there is often a need for the model to be able to retain information over time, such as earlier parts of a sentence or audio snippet given as input. Such a scenario would once again be sub-optimally modelled by an MLP. In principle, this can be achieved by a CNN that has a receptive field up to the required temporal distance. However, once a CNN is trained and its receptive field length is fixed and this property would no longer be satisfied for longer inputs. Another NN architecture that has a theoretically infinite receptive field is the recurrent neural network (RNN, Rumelhart et al., 1986; Goodfellow et al., 2016). The basic schematic of an RNN is shown in Figure 3.5, and can be summarised as

$$\begin{aligned} \mathbf{h}_t &= g(\mathbf{h}_{t-1}\mathbf{W}_h + \mathbf{x}_t\mathbf{W}_x) \\ \mathbf{y}_t &= \mathbf{W}_o\mathbf{h}_t \end{aligned} \tag{3.23}$$

where $\mathbf{W}_h$, $\mathbf{W}_x$ and $\mathbf{W}_o$ are the parameters of the RNN, $\mathbf{h}_t$, $\mathbf{x}_t$ and $\mathbf{y}_t$, the hidden state, inputs, and outputs of the RNN respectively at time t, and g is the activation function.

The basic RNN architecture described above faces a problem during back-propagation-based training, since the gradients coming from the end of the signal decay quickly in time and may therefore become negligible before reaching the earlier parts of the input sequence. This is because the

gradients are calculated as a series of products of gradients (see Equation 3.22) from the end of the network to the beginning. When the gradients are lower than 1, this series of products becomes very close to zero. This problem is referred to as the *vanishing gradient problem*. Because of this issue, RNNs take a very long time to train. A long-short term memory network (LSTM, Hochreiter and Schmidhuber, 1997; Gers et al., 1999; Goodfellow et al., 2016), an extension of the RNN, handles the vanishing gradient problem by having a separate cell state in addition to the hidden state, $\mathbf{h}_t$, through which information runs through time with only small modifications being made with separate gates called the *forget gate* and *add gate* respectively. Other extensions such as gated recurrent unit networks (GRUs, Cho et al., 2014; Chung et al., 2014; Goodfellow et al., 2014) avoid the use of a separate cell state by using a similar gating mechanism as the LSTMs directly on the hidden state, $\mathbf{h}_t$.

Bidirectional LSTMs, a combination of two LSTM layers that operate in both the forward and backward directions, were used for the word count estimation (WCE) task in Publication III. LSTMs are also used in the final layer of the novel SylNet architecture used in the syllable count estimation (SCE) task in Publication IV.

3.2.1 End-to-end models

The traditional approach to solving complicated problems with MLMs would be to split the task into smaller easier-to-model sub-tasks. These sub-tasks are created by leveraging domain-specific expertise and making a set of assumptions between multiple modules of a larger system, such as intermediate phone-level representations used in classical ASR systems. A separate set of training data and training criteria are then required to train each of these sub-tasks. The advantage of this approach is that the individual sub-tasks often do not necessarily require as much data to train as it would take to train the entire task as a whole, or some sub-tasks may potentially have more training data available than others which would be beneficial to utilise fully. Let us consider the example of an ASR system. The task of ASR can be summarised as finding the best sequence of words $\mathbf{j}$, given audio input, $\mathbf{x}$, that maximises the probability $p(\mathbf{j}|\mathbf{x})$ as

$$\begin{aligned} \mathbf{j}^* &= \arg\max_{\mathbf{j}} p(\mathbf{j}|\mathbf{x}) \\ &= \arg\max_{\mathbf{j}} p(\mathbf{x}|\mathbf{j})p(\mathbf{j}) \end{aligned} \tag{3.24}$$

Equation 3.24 can be re-written based on the assumption that each of the words in $\mathbf{j}$ can be broken up into a sequence of sub-word units (such as phones or context dependent units such as tri-phones)

$$\mathbf{j}^* \approx \arg\max_{\mathbf{j},\mathbf{l}} p(\mathbf{x}|\mathbf{l})p(\mathbf{l}|\mathbf{j})p(\mathbf{j}) \tag{3.25}$$

where **l** is the sequence of sub-word units. The simplified pipeline of an ASR system is shown in Figure 3.6.

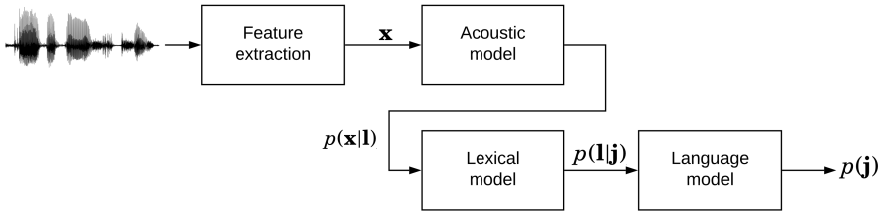


Figure 3.6. A simplified schematic of an ASR pipeline, where the acoustic model, the lexical model (also called the pronunciation model) and language model (also called the grammar model) model the terms $p(\mathbf{x}|\mathbf{l})$, $p(\mathbf{l}|\mathbf{w})$ and $p(\mathbf{w})$ respectively.

The disadvantage of the sub-task approach is mainly the level of domain expertise required. Expertise is required first and foremost to be able to split the complicated task into sub-tasks without a loss of relevant information in the interfacing of the task-specific modules, and to determine the MLMs and cost functions that best solve these sub-tasks. Furthermore, each of the sub-tasks need to be trained independently and hence require training data for each of them, which once again requires expert annotation which is expensive and often very domain-specific. For instance the training data for the lexical model shown in Figure 3.6 should contain the possible sub-word (phoneme) sequences for different pronunciations of all the words in a particular language/dialect, whereas training of the acoustic model would require audio data with aligned phonetic transcriptions.

The back-propagation-based training of DNNs, as well as the access to greater computing resources and larger datasets, gives us an opportunity to model the entire complicated task with a single DNN. For an ASR model, this would correspond to the schematic shown in Figure 3.7, where the acoustic model, lexical model, and the language model are replaced by a single DNN-based end-to-end model. The input to such a model would just be the acoustic waveform (or sometimes spectral features extracted from the waveform) and the output would consist of the most likely string of words, given the input. Such an approach can also potentially lead to better results as no assumptions about the internal processing steps and representations of the system need to be made beyond specifying the overall network architecture. In this context, end-to-end modelling has the benefit of not requiring phone level annotations of speech used to train the model. The disadvantages, however, are the need for much larger datasets, computational cost, and, difficulty to adapt to domains differing from that of the training data, and, naturally, the definition of network architectures that are suitable to solve the problem at hand.

Several end-to-end modelling alternatives were explored for the SCE task in Publication IV, including bidirectional LSTMs and the SylNet.

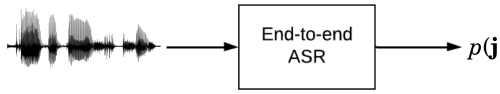


Figure 3.7. Schematic of an end-to-end ASR system.

3.2.2 Generative adversarial networks (GANs)

As explained in the beginning of Section 3, cost functions do the job of judging how well an MLM fits the data during training. For the task of sample generation, the cost function should capture the distance between the training data distribution, $\mathcal{P}_{\text{Data}}$, and the generated data distribution, $\mathcal{P}_{\text{MLM}}$, from the generated samples, $\mathbf{x} \sim \mathcal{P}_{\text{MLM}}$. Explicit cost functions such as the L1, L2, binary cross entropy etc. discussed earlier work well for simple regression and classification tasks. However, these are often too simplistic for the task of sample generation, especially when dealing with complex multi-modal distributions such as speech and images. One simple idea is to replace the explicit cost function with another DNN to judge the quality of the solution, which is exactly what a generative adversarial network (GAN, Goodfellow et al., 2014) does.

The basic structure of a GAN is shown in Figure 3.8. An NN called the generator, G, is fed with samples $\boldsymbol{\omega}$ drawn from a known distribution (such as the standard Gaussian e.g., $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$), which it transforms into samples $\hat{\mathbf{x}}$, as $\hat{\mathbf{x}} = G(\boldsymbol{\omega})$, that should be from the same distribution as that of the true data $\mathbf{x} \sim \mathcal{P}_{\text{Data}}$. Another NN, called the discriminator, D, does the job of classifying the *real*, $\mathbf{x}$ and *fake* samples, $\hat{\mathbf{x}}$.

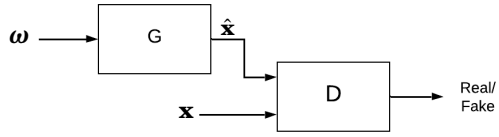


Figure 3.8. Basic schematic of the generative adversarial network (GAN).

The standard GAN (as introduced in Goodfellow et al., 2014) can be trained using the minimax equations as shown below, where the discriminator D has a *sigmoid* unit in the last layer. For a fixed generator G, this corresponds to the Jenson Shannon divergence (JSD, Lin, 1991)— a symmetric extension of the KLD— between the true ($\mathcal{P}_{\text{Data}}$) and generated ($\mathcal{P}_{\text{MLM}}$) data distributions:

$$\begin{aligned} D^*_{\text{opt}} &= \arg\max_D \mathbb{E}_{\mathbf{X}}[\log(D(\mathbf{x}))] + \mathbb{E}_G[\log(1 - D(\hat{\mathbf{x}}))] \\ G^*_{\text{opt}} &= \arg\min_G \mathbb{E}_G[\log(1 - D(\hat{\mathbf{x}}))] \end{aligned} \tag{3.26}$$

The problem with the JSD is that it is constant when the two distributions do not have overlapping support, which is often the case in the early part of

the training. This slows down the training and reduces the performance of GANs. Several studies have therefore explored the use of a variety of other loss functions based on different metrics to measure the distance between the true data $\mathcal{P}_{\text{Data}}$ and generated data $\mathcal{P}_{\text{MLM}}$ distributions. These include approaches such as the least squares GAN (LSGAN, Mao et al., 2017) that minimises the Pearson $\mathcal{X}^2$ divergence (Pearson, 1900) and the Wasserstein GAN (WGAN, Arjovsky et al., 2017) that minimises the Wasserstein distance (Kolouri et al., 2017; Villani, 2003). The WGAN, for instance, requires that the discriminator D is K-Lipschitz continuous (O'Searcoid, 2006). In order to enforce this constraint, different regularisation schemes have been proposed, such as weight clipping (Arjovsky et al., 2017), gradient penalty (Gulrajani et al., 2017), and spectral normalisation (Miyato et al., 2018). The usage of GAN-based models in style conversion is briefly explained below.

Style conversion using GANs

Style conversion aims to train a mapping function for data from one style ("domain") to another. The kind of data available for training has a major impact on the approach used to tackle this problem. *Parallel* training data is data that is available from both the source and target domains with all other modalities of variation being constant. For example, in the task of speaking style conversion from normal to Lombard styles, parallel data would include utterances from both styles spoken by the same person with the same linguistic content and signal quality. This kind of data is relatively tedious to collect. In contrast, *non-parallel* data does not impose this restriction and is hence much easier to come by. A style conversion system can be trained relatively simply using parallel data, but training on non-parallel data requires a more complicated approach (see Erro et al., 2009a for a commonly used approach to training style conversion systems using non-parallel data).

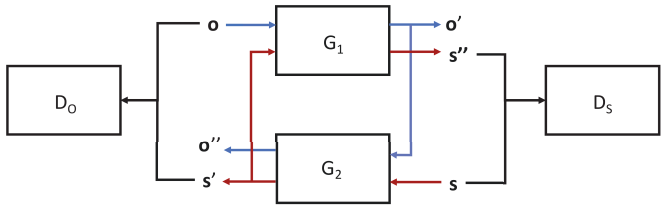


Figure 3.9. Block diagram of a CycleGAN with mapping functions G_1 and G_2 , and discriminators D_O and D_S.The forward cycle, backward cycle, and identity mapping are indicated with red and blue respectively.

GAN-based training can be used for style conversion with non-parallel data by using models called cycle-consistent adversarial networks (CycleGAN, Zhu et al., 2017). Let us consider domains O and S (e.g., utterances in normal and Lombard-style speech, respectively), given non-aligned

training samples $\mathbf{o} \sim p(O)$ and $\mathbf{s} \sim p(S)$. The basic structure of a CycleGAN is shown in Figure 3.9. A CycleGAN contains two mapping functions G_1 and G_2 that map the training data from style domains $O \rightarrow S$ and $S \rightarrow O$ respectively. It also contains two discriminators D_O and D_S, which determine whether the data is from the true distributions $P(O)$ and $P(S)$, respectively. During training, data flows in two directions: the forward cycle $\mathbf{o} \xrightarrow{G_1} \mathbf{s}' \xrightarrow{G_2} \mathbf{o}''$ and the backward cycle $\mathbf{s} \xrightarrow{G_2} \mathbf{o}' \xrightarrow{G_1} \mathbf{s}''$ as indicated by the blue and red arrows, respectively, in Figure 3.9.

The loss function of a CycleGAN consists mainly of two terms. The first one, adversarial loss, measures distance of the mapped data to the true target distribution and can be represented as in Equation 3.26. A cyclic reconstruction loss term is also defined to ensure that data passing through both G_1 and G_2 (or in the backward direction) results in an identity mapping, denoted as

$$\begin{aligned} \mathcal{L}_{cyc}(G_1, G_2, O, S) = & \mathop{\mathbb{E}}_{\mathbf{o} \sim p(O)} [||G_2(G_1(\mathbf{o})) - \mathbf{o}||_1] \\ & + \mathop{\mathbb{E}}_{\mathbf{s} \sim p(S)} [||G_1(G_2(\mathbf{s})) - \mathbf{s}||_1] \end{aligned} \tag{3.27}$$

Other GAN-based architectures such as the DiscoGAN (Kim et al., 2017), DualGAN (Yi et al., 2017) and XGAN (Royer et al., 2020) have also been used for style conversion with non-parallel data. CycleGANs were used for modelling the mapping function in an SSC system in Publication VI. Augmented CycleGANs, another variant of the CycleGAN, simultaneously learn a latent space through which the degree of mapping from one domain to another can be controlled. Augmented CycleGANs are explored in Publication VII.

4. Zero resource speech processing

Human infants do not learn speech and language with annotated datasets. Motivated by this observation, zero-resource speech processing (Zero resource speech processing initiative[1], Glass, 2012; Versteegh et al., 2015; Dunbar et al., 2017, 2019) systems aim to unsupervisedly learn structural representations of speech without access to annotated data. This is a very challenging problem. Firstly, speech signals have enormous acoustic variability, arising from factors other than just the linguistic variation, such as speaker identity, speaking styles, background noises, and various other factors. This means, for example, that the same word never occurs twice in exactly the same acoustic form. Moreover, human infants have access to other modalities of information such as visual, touch etc., and not just the speech inputs (Lerner et al., 2015; Johnson, 2010). Despite this difference, breakthroughs in ZS technology could potentially shed some light on infant language acquisition (Dupoux, 2018; Räsänen, 2012), as well as produce algorithms for low-resource speech processing scenarios (Kamper et al., 2016, 2017). ZS research[1] has so far primarily focused on the unsupervised discovery of linguistic units from raw speech in an unknown language. This task can be addressed from two perspectives of linguistic structure: subword representations and word (spoken term) units, respectively. Another avenue of ZS research has been the so-called text-to-speech without text task (Dunbar et al., 2019). This task explores the capability of discrete unsupervisedly learned subword unit representations in speech synthesis that operates without transcribed training data or input texts. These research tracks are discussed below.

- *Unsupervised sub-word modelling (USM):* This task aims to construct a phonemic representation of speech sounds which support word identification while being robust to within- and between-talker variation. This problem can be approached from a purely signal processing perspective or based on clustering (or some extension) of frame-level speech features using MLMs. Some examples of the latter include,

[1] http://zerospeech.com/

for example, Chen et al., 2015 who used a DPGMM for unsupervised modelling of speech frames. Badino et al., 2015, on the other hand, used binarized autoencoders (Goodfellow et al., 2016) and hidden Markov models (HMMs) to obtain a 1-of-K representation for each frame of speech. Kamper et al., 2015; Renshaw et al., 2015 used a DNN architecture known as the correspondence autoencoder to learn nonlinear mappings from MFCCs to latent distributed feature representations. Thiolliere et al., 2015 used DTW to find clusters of repeating fragments in speech, and a Siamese DNN architecture (Bromley et al., 1994) was then trained to find a vector representation of the speech sounds from the frames of the DTW-aligned fragments.

- *Unsupervised spoken term discovery (UTD):* This task deals with the unsupervised discovery of recurring patterns in speech, corresponding to, for example, syllables, words or phrases. This process can be broken down into three steps – 1) *parsing*, where matching pairs of speech fragments are found on the basis of their global similarity, 2) *clustering* of these speech fragments, thereby building a library of classes with potentially many instances, and, finally, 3) *matching* where the acquired classes are used to parse the speech into candidate tokens and boundaries. UTD has been primarily tackled in existing research using DTW-based approaches (for example Park and Glass, 2006; ten Bosch and Cranen, 2007; Aimetti, 2009; Jansen and Van Durme, 2011) and acoustic word embedding (AWE)-based approaches. The latter approach maps speech fragments into an embedding space where pattern clustering is then performed (for example (Räsänen et al., 2015; Kamper et al., 2017; Kamper et al., 2017)). In addition, there are recent models that attempt to solve both the USM and UTD tasks at the same time (Chen et al., 2019) or in a sequence (Last et al., 2020).

- *Text-to-speech without text:* This task aims to build a speech synthesiser without any text or phonetic labels, but using unsupervisedly discovered discrete sub-word units (similar to the USM task). Other modalities that are necessary for synthesis, but which are not encoded by the the discovered sub-word units, such as speaker identity and speaking style, are fed directly to the synthesis module (see Liu et al., 2019; Tjandra et al., 2019; Karthik and Murthy, 2019; Eloff et al., 2019; Yusuf et al., 2019; Feng et al., 2019).

In this thesis, Publication I addresses the problem of unsupervised discovery of word units using syllable-like representations as an intermediate step. Syllabic units provide a good starting point for ZS systems as the characteristics of these units are largely determined by the rhythmic properties of speech (e.g., Räsänen et al., 2015; Lyzinski et al., 2015). Therefore, they

could potentially be extracted, reliably and unsupervisedly, in a language independent manner.

5. Automatic word and syllable count estimation

Automatic syllable and word count estimation (SCE and WCE) are the technologies of quantifying the amount of linguistic activity present in realistic daylong recordings such as those from a wearable microphone (Ziaei et al., 2013). This technology can be used to investigate vocal activity and social interaction as a function of the recording time and location (Ziaei et al., 2015, 2016). Such methodologies are also useful in the field of child language acquisition, where researchers investigate the language experiences of children using child-centred daylong audio recordings (Bergelson et al.). Such tools are required to analyse large-scale datasets by providing the necessary information to answer questions such as 1) how much speech do children hear in their lives in different contexts (Bergelson et al., 2019), and 2) how does this speech input map to developmental outcomes (Weisleder and Fernald, 2013; Ramírez-Esparza et al., 2014). A closely related problem is speaking rate estimation (Morgan and Fosler-Lussier, 1998; Wang and Narayanan, 2007), where the goal is to estimate the amount of linguistic content per unit time.

The task of WCE and SCE would be trivial if there were an ASR system that was language independent and robust to the heavy noise and cross-speaker talk present in realistic daylong recordings. Since training such high-quality ASR models requires a large quantity of training data that is not possible for all languages, a separate set of techniques needs to be explored for the task of WCE and SCE.

The problem of quantification of linguistic content in speech could be approached from at least two different angles—a segmental approach that looks for phonemic units (cf., language-independent acoustic models) and a suprasegmental approach that focuses on rhythmic units such as syllables. As an example of the segmental approach, the popular LENA^{TM} system[1] uses a pre-trained English ASR model to track the number of vowels and consonants (Gilkerson and Richards, 2009; Xu et al., 2008). The works by Morgan and Fosler-Lussier, 1998; Wang and Narayanan, 2007; Yarra et al.,

[1]The LENA^{TM} system is developed by the LENA research organisation (`http://www.lena.org`).

2016 are examples of the suprasegmental approach that keep track of the rhythmic envelope.

The LENA system has been the go-to option for language researchers as it is designed for infant directed audio, and it includes a compact wearable recorder worn by the child and proprietary software that analyses various aspects of the recorded audio. Apart from WCE, this system also measures other aspects of the recorded audio, including conversational turns and segmentation of adult speech and infant vocalisations. Despite being the previous state of the art, LENA as a software solution has a few issues. Firstly, the software is proprietary and expensive. Secondly, only audio captured with the LENA recorder can be analysed with the software, that is, other audio files cannot be run through the same software. Moreover, the included algorithms are likely to be outdated (the basic building blocks having been introduced nearly 10 years ago e.g., Xu et al., 2008, see also Cristia et al., 2020 for an evaluation of the performance of LENA). Finally, LENA speech processing algorithms, including the WCE module, have been optimised for American English, causing its accuracy for different populations and languages to be inconsistent. Given this background, there is an increasing demand from the research community to develop an alternative to LENA that would be 1) open source and free of charge, 2) compatible with audio data obtained using a variety of recorders, and 3) robustly applicable to a variety of languages.

A collaborative project called Analysing Child Language Experiences Around the World (ACLEW[2]) aims at developing an open-source software package that would address the mentioned shortcomings of LENA. As a part of the work carried out in that project, this thesis explores data-driven syllable-based approaches to the SCE and WCE problems using recent developments in RNNs and end-to-end MLMs. Publication III focuses on the problem of WCE using LSTM-based models that are trained on datasets with phone level annotations. Publication IV introduces a novel end-to-end NN model for the task of SCE.

[2] https://sites.google.com/view/aclewdid/home

6. Speaking style conversion

The paralinguistic conversion of speech is the technology of converting one or more modalities of paralinguistic information in natural speech utterances, such as speaker identity (Stylianou, 2009; Lorenzo-Trueba et al., 2018; Toda et al., 2016) and style. It is important that the linguistic information, signal quality as well as the rest of the paralinguistic modalities of information of the original speech signal remain unmodified in this process. Speaking style conversion (SSC), then, is the technology of converting utterances from one style to another. SSC is related to other areas of speech technology such as statistical parametric speech synthesis (SPSS) (Zen et al., 2009), voice/speaker identity conversion (VC) (Stylianou, 2009), and speech intelligibility enhancement in speech transmission (Loizou, 2013). However, SSC can be considered as a distinct area of research as it differs in one way or another to those mentioned. For instance, there is no linguistic-to-acoustic mapping as there is in SPSS. Intelligibility enhancement in speech transmission (ITU-T, 2003), on the other hand, has strict latency requirements of speech which are not necessarily present in SSC, where offline processing is also possible for several potential use scenarios.

The speaking style of an utterance itself includes several modalities of paralinguistic information including emotion (e.g., Inanoglu and Young, 2009; Erro et al., 2009b; Wang et al., 2012), and vocal effort (e.g., Konno et al., 2016; Meenakshi and Ghosh, 2018 who studied whispered to normal conversion and Nathwani et al., 2017; Calzada and Socoró, 2011; Gentet et al., 2020 who focused on other aspects of vocal effort-based SSC). Certain aspects of style information in speech can be directly converted from the original speech waveform using a mapping function derived from signal processing theory. For example, Nordstrom et al., 2008 used adaptive pre-emphasis linear prediction to transform the speech utterance in terms of its vocal effort and breathiness. Alternatively, in the data-driven approach (e.g., Meenakshi and Ghosh, 2018) the mapping function is a learnt MLM from training data that includes the style variation that we wish to capture. Figure 6.1 shows a basic schematic of an SSC system.

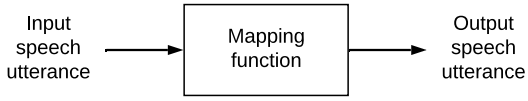


Figure 6.1. Basic schematic of an SSC system. The mapping function could be based on signal processing operations or an MLM trained on data.

This thesis focuses on vocal effort-based SSC, such as the conversion between normal, Lombard, whispered or shouted speech. Vocal effort-based SSC has multiple potential applications where it can be used to personalise speech to the needs of the end-listener. For instance, enhancement and maintenance of speech intelligibility is an important topic in speech technology (see, e.g., Loizou, 2013; Tang et al., 2018; PV et al., 2018). In this context, SSC could be used to adapt the signal in such a way that the signal becomes more intelligible in adverse listening conditions. While there has already been work on whispered-to-normal speech conversion (e.g., Konno et al., 2016; Tao et al., 2010; Meenakshi and Ghosh, 2018; Janke et al., 2014; Morris and Clements, 2002), SSC for other aspects of vocal effort has only been studied in a small number of previous works (Nathwani et al., 2017; Calzada and Socoró, 2011; Huang et al., 2010; Nordstrom et al., 2008; d'Alessandro and Doval, 1998). Another factor to consider for an MLM-based SSC system is the amount of data required to train the mapping function. For speaking styles such as Lombard speech or shouting, the collection of a large quantity of data is laborious and potentially injurious to the health of the speakers. This thesis introduces a parametric speaking style conversion system. In such a system, rather than working directly on the speech signal, a parametric vocoder (see Section 2.3.2) could be used to extract speech features. A subset of these features that are considered to be important to the mapping are then mapped using the mapping function. The vocoder can then be used to synthesise the speech waveform in the target style from the mapped and unmodified features. Figure 6.2 shows a basic schematic of a parametric SSC system. When the mapping function of a parametric SSC system is trained using an MLM the amount of data required is considerably lower.

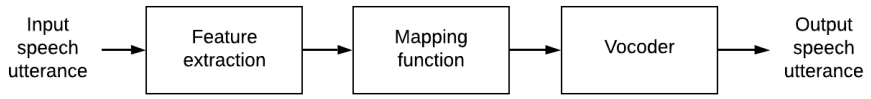


Figure 6.2. Basic schematic of a parametric SSC system.

In this thesis, Publication V explains the parametric SSC systems and compares several vocoders (see Section 2.3.2) and MLMs (see Section 3). Publications VI and VII focus on GAN-based (see Section 3.2.2) solutions for the same task.

7. Summary of publications

7.1 Publication I: "Comparison of non-parametric Bayesian mixture models for syllable clustering and zero-resource speech processing"

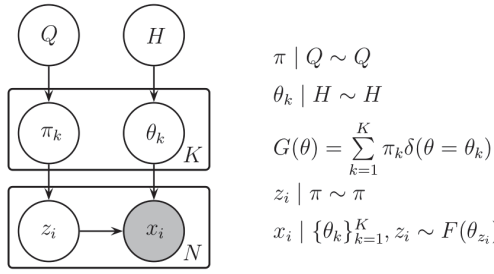


Figure 7.1. The Bayesian graph of the generative process of a BMM G with K mixture components. The kth component, with $k \in [1,K]$, has weights π_k and parameters θ_k that are sampled from the prior distributions Q and H respectively. The mixture component that the observed data x_i, with $i \in [1,N]$, is associated with is represented by the latent variable z_i. Finally, x_i is then sampled from the model $F(\theta_{z_i})$.

Zero-resource speech processing (ZS) systems aim to unsupervisedly learn structural representations of speech to build speech technology applications without labelled training data. This conference article tackles the ZS problem of learning recurring linguistic patterns in speech. One previously proposed approach to tackle this problem has been the extraction and clustering of syllabic units. Simple N-gram models built on these clustered units are then tested against typical units of linguistic content such as words or phonemes. Traditional clustering algorithms require a hyperparameter to be set which is the number of clusters to be learnt, which is highly dataset dependent for the present task. This paper explores the use of non-parametric Bayesian (NPB) methods using Bayesian mixture models (BMMs) to cluster the syllabic units represented as unit-normalised

MFCC-based features. Figure 7.1 shows the Bayesian graph for a BMM. These models are capable of learning the cluster models as well as their number based on the properties of a dataset. The article compares several BMM variants of priors over the weights such as the Dirichlet distribution (DD), the Dirichlet process (DP) and the Pitmann-Yor process (PYP), with the latter two being designed for exponential and power-law-based distributions respectively. Since the distribution of words in different languages follows a Zipfian (power law) distribution (Piantadosi, 2014), and since syllables are expected to follow a similar distribution, a PYP prior is more theoretically motivated than DP for the syllable clustering problem. Also explored are the variants of the cluster component models such as Gaussian and Von-Mises Fischer (VMF), which are designed for Euclidean and unit-nonnmalised data respectively. Since the syllabic features used here are unit-normalised, the VMF prior is theoretically more motivated. These methods are studied using conversational speech from several languages. The models are first evaluated in a separate syllable clustering task and then as a part of a full ZS system. The experiments show that non-parametric methods clustering syllable-rhythmic units perform consistently across several data sets for the ZS task. The PYPVMM gave the best performance among the methods compared, as expected from the theoretical considerations.

7.2 Publication II: "Dirichlet process mixture models for clustering i-vector data"

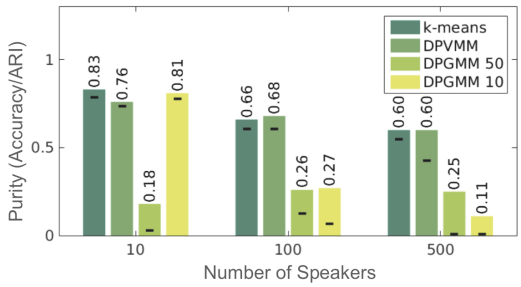


Figure 7.2. Speaker clustering performance for different numbers of speakers based on accuracy (shown as bars) and adjusted rand index (ARI) (shown as -).

The methodological framework of NPB methods used in Publication I was also explored for another task, namely, speaker clustering that involves suprasegmental speech processing in Publication II. This conference article explores the problem of clustering utterances based on the speaker identity when the number of speakers is not known. As in Publication I, the applicability of NPB methods is explored for their ability to implicitly infer the number of clusters. The utterances are represented as i-vectors,

a standard in speaker diarisation/verification—corresponding to a GMM mean supervector mapped into a lower-dimensional factor domain, capturing differences between speakers in the vector direction. These features are typically unit-normalised and are hence best compared using measures such as the cosine distance. This article explores the use of the VMF distribution, which is designed to work with unit-normalised data, as the component distribution of the BMM. It is compared to the much more commonly used Gaussian, which is designed to work with Euclidean data. A DP is used as a prior over the weights of the mixture models. The results in terms of cluster purity are shown in Figure 7.2. It can be seen that the Dirichlet process von Mises-Fisher mixture model (DPVMM) can produce more accurate speaker clusters than the Dirichlet process Gaussian mixture model (DPGMM), demonstrating the importance of choosing the correct distributions and distance metrics for the given data and problem.

7.3 Publication III: "Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech"

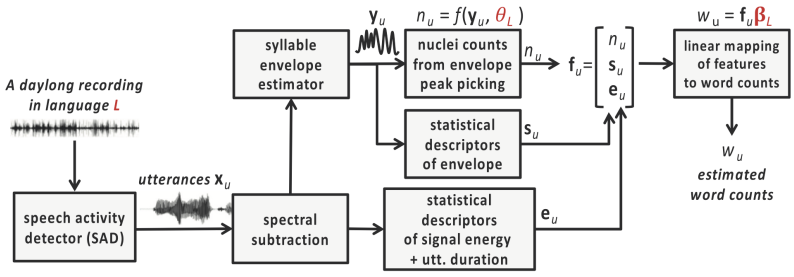


Figure 7.3. Block diagram of the WCE system. Input audio is first passed through a speech activity detector (SAD), followed by speech enhancement. A syllabification algorithm is then used to calculate syllable counts using peak picking from a syllable envelope. These syllable counts along with a number of other statistical descriptors are fed to a linear mapping for an estimate of the word counts.

Automatic word count estimation (WCE) is the technology of quantifying the amount of linguistic content present in realistic audio data in the form of word counts. WCE can be used on datasets such as recordings from wearable sensors to investigate vocal activity and social interaction as a function of the recording time and location. In the context of child language acquisition, WCE is an essential tool for answering questions such as how language exposure varies between families with different socioeconomic and cultural environments, and how this language input maps to later developmental outcomes. This journal article presents an open source WCE pipeline that is based on language-independent syllabification of speech, followed by a language-dependent mapping from several suprasegmental

speech features such as syllable counts to the corresponding word count estimates. A block diagram of the system is shown in Figure 7.3. The previous state-of-the-art solution, the LENA system, was based on proprietary software and has only been optimised for American English, limiting its applicability. In the experiments, the proposed WCE system was shown to have consistent accuracy across multiple corpora and languages and compares well to the LENA system.

7.4 Publication IV: "SylNet: An adaptable end-to-end syllable count estimator for speech"

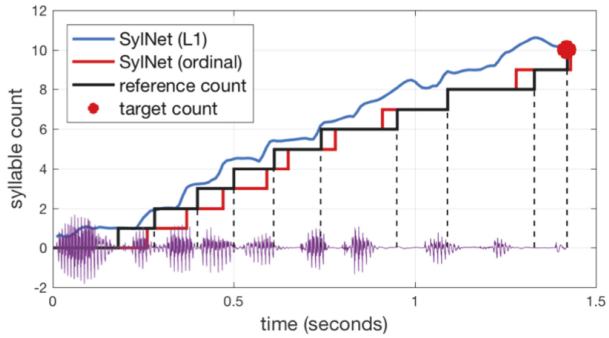


Figure 7.4. An example of SylNet PostNet output accumulation.

The WCE pipeline from Publication III relied on the use of syllable counts as one of the features used for word count prediction. The majority of the previously studied syllable count estimators (SCE) (also those used in Publication III) are based on a syllable envelope detector, followed by peak picking and counting. These SCEs have mainly relied on heuristic digital signal processing (DSP) methods, and only a small number of these use mordern data-driven machine learning approaches. Publication IV proposes a novel data-driven end-to-end method neural network model called SylNet for the SCE task. This model directly optimises the syllable counts without the need for training data with annotations aligned at the syllable level. Experiments on several languages reveal that SylNet generalises to languages beyond its training data and further improves with adaptation. It also outperforms several previously proposed methods for syllabification and end-to-end BLSTMs tested in the study. Figure 7.4 shows an example of the output of SylNet with different loss functions. It can be seen that even though the only information fed during training is the total number of syllables in each utterance, SylNet is still capable of detecting the syllable boundaries in addition to estimating the total syllable counts in the input utterances.

7.5 Publication V: "Vocal effort based speaking style conversion using vocoder features and parallel learning"

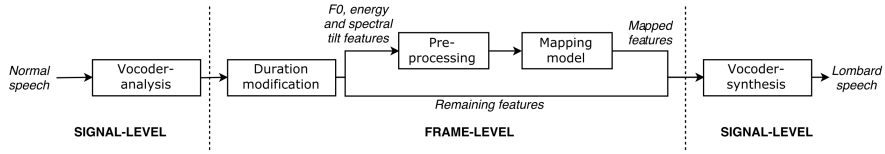


Figure 7.5. Block diagram of the normal-to-Lombard SSC system. Prior to the conversion, the mapping models are trained using DTW-aligned pairs of normal and Lombard speech utterances.

Publication V proposes a parametric speaking style conversion (SSC) system for converting speech utterances from one style to another with varying vocal effort. This study focuses on normal-to-Lombard conversion as a case study of this problem. Figure 7.5 shows a block diagram of the normal-to-Lombard SSC system. The proposed parametric model uses a vocoder to extract frame-level speech features from the input normal-style utterance. These features are then mapped using parallelly trained frame-level machine learning models (MLMs) to the corresponding features of the target Lombard speech. Finally, the mapped features are converted to a Lombard speech waveform with the same vocoder. A total of three vocoders—GlottDNN, STRAIGHT, and Pulse model in log domain (PML); and three MLMs—standard GMM, Bayesian GMM, and feed-forward DNN, were compared in the proposed normal-to-Lombard style conversion system. The SSC system was evaluated using subjective listening tests that measured the perceived Lombardness and quality of the converted speech utterances. An instrumental measure called speech intelligibility in bits (SIIB) was also used to evaluate the intelligibility of the converted Lombard utterances under various noise conditions. The results show that the system is able to convert normal speech into Lombard speech. While comparing the different design choices used, we see that there is a trade-off between the quality and Lombardness of the mapped utterances.

7.6 Publication VI: "Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion"

The MLMs used in Publication V are based on parallel learning mechanisms. SSC systems trained with these models are restricted to training data that have linguistically identical utterances from both the source and target styles from the same speaker. This kind of data is especially difficult to obtain for speech produced with a varied vocal effort, especially in large quantities. This conference article uses a DNN-based non-parallel learning scheme called the cycle-consistent adversarial network (CycleGAN) for the

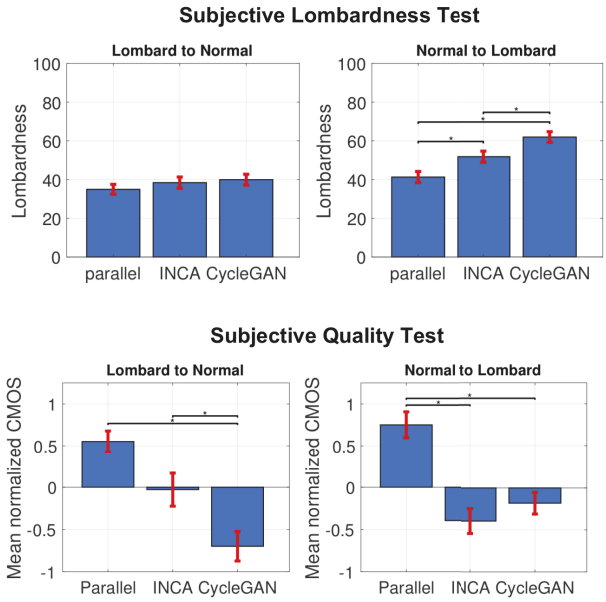


Figure 7.6. Results of the subjective Lombardness (top) and quality (bottom) tests for the Lombard-to-normal (left) and normal-to-Lombard (right) style conversions. For the Lombardness tests a higher value indicates a greater degree of Lombardness and for the quality tests a lower value indicates a better quality. Standard errors are shown in red. Significant difference values as measured using the Student's t-test and the Mann-Whitney U-test for the Lombardness tests and the quality tests respectively with Bonferroni correction are highlighted.

task of SSC. The block diagram of the CycleGAN is shown in Figure 3.9. The same parametric system used in Publication V was used here with the PML as the vocoder. However, a more descriptive set of features were used here (F0, voicing decisions (V/UV), and the first 10 MGC coefficients) as opposed to Publication V (which included F0, energy, and spectral tilt), since the amount of data available for training the MLMs was larger. The CycleGAN is compared with the parallelly trained standard GMM (one of the MLMs used in publication V) as a baseline as well as the iterative combination of a nearest neighbour search step and a conversion step alignment method (INCA), which is a standard technique used in non-parallel learning for related problems such as voice conversion. The systems were evaluated for both normal-to-Lombard as well as Lombard-to-normal conversions using subjective listening tests for quality and Lombardness (the same as used in Publication V). The results are shown in Figure 7.6. It can be seen in the normal-to-Lombard conversion that the CycleGAN produces the highest Lombardness, whereas both INCA and CycleGAN both performed well in terms of quality in comparison to the parallel GMM. For the Lombard-to-normal mapping, the Lombardness of the three methods is almost indistinguishable, with the CycleGAN being the best in terms of quality.

7.7 Publication VII: "Augmented CycleGANs for continuous scale normal-to-Lombard speaking style conversion"

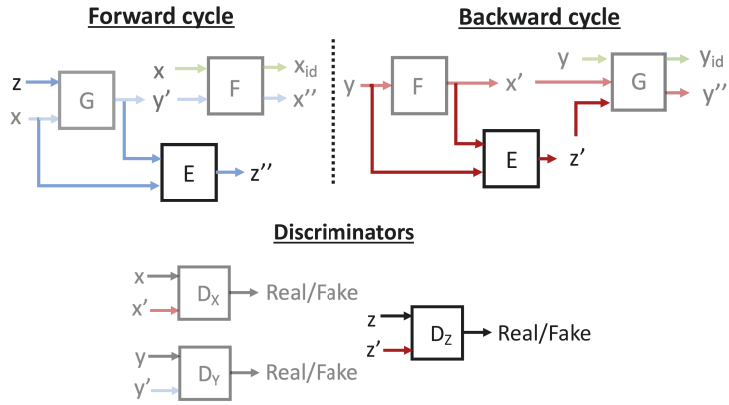


Figure 7.7. Block diagrams showing the augmented CycleGAN with mapping functions G, F and E, and discriminators D_X, D_Y and D_Z. The forward cycle, backward cycle, and identity mapping are indicated in blue, red, and green respectively. The parts shown in the lighter shade form the CycleGAN (used in Publication VI), and the parts shown in the darker shade are additional blocks used for the augmented CycleGAN.

The CycleGAN used in Publication VI is only capable of a deterministic mapping from one style to another, such as in the normal-to-Lombard case studied in the publication. However, increasing the level of Lombardness beyond what is required in a given environment may result in an undesired mismatch between communicative expectations and the resulting vocal expression in the given situation. In contrast, not increasing the level of Lombardness enough would result in an insufficient improvement in intelligibility. Hence, it would be desirable to have an SSC system capable of converting speech from one style to another where the degree of conversion is controllable. In this conference article, we propose the use of recently developed augmented CycleGAN for conversion from normal-to-Lombard speech. The block diagram of the augmented CycleGAN is shown in Figure 7.7. This MLM unsupervisedly learns a latent space that encodes missing information about the target style utterance that is not present in the source style utterance. As in Publication VI, the PML is used as the vocoder. The augmented CycleGAN model is trained on multi-language data. The effectiveness of the model is compared to a standard CycleGAN using subjective experiments for the intelligibility and quality of mapped speech, as well as using the instrumental SIIB intelligibility score.

8. Conclusions

This thesis set out to explore the application of state-of-art ML methods to selected applications in the field of speech processing. The recurring theme in the studies was that the methodology focused on analysis or conversion of speech at the level of suprasegmental units. The main aim of the work was to improve performance in the existing tasks of zero resource speech processing and word/syllable count estimation with new ML-based techniques (Publications I–IV) and to introduce a new parametric system—a system that is based on the extraction and transformation of key features followed by synthesis using vocoders—for the task of speaking style conversion and to propose a number of technical solutions to tackle the transformation problem (Publications V–VII).

In more detail, Publication I explored the use of NPB mixture models for the task of clustering syllable-rhythmic units that correspond to linguistic units in the ZS setting. Several families of prior distributions were compared. It was shown that NPB methods performed consistently across several data sets without requiring manual specification of the total number of clusters. Comparing the priors used, it was found that the Pitman-Yor process—a model that best fits the assumed distributional structure of linguistic data (i.e., Zipf's law)—was the best performing prior distribution among all compared variants. Additionally, the Von Mises-Fisher distribution, which is based on the cosine distance, was found to be more suited for clustering unit-normalised feature vectors than the widely utilised Gaussian distribution, which is based on the Euclidean distance. Publication II further explored the applicability of NPB clustering methods to the task of speaker clustering, in which the utterances were represented using i-vector features. Once again, in line with the theory that the Von Mises-Fisher distribution is more suited for clustering unit-normalised units than the Gaussian, it was observed that the DPVMM had better performance than the DPGMM. The performance of the DPVMM was also close to that of parametric methods when the true number of clusters is known by those models. Overall, Publications I and II show that NPB methods provide a potential alternative to more traditional parametric

methods for speech processing applications where data clustering is required, especially when the number of clusters is not known. Practically, however, some level of task-specific tuning is required to find the best set of hyper-parameters for the NPB models. Further, Publications I and II show that the influence of the choice of the priors for these models can largely be predicted based on some theoretical considerations related to the data, such as the Zipfian distributional properties of lexical data (Piantadosi, 2014).

Publication III dealt with the task of WCE in the context of realistic day-long infant-directed recordings from wearable microphones. In this context, the automatic syllabification of speech was used as a processing step to obtain estimated word counts in different languages. A number of unsupervised DSP-based methods and a supervised RNN-based ML model were compared for this purpose. The RNN-based model was found to be the best performing alternative over several realistic data sets. However, a disadvantage of this approach is the requirement of phone-level annotations for training the syllable envelope detector within the model. This shortcoming was addressed in Publication IV with an end-to-end model called SylNet that directly predicts the syllable count from input speech. The training of the SylNet only requires the true syllable count per utterance, but not any alignment between the linguistic information and the underlying speech data. Publication IV compared the SylNet with the envelope detection-based methods used in Publication III, as well as an alternative RNN-based end-to-end model. It was found that the SylNet had the best performance across the different test datasets in different languages and recording conditions, all differing from the training set. Moreover, it was shown that SylNet was easily adaptable to new domains, with the performance of the model substantially improving with just a few minutes of adaptation data from the target domain.

The last three publications focused on the problem of automatic speaking style conversion. Publication V laid down the basic problem formulation and proposed a parametric system for vocal effort-based SSC. Several parametric vocoders and ML methods were compared for the task of normal to Lombard conversion in a parallel learning setting. Based on subjective listening tests and instrumental measurements, it was observed that all methods achieved a significant shift in speaking style towards Lombard speech, however, with some degradation in the perceived quality. In order to better use non-parallel data, which, is more readily available, Publication VI compared a GAN-based non-parallel learning scheme called the CycleGAN with a more standard approach for non-parallel learning called INCA as well as the parallel learning methods. Listening tests indicated that the CycleGAN produced encouraging results compared to the other methods. It has to be noted that the results were obtained with the CycleGAN trained with a combination of parallel and non-parallel

data. Therefore, the performance of the SSC system studied might drop slightly if the training of CycleGAN was conducted only using non-parallel data. Publication VII extended the CycleGAN approach to a method called augmented CycleGAN that allows simultaneous learning of a latent space which enables the parameterisation of information about the target style not present in the source style. This latent space can then be utilised during inference to achieve a controllable degree of style conversion, thereby allowing the adaptation of the conversion process to different use cases and listening conditions.

To summarise, the thesis shows that ML-based solutions are capable of tackling and providing state-of-the-art performance in a number of problems focused on the broad topics of speech analysis and conversion. In addition, some of these methods help to lessen the amount of a priori assumptions required for successful modelling of the data (such as in the case of NPB models for data clustering), or reduce the requirements for the type of data required for model training (such as enabling the use of non-parallel data for speaking style conversion or the need for phonetic annotations for syllable count estimation). This thesis shows that NPB models can be used for clustering in the ZS task eliminating the need for model selection. A novel end-to-end architecture developed for the WCE/SCE task outperformed the state-of-the art LENA solution, and is available as an open-source solution. The topic of speaking style conversion was also studied in the thesis in more depth. The thesis introduced a new parametric system for SSC as well as a number of technical solutions to tackle the SSC problem. The systems developed in this thesis enable the conversion of speech utterances from normal to Lombard style with a controllable degree of conversion. Overall, the thesis work has shown that state-of-the-art ML models can be used to tackle a variety of speech processing problems with a competitive performance when focusing on suprasegmental signal structures, and that good quality speaking style conversion can be achieved with both parallel and non-parallel datasets.

References

G. Aimetti. Modelling early language acquisition skills: Towards a general statistical learning mechanism. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 1–9, Athens, Greece, 2009. Association for Computational Linguistics.

M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku. GlottDNN-A full-band glottal vocoder for statistical parametric speech synthesis. In *Proceedings of Interspeech 2016*, pages 2473–2477, San Francisco, CA, USA, 2016. ISCA.

M. B. Akçay and K. Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, Berlin, Germany, 1985. Springer.

P. Alku, J. Vintturi, and E. Vilkman. Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. *Speech Communication*, 38(3-4):321–334, 2002.

P. Alku, M. Airas, E. Björkner, and J. Sundberg. An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. *The Journal of the Acoustical Society of America*, 120(2): 1052–1062, 2006.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML-2017)*, pages 214–223, Sydney, Australia, 2017. PMLR.

B. S. Atal. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4):460–475, 1976.

L. Badino, A. Mereta, and L. Rosasco. Discovering discrete subword units with binarized autoencoders and hidden-Markov-model encoders. In *Proceedings of Interspeech 2015*, pages 3174–3178, Dresden, Germany, 2015. ISCA.

R. Baker and V. Hazan. Acoustic-phonetic characteristics of naturally-elicited clear speech in British English. *The Journal of the Acoustical Society of America*, 125(4):2729–2729, 2009.

E. Bergelson, A. Warlaumont, A. Cristia, M. Casillas, C. Rosemberg, M. Soderstrom, C. Rowland, S. Durrant, and J. Bunce. Starter-ACLEW, Databrary. URL https://nyu.databrary.org/volume/390. DOI:10.17910/B7.390, Accessed: 2020-04-25.

E. Bergelson, M. Casillas, M. Soderstrom, A. Seidl, A. S. Warlaumont, and A. Amatuni. What do north american babies hear? A large-scale cross-corpus analysis. *Developmental Science*, 22(1):e12724, 2019.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, New York, NY, USA, 2006.

D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

A. R. Bradlow and T. Bent. The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, 112(1):272–284, 2002.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "Siamese" time delay neural network. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS-1994)*, pages 737–744, Denver, CO, USA, 1994. MIT Press.

À. Calzada and J. C. Socoró. Vocal effort modification through harmonics plus noise model representation. In *Proceedings of the 5th International Conference on Nonlinear Speech Processing*, pages 96–103, Berlin, Heidelberg, 2011. Springer.

J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9): 1437–1462, 1997.

A. Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comptes rendus des séances de l'Académie des sciences*, 25:536–538, 1847.

H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li. Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In *Proceedings of Interspeech 2015*, pages 3189–3193, Dresden, Germany, 2015. ISCA.

N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

Y. Chen, S. Huang, H. Lee, Y. Wang, and C. Shen. Audio word2vec: Sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(9):1481–1493, 2019.

L. Chittka and A. Brockmann. Perception space—the final frontier. *PLoS Biology*, 3(4), 2005.

C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2018)*, pages 4774–4778, Calgary, Canada, 2018. IEEE.

K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, pages 103–111, Doha, Qatar, 2014. Association for Computational Linguistics.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS-2014)*, pages 2672–2680, Montréal, Canada, 2014. Curran Associates, Inc.

M. Cooke and Y. Lu. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *The Journal of the Acoustical Society of America*, 128(4):2059–2069, 2010.

M. Cooke, C. Mayo, and J. Villegas. The contribution of durational and spectral changes to the lombard speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 135(2):874–883, 2014.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

A. Cristia, M. Lavechin, C. Scaff, M. Soderstrom, C. Rowland, O. Räsänen, J. Bunce, and E. Bergelson. A thorough evaluation of the language environment analysis (lena) system. *Behavior Research Methods*, pages 1–20, 2020.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2:303–314, 1989.

C. d'Alessandro and B. Doval. Experiments in voice quality modification of natural speech signals: The spectral approach. In *Proceedings of the ESCA/COCOSDA Workshop on Speech Synthesis*, pages 277–282, Blue Mountains, Australia, 1998. ISCA.

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

G. Degottex, P. Lanchantin, and M. Gales. A log domain pulse model for parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):669–679, 2016.

L. Dilley, T. Morrill, and E. Banzina. New tests of the distal speech rate effect: examining cross-linguistic generalization. *Frontiers in Psychology*, 4:1002, 2013.

J. J. Dreher and J. O'Neill. Effects of ambient noise on speaker intelligibility for words and phrases. *The Journal of the Acoustical Society of America*, 29(12): 1320–1323, 1957.

E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux. The zero resource speech challenge 2017. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE, 2017.

E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, L. Besacier, S. Sakti, and E. Dupoux. The zero resource speech challenge 2019: TTS without T. In *Proceedings of Interspeech 2019*, pages 1088–1092. ISCA, 2019.

E. Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.

R. Eloff, A. Nortje, B. van Niekerk, A. Govender, L. Nortje, A. Pretorius, E. van Biljon, E. van der Westhuizen, L. van Staden, and H. Kamper. Unsupervised Acoustic Unit Discovery for Speech Synthesis Using Discrete Latent-Variable Neural Networks. In *Proceedings of Interspeech 2019*, pages 1103–1107, Graz, Austria, 2019. ISCA.

D. Erro, A. Moreno, and A. Bonafonte. Inca algorithm for training voice conversion systems from nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):944–953, 2009a.

D. Erro, E. Navas, I. Hernáez, and I. Saratxaga. Emotion conversion based on prosodic unit selection. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):974–983, 2009b.

D. Erro, I. Sainz, E. Navas, and I. Hernaez. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):184–194, 2013.

G. Fant. *Acoustic theory of speech production*. Mouton & Co. N. V. Publishers, The Hague, Paris, France, 2nd edition, 1970.

S. Feng, T. Lee, and Z. Peng. Combining Adversarial Training and Disentangled Speech Representation for Robust Zero-Resource Subword Modeling. In *Proceedings of Interspeech 2019*, pages 1093–1097, Graz, Austria, 2019. ISCA.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

J. L. Flanagan. *Speech analysis synthesis and perception*. Springer-Verlag, Berlin, Germany, 1972.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM, 1993.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, Boca Raton, FL, USA, 3rd edition, 2013.

E. Gentet, B. David, S. Denjean, G. Richard, and V. Roussarie. Neutral to Lombard speech conversion with deep learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2020)*, pages 7739–7743, Barcelona, Spain, 2020. IEEE.

F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with LSTM. In *Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN-99)*, pages 850–855, Edinburgh, UK, 1999.

S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

J. Gilkerson and J. A. Richards. The LENA natural language study. Boulder, CO, USA, 2009. LENA Foundation.

B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

J. Glass. Towards unsupervised speech processing. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1–4. IEEE, 2012.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS-2014)*, pages 2672–2680, Montréal, Canada, 2014. Curran Associates, Inc.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Boston, MA, USA, 2016. `http://www.deeplearningbook.org`.

P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins. Relationship between changes in voice pitch and loudness. *Journal of voice*, 2 (2):118–126, 1988.

S. Granlund, V. Hazan, and R. Baker. An acoustic–phonetic comparison of the clear speaking styles of finnish–english late bilinguals. *Journal of Phonetics*, 40(3):509–520, 2012.

J. Grynpas, R. Baker, and V. Hazan. Clear speech strategies and speech perception in adverse listening conditions. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS-2011)*, Hong Kong, 2011. International Phonetic Association.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS-2017)*, pages 5767–5777, Long Beach, CA, USA, 2017. Curran Associates, Inc.

J. H. Hansen and V. Varadarajan. Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2): 366–378, 2009.

V. Hazan, O. Tuomainen, J. Kim, C. Davis, B. Sheffield, and D. Brungart. Clear speech adaptations in spontaneous speech produced by young and older adults. *The Journal of the Acoustical Society of America*, 144(3):1331–1346, 2018.

S. Hertegård, J. Gauffin, and P.-Å. Lindestad. A comparison of subglottal and intraoral pressure measurements during phonation. *Journal of voice*, 9(2): 149–155, 1995.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK, 2010.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

F. M. Hoppe. Pólya-like urns and the Ewens' sampling formula. *Journal of Mathematical Biology*, 20(1):91–94, 1984.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

K. Hornik, M. Stinchcombe, H. White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Cao, and Y. Wang. Hierarchical generative modeling for controllable speech synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR-2019)*, New Orleans, LO, USA, 2019. OpenReview.net.

Q. Hu, Y. Stylianou, R. Maia, K. Richmond, J. Yamagishi, and J. Latorre. An investigation of the application of dynamic sinusoidal models to statistical parametric speech synthesis. In *Proceedings of Interspeech 2014*, pages 780–784, Singapore, 2014.

D.-Y. Huang, S. Rahardja, and E. P. Ong. Lombard effect mimicking. In *Proceedings of the Speech synthesis workshop (SSW-2010)*, pages 258–263, Kyoto, Japan, 2010.

Z. Inanoglu and S. Young. Data-driven emotion conversion in spoken English. *Speech Communication*, 51(3):268–283, 2009.

N. Isshiki. Regulatory mechanism of voice intensity variation. *Journal of speech and hearing research*, 7(1):17–29, 1964.

ITU-T. One-way transmission time. Rec. g.114, International Telecommunication Union, Geneva, Switzerland, May 2003.

M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad. Fundamental frequency generation for whisper-to-audible speech conversion. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2014)*, pages 2579–2583, Florence, Italy, 2014. IEEE.

A. Jansen and B. Van Durme. Efficient spoken term discovery using randomized algorithms. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 401–406, Waikoloa, HI, USA, 2011. IEEE.

S. P. Johnson. How infants learn about the visual world. *Cognitive Science*, 34(7): 1158–1184, 2010.

S. Kakouros and O. Räsänen. 3PRO–An unsupervised method for the automatic detection of sentence prominence in speech. *Speech Communication*, 82:67–84, 2016.

N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu. Efficient neural audio synthesis. In *Proceedings of Machine Learning Research*, pages 2410–2419, Stockholm Sweden, 2018. PMLR.

O. Kalinli and S. Narayanan. Prominence detection using auditory attention cues and task-dependent high level information. *IEEE Transactions on audio, Speech, and language processing*, 17(5):1009–1024, 2009.

H. Kamper, M. Elsner, A. Jansen, and S. Goldwater. Unsupervised neural network based feature extraction using weak top-down constraints. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2015)*, pages 5818–5822, Brisbane, Australia, 2015. IEEE.

H. Kamper, A. Jansen, and S. Goldwater. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):669–679, 2016.

H. Kamper, A. Jansen, and S. Goldwater. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174, 2017.

H. Kamper, K. Livescu, and S. Goldwater. An embedded segmental K-means model for unsupervised segmentation and clustering of speech. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2017)*, pages 719–726, Okinawa, Japan, 2017. IEEE.

P. D. S. Karthik and H. A. Murthy. Zero Resource Speech Synthesis Using Transcripts Derived from Perceptual Acoustic Units. In *Proceedings of Interspeech 2019*, pages 1113–1117, Graz, Austria, 2019. ISCA.

H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207, 1999.

J. Kiefer, J. Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. Proceedings of Machine Learning Research, pages 1857–1865, Sydney, Australia, 2017. PMLR.

S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

H. Konno, M. Kudo, H. Imai, and M. Sugimoto. Whisper to normal speech conversion using pitch estimated from spectrum. *Speech Communication*, 83: 10–20, 2016.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

P. Ladefoged and N. P. McKinney. Loudness, sound pressure, and subglottal pressure in speech. *The journal of the Acoustical Society of America*, 35(4): 454–460, 1963.

H. Lane and B. Tranel. The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14(4):677–709, 1971.

P. Last, H. A. Engelbrecht, and H. Kamper. Unsupervised feature learning for speech using correspondence and siamese networks. *IEEE Signal Processing Letters*, 27:421–425, 2020.

Y. LeCun, J. S. Denker, and S. A. Solla. Optimal brain damage. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS-1990)*, pages 598–605, Denver, CO, USA, 1990. Morgan-Kaufmann.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

R. M. Lerner, L. S. Liben, and U. Mueller. *Handbook of child psychology and developmental science: Volume 2, Cognitive processes*. John Wiley & Sons, Hoboken, NJ, USA, 7th edition, 2015.

P. Lieberman and R. McCarthy. The evolution of speech and language. In *Handbook of Paleoanthropology*, pages 1–41. Springer-Verlag, Berlin, Germnay, 2nd edition, 2015.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

B. Lindblom. Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling*, pages 403–439. Kluwer Academic Publishers, London, UK, 1990.

A. T. Liu, P. chun Hsu, and H.-Y. Lee. Unsupervised End-to-End Learning of Discrete Linguistic Units for Voice Conversion. In *Proceedings of Interspeech 2019*, pages 1108–1112, Graz, Austria, 2019. ISCA.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

P. C. Loizou. *Speech enhancement: Theory and practice*. CRC press, Boca Raton, FL, USA, 2nd edition, 2013.

E. Lombard. Le signe de l'elevation de la voix. *Annales des Maladies de L'Oreille et du Larynx*, 37:101–119, 1911.

J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. In *Proceedings of Odyssey 2018, The Speaker and Language Recognition Workshop*, pages 195–202, Les Sables d'Olonne, France, 2018. ISCA.

Y. Lu and M. Cooke. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51 (12):1253–1262, 2009a.

Y. Lu and M. Cooke. Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *The Journal of the Acoustical Society of America*, 126(3):1495–1499, 2009b.

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS-2017)*, pages 6231–6239, Long Beach, CA, USA, 2017. Curran Associates, Inc.

V. Lyzinski, G. Sell, and A. Jansen. An evaluation of graph clustering methods for unsupervised term discovery. In *Proceedings of Interspeech 2015*, Dresden, Germany, 2015. ISCA.

S. Y. Manuel. Vowel reduction and perceptual recovery in casual speech. *The Journal of the Acoustical Society of America*, 91(4):2388–2388, 1992.

X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, Venice, Italy, 2017. IEEE.

G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York, NY, USA, 1988.

G. N. Meenakshi and P. K. Ghosh. Whispered speech to neutral speech conversion using bidirectional lstms. In *Proceedings of Interspeech 2018*, pages 491–495, Hyderabad, India, 2018. ISCA.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR-2018)*, Vancouver, Canada, 2018. OpenReview.net.

S. H. Mohammadi and A. Kain. An overview of voice conversion systems. *Speech Communication*, 88:65–82, 2017.

B. C. Moore. *An introduction to the psychology of hearing*. Emerald Group Publishing, Bingley, UK, 6th edition, 2012.

B. C. Moore and B. R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-98)*, pages 729–732, Seattle, WA, USA, 1998. IEEE.

M. Morise, F. Yokomori, and K. Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, 99(7):1877–1884, 2016.

R. W. Morris and M. A. Clements. Reconstruction of speech from whispers. *Medical Engineering & Physics*, 24(7-8):515–520, 2002.

K. P. Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, The University of British Columbia, Vancouver, Canada, 2007.

K. P. Murphy. *Machine learning: A probabilistic perspective*. MIT press, Cambridge, MA, USA, 2012.

N. Narendra and P. Alku. Dysarthric speech classification from coded telephone speech using glottal features. *Speech Communication*, 110:47–55, 2019.

K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie. Speech intelligibility improvement in car noise environment by voice transformation. *Speech Communication*, 91:17–27, 2017.

K. I. Nordstrom, G. Tzanetakis, and P. F. Driessen. Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction. *IEEE Transactions on Audio, Speech, and Language processing*, 16(6):1087–1096, 2008.

M. O'Searcoid. *Metric spaces*. Springer Science & Business Media, Berlin, Germany, 2006.

D. O'Shaughnessy. *Speech Communications: Human And Machine*. IEEE Press, New York, NY, USA, 2nd edition, 1987.

A. Park and J. R. Glass. Unsupervised word acquisition from speech using pattern discovery. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2006)*, pages 409–412, Toulouse, France, 2006. IEEE.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

S. T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130, 2014.

J. M. Pickett. Effects of vocal force on the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 28(5):902–905, 1956.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

J. Pitman et al. Combinatorial stochastic processes. Technical report, UC Berkeley, Berkeley, CA, USA, 2002.

S. C. Poole. *An introduction to linguistics*. Palgrave, New York, NY, USA, 7th edition, 1999.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *Proceedings of the workshop on automatic speech recognition and understanding (ASRU-2011)*, pages 4774–4778, Hawaii, USA, 2011. IEEE.

R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2019)*, pages 3617–3621, Brighton, UK, 2019. IEEE.

M. S. PV, V. Tsiaras, and Y. Stylianou. Speech intelligibility enhancement based on a non-causal Wavenet-like model. In *Proceedings of Interspeech 2018*, pages 1868–1872, Hyderabad, India, 2018. ISCA.

T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio, Speech, and Language processing*, 19(1):153–165, 2010.

T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku. Analysis and synthesis of shouted speech. In *Proceedings of Interspeech 2013*, pages 1544–1548, Lyon, France, 2013. ISCA.

N. Ramírez-Esparza, A. García-Sierra, and P. K. Kuhl. Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, 17(6):880–891, 2014.

O. Räsänen. Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, 54(9):975–997, 2012.

O. Räsänen, G. Doyle, and M. C. Frank. Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Proceedings of Interspeech 2015*, pages 3204–3208, Dresden, Germany, 2015. ISCA.

D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater. A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. In *Proceedings of Interspeech 2015*, pages 3199–3203, Dresden, Germany, 2015. ISCA.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

A. E. Rosenberg. Automatic speaker verification: A review. *Proceedings of the IEEE*, 64(4):475–487, 1976.

T. Rossing, F. Moore, and P. Wheeler. *The Science of Sound*. Pearson Education, London, UK, 3rd edition, 2002.

D. Rostolland. Intelligibility of shouted voice. *Acta Acustica united with Acustica*, 57(3):103–121, 1985.

A. Rouhe, T. Kaseva, and M. Kurimo. Speaker-aware training of attention-based end-to-end speech recognition using neural speaker embeddings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2020)*, pages 7064–7068, Barcelona, Spain, 2020. IEEE.

A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, and K. Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. In R. Singh, M. Vatsa, V. M. Patel, and N. Ratha, editors, *Domain Adaptation for Visual Understanding*, pages 33–49. Springer International Publishing, New Delhi, India, 2020.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

L. Schmidt, M. Sharifi, and I. Lopez-Moreno. Large-scale speaker identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2014)*, pages 1650–1654, Florence, Italy, 2014. IEEE.

J. Schnupp, I. Nelken, and A. King. *Auditory neuroscience: Making sense of sound*. MIT press, Cambridge, MA, USA, 2011.

B. W. Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.

G. Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4 (2):639–650, 1994.

K. Smith and S. Kirby. Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3591–3603, 2008.

S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

Y. Stylianou. Voice transformation: A survey. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2009)*, pages 3585–3588, Taipei, Taiwan, 2009. IEEE.

Y. Stylianou. Voice transformation: A survey. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2009)*, pages 3585–3588, Taipei, Taiwan, 2009. IEEE.

Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing*, 6(2): 131–142, 1998.

W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

Y. Tang, B. M. Fazenda, and T. J. Cox. Automatic speech-to-background ratio selection to maintain speech intelligibility in broadcasts using an objective intelligibility metric. *Applied Sciences*, 8(1):59, 2018.

Z. Tao, X.-D. Tan, T. Han, J.-H. Gu, Y.-S. Xu, and H.-M. Zhao. Reconstruction of normal speech from whispered speech based on RBF neural network. In *Proceedings of Intelligent Information Technology and Security Informatics (IITSI-2010)*, pages 374–377, Jinggangshan, China, 2010. IEEE.

V. C. Tartter, H. Gomes, and E. Litwin. Some acoustic effects of listening to noise on speech production. *The Journal of the Acoustical Society of America*, 94(4): 2437–2440, 1993.

P. Taylor. *Text-to-speech synthesis*. Cambridge university press, Cambridge, UK, 2009.

L. ten Bosch and B. Cranen. A computational model for unsupervised word discovery. In *Proceedings of Interspeech 2007*, pages 1481–1484, Antwerp, Belgium, 2007. ISCA.

R. Thiolliere, E. Dunbar, G. Synnaeve, M. Versteegh, and E. Dupoux. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *Proceedings of Interspeech 2015*, pages 3179–3183, Dresden, Germany, 2015. ISCA.

A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura. VQVAE Unsupervised Unit Discovery and Multi-Scale Code2Spec Inverter for Zerospeech Challenge 2019. In *Proceedings of Interspeech 2019*, pages 1118–1122, Graz, Austria, 2019. ISCA.

T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi. The voice conversion challenge 2016. In *Proceedings of Interspeech 2016*, pages 1632–1636, San Francisco, CA, USA, 2016. ISCA.

K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP-94)*, pages 1043–1046, Yokohama, Japan, 1994. ISCA.

J.-M. Valin and J. Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2019)*, pages 5891–5895, Brighton, UK, 2019. IEEE.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125, Sunnyvale, USA, 2016. ISCA.

A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In *Proceedings of Machine Learning Research*, pages 3918–3926, Stockholm Sweden, 2018.

M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux. The zero resource speech challenge 2015. In *Proceedings of Interspeech 2015*, pages 3169–3173, Dresden, Germany, 2015. ISCA.

C. Villani. *Topics in optimal transportation*. American Mathematical Society, Providence, RI, USA, 2003.

D. Wang and S. S. Narayanan. Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8): 2190–2201, 2007.

M. Wang, M. Wen, K. Hirose, and N. Minematsu. Emotional voice conversion for Mandarin using tone nucleus model–small corpus and high efficiency. In *Proceedings of Speech Prosody 2012*, Shanghai, China, 2012. ISCA.

X. Wang, S. Takaki, and J. Yamagishi. Neural source-filter-based waveform model for statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2019)*, pages 5916–5920, Brighton, UK, 2019. IEEE.

Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning (ICML-2018*, pages 5180–5189, Stockholm, Sweden, 2018. PMLR.

A. Weisleder and A. Fernald. Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24 (11):2143–2152, 2013.

D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. Hansen. Signal processing for young child speech language development. In *Proceedings of the 1st Workshop on Child, Computer and Interaction (WOCCI2008)*, Crete, Greece, 2008. ISCA.

C. Yarra, O. D. Deshmukh, and P. K. Ghosh. A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection. *Speech Communication*, 78:62–71, 2016.

Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision (ICCVW-2017)*, pages 2849–2857, Venice, Italy, 2017. IEEE.

B. Yusuf, A. Gök, B. Gundogdu, O. D. Kose, and M. Saraclar. Temporally-Aware Acoustic Unit Discovery for Zerospeech 2019 Challenge. In *Proceedings of Interspeech 2019*, pages 1098–1102, Graz, Austria, 2019. ISCA.

H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *speech communication*, 51(11):1039–1064, 2009.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV-2017)*, pages 2223–2232, 2017.

A. Ziaei, A. Sangwan, and J. H. Hansen. Prof-life-log: Personal interaction analysis for naturalistic audio streams. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2013)*, pages 7770–7774, Vancouver, Canada, 2013. IEEE.

A. Ziaei, A. Sangwan, L. Kaushik, and J. H. Hansen. Prof-life-log: Analysis and classification of activities in daily audio streams. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2015)*, pages 4719–4723, South Brisbane, Australia, 2015. IEEE.

A. Ziaei, A. Sangwan, and J. H. Hansen. Effective word count estimation for long duration daily naturalistic audio recordings. *Speech Communication*, 84:15–23, 2016.

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

Errata

Publication VII

In Figure 1, for the forwards cycle the mapping F should be from y' to x'' and not from y' to y''. Similarly, for the backward cycle, the mapping G should be from x' to y'' and not from x' to x''.

Speech technology is a field of technological research focusing on methods to process spoken language. Work in the area has largely relied on a combination of domain-specific knowledge and digital signal processing algorithms, often also combined with statistical (parametric) models to develop applications. In this context, machine learning (ML) has played a central role in estimating the parameters of such models. The goal of this thesis is to investigate the applicability of recent state-of-the-art developments in ML to the modelling and processing of speech at the so-called suprasegmental level to tackle the following topical problems in speech research: 1) zero-resource speech processing, where the aim is to learn language patterns from speech without access to annotated datasets, 2) automatic word and syllable count estimation which focus on quantifying the amount of linguistic content in audio recordings, and 3) speaking style conversion, which deals with the conversion of the speaking style of an utterance while retaining the linguistic content, speaking identity and quality.

Aalto-DD 201/2020

ISBN 978-952-64-0166-9 (printed)
ISBN 978-952-64-0167-6 (pdf)
ISSN 1799-4934 (printed)
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