Department of Signal Processing and Acoustics

Statistical methods for incomplete speech data

Ulpu Remes

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 149/2016

Statistical methods for incomplete speech data

Ulpu Remes

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall S1 of the school on 2 September 2016 at 12.

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

Supervising professor
Professor Mikko Kurimo

Thesis advisor
Docent Kalle Palomäki

Preliminary examiners
Dr. Jon Barker, University of Sheffield, UK
Dr. Ramon Fernandez Astudillo, INESC-ID, Portugal

Opponent
Professor Hugo Van hamme, KU Leuven, Belgium

Aalto University publication series
DOCTORAL DISSERTATIONS 149/2016

ISBN 978-952-60-6936-4 (printed)
ISBN 978-952-60-6937-1 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6937-1

Unigrafia Oy
Helsinki 2016

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Ulpu Remes

Name of the doctoral dissertation
Statistical methods for incomplete speech data

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL DISSERTATIONS 149/2016

Field of research Speech and Language Technology

Manuscript submitted 7 March 2016 **Date of the defence** 2 September 2016

Permission to publish granted (date) 19 May 2016 **Language** English

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

Abstract

Speech can be represented as an observation matrix where each node corresponds to a certain speech feature. However when speech is mixed with environmental sounds, some features cannot be observed and the observation matrix remains incomplete. The missing values are a problem because incomplete observations can support incorrect conclusions and because most applications cannot process incomplete data. Methods that are used to handle incomplete observations are called missing-data methods.

This thesis presents on overview on missing-data methods and discusses their application in noise-robust automatic speech recognition. Hence we assume that the speech observations are incomplete due to environmental sounds. The methods studied in this work substitute unobserved feature values with estimates calculated based on the incomplete observations and statistical dependencies between the observed and unobserved features. This is called missing-data imputation. The main research directions include imputation methods that utilise temporal dependencies between observations and imputation methods that associate feature estimates with uncertainties. The experiments conducted in this work indicate that temporal dependencies and imputation uncertainties improve automatic speech recognition performance when speech is corrupted with environmental noise.

The thesis also discusses narrowband telephone speech and bandwidth extension. Narrowband speech can be considered incomplete since observations associated with certain features are not included in the narrowband transmission. Bandwidth extension means that the narrowband observations are converted into wideband observations which include more features. The bandwidth extension methods evaluated in this work estimate wideband observations based on narrowband observations and statistical dependencies between narrowband and wideband features.

Keywords automatic speech recognition, missing-data methods, noise robustness, observation uncertainties

ISBN (printed) 978-952-60-6936-4 **ISBN (pdf)** 978-952-60-6937-1

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2016

Pages 167 **urn** http://urn.fi/URN:ISBN:978-952-60-6937-1

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Ulpu Remes

Väitöskirjan nimi
Tilastollisia menetelmiä osittain havaitun puheen käsittelyyn

Julkaisija Sähkötekniikan korkeakoulu

Yksikkö Signaalinkäsittelyn ja akustiikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 149/2016

Tutkimusala Puhe- ja kieliteknologia

Käsikirjoituksen pvm 07.03.2016 **Väitöspäivä** 02.09.2016

Julkaisuluvan myöntämispäivä 19.05.2016 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

Puhe voidaan esittää havaintomatriisina, missä yksittäiset havainnot vastaavat puheen eri ominaisuuksia tai piirteitä eri ajanhetkillä. Kun puheeseen sekoittuu muita ääniä, kaikkien piirteiden havaitseminen ei kuitenkaan onnistu. Tällöin havaintomatriisista ei tule kokonaista. Puuttuvat havainnot hankaloittavat puheaineiston käyttöä, koska osittaisen tiedon perusteella saatetaan tehdä vääriä päätelmiä. Useimmat sovellukset eivät myöskään käsittele osittaista havaintoaineistoa. Osittaisen havaintoaineiston käsittelyyn soveltuvia menetelmiä kutsutaan puuttuvan tiedon menetelmiksi.

Tässä väitöskirjassa tutustutaan puuttuvan tiedon menetelmiin ja menetelmien käyttöön melusietoisessa automaattisessa puheentunnistuksessa. Työssä tutkittava puhe on siis taustamelun takia osittain havaitsematta. Puheen käsittelyyn käytetään puuttuvan tiedon paikkaamiseen eli imputointiin perustuvia menetelmiä. Imputointimenetelmät käyttävät osittaista havaintomatriisia sekä havaintomatriisin piirteiden välisiä tilastollisia riippuvuuksia puuttuvan osuuden estimointiin. Tässä työssä tutkitaan erityisesti aikariippuvuuksien käyttöä puuttuvan havaintotiedon estimoinnissa sekä estimoinnin luotettavuuden arviointia. Estimoinnin luotettavuuden arviointia kutsutaan havainnon epävarmuuden mallintamiseksi. Puheentunnistuskokeet osoittavat sekä aikariippuvuuksien että epävarmuuksien käytön parantavan puheentunnistustarkkuutta, kun havaittuun puheeseen on sekoittunut taustamelua.

Melusietoisen puheentunnistuksen lisäksi väitöskirjassa käsitellään puheen taajuuskaistan keinotekoista laajentamista. Esimerkiksi kapeakaistaisen puhelinverkon välittämää puhetta voidaan pitää osittain havaittuna, koska kapeakaistaisessa tiedonsiirrossa välitetään ainostaan osa puheen piirteistä. Taajuuskaistan laajentamisella tarkoitetaan kapeakaistaisen puheen muuntamista laajakaistaiseksi. Kapeakaistaista puhetta kuvaavaan havaintomatriisin lisätään tällöin laajakaistaista puhetta kuvaavia piirteitä. Piirteiden arvot estimoidaan havaitun kapeakaistaisen puheen perusteella käyttäen hyväksi tilastollisia riippuvuuksia kapeakaistaisen ja laajakaistaisen puheen välillä.

Avainsanat automaattinen puheentunnistus, havaintojen epävarmuus, melusietoisuus, puuttuvan tiedon menetelmät

ISBN (painettu) 978-952-60-6936-4 **ISBN (pdf)** 978-952-60-6937-1

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2016

Sivumäärä 167 **urn** http://urn.fi/URN:ISBN:978-952-60-6937-1

Preface

The research presented in this thesis was carried out at Aalto University and its predecessor Helsinki University of Technology between years 2007–2015. The work started at the Department of Information and Computer Science and continued at the Department of Signal Processing and Acoustics. The work was financially supported by the Academy of Finland, Helsinki Graduate School in Computer Science and Engineering, and Tekes. I have also received a personal grant from the Emil Aaltonen Foundation.

I am most indebted to my supervisor and previous instructor Professor Mikko Kurimo and my instructor Docent Kalle Palomäki. This work would never have been completed without their support and guidance. I would also like to extend my deepest gratitude to my previous supervisor Professor Emeritus Erkki Oja who provided me valuable support and advice.

I would like to express my sincere appreciation and gratitude to the speech recognition group members and everyone else I had the fortune to work with. I would like to thank Bert Cranen and Jort Gemmeke for the chance to participate in the sparse imputation work, Paavo Alku and Hannu Pulakka for the chance to participate in the bandwidth extension work, and Keiichi Tokuda, Yoshihiko Nankaku, Tapani Raiko, Antti Honkela, and Ana Ramírez López for their contributions in the other publications in this thesis. In addition I would like to thank Emma Jokinen, Heikki Kallasjoki, Reima Karhila, and Sami Keronen for the chance to participate in their work and Guy Brown and Lauri Juvela for their contributions in the perceptual restoration work. I would also like to thank Mari-Sanna Paukkeri and Okko Räsänen who helped me write and improve this thesis.

I would also like to thank the preliminary examiners Dr. Jon Barker and

Dr. Ramon Fernandez Astudillo whose constructive comments helped me improve the work.

Finally I would like to thank my friends and family whose continued support has been most precious to me.

Espoo, July 29, 2016,

Ulpu Remes

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 Jort F. Gemmeke, Bert Cranen and Ulpu Remes. Sparse imputation for large vocabulary noise robust ASR. *Computer Speech & Language*, volume 25, issue 2, pp. 462–479, April 2011.

II Jort F. Gemmeke, Ulpu Remes and Kalle J. Palomäki. Observation uncertainty measures for sparse imputation. In *INTERSPEECH*, Makuhari, Chiba, Japan, pp. 2262–2265, September 2010.

III Ulpu Remes, Kalle J. Palomäki, Tapani Raiko, Antti Honkela and Mikko Kurimo. Missing-feature reconstruction with a bounded nonlinear state-space model. *IEEE Signal Processing Letters*, volume 18, issue 10, pp. 563–566, October 2011.

IV Ulpu Remes, Yoshihiko Nankaku and Keiichi Tokuda. GMM-based missing-feature reconstruction on multi-frame windows. In *INTERSPEECH*, Florence, Italy, pp. 1665–1668, August 2011.

V Ulpu Remes. Bounded conditional mean imputation with an approximate posterior. In *INTERSPEECH*, Lyon, France, pp. 3007–3011, August 2013.

VI Ulpu Remes, Ana Ramírez López, Kalle Palomäki and Mikko Kurimo. Bounded conditional mean imputation with observation uncertainties

and acoustic model adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 23, issue 7, pp. 1198–1208, July 2015.

VII Hannu Pulakka, Ulpu Remes, Kalle Palomäki, Mikko Kurimo and Paavo Alku. Speech bandwidth extension using Gaussian mixture model-based estimation of the highband mel spectrum. In *ICASSP*, Prague, Czech Republic, pp. 5100–5103, May 2011.

VIII Hannu Pulakka, Ulpu Remes, Santeri Yrttiaho, Kalle Palomäki, Mikko Kurimo and Paavo Alku. Bandwidth extension of telephone speech to low frequencies using sinusoidal synthesis and Gaussian mixture model. *IEEE Transactions on Audio, Speech, and Language Processing*, volume 20, issue 8, pp. 2219–2231, October 2012.

Author's Contribution

Publication I: "Sparse imputation for large vocabulary noise robust ASR"

The author conducted the cluster-based imputation experiments and wrote the sections on cluster-based imputation and baseline automatic speech recognition system (Section 3, Section 5.1, Section 5.5).

Publication II: "Observation uncertainty measures for sparse imputation"

The author contributed in planning and conducting the experiments. The paper was written in collaboration with the coauthors.

Publication III: "Missing-feature reconstruction with a bounded nonlinear state-space model"

The author conducted most experiments and was the main writer. The experiments were planned and conclusions were made in collaboration with the coauthors.

Publication IV: "GMM-based missing-feature reconstruction on multi-frame windows"

The author conducted the experiments and was the main writer. The experiments were planned in collaboration with the coauthors.

Publication V: "Bounded conditional mean imputation with an approximate posterior"

The author planned and conducted the work.

Publication VI: "Bounded conditional mean imputation with observation uncertainties and acoustic model adaptation"

The author planned and conducted most experiments and was the main writer.

Publication VII: "Speech bandwidth extension using Gaussian mixture model-based estimation of the highband mel spectrum"

The author trained a regression model that was used in the experiments and wrote the model description (Section 2.2).

Publication VIII: "Bandwidth extension of telephone speech to low frequencies using sinusoidal synthesis and Gaussian mixture model"

The author trained a regression model that was used in the experiments and wrote the model description part in Section II-C.

List of symbols

i	hidden state index
k	variable index
$\boldsymbol{m}$	partition to recorded and unrecorded variables
n	observation index
$\boldsymbol{n}$	noise variables
N	number of observations
T	number of observations in a window
$\boldsymbol{u}$	unobserved variables
W	word sequence
$\boldsymbol{x}$	data variables
$\boldsymbol{x}_r$	recorded variables
$\boldsymbol{x}_u$	unrecorded variables
$\hat{\boldsymbol{x}}$	point estimate
$\boldsymbol{x}'$	observation vector in acoustic-model domain
X	observation vector sequence in acoustic-model domain
$\boldsymbol{y}$	observed variables
Z	acoustic model state sequence
$\boldsymbol{z}$	hidden variables
ζ	acoustic model state
θ	outcome variables
$\boldsymbol{\mu}$	model mean
ν	acoustic model component
$\boldsymbol{\sigma}$	model variance
$\hat{\boldsymbol{\sigma}}$	observation variance
ϕ	observation model parameters

List of Abbreviations

ABE	artificial bandwidth extension
ASR	automatic speech recognition
BCMI	bounded conditional mean imputation
CI	cluster-based imputation
CMLLR	constrained maximum likelihood linear regression
GMM	Gaussian mixture model
HMM	hidden Markov model
LER	letter error rate
LVCSR	large-vocabulary continuous speech recognition
MAP	maximum a posteriori
MFA	mixture of factor analysers
MI	multiple imputation
MMSE	minimum mean square error
NSSM	nonlinear state–space model
SI	sparse imputation
SNR	signal-to-noise ratio
TPMI	truncated posterior mean imputation

1. Introduction

1.1 Background and research environment

The work presented in this thesis relates to automatic speech recognition (ASR) research. ASR systems transcribe speech to text. The system used in this work assumes that the speech observations are recorded in a quiet environment. However, since quiet environments are not common, most speech observations contain environmental noise when the observation conditions are not constrained. Transcriptions estimated based on noise-corrupted observations are not as accurate as transcriptions estimated based on clean speech observations that do not contain noise. However, it is possible to use a noise compensation method to improve transcription accuracies in unconstrained ASR.

Noise-corrupted speech data can be modelled as a speech and noise mixture. How speech and noise contribute to the noise-corrupted observations can be represented with an interaction model. The methods studied in this thesis model noise-corrupted observations as incomplete clean speech data [1, 2]. The methods assume that noise-corrupted observations can be divided into speech-dominated and noise-dominated spectrotemporal components so that the speech-dominated components correspond to clean speech observations and noise-dominated components correspond to environmental noise. Hence we do not observe the speech data in the noise-dominated components, and noise causes information loss rather than distortion.

The missing-data methods used in ASR include marginalisation and imputation-based approaches. Marginalisation and bounded marginalisation approaches calculate acoustic model likelihoods based on the incomplete observations, whereas conventional imputation approaches sub-

stitute the noise-corrupted components with clean speech estimates [1, 2]. The imputation methods used in ASR include front-end imputation methods and class-conditioned imputation methods. The class-conditioned imputation methods use an imputation model that has the same classes as the acoustic model used in ASR, whereas front-end imputation methods do not assume the same class structure. The methods studied in this thesis are front-end imputation methods.

ASR and missing-data imputation are the main research topic in this thesis. Another topic is spectral envelope extension. Envelope extension is used in artificial bandwidth extension (ABE) methods. These can be used in mobile phones to enhance speech quality for human listeners. The experiments reported in this work evaluate bandwidth extension on mobile telephone speech.

The research presented in this work has been conducted at Aalto University between 2007–2015 and continues previous work on automatic speech recognition and bandwidth extension. The previous work on automatic speech recognition includes research on phonetic classification using neural network methods [3, 4] and large-vocabulary continuous speech recognition (LVCSR) in Finnish [5, 6, 7]. Applications that have been studied include spoken document retrieval [8] and medical dictation [9]. Recent research directions also include robustness to environmental noise [10, 11]. Previous work on bandwidth extension includes the research presented in [12, 13].

The imputation methods studied in this thesis relate to the research presented in [10, 11]. Keronen [10] studies mask estimation methods that partition noise-corrupted observations into speech-dominated and noise-dominated components. Mask estimation thus determines the incomplete observations that can be reconstructed with imputation methods. Kallasjoki [11] studies sparse separation which is alternative to imputation approaches. While imputation methods model noise-corrupted speech data as incomplete clean speech, sparse separation models the observations as linear speech and noise mixtures. Hence sparse separation calculates clean speech estimates based on speech and noise codebooks whereas imputation methods calculate clean speech estimates based on clean speech data.

1.2 Objectives and scope

The problems studied in this thesis include missing-data imputation in ASR and spectral envelope extension in ABE. ASR uses speech data to calculate acoustic model likelihoods which are converted into text. The experiments conducted in this work evaluate ASR on noise-corrupted observations that are modelled as incomplete clean speech data and reconstructed with missing-data imputation. The research aims to find imputation methods that best remove errors in acoustic model likelihood calculation and improve transcription accuracies. A further aim was to ensure that the approaches are compatible with speaker-based adaptation. This is because it is common to use speaker-based adaptation to improve transcription accuracies in LVCSR.

The work presented in this thesis focusses on single-channel speech data and additive noise corruption. Additive noise includes environmental sounds such as announcement chimes and chatter in public environments or the hum inside a car. Reverberation and other convolutive noises are not considered in this work. The work concentrates on large vocabulary speech recognition while previous works on imputation had focussed on systems that use limited vocabularies.

The experiments conducted in this work evaluate imputation methods as noise compensation approaches. However other noise compensation approaches, such as model-based approaches or feature-based approaches that do not assume the same interaction model as imputation, are not studied. The chosen imputation approaches are front-end based methods that do not utilise the acoustic model used in ASR. The alternative imputation approaches would have constrained feature selection and complicated speaker-based adaptation.

The bandwidth extension experiments reported in this work focussed on narrowband telephone speech and bandwidth extension methods that estimate the wideband spectrum in unrecorded and attenuated frequencies based on the narrowband observations.

1.3 Research approach

The experiments conducted in this work evaluated and compared imputation methods under diverse noise conditions. The baseline imputation method considered in this work was cluster-based imputation [2]. The re-

search approach was to test imputation approaches that (1) use temporal dependencies between consecutive observations and (2) model the uncertainties associated with estimated values. Related to temporal dependencies, experiments were conducted to compare window widths in window-based sparse imputation and cluster-based imputation, and to evaluate an imputation method that uses a hidden source variable to model the dependencies between consecutive observations. The uncertainties associated with estimated values were calculated based on the imputation model and introduced in acoustic model likelihood calculation as so called observation uncertainties [14, 15]. The uncertainties can conceal estimation errors.

Experimental evaluation was conducted with a hidden Markov model (HMM) based LVCSR system developed at Aalto University. The system uses morpheme-like subword units which allow it to recognise all words and word forms [16, 17]. The system was evaluated on Finnish SPEECON data [18]. The experiments used real and simulated noise-corrupted speech data and evaluation was based on standard evaluation measures.

The bandwidth extension experiments were conducted with methods that estimate wideband observations based on narrowband observations. The current thesis focusses on the envelope extension part where the approach was to train a regression model based on parallel narrowband and wideband data.

1.4 Research process and thesis structure

This thesis includes an introduction part and publications PI–PVIII. Publications I–VI evaluate missing-data imputation as a noise compensation method in LVCSR. The imputation approaches studied in Publications I–VI include sparse imputation [19] and model-based imputation approaches. Publications I–II focus on sparse imputation. Publication I compares window-based sparse imputation [20] and cluster-based imputation and studies temporal dependencies in sparse imputation. Publication II presents heuristic measures to estimate the uncertainties in observations reconstructed with sparse imputation. The research extends co-authors' previous work on sparse imputation [19, 20].

Publications III–VI focus on imputation methods that estimate the complete observations based on a statistical distribution model. Publications

III–IV present model-based imputation methods that use temporal dependencies while Publications V–VI evaluate model-based imputation and estimation uncertainties. Temporal dependencies between baseline observation vectors were modelled as nonlinear dependencies between hidden state variables (PIII) or the baseline observations were extended to include temporal context and hidden state variables were used to reduce model dimension (PIV). The research presented in PIII is based on co-authors' previous work on nonlinear model-based imputation [21] and the research presented in PIV continues co-authors' previous work on factor analysis in speech applications [22].

Publications VII–VIII evaluate bandwidth extension methods. The research continues co-authors' previous work on bandwidth extension for narrowband telephone speech [23, 24].

The introduction part in this thesis provides an overview on the research topic and contributions. Chapter 2 discusses the theoretical foundations that motivate the imputation approaches studied in this work. Chapter 3 introduces ASR and ABE, and discusses previous work on imputation-based ASR and spectral envelope extension. Chapter 4 introduces the publications included in this thesis and summarises the research contribution in each publication. The contributions are discussed in Chapter 5 and conclusions are presented in Chapter 6.

2. Missing-data methods

2.1 Motivation

Missing values are a common occurrence in studies that involve data collection. Missingness can relate to various causes such as nonresponse in questionnaires, transmission errors, or environmental disturbances. For example, when a satellite is used to observe land conditions, the observed land-area data is incomplete when clouds occlude the land area under observation.

The current section discusses parameter estimation and prediction based on incomplete data. Parameter estimation is most related to observational studies conducted to estimate population statistics or causal dependencies such as response to medication, while accurate predictions are needed in various applications. Examples include automated control in industrial processes and speech applications such as ASR and ABE. ASR calculates an estimated transcription based on speech observations and ABE calculates wideband speech estimates based on narrowband speech observations.

Parameter estimation and prediction aim to minimise the estimation or prediction error. For example, parameter estimation aims to minimise the error between a parameter value θ and a parameter estimate $\hat{\theta}$ that is calculated based on N observation vectors $\boldsymbol{x}(n)$. The error includes bias and variance components as illustrated in Figure 2.1. Here we separate between $\hat{\theta}$ that indicates a parameter estimate calculated based on N observations $\{\boldsymbol{x}(n)\}$ and $E\{\hat{\theta}\}$ that indicates an expectation value when the parameter estimate is calculated based on an N-observation dataset that is modelled as a random variable. To minimise the estimation error, parameter estimation must aim to minimise the bias that is calculated as

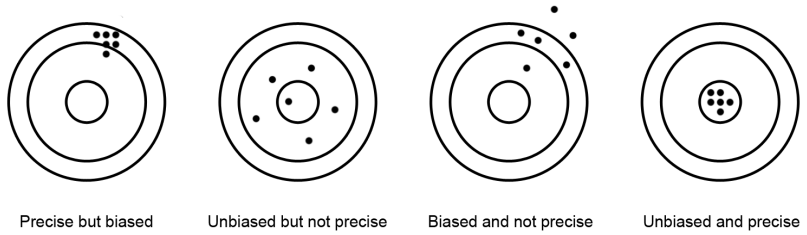

Figure 2.1. Estimation errors can be decomposed into bias and variance components. Unbiased estimates are centred around the correct value and precise estimates have low variance.

the error between $E\{\hat{\theta}\}$ and θ and the variance that is calculated as the error between $\hat{\theta}$ and $E\{\hat{\theta}\}$.

Estimated outcome values can be biased due to errors in estimation process or when the observations $\{\boldsymbol{x}(n)\}$ do not have the same statistical properties as the complete population. For example, when incomplete observations arise due to a selective process where observations with certain properties are removed or become incomplete, complete and incomplete observations do not have the same statistical properties, and estimates calculated based on incomplete data can be biased. The variance or standard error measures the variation between estimates calculated based on individual datasets, which means that in most cases, an increase in sample size N decreases the variance. Since an incomplete dataset contains less data than a complete dataset, estimates calculated based on N incomplete observations are less accurate than estimated calculated based on N complete observations.

Most parameter estimation and prediction methods cannot process incomplete data. Approaches that overcome the issue are called missing-data methods. Conventional methods include complete-case approaches that remove incomplete observations prior to outcome estimation and imputation approaches that substitute missing values with reasonable estimates [25, 26]. Hence the methods produce a reduced or reconstructed dataset that does not contain incomplete observations and can be processed with standard parameter estimation or prediction methods. Alternatives include approaches that calculate parameter estimates or predictions based on incomplete data. For example, parameter estimation approaches based on expectation–maximisation [27] and data augmentation [28] are discussed in [25, 29] and pattern classification based on incomplete data is discussed in [30].

The current section focusses on imputation methods and method selection when the reconstructed dataset is used in parameter estimation or prediction. The methods utilise an incomplete-data model that is derived based on a complete-data model and dependencies between complete and incomplete observations. While parameter estimation and prediction assume certain complete-data model, imputation methods can use a complete-data model that does not correspond to the complete-data model used in the end application. An appropriate model depends on the incomplete observations and the complete-data model used in the end application. The aim in model selection is to ensure accurate outcome estimates.

2.2 Missing-data pattern

The complete and incomplete observations considered in this thesis can be modelled as a multivariate random variable $\boldsymbol{x}$ whose realisations are denoted as $\boldsymbol{x}(n)$. The variables represented in $\boldsymbol{x}$ are indexed with k and we assume that each variable has a realisation $x(n,k)$ in each observation n. Hence there is no structural missingness, but incomplete observations arise due to an observation or data collection process where some realisations are not recorded.

We assume that each observation vector $\boldsymbol{x}(n)$ is partitioned into recorded values $\boldsymbol{x}_r(n)$ and unrecorded values $\boldsymbol{x}_u(n)$. The partition can be represented as a binary indicator vector $\boldsymbol{m}(n)$ whose components indicate whether a variable is recorded or unrecorded in observation n. We note that when unrecorded values are substituted with erroneous data or when observations are censored, the partition has to be estimated based on the available observations and context information. Examples on error detection based on domain expertise and distributional anomalies are discussed in [31]. Error detection divides observations into reliable values that are interpreted as recorded data and unreliable values that indicate unrecorded data.

The partitions $\{\boldsymbol{m}(n)\}$ that exist in an incomplete dataset $\{\boldsymbol{x}(n)\}$ determine a missing-data pattern. The patterns can be monotone or nonmonotone. A pattern is monotone when the variables represented in $\boldsymbol{x}$ can be reordered and indexed so that (1) when variable k is unrecorded, variables $l > k$ are unrecorded, and (2) when variable l is recorded, variables $k < l$ are recorded. Examples include data collected in longitudinal panel studies where incomplete observations arise when participants drop out

[32]. Monotone patterns are also common in studies that utilise planned missingness to reduce the costs associated with data collection [33]. The common patterns include a condition where observation is either complete or incomplete with a constant partition to recorded and unrecorded variables. For example, when predictive model parameters are estimated based on annotated and unannotated data, missingness is constrained to the outcome variables which are recorded in the annotated data and unrecorded in the unannotated data [34, 35].

Nonmonotone patterns include unconstrained and constrained patterns. When missingness follows a monotone pattern or a constrained nonmonotone pattern, the incomplete data distribution can be factorised into complete and incomplete data distributions with distinct parameters. This means that certain parameters can be estimated with complete-data methods [36, 37], which is not possible when missingness follows an unconstrained pattern where each observation $x(n, k)$ can be recorded or unrecorded. The current work focusses on incomplete data where the unrecorded values arise in an unconstrained pattern.

2.3 Missing-data mechanism

Rubin [38] proposed to model the indicator vector $\boldsymbol{m}$ as a random variable whose joint distribution with variables $\boldsymbol{x}$ can be factorised into a complete-data model $p(\boldsymbol{x})$ and missing-data mechanism $p(\boldsymbol{m}|\boldsymbol{x})$. The complete-data model aims to explain how the data vectors are generated while the missing-data mechanism explains the observation process. While we cannot assume that the observation process can be modelled, we can determine conditions in which the missing-data mechanism $p(\boldsymbol{m}|\boldsymbol{x})$ is ignorable and the incomplete observations can be used in parameter estimation or prediction without the need to model the observation process. Alternative factorisations which are not considered in the current work include pattern-mixture models [39] and shared-parameter models [40].

The missing-data mechanism is considered ignorable when (1) missingness $\boldsymbol{m}$ is uninformative about the unrecorded values and (2) the parameters ϕ that control missingness are uninformative about the outcomes θ that the incomplete data is used to estimate [38]. When the outcomes are assumed deterministic and maximum-likelihood estimation is used, the second condition means that the parameters ϕ must not limit the possible parameter values or responses. When the outcomes θ are modelled as

random variables, the parameters ϕ must not provide prior information on the outcomes. However the second condition is seldom decisive since statistical dependencies between ϕ and θ are not common when missingness is uninformative about the unrecorded values. Thus in practice, we can assume that the statistical dependencies between missingness and unrecorded variables determine if the missing-data mechanism is ignorable.

When the partition $\boldsymbol{m}(n)$ is predetermined or selected at random, missingness depends neither on the recorded nor the unrecorded variables. The condition can be expressed as $p(\boldsymbol{m}|\boldsymbol{x}) = p(\boldsymbol{m})$. Since missingness is not selective, the complete and incomplete data vectors have the same distribution $p(\boldsymbol{x})$ and the missing-data mechanism is ignorable. In contrast, when missingness probabilities depend on recorded or unrecorded values, the observation process biases the incomplete data distribution. However, on the condition that missingness probabilities do not depend on the unrecorded variables when conditioned on the recorded variables, the dependencies between the recorded and unrecorded variables remain constant. Thus the dependencies between variables $\boldsymbol{x}_r$ whose values are recorded in observation n and variables $\boldsymbol{x}_u$ whose values are unrecorded in observation n are the same as dependencies between variables $\boldsymbol{x}_r$ and $\boldsymbol{x}_u$ in complete observations, and the variables in an incomplete observation n have the same conditional distribution $p(\boldsymbol{x}_u|\boldsymbol{x}_r)$ as variables in complete observations.

When the dependencies between missingness and unrecorded variables cannot be modelled as dependencies between missingness and recorded variables, the missing-data mechanism is considered nonignorable. Examples include data where observations $x(n,k)$ that do not exceed certain threshold values $c(n,k)$ are censored. This means that the variable k in observation n is recorded if $x(n,k) \geq c(n,k)$ and unrecorded otherwise. Examples on censored and coarsened data are discussed in [41]. Since missingness depends on the variable whose values are unrecorded, a statistical model estimated based on complete observation vectors cannot describe the incomplete data distribution and the observation process must be modelled in order to determine the incomplete data distribution. For example, in case the observation boundaries $c(n,k)$ are known and we know that the observations have been censored, the observation process can be modelled and the incomplete data distribution can be determined based on the complete-data model $p(\boldsymbol{x})$ as a truncated distribution model

$Tp(\boldsymbol{x})$.

The conditions in which the observation process can be modelled are not common. In practice, the model must be available as a priori information or additional data must be collected as discussed in [42]. This is because it is not possible to estimate the parameters ϕ based on incomplete observations. Indeed, incomplete observations cannot even indicate if missingness probabilities depend on the recorded or unrecorded values [43]. For example, assume we have a recommendation dataset where the recorded values are items, such as movies, that the users have rated and unrecorded values are items that the users have not rated. The available observations in this dataset do not indicate whether a user has not rated a movie because he has not seen it or because his opinion about it was such that he chose not to rate it. A comparison between recommendation datasets and control data is presented in [44].

Since it is not common that the observation process can be modelled, most missing-data methods choose to assume that missingness does not depend on the unrecorded variables when conditioned on the recorded variables. The assumption can seem unintuitive when missingness follows an unconstrained pattern, but since the conditional probabilities denote statistical rather than causal dependencies, conditional independence $p(\boldsymbol{m}|\boldsymbol{x}) = p(\boldsymbol{m}|\boldsymbol{x}_r)$ does not mean that the partition $\boldsymbol{m}(n)$ is decided based on the recorded values in observation n. Instead, conditional independence means that the dependencies between missingness and unrecorded variables do not need to be modelled because unrecorded values can be predicted based on the recorded values. To improve prediction accuracies, the data vectors $\boldsymbol{x}$ can be extended to include auxiliary variables that are informative about the variables needed in parameter estimation or prediction [45].

2.4 Missing-data imputation

This thesis focusses on missing-data methods that substitute unrecorded values with reasonable estimates. The process is called imputation or reconstruction. The approaches used in imputation or reconstruction include approaches that substitute unrecorded values with point estimates and approaches that model the uncertainties introduced in the reconstruction process. Alternatives to missing-data imputation include complete-case approaches that calculate parameter estimates based on a re-

duced dataset, and approaches that calculate parameter estimates or predictions based on incomplete data [25, 26].

The imputation methods considered in this work estimate unrecorded values based on the recorded data and statistical dependencies between recorded and unrecorded variables. The methods include model-based approaches that encode the dependencies between recorded and unrecorded data in a predictive model and instance-based approaches that substitute unrecorded values with the recorded values observed in another data vector. Alternatives that are not discussed in this work include zero imputation and mean imputation approaches that substitute unrecorded values with a constant that does not depend on the recorded values in observation n. Moreover the discussion on model-based approaches is limited to approaches that assume a statistical distribution model, and imputation approaches based on neural representations, matrix completion, or other alternatives are not discussed in this work.

2.4.1 Model selection

The statistical dependencies between recorded and unrecorded variables are captured in a complete-data model. The complete-data model that is used in imputation determines the qualities that are preserved in the reconstructed data. Thus when reconstructed data is passed on to an end application, inconsistencies between the imputation and end application models can introduce bias in the application outcomes [46, 47]. Moreover, if conditional independence is assumed between missingness and unrecorded values and the missing-data mechanism is ignored, the imputation model must capture the dependencies between recorded and unrecorded variables so that missingness does not provide additional information about the unrecorded values and conditional independence can be assumed.

The complete observations assumed in imputation can include auxiliary variables that are not used in the end application [45]. For example, assume we have an application that operates on optical satellite data where incomplete observations arise due to cloud cover. While this end application estimates outcomes based on optical data, imputation can utilise observations that are extended with microwave data [48]. Auxiliary variables can enhance precision and reduce bias in case a nonignorable missing-data mechanism is assumed ignorable [45]. However redundancies between variables can introduce noise in the predictive model

parameters and the variables also increase data dimension and computational cost, and there are conditions in which the variables induce bias in parameter estimates [49].

The model-based imputation approaches considered in this work model the recorded and unrecorded values in observation n as a random variables $\boldsymbol{x}_r$ and $\boldsymbol{x}_u$ and estimate the unrecorded values based on posterior predictive distribution $p(\boldsymbol{x}_u|\boldsymbol{x}_r, \boldsymbol{m})$. When missingness is uninformative about the unrecorded values, the posterior predictive distribution reduces to conditional distribution $p(\boldsymbol{x}_u|\boldsymbol{x}_r)$. The conditional distribution is calculated based on a complete-data model $p(\boldsymbol{x})$ whose parameters are estimated based on the complete data vectors. When observations have been censored, the posterior predictive distribution is calculated as a truncated conditional distribution $Tp(\boldsymbol{x}_u|\boldsymbol{x}_r)$.

The statistical distributions models evaluated in this work are based on the multivariate normal distribution. A multivariate normal distribution $N(\boldsymbol{x})$ is practical because its conditional distributions are multivariate normal distributions, and when observations are censored, the incomplete data distribution is a truncated multivariate normal distribution whose conditional distributions are truncated multivariate normal distributions [50]. When a multivariate normal distribution is too restrictive, observations can be associated with a hidden variable $\boldsymbol{z}$ so that the conditional distributions $p(\boldsymbol{x}|\boldsymbol{z})$ can be modelled as multivariate normal distributions $N(\boldsymbol{x}|\boldsymbol{z})$ while the complete observation model has more structure. The statistical distribution models evaluated in this work use such hidden variables.

The hidden variable $\boldsymbol{z}$ can be interpreted as a model state or source representation. A discrete hidden variable can be introduced to model clustered data while continuous hidden variables are used to remove redundancies and summarise covariance structures [51]. Hence a continuous hidden variable that corresponds to a low-dimensional source representation can reduce the model dimension and limit the increase in parameter count when observations $\boldsymbol{x}$ are extended to include auxiliary variables. When the hidden variable follows a multivariate normal distribution, nonlinearities can be modelled as a nonlinear transformation between the observed and hidden variables. Furthermore, sequential or hierarchical structures can be modelled as dependencies between the hidden variables (Figure 2.2).

Alternative to model-based imputation are instance-based approaches

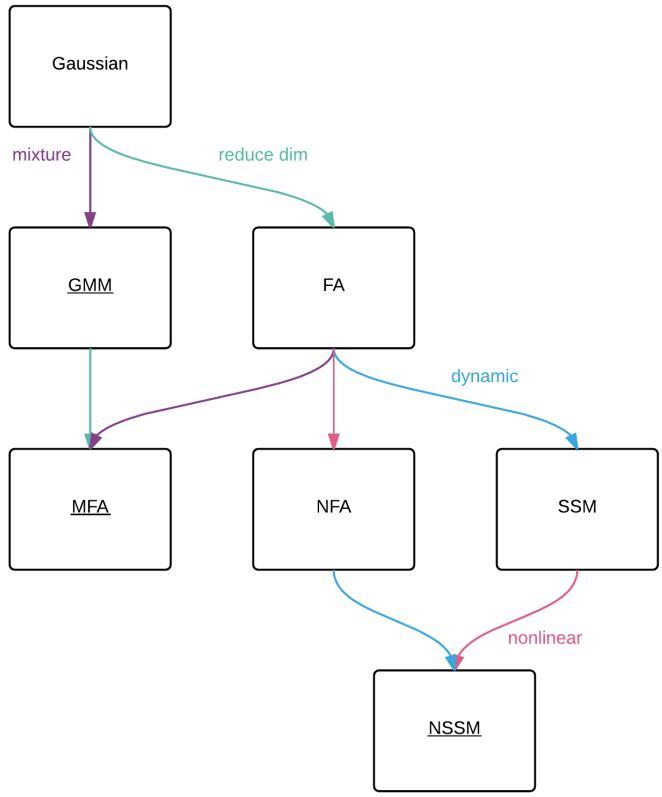

Figure 2.2. The model-based imputation approaches considered in this work (underlined) assume that complete observations follow a multivariate normal distribution when conditioned on a hidden variable. Gaussian mixture model (GMM) associates observations with a discrete hidden variable that partitions observations into clusters, whereas linear factor analysis (FA) associates observations with a representation in low-dimensional source domain. Dependencies between observations and sources are assumed linear and dependencies between source variables are modelled with a multivariate normal distribution. GMM and FA can be extended into a mixture model that partitions observations into clusters and associates each cluster with a low-dimensional source domain. Nonlinearities can be modelled with nonlinear factor analysis (NFA) and time series data can be modelled with state–space models (SSM). Furthermore, a nonlinear state–space model (NSSM) assumes nonlinear dependencies between the observation and source vectors and between consecutive source vectors. The visualisation is based on a more extensive chart presented in [52].

that substitute unrecorded values with recorded values available in complete data vectors [53]. When the missing-data mechanism is ignorable, substitutes are searched within complete data vectors that are considered close to the incomplete observation. The complete data vectors thus comprise a search base or a codebook that is narrowed down to one substitute vector or an active subset based on the recorded values. The search base is comparable to a complete-data model whose size can be controlled to reduce the computational cost associated with the search operation [54]. Moreover when the observations have been censored, the search base can be constrained to complete data vectors where variables x_u are within the censored interval.

Instance-based approaches model the assumed dependencies between recorded and unrecorded variables in the search criteria that determine the active subset and the estimates that substitute the unrecorded values. Since the methods do not impose a statistical distribution model on the data, imputation can preserve non-standard and non-smooth data distributions.

2.4.2 Estimation

The posterior predictive distribution calculated based on a complete-data model and the recorded values in observation n does not associate the unrecorded values in observation n with point estimates. The estimates that can substitute the unrecorded values are calculated based on the posterior predictive distribution and a cost function. Estimates include minimum mean square error (MMSE) estimates that are calculated as the posterior distribution mean and maximum a posteriori (MAP) estimates that are calculated as the posterior distribution mode. MMSE estimation minimises a quadratic cost function where estimation errors are associated with a cost that is proportional to the squared estimation error whereas MAP estimation minimises a cost function where estimation errors have a fixed cost.

Estimates that minimise an expected reconstruction error are appropriate in various applications but are not guaranteed to minimise estimation errors in derivatives calculated based on the reconstructed data. For example, when unrecorded values are substituted with the predictive distribution mean that minimises the expected reconstruction error, correlation between variables is overestimated and standard errors are underestimated in statistics calculated based on the reconstructed data. To

preserve variation between observations, unrecorded values can be substituted with a random draw rather than distribution mean, or the predictive distribution mean can be used as a search criteria in instance-based imputation [55, 46]. The restored variation removes the error in correlation estimates and increases standard errors, but the standard errors remain underestimated as the issue cannot be solved with conventional imputation approaches.

Parameters estimated based on reconstructed data are associated with underestimated standard errors because uncertainties related to the imputation process are not taken into account. Rubin [56] proposed to overcome the issue with multiple imputation (MI). MI assumes that the unrecorded values and predictive distribution parameters are modelled as random variables whose posterior distributions are simulated with numerical methods. The simulation process is repeated m times to associate unrecorded values with m alternative imputations calculated based on m alternative distribution parameters [25, 29]. When parameter estimation is conducted on the reconstructed data, variation between parameter estimates calculated based on the m reconstructed datasets captures the uncertainties introduced in imputation. Alternatives include resampling approaches where the m parameter estimates are calculated based on reconstructed datasets that corresponds to m resampled incomplete datasets [57, 58] and parameter estimation that takes into account the posterior predictive distribution mean and variance [59].

The uncertainties considered in this work correspond to the posterior predictive distribution variance. We note that imputation uncertainties calculated as posterior predictive distribution variances are approximate because the distribution parameters are estimated based on the complete data vectors and uncertainties introduced in parameter estimation are not modelled. However the uncertainties related to distribution parameters are not expected to have a notable contribution in the imputation uncertainties when the predictive distribution parameters are estimated based on a sufficiently large dataset [59]. We also note that while conventional imputation and multiple imputation approaches produce reconstructed datasets that can be processed with standard parameter estimation and classification methods, when the unrecorded values are substituted with a mean and variance, the end application needs to operate on uncertain observations.

2.5 Summary

The current section discussed incomplete observations and missing-data imputation. The incomplete observation were associated with a missing-data pattern and missing-data mechanism. The missing-data pattern describes partitions between the recorded and unrecorded values whereas the missing-data mechanism describes dependencies between missingness and the recorded and unrecorded variables. The research presented in this work studies speech data represented in a compressed spectral domain where missingness can relate to speech transmission or environmental noise. The research focusses on noise-corrupted speech data where the unrecorded values arise in an unconstrained pattern and correspond to censored data.

When the reconstructed observations are used in parameter estimation or prediction, imputation should aim to minimise the estimation or prediction error and the end application should be taken into account when the imputation model and estimation method are selected. However, the complete data model used in imputation is not constrained to model the dependencies used in end application. Instead, imputation can utilise additional dependencies to reduce reconstruction error or exclude dependencies to reduce noise. The experiments conducted in this work evaluate whether temporal dependencies between consecutive observations improve imputation. The experiments also evaluate whether imputation uncertainties improve transcription accuracies when imputation is used as a noise compensation method in LVCSR. The applications and imputation methods studied in this work are introduced in more detail in the next section.

3. Applications

3.1 Automatic speech recognition

The current section introduces the automatic speech recognition system used in PI–PVI and discusses previous work on missing-data imputation in ASR.

3.1.1 System overview

ASR converts speech into a written transcription. The process can be partitioned into front-end operations that convert speech into acoustic observation vectors $\boldsymbol{x}'(n)$ and back-end operations that convert observations into words. The conversion can utilise statistical dependencies learned based on speech and text data. Dependencies between observations and words are represented in an acoustic model while dependencies between consecutive words are represented in a language model. Since the dependencies between consecutive words are calculated based on text data and are not affected when observations are corrupted with environmental noise, the research presented in this thesis focusses on the acoustic model. The current section introduces the acoustic observation vectors and acoustic model used in PI–PVI.

The acoustic observation vectors $\boldsymbol{x}'(n)$ contain features calculated based on speech data that is captured with a microphone and processed in short time windows. The main operations in common feature calculation front-ends include spectral envelope estimation and normalisation. Envelope estimation means that the observations in time frame n are modelled in spectral domain and associated with an envelope representation $\boldsymbol{x}(n)$. The envelope representation used in PI–PVI is a smoothed spectrum that is calculated based on mel-scale sub-bands and log-compressed.

The features $\boldsymbol{x}'(n)$ used in acoustic likelihood calculation in PI–PVI are calculated based on the envelope estimates $\boldsymbol{x}(n)$. The envelope estimates that model speech data in a compressed spectral domain are mapped into mel-scale cepstrum [60] and time normalised with window-based cepstral mean normalisation [61]. The feature vectors $\boldsymbol{x}'(n)$ are also extended with differential features [62] and decorrelated with a maximum likelihood linear transformation [63]. Alternative feature representations are discussed in [64].

The acoustic observation vectors $\boldsymbol{x}'(n)$ calculated at consecutive time indices n constitute an observation sequence that corresponds to an unobserved word sequence. The acoustic observation sequence is modelled as random variable X and the transcription as random variable W. Furthermore, X and W are extended with an intermediate representation Z that associates each observation $\boldsymbol{x}'(n)$ with a discrete state variable $\zeta(n)$. The system evaluated in PI–PVI aims to determine the transcription that maximises

$$\hat{W} = \arg\max_{W}\{\max_{Z} P(W)P(Z|W)P(X|Z)\}, \tag{3.1}$$

where $P(W)$ associates transcriptions with a priori probabilities while $P(Z|W)$ associates transcriptions with state sequences Z and $P(X|Z)$ associates state sequences with acoustic observations X [65]. Alternatives include approaches that estimate the transcription $\hat{W}$ based on a discriminative model $P(W|X)$ [66].

The state transcription probabilities $P(Z|W)$ and acoustic observation probabilities $P(X|Z)$ are modelled in the acoustic model. The acoustic model used in PI–PVI encodes the state and observation probabilities in a HMM [67, 68]. The word-level transcriptions W are converted into phoneme-level transcriptions where the units are context-dependent phonemes [69]. The conversion is based on pronunciation rules. The phonetic units are associated with acoustic observations and the observation sequences associated with each unit are modelled as HMMs whose states are tied across units [70]. The HMMs associated with the phonetic units can be concatenated to determine the HMM representation associated with each transcription W, and the probabilities $P(Z|W)$ and $P(X|Z)$ can be calculated based on the state sequence and acoustic observation probabilities that relate observations and phonetic units. A comprehensive introduction into HMM-based ASR is provided in [71].

HMM-based acoustic models associate phonetic units with a state structure, state transition probabilities, and state-conditioned observation prob-

abilities. The dependencies between consecutive states ζ are assumed to capture the dependencies between consecutive observation vectors so that the observation probabilities

$$P(X|Z) = \prod_n p(\boldsymbol{x}'(n)|\zeta(n)), \tag{3.2}$$

where $p(\boldsymbol{x}'|\zeta)$ denotes observation probabilities conditioned on HMM state ζ. The state-conditioned observation probabilities are modelled as Gaussian mixture densities whose parameters include state-conditioned component probabilities and component-conditioned distribution parameters. The component-conditioned distribution parameters include the mean vector and covariance matrix that determine a multivariate normal distribution. The covariance matrices are assumed to be diagonal due to the decorrelation included in feature normalisation. Alternatives to GMM-based acoustic models include neural network representations [72, 73, 74] and instance-based representations [75, 76].

The acoustic model parameters are learned based on acoustic observation sequences $\{X\}$ and transcriptions $\{W\}$ which are converted into phoneme-level transcriptions that utilise context-dependent phonemes. Since the model parameters are learned in a data-driven manner, the observation sequences $\{X\}$ that we use in parameter estimation determine the acoustic patterns accepted within each state ζ. The acoustic model parameters in PI–PVI are estimated based on speech data that is recorded in quiet conditions so that the state-conditioned distribution parameters do not model environmental noise. Alternatives include multicondition data where various noise conditions are represented to ensure that the state-conditioned distributions capture environmental variation and activate based on noise-corrupted observations. Advanced parameter estimation methods use multicondition data with environmental noise compensation [77, 78].

To summarise, we have discussed the feature calculation front-end and acoustic model used in PI–PVI. The HMM-based acoustic model determines (1) state-conditioned observation probabilities that are used in acoustic likelihood calculation and (2) state transition probabilities that are needed to determine the most probable state sequence $\hat{Z}$ and word sequence $\hat{W}$. The search operation that determines the state and word transcriptions also takes into account word sequence probabilities $P(W)$ (Equation 3.1). To evaluate word sequence probabilities, the system used in PI–PVI decomposes words into morpheme-like subword units. The

units are learned based on text data in an unsupervised manner [79] and the dependencies between consecutive subword units are modelled as n-gram probabilities [80]. The decoder that searches the most probable transcription is presented in [81] and alternative approaches are discussed in [82].

3.1.2 Environmental noise

The experiments reported in PI–PVI evaluate ASR in conditions that do not correspond to the conditions represented in the clean speech dataset based on which the acoustic model parameters have been estimated. Hence the environmental conditions introduce unwanted variation in the evaluation data. This variation is called noise. Noise corruption can be additive or convolutive in the time domain. Additive noise includes environmental sounds whereas convolutive noise includes reverberation and distortions introduced in the speech transmission channel.

The current section discusses acoustic likelihood calculation when speech is corrupted with additive noise. The noise-corrupted speech data is represented in the compressed spectral domain as $\boldsymbol{y}(n)$ and in the acoustic-model domain as $\boldsymbol{y}'(n)$. Since the noise-corrupted observations $\boldsymbol{y}'(n)$ do not have the same statistical properties as clean speech observations $\boldsymbol{x}'(n)$, the acoustic model parameters estimated based on clean speech observations do not model the noise-corrupted observations. To calculate state-conditioned likelihood scores based on an observation $\boldsymbol{y}'(n)$, the unobserved clean speech representation $\boldsymbol{x}'(n)$ is introduced in the likelihood calculation as a hidden variable,

$$p(\boldsymbol{y}'(n)|\zeta) = \int p(\boldsymbol{y}'(n)|\boldsymbol{x}'(n), \zeta) p(\boldsymbol{x}'(n)|\zeta) d\boldsymbol{x}'(n), \tag{3.3}$$

where $p(\boldsymbol{y}'|\boldsymbol{x}', \zeta)$ models statistical dependencies between acoustic observation vectors $\boldsymbol{y}'(n)$ and $\boldsymbol{x}'(n)$ that are associated with acoustic model state ζ.

The observation model $p(\boldsymbol{y}'|\boldsymbol{x}', \zeta)$ can be estimated based on parallel clean and noise-corrupted speech data or derived based on a model that describes interaction between speech and noise. Feature-based approaches [83, 84, 85] assume that the dependencies between observations $\boldsymbol{y}'(n)$ and $\boldsymbol{x}'(n)$ do not depend on the acoustic model state, whereas model-based approaches [84, 86, 87] estimate state-conditioned dependencies. Alternatives include approaches that marginalise over the acoustic model parameters associated with noise-corrupted features [88, 89] and approaches

that calculate expected state-conditioned likelihood scores [90, 91]. The latter include front-end noise compensation approaches extended with observation uncertainties [14, 15]. For a comprehensive review and discussion, see [92].

The experiments discussed in this work use noise compensation extended with observation uncertainties [14, 15]. Observation uncertainties can be used when the observations $\boldsymbol{y}'(n)$ are calculated based on enhanced or reconstructed data. The observations $\boldsymbol{y}'(n)$ are considered clean speech estimates, and the dependencies between observations $\boldsymbol{y}'(n)$ and clean speech values $\boldsymbol{x}'(n)$ are modelled as $\boldsymbol{y}'(n) = \boldsymbol{x}'(n) + \boldsymbol{e}(n)$, where $\boldsymbol{e}(n)$ denotes estimation error. The error is modelled as a random variable whose distribution is a multivariate normal distribution with mean $E\{\boldsymbol{e}(n)\} = 0$ and variance $\Sigma\{\boldsymbol{e}(n)\} = \hat{\boldsymbol{\sigma}}'(n)$. When estimation errors are interpreted as observation uncertainties, we assume that the errors do not depend on the clean speech values $\boldsymbol{x}'(n)$ or clean speech estimates $\boldsymbol{y}'(n)$. Hence the observation model $p(\boldsymbol{y}'(n)|\boldsymbol{x}'(n))$ calculated based on the observation uncertainties model is a multivariate normal distribution with mean $\boldsymbol{x}'(n)$ and variance $\hat{\boldsymbol{\sigma}}'(n)$ and the evidence model $p(\boldsymbol{x}'(n)|\boldsymbol{y}'(n))$ a multivariate normal distribution with mean $\boldsymbol{y}'(n)$ and variance $\hat{\boldsymbol{\sigma}}'(n)$.

The system evaluated PI–PVI calculates the state-conditioned likelihoods $p(\boldsymbol{x}'(n)|\zeta)$ based on component-conditioned likelihoods $p(\boldsymbol{x}'(n)|\nu)$ that are modelled as multivariate normal distributions with mean $\boldsymbol{\mu}(\nu)$ and variance $\boldsymbol{\sigma}(\nu)$. The likelihood that mixture component ν produced the clean speech observation $\boldsymbol{x}'(n)$ is calculated as

$$p(\boldsymbol{x}'(n)|\nu) = N(\boldsymbol{x}'(n); \boldsymbol{\mu}(\nu), \boldsymbol{\sigma}(\nu)). \tag{3.4}$$

When the feature vectors $\boldsymbol{y}'(n)$ are clean speech estimates associated with observation uncertainties $\hat{\boldsymbol{\sigma}}'(n)$, the component-conditioned likelihoods are calculated as

$$p(\boldsymbol{y}'(n)|\nu) = \int N(\boldsymbol{x}'(n); \boldsymbol{\mu}(\nu), \boldsymbol{\sigma}(\nu)) N(\boldsymbol{x}'(n); \boldsymbol{y}'(n), \hat{\boldsymbol{\sigma}}'(n)) d\boldsymbol{x}'(n), \tag{3.5}$$

where $N(\boldsymbol{x}'; \boldsymbol{y}', \hat{\boldsymbol{\sigma}}') = N(\boldsymbol{y}'; \boldsymbol{x}', \hat{\boldsymbol{\sigma}}')$ is the observation or evidence model derived based on the observation uncertainties model. The component-conditioned likelihood has a closed-form solution,

$$p(\boldsymbol{y}'(n)|\nu) = N(\boldsymbol{y}'(n); \boldsymbol{\mu}(\nu), \boldsymbol{\sigma}(\nu) + \hat{\boldsymbol{\sigma}}'(n)). \tag{3.6}$$

We observe that the distributions $p(\boldsymbol{x}'|\nu)$ and $p(\boldsymbol{y}'|\nu)$ are centred at the same location, but the distributions associated with uncertain observations $\boldsymbol{y}'$ have an increased variance. The estimation error $\hat{\boldsymbol{\sigma}}'(n)$ varies

between observations $\boldsymbol{y}'(n)$ so that the component-conditioned likelihood $p(\boldsymbol{y}'(n)|\nu) = N(\boldsymbol{y}'(n); \boldsymbol{\mu}(\nu), \boldsymbol{\sigma}(\nu))$ when the observations are considered reliable and estimation error $\hat{\boldsymbol{\sigma}}'(n) = 0$ and $p(\boldsymbol{y}'(n)|\nu) \to 0$ when estimation error $\hat{\boldsymbol{\sigma}}'(n) \to \infty$. The uncertainties can also be taken into account in acoustic model parameter estimation [93, 94].

Equations (3.4)–(3.6) indicate that observation uncertainties can mitigate the effect of estimation errors introduced in front-end noise compensation. The decrease in component-conditioned likelihoods (Equation 3.6) decreases and smoothens the state-conditioned likelihoods $p(\boldsymbol{y}'(n)|\zeta)$ and pushes the decoder to emphasise state transition probabilities and word sequence probabilities when the acoustic observation vectors are calculated based on enhanced or reconstructed data. The uncertainties can be assumed proportional to heuristic measures or calculated based on the noise compensation model as posterior variances [95, 96]. The uncertainties can also be optimised based on test conditions to improve transcription accuracies [97, 98, 99]. Estimation approaches have been compared in [100]. Alternatives to observation uncertainties include likelihood calculation based on multiple clean speech estimates [101] and using frame-based uncertainties to post-process likelihoods calculated based on estimated features [102].

3.1.3 Missing-data methods

When speech is observed under environmental noise, the observation vectors do not match the acoustic classes trained on speech data recorded in quiet conditions. The experiments reported in PI–PVI assume that the noise-corrupted observations $\boldsymbol{y}(n)$ can be interpreted as incomplete clean speech data $\boldsymbol{x}(n)$. The motivation to interpret noise-corrupted observations as incomplete relates to the observation that human perception restores incomplete speech data when the unrecorded components are masked with environmental noise [103, 104, 105].

Since the observations are converted into state-conditioned acoustic likelihoods (Section 3.1.1), method selection is not limited to conventional imputation approaches that calculate reconstructed observations $\hat{\boldsymbol{x}}(n)$. For example, marginalisation or bounded marginalisation can be used in acoustic model likelihood calculation to determine the likelihoods based on incomplete data [1]. Furthermore, imputation approaches used in automatic speech recognition include class-conditioned imputation and front-end based imputation.

An imputation approach is called class-conditioned when the unrecorded values are estimated based on the acoustic model state-conditioned or component-conditioned distributions [106, 107]. Front-end based imputation approaches assume a separate imputation model and estimate reconstructed observations prior to acoustic model likelihood calculation [2]. Marginalisation and class-conditioned imputation do not have the model selection issue associated with conventional imputation methods, but require modifications in acoustic model likelihood calculation and constrain feature selection.

The current work focuses on front-end-based imputation approaches. Missing-data imputation under environmental noise requires error detection or mask estimation methods to determine the partition into recorded and unrecorded variables $\boldsymbol{m}(n)$ (Section 2.2). The unrecorded values are estimated based on the recorded values and a complete-data model $p(\boldsymbol{x})$, and the reconstructed data vectors are passed on to feature normalisation and acoustic likelihood calculation. The current subsection introduces common mask estimation and reconstruction approaches. Alternative approaches to environmental noise compensation are reviewed in [108, 109].

Mask estimation

Mask estimation approaches find erroneous observations and partition the noise-corrupted observations $\boldsymbol{y}(n)$ into reliable and unreliable components. The current work focusses on single-channel speech data and additive noise corruption. We assume that the observations are represented in a compressed spectral domain where interaction between speech and additive noise can be modelled with the max approximation [110, 111]. The observations $\boldsymbol{y}(n)$ are modelled as

$$\boldsymbol{y}(n) \approx \max\{\boldsymbol{x}(n), \boldsymbol{n}(n)\}, \tag{3.7}$$

where $\boldsymbol{x}(n)$ denotes the speech and $\boldsymbol{n}(n)$ the additive noise features time frame n, and we assume a component-wise $\max$ operation. Alternative interaction models are discussed and the max approximation model is derived in [112].

Since we assume that observations $y(n,k)$ correspond to clean speech features when $x(n,k) > n(n,k)$, speech-dominated components are considered reliable, whereas noise-dominated components are unreliable. Since $y(n,k)$ represents noise when $x(n,k) < n(n,k)$, it is not possible to determine the clean speech value $x(n,k)$ based on a noise-dominated observation $y(n,k)$. Hence the clean speech vectors $\boldsymbol{x}(n)$ derived based on

noise-corrupted observations $\boldsymbol{y}(n)$ are incomplete.

When speech and noise features are available as a priori information, an observation $y(n,k)$ can be marked reliable when $x(n,k) - n(n,k) > \gamma$, where γ is a parameter which can be hand-tuned to optimise system performance. The partition into reliable and unreliable components can be represented as a mask vector $\boldsymbol{m}(n)$. The partition $\boldsymbol{m}(n)$ is called an oracle mask because we calculate it based on speech features $x(n,k)$ and noise features $n(n,k)$ that are not available when methods are tested on real noise-corrupted speech data. Experiments are conducted with oracle masks to compare missing-data methods in ideal conditions and to evaluate methods' robustness to mask estimation errors.

Mask estimation approaches that operate on noise-corrupted observations include methods based on noise estimates, classification, and perceptual criteria [1, 113, 114]. A comprehensive review of single-channel mask estimation approaches is provided in [115]. Experiments reported in the current work calculate mask estimates $\hat{\boldsymbol{m}}(n)$ based on noise estimates $\hat{\boldsymbol{n}}(n)$. An observation $y(n,k)$ is marked reliable when $y(n,k) - \hat{n}(n,k) > \gamma$, and the estimated mask is post-processed to remove reliable areas that do not exceed a minimum size. The minimum size condition is intended to simulate human speech perception [116].

We note that the mask estimation and imputation approaches evaluated in this work do not take into account uncertainties related to mask estimation. The estimated masks are assumed accurate, even if a mask calculated based on the observed features and noise estimates is never error-free due to inaccuracies in noise estimation and unaccounted phase differences between speech and noise [117]. The uncertainties that arise in mask estimation can be modelled in missing-data reconstruction if estimated masks $\hat{\boldsymbol{m}}(n)$ are substituted with mask probabilities $p(\boldsymbol{m}|\boldsymbol{y}(n))$ [118, 119]. The probabilities can be calculated prior to reconstruction and classification, or mask estimation can be coupled with the reconstruction or classification tasks [120, 121].

Reconstruction

Front-end imputation methods estimate unrecorded values and calculate reconstructed features $\hat{\boldsymbol{x}}(n)$ based on a clean speech model that is not the acoustic model. The reconstructed features are calculated in the compressed spectral domain where the observed feature vector $\boldsymbol{y}(n)$ has reliable components represented as subvector $\boldsymbol{y}_r(n)$ and unreliable compo-

nents represented as $\boldsymbol{y}_u(n)$. Since reconstructed features $\hat{\boldsymbol{x}}(n)$ can be processed like complete observations, the features can be mapped into acoustic observation vectors $\hat{\boldsymbol{x}}'(n)$ with the standard feature calculation front-end and used in acoustic model likelihood calculation without modifications. However, the acoustic class probabilities calculated based on reconstructed data are overestimated unless the uncertainties introduced in imputation are modelled.

The most common model-based imputation methods assume that the clean speech features in time frame n are modelled as a random variable $\boldsymbol{x}$ and that statistical dependencies between features are modelled as a GMM with component-conditioned distributions $N(\boldsymbol{x}|i)$, where i denotes the mixture component. In the case that the missing-data mechanism is ignorable, the conditional distribution $p(\boldsymbol{x}_u|\boldsymbol{x}_r)$ that corresponds to the posterior predictive distribution can be calculated based on a GMM distribution in closed form. However, when incomplete observations arise due to additive-noise corruption, we know that the unrecorded values cannot exceed the observed features, $\boldsymbol{x}_u \leq \boldsymbol{y}_u(n)$. Thus the incomplete observations correspond to censored data and the posterior predictive distribution is a truncated conditional distribution $Tp(\boldsymbol{x}_u|\boldsymbol{x}_r)$ and the component-conditioned posterior predictive distributions are truncated normal distributions $TN(\boldsymbol{x}_u|\boldsymbol{x}_r, i)$.

The truncated conditional distribution parameters have closed-form solutions and the unrecorded values can be substituted with the truncated conditional distribution mean when the component-conditioned distributions $N(\boldsymbol{x}|i)$ have diagonal covariance matrices [119]. Furthermore, when component-conditioned distributions have diagonal covariance matrices, truncated conditional mean estimates can be calculated based on probabilistic masks. However, a diagonal-covariance model requires numerous components to capture the statistical dependencies between recorded and unrecorded data. Diagonal-covariance models are not evaluated in PI–PVI.

When component-conditioned distributions $N(\boldsymbol{x}|i)$ have full covariance matrices, component-conditioned truncated distribution parameters do not have closed-form solutions. The most common reconstruction method used in this condition is cluster-based imputation [2]. The method substitutes unrecorded values with approximate MAP estimates calculated based on the observed features and the complete-data model. The truncated distribution parameters are not needed since MAP estimation can

be formulated as a constrained optimisation problem and component-conditioned MAP estimates can be calculated based on the observed features and component-conditioned distribution parameters [2]. The experiments conducted in this work use cluster-based imputation as a baseline method in PI and PIII–PVI.

Alternatives to cluster-based imputation include bounded conditional mean imputation (BCMI) approaches that calculate approximate truncated distribution parameters and substitute unrecorded values with the truncated conditional distribution mean. BCMI approaches proposed in previous works calculate the approximate component-conditioned truncated conditional distribution parameters based on the component-conditioned conditional distribution parameters with axis-parallel approximation or diagonal-covariance approximation [122, 123]. BCMI approaches do not need iterative optimisation, and when unrecorded values are substituted with a truncated distribution mean, the uncertainties related to reconstructed features can be calculated as the distribution variance. The experiments conducted in this work compare cluster-based imputation and BCMI in PV–PVI.

Alternatives to model-based imputation include imputation approaches that model the features $\boldsymbol{x}(n)$ with a sparse representation [19, 124]. We assume that the complete features are represented as a linear combination

$$\boldsymbol{x}(n) = \sum_m a(n,m)\boldsymbol{b}(m), \tag{3.8}$$

where $\boldsymbol{a}(n)$ denotes a sparse activation vector in time frame n and $\boldsymbol{b}(m)$ are basis vectors. The experiments conducted in this work evaluate sparse imputation and compare sparse imputation and model-based imputation approaches in PI–PIV. The sparse imputation method evaluated in this work uses complete data vectors as basis vectors $\boldsymbol{b}(m)$ and estimates the sparse activation vector based on l_1-minimisation [19, 125]. The method does not model the observation process that constrains unrecorded values not to exceed the observed values. However, observations are processed in windows that span multiple time frames, which improves prediction accuracies, and the reconstructed observations are also post-processed not to exceed the observed values.

Most cluster-based imputation and bounded conditional mean imputation experiments have been conducted on observation vectors $\boldsymbol{x}(n)$ that include the compressed spectral envelope in time frame n. However, frame-based imputation is unreliable when recorded values are scarce, and mod-

elling temporal dependencies can improve imputation performance. To utilise temporal dependencies between consecutive time frames, model-based imputation methods can process the speech data in windows like sparse imputation or model the data as a time series. We evaluate window-based sparse imputation in PI and window-based model-based imputation in PIV and PVI. The window-based approaches use windows that span T complete time frames as proposed in [20]. Alternative approaches to window-based imputation include correlation-based imputation [2] and selective frame and channel-based imputation [126].

Alternatives to window-based imputation include recursive neural networks [127] and state–space approaches that model the observations as time series data and model the dependencies between consecutive observation vectors as dependencies between hidden state variables. For example, HMM-based imputation methods [128, 123] use state transition probabilities to model dependencies between discrete hidden states i. The state probabilities depend on state transition probabilities and observation probabilities in consecutive states, whereas state-conditioned posterior predictive distribution parameters are assumed constant [128] or calculated based on the observation vector in time frame n [123]. Alternatives include imputation based on a nonlinear state–space model (NSSM) [21] where state information is represented with a continuous hidden variable $\boldsymbol{z}$ and the dependencies between variables are assumed nonlinear. The experiments conducted in this work evaluate NSSM-based imputation in PIII.

We note that recent advances in front-end imputation methods are not limited to approaches that model temporal dependencies. For example, model-based imputation has been evaluated on distributions conditioned on voicedness information [129] while sparse imputation has been extended with regularisation and evaluated on clustered search bases [130, 131, 132]. Advances also include imputation uncertainties. The uncertainties do not improve reconstruction accuracies, but reduce the variance bias in parameters estimated based on reconstructed observations.

Observation uncertainties

The reconstructed features $\hat{\boldsymbol{x}}(n)$ can be processed as complete observations so that feature calculation or acoustic model likelihood calculation do not need to be modified. This can have immense practical value. However, transcription accuracies are expected to improve when the recon-

structed values are associated with imputation uncertainties $\hat{\boldsymbol{\sigma}}(n)$ and the uncertainties are introduced in acoustic model likelihood calculation. When reconstructed values are calculated as the posterior predictive distribution mean, uncertainties associated with the reconstructed values can be calculated as the posterior predictive distribution variance. The uncertainties are calculated as the posterior predictive distribution variance in PV–PVI. Alternatives include estimation uncertainties calculated based on heuristic rules. Heuristic uncertainties are used in PII.

The uncertainties associated with reconstructed observations can be introduced into acoustic model likelihood calculation as observation uncertainties [14, 15] or converted into frame-based confidence measures [102]. Previous work on environmental noise compensation with imputation and uncertainties includes cluster-based imputation evaluated with observation uncertainties [133, 134] and bounded conditional mean imputation evaluated with frame-based uncertainties $0 \leq \alpha(n) \leq 1$ [135]. Observation uncertainties have also been used with conditional mean imputation in distributed speech recognition where the incomplete observations arise due to transmission errors [136].

The experiments conducted in this work use observation uncertainties. Since observation uncertainties are used to update acoustic model parameters, we need to calculate uncertainties associated with the representation that is used in acoustic likelihood calculation. The observations $\hat{\boldsymbol{x}}'(n)$ that are calculated based on the reconstructed observations $\hat{\boldsymbol{x}}(n)$ are associated with uncertainties $\hat{\boldsymbol{\sigma}}'(n)$ that are calculated based on the imputation uncertainties $\hat{\boldsymbol{\sigma}}(n)$ or complete posterior predictive distribution $p(\boldsymbol{x}_u(n)|\boldsymbol{x}_r(n))$.

The process in which the uncertainties associated with source variables $\boldsymbol{x}$ are converted into uncertainties associated with variables $f(\boldsymbol{x})$ is called error propagation. When missing-data imputation and acoustic model likelihood calculation operate in separate feature domains, error propagation is needed to calculate the uncertainties $\hat{\boldsymbol{\sigma}}'(n)$ associated with observations $\hat{\boldsymbol{x}}'(n)$. Error propagation methods include model-driven approaches [96, 137] and data-driven approaches [133, 134]. Estimation is model-driven when the transformation between observations in the compressed spectral domain and acoustic-model domain is modelled as M sequential transformations $\boldsymbol{x}' = f_1(f_2 \cdots (f_M(\boldsymbol{x})))$ and each transformation is associated with approximate variance transformation rules. This approach is used in PV–PVI. The data-driven alternative means that a statisti-

cal model is trained to convert uncertainties estimated in the compressed spectral domain into observation uncertainties in the acoustic-model domain. The model parameters are estimated based on estimation uncertainties $\hat{\boldsymbol{\sigma}}(n)$ and oracle uncertainties that are calculated based on parallel clean and noise-corrupted speech data in the acoustic-model domain. This approach is used in PII.

3.2 Artificial bandwidth extension

Bandwidth extension is used in telecommunication where speech waves are transmitted over a distance. Telecommunication systems limit the speech bandwidth and encode the speech wave to minimise transmission size and optimise transmission speed. The experiments reported in PVII–PVIII evaluate ABE in mobile telecommunication where wideband transmissions include frequencies 50–7000 Hz whereas most narrowband transmissions approximate the conventional telephone band 300–3400 Hz. The exact passband varies between transmission systems and terminal devices, but frequencies above 4000 Hz are not included since the narrowband speech wave is sampled at 8000 Hz. ABE can include lowband extension to restore frequencies below the passband and highband extension to restore frequencies above the passband. Bandwidth extension can enhance naturalness and reduce the perceived difference between narrowband and wideband connections.

The approaches evaluated in PVII–PVIII model bandwidth extension as separate envelope extension and excitation extension processes. We assume that speech data is processed in short windows that can be represented with a spectrum. The spectrum can be decomposed into a component that determines the envelope and a component that determines the fine structure. The component that determines the fine structure is called excitation. The envelope and excitation components are assumed independent so that envelope extension and excitation extension can be implemented as separate processes. The current section focusses on envelope extension. Methods that calculate the extended excitation based on narrowband excitation include modulation techniques [138] and nonlinear processes [139]. Alternatives include sinusoidal synthesis that is common in lowband extension [140] and modulated noise that is common in highband extension [141].

ABE approaches estimate extension-band envelopes $\boldsymbol{u}(n)$ based on nar-

rowband observations $\boldsymbol{y}(n)$. The envelope extension method evaluated in PVII–PVIII models the extension-band envelopes and narrowband features as random variables, as proposed in [142]. The conditional distribution $p(\boldsymbol{u}|\boldsymbol{y})$ is modelled as a GMM [143]. Alternatives include codebook-based approaches [144, 145, 146], linear prediction approaches [147, 148], HMM-based approaches [149, 150, 151], and neural network approaches [152, 153]. The input variables $\boldsymbol{y}(n)$ can include a narrowband envelope representation and other variables calculated based on the narrowband data [154]. Moreover the narrowband observation vector in window n can include variables calculated based on narrowband data in previous windows [155].

The envelope extension method evaluated in PVII–PVIII estimates the statistical dependencies between the input and output variables $\boldsymbol{y}$ and $\boldsymbol{u}$ based on parallel narrowband and wideband data. The dependencies are encoded in a conditional distribution model $p(\boldsymbol{u}|\boldsymbol{y})$ and when envelope extension is used, estimates $\hat{\boldsymbol{u}}(n)$ are calculated based on the conditional distribution model and narrowband observations $\boldsymbol{y}(n)$. The approaches evaluated PVII–PVIII calculate extension-band envelopes as MMSE estimates. MMSE estimates do not minimise the expected perceived error between wideband speech and extended narrowband speech since listeners are more sensitive to overestimation errors $\hat{\boldsymbol{u}}(n) > \boldsymbol{u}(n)$ [139]. However MMSE estimation is common in envelope extension, and the estimates can be post-processed to reduce occasional estimation errors [156]. Alternative approaches include estimates derived based on a cost function that penalises overestimation errors [157].

4. Research contribution

4.1 Sparse imputation

Publication I studies window-based sparse imputation and compares sparse imputation and cluster-based imputation in LVCSR. The window-based sparse imputation method [20] processed speech data in windows that span multiple observations $\boldsymbol{x}(n)$ whereas cluster-based imputation [2] modelled the observation vectors as $i.i.d$ random variables. The imputation methods were evaluated on real and simulated noise-corrupted speech data. The real noise-corrupted speech data means utterances that were recorded in public and car environments, and the simulated noise-corrupted speech data means clean speech utterances that were mixed with additive noise. The experiments conducted in PI used clean speech data mixed with babble noise.

The experiments conducted with sparse imputation indicated that the additional time context introduced in multi-frame windows improves transcription accuracies when sparse imputation is used as a noise compensation method in LVCSR. Furthermore, the comparison between sparse imputation and cluster-based imputation indicated that sparse imputation is better than or comparable to cluster-based imputation when sparse imputation is evaluated on windows that span $T \geq 5$ frames. The complete evaluation is presented in PI. The examples presented in this section pertain to sparse imputation applied on windows that span $T = 15$ frames. The same window width is used in the sparse imputation experiments reported in PII–PIV.

The comparison between sparse imputation ($T = 15$) and cluster-based imputation on the noise-corrupted data recorded in public and car environments is presented in Table 4.1. The transcription accuracies obtained

Table 4.1. Letter error rates calculated on noise-corrupted speech data recorded in public environments $P0$–$P2$ and car environments $C0$–$C2$. Experiments compared cluster-based imputation (CI) and sparse imputation (SI).

	P0	P1	P2	C0	C1	C2
Baseline	3.4	22.1	38.3	4.2	33.7	67.3
CI	3.3	13.3	24.6	3.5	19.7	44.3
SI	3.6	13.3	21.2	3.7	15.9	33.3

with the uncompensated baseline system, cluster-based imputation, and sparse imputation are reported in letter error rate. $P0$–$P2$ and $C0$–$C2$ indicate the environmental conditions. $P0$ and $C0$ indicate letter error rates calculated on headset-recorded data that is close to clean speech. $P1$ and $C1$ indicate letter error rates calculated on speech data recorded with a lavalier microphone that captures more environmental noise. $P2$ indicates letter error rates calculated on far-recorded data recorded with a microphone at a medium distance and $C2$ letter error rates calculated on far-recorded data recorded with a hands-free microphone that is mounted on the rear-view mirror. Sparse imputation and cluster-based imputation substitute noise-corrupted features with clean speech estimates and improve transcription accuracies when evaluated on the noise-corrupted speech data in environmental conditions $P1$–$P2$ and $C1$–$C2$. The experiments indicate that sparse imputation is better than cluster-based imputation when environmental noise is severe. Since reliable features are scarce under loud additive noise, we believe that the outcome is in part due to the additional context available for sparse imputation in the windows that span multiple time frames.

Experiments conducted on simulated noise-corrupted speech data evaluate imputation methods under several signal-to-noise ratios (SNR). Letter error rates obtained with the uncompensated baseline system, cluster-based imputation, and sparse imputation ($T = 15$) are reported in Table 4.2. Evaluations conducted with (a) estimated masks and (b) oracle masks indicate that the difference between sparse imputation and cluster-based imputation becomes more notable when the methods are evaluated on more accurate masks. Here sparse imputation introduced 16–84% relative error reduction compared to cluster-based imputation when oracle masks were used (Table 4.2 (b)). Moreover, we note that when oracle masks were used, sparse imputation performed better than cluster-based imputation even when $T = 1$.

Table 4.2. Letter error rates calculated on speech data mixed with babble noise at signal-to-noise ratios (SNR) 0–15 dB. Experiments compared cluster-based imputation (CI) and sparse imputation (SI) with (a) estimated masks and (b) oracle masks.

(a)

	SNR 15	SNR 10	SNR 5	SNR 0
Baseline	13.6	43.4	77.9	94.5
CI	7.6	20.4	55.0	76.6
SI	7.7	18.3	45.2	73.3

(b)

	SNR 15	SNR 10	SNR 5	SNR 0
Baseline	13.6	43.4	77.9	94.5
CI	4.9	10.7	30.3	68.3
SI	4.1	6.7	7.5	12.3

4.2 Sparse imputation and observation uncertainties

In Publication II, sparse imputation was evaluated with observation uncertainties. Since sparse imputation does not determine an explicit predictive distribution model, the reconstructed features were associated with uncertainties estimated with heuristic measures. The proposed measures included measures calculated based on a priori information in the observed features and missing-data mask, and measures calculated based on the reconstructed features or activations. In addition, we evaluated sparse imputation with oracle uncertainties calculated as a squared error between the reconstructed and clean speech features. The uncertainties were mapped to the acoustic-model domain with a linear regression model that was trained on oracle uncertainties calculated in the acoustic-model domain.

The proposed system with sparse imputation and observation uncertainties was evaluated on the simulated noise-corrupted speech data that includes clean speech utterances mixed with babble noise. Experiments conducted with oracle and estimated masks indicated that observation uncertainties improve transcription accuracies when estimated masks are used. The best measures introduced 8–15% relative error reductions compared to sparse imputation baseline when evaluated on the babble-noise corrupted data with estimated masks.

4.3 Nonlinear state–space model

Publication III evaluated NSSM-based imputation [21] on speech data mixed with environmental noise. NSSM associates the observed feature vectors $\boldsymbol{x}(n)$ with hidden source vectors $\boldsymbol{z}(n)$. Dependencies between the observation vectors and the lower-dimensional source vectors as well as dependencies between the source vectors in consecutive time frames, $\boldsymbol{z}(n)$ and $\boldsymbol{z}(n-1)$, were assumed nonlinear and modelled with multilayer perceptron (MLP) networks. The model parameters were estimated based on complete clean speech vectors, and reliable observations were used to estimate the source vector distributions that determine the posterior predictive distributions $p(\boldsymbol{x}(n)|\boldsymbol{z}(n))$. The model parameters and source distributions were calculated based on the cost function proposed in [158]. The source distributions were calculated with the total derivatives approach proposed in [21]. To encode the assumption that missingness is due to additive noise, we constrained the predictive distribution means not to exceed the observed values $\boldsymbol{y}(n)$.

NSSM-based imputation was evaluated on clean speech data mixed with babble or impulsive noise. The clean speech data and the babble noise are the same that were used in PI–PII, whereas the impulsive-noise condition was introduced as a new condition in PIII. The impulsive-noise condition was introduced to compare and evaluate imputation methods on data where complete data vectors are unrecorded and temporal dependencies are needed in reconstruction. The clean speech utterances mixed with babble and impulsive noise are the same, so transcription accuracies can be compared across conditions. The accuracies cannot be compared to transcription accuracies presented in PI since the baseline systems used in PI and PIII are not identical. However comparison between imputation methods is possible because cluster-based imputation and sparse imputation were evaluated in PIII.

The experiments reported in PIII evaluated NSSM-based imputation with and without the constraint on predictive distribution mean. Since constrained optimisation improved NSSM-based imputation in both babble and impulsive noise conditions, transcription accuracies reported in this section pertain to constrained NSSM. Letter error rates obtained with constrained NSSM-based imputation, cluster-based imputation, and sparse imputation in (a) babble noise and (b) impulsive noise condition are reported in Table 4.3. NSSM-based imputation outperformed cluster-

Table 4.3. Letter error rates calculated on speech data mixed with (a) babble noise and (b) impulsive noise. Experiments compared nonlinear state–space model (NSSM) based imputation, cluster-based imputation (CI), and sparse imputation (SI).

(a)

	SNR 15	SNR 10	SNR 5	SNR 0
NSSM	8.2	20.7	54.3	76.9
CI	8.3	21.6	54.9	78.5
SI	7.1	16.4	43.3	72.3

(b)

	SNR 5	SNR 0	SNR -5	SNR -15
NSSM	12.8	17.4	25.0	44.8
CI	25.6	30.1	35.9	46.7
SI	18.7	24.7	32.9	46.2

based imputation and sparse imputation in the impulsive-noise condition (Table 4.3 (b)), whereas in the babble-noise condition, NSSM-based imputation performance was comparable to cluster-based imputation performance and sparse imputation resulted in the best transcription accuracies (Table 4.3 (a)).

4.4 Window-based cluster-based imputation

Publication IV continued experiments on temporal dependencies and clean speech data mixed with babble or impulsive noise. The imputation approaches evaluated in this work processed speech in multi-frame windows same as sparse imputation (PI–PII). However statistical dependencies between feature components were modelled as a full-covariance GMM and reconstructed features were calculated with a cluster-based imputation method where the component-conditional reconstructions were estimated with conditional mean imputation and post-processed not to exceed the observed value. GMM distribution was also compared with a mixture of factor analysers (MFA) [159]. MFA prevents possible overparametrisation when the window width and feature dimension increase, since the covariance structure in each mixture component is modelled in a lower-dimensional factor domain.

The cluster-based imputation method was evaluated on $T = \{1, 5, 10\}$ frame windows. The experiments were conducted on clean speech data

Table 4.4. Letter error rates calculated on speech data mixed with (a) babble noise and (b) impulsive noise. Experiments evaluated a window-based cluster-based imputation method on $T = \{1, 5, 10\}$ frame windows.

(a)

T	SNR 15	SNR 10	SNR 5	SNR 0
1	8.5	23.2	56.2	78.8
5	7.7	19.5	48.9	74.5
10	8.4	22.2	54.9	79.4

(b)

T	SNR 5	SNR 0	SNR -5	SNR -15
1	26.1	30.7	36.8	48.9
5	19.3	25.7	33.1	47.1
10	12.8	19.5	28.8	45.4

mixed with (a) babble noise and (b) impulsive noise as proposed in PIII. The comparison between frame-based and window-based cluster-based imputation indicated that window-based imputation improves transcription accuracies in both conditions (Table 4.4). Moreover comparison between $T = 5$ and $T = 10$ frame windows indicated that $T = 5$ frame windows are better when cluster-based imputation is evaluated in the babble-noise condition (Table 4.4 (a)), whereas in the impulsive-noise condition, transcription accuracies improved when cluster-based imputation was evaluated on $T = 10$ frame windows (Table 4.4 (b)). We believe that the added temporal context improved transcription accuracies in the impulsive-noise condition because impulsive noise is localised in time, which means that an increase in window width increases the ratio between reliable and unreliable features within windows that contain unreliable values. The ratio between reliable and unreliable features is expected to remain constant in the babble-noise condition.

The window-based cluster-based imputation experiments were repeated with MFA clean speech models. Letter error rates obtained in the condition with (a) babble noise and (b) impulsive noise are reported in Table 4.5. A comparison between Table 4.4 and Table 4.5 indicates that the factor domain approach improves the window-based cluster-based imputation performance compared to models that have full-rank covariance matrices. The transcription accuracies obtained with window-based cluster-based imputation can also be compared to accuracies reported in PIII. The transcription accuracies obtained are comparable to transcription ac-

Table 4.5. Letter error rates evaluated on speech data mixed with (a) babble noise (b) impulsive noise. Experiments evaluated a window-based cluster-based imputation method on $T = \{5, 10\}$ frame windows and a clean speech model trained in a low-dimensional factor domain.

(a)

T	SNR 15	SNR 10	SNR 5	SNR 0
5	7.4	18.1	47.0	74.3
10	7.9	21.0	53.5	79.0

(b)

T	SNR 5	SNR 0	SNR -5	SNR -15
5	18.4	24.5	32.3	45.7
10	12.5	18.0	27.0	43.6

curacies obtained with NSSM-based imputation in the impulsive noise condition (Table 4.3 (b)) and better than transcription accuracies obtained with NSSM-based imputation in the babble noise condition (Table 4.3 (a)). Moreover the transcription accuracies are better than transcription accuracies obtained with sparse imputation in the impulsive-noise condition (Table 4.3 (b)) whereas sparse imputation resulted in better transcription accuracies in the babble-noise condition (Table 4.3 (a)).

4.5 Bounded conditional mean imputation

Publication V presented a bounded conditional mean imputation method that estimates unrecorded features based on an approximate posterior distribution. Approximation is needed since the posterior predictive distribution $Tp(\boldsymbol{x}_u|\boldsymbol{x}_r)$ does not have a closed-form solution when the prior distribution $p(\boldsymbol{x})$ is modelled as a full-covariance GMM and features $\boldsymbol{x}$ are constrained not to exceed the observed values $\boldsymbol{y}(n)$. The proposed method approximates the unconstrained posterior used in conditional mean imputation with a diagonal-covariance normal distribution so that the approximate posterior distribution in each channel can be constrained. The approximate posterior distribution in each channel becomes a truncated normal distribution whose mean and variance have closed-form solutions. Bounded conditional mean imputation with the proposed approximate posterior is called truncated posterior mean imputation (TPMI).

The proposed reconstruction method was evaluated on the speech data recorded in public and car environments, and the letter error rates were

Table 4.6. Baseline system evaluated with and without acoustic model adaptation. Adaptation was conducted with constrained maximum likelihood linear regression (CMLLR). Evaluation is based on letter error rates and conducted on noise-corrupted speech data recorded in public environments $P0$–$P2$ and car environments $C0$–$C2$.

	P0	P1	P2	C0	C1	C2
Baseline	3.4	22.2	38.1	4.2	33.7	66.9
CMLLR	2.7	12.7	25.5	2.5	11.5	52.0

compared to letter error rates obtained with cluster-based imputation and conditional mean imputation with estimates post-processed not to exceed the observed values. TPMI outperformed conditional mean imputation and cluster-based imputation when evaluated on far-recorded data, and its performance further improved when the posterior distribution variances were used as observation uncertainties in acoustic model likelihood calculation.

4.6 Bounded conditional mean imputation and acoustic model adaptation

Publication VI extended previous work on bounded conditional mean imputation and observation uncertainties. Since the experiments conducted in previous work indicated that observation uncertainties improve frame-based imputation, experiments conducted in PVI focussed on (1) observation uncertainties and window-based imputation and (2) observation uncertainties and acoustic model adaptation. TPMI was evaluated on multi-frame windows and observation uncertainties were introduced into acoustic model adaptation conducted with constrained maximum likelihood linear regression (CMLLR) [160].

CMLLR uses a linear transformation to minimise distortion between adaptation data and acoustic model parameters. CMLLR parameters are calculated based on acoustic observation vectors $\{X\}$ and estimated state sequences $\{\hat{Z}\}$. Hence unsupervised adaptation experiments are conducted in two passes. The state sequence estimates are calculated in the first pass and the transcriptions that are passed on to evaluation are estimated in the second pass. Transcription accuracies obtained with adaptation and the uncompensated baseline system are reported in Table 4.6. The experiments were conducted on the noise-corrupted speech

Table 4.7. Truncated posterior mean imputation evaluated (a) without uncertainties and (b) with uncertainties. Truncated posterior mean imputation was evaluated on $T = \{1, 5\}$ frame windows. Evaluation is based on letter error rates and conducted on noise-corrupted speech data recorded in public environments $P0$–$P2$ and car environments $C0$–$C2$.

(a)

T	P0	P1	P2	C0	C1	C2
1	3.8	12.9	22.9	3.7	18.1	32.6
5	3.7	11.7	20.4	3.6	15.2	28.2

(b)

T	P0	P1	P2	C0	C1	C2
1	3.4	11.3	20.4	3.4	13.1	23.3
5	3.5	10.8	19.9	3.4	13.8	25.6

data recorded in public and car environments and the transcription accuracies are reported in letter error rate. A pairwise comparison across test conditions indicates that adaptation introduces 21–43% relative error reductions in test conditions $P0$–$P2$ and 22–65% relative error reductions in test conditions $C0$–$C2$.

Letter error rates obtained with frame-based and window-based TPMI (a) without observation uncertainties and (b) with observation uncertainties are reported in Table 4.7. The experiments were conducted with the uncompensated baseline system without acoustic model adaptation. We note that the experiments were conducted on the same data that was used in PI, but the transcription accuracies reported in PI and PVI cannot be compared because the experiments were not conducted with the same acoustic model or the same mask estimation method. The experiments conducted without observation uncertainties indicate that window-based imputation is better than frame-based imputation (Table 4.7 (a)). However when uncertainties were introduced in acoustic model likelihood calculation, frame-based and window-based imputation resulted in comparable letter error rates (Table 4.7 (b)). We believe that the window-based imputation did not improve transcription accuracies compared to frame-based imputation when uncertainties were used because window-based imputation and observation uncertainties function as alternative approaches to reduce classification errors in acoustic model likelihood calculation.

Letter error rates obtained with adaptation and imputation (a) without observation uncertainties and (b) with observation uncertainties are re-

Table 4.8. Truncated posterior mean imputation and adaptation evaluated (a) without uncertainties and (b) with uncertainties. Truncated posterior mean imputation was evaluated on $T = \{1, 5\}$ frame windows. Evaluation is based on letter error rates and conducted on noise-corrupted speech data recorded in public environments $P0$–$P2$ and car environments $C0$–$C2$.

(a)

T	P0	P1	P2	C0	C1	C2
1	3.1	12.2	22.5	2.5	7.8	28.3
5	3.0	10.7	18.6	2.4	5.9	20.9

(b)

T	P0	P1	P2	C0	C1	C2
1	3.0	9.7	16.6	2.4	5.8	17.4
5	2.9	9.0	15.9	2.4	5.3	16.8

ported in Table 4.8. The results indicate that adaptation improves the transcription accuracies and that observation uncertainties increased the relative error reduction introduced in acoustic model adaptation. For example, comparison between baseline and CMLLR experiments reported in Table 4.7 (a) and Table 4.8 (a) indicates that CMLLR introduced under 10% relative error reductions in test conditions $P1$–$P2$ when observation uncertainties were not used, and comparison between baseline and CMLLR experiments reported in Table 4.7 (b) and Table 4.8 (b) indicates that the relative error reductions in test conditions $P1$–$P2$ increased to 10–41% when observation uncertainties were used in acoustic model likelihood calculation and adaptation. In contrast to the experiments conducted without adaptation, the experiments conducted with adaptation indicate that window-based imputation is better than frame-based imputation also when uncertainties are used (Table 4.8 (b)). We believe this is because window-based imputation reduces reconstruction errors whereas uncertainties make acoustic model likelihood calculation and CMLLR parameter estimation more conservative.

4.7 Artificial bandwidth extension

Publications VII–VIII studied narrowband telephone speech and ABE. The highband extension method (HB-ABE) evaluated in PVII and the lowband extension method (LB-ABE) evaluated in PVIII estimated wideband spectral envelopes based on narrowband spectral envelopes in the current and previous time frames. The narrowband and wideband envelopes

were represented as log-compressed sub-band spectra and the statistical dependencies between the narrowband and wideband envelopes were modelled as a GMM. The wideband envelopes were calculated as MMSE estimates and post-processed to reduce estimation errors. Highband extension was based on a filter-bank method that used an upsampled linear prediction residual and modulated noise excitation whereas lowband extension was based on sinusoidal synthesis. Since narrowband transmission does not remove the lowband channel, as assumed in envelope extension, but attenuation varies between devices and connections, reconstruction amplitudes calculated based on the estimated lowband envelope were adapted based on the observed lowband amplitudes (PVIII).

HB-ABE and LB-ABE were evaluated on clean speech utterances that were processed with a modified mobile station input (MSIN) filter and adaptive multi-rate (AMR) narrowband codec to simulate a narrowband telephone connection. HB-ABE was compared to a reference method [23] that has been used in mobile phones [161]. While HB-ABE sometimes introduced audible distortion, listeners preferred highband extension over narrowband speech and HB-ABE over the reference method (PVII). LB-ABE was evaluated in combination with the highband extension method proposed in [24]. LB-ABE did not improve the perceived speech quality compared to highband extension without lowband extension, but listeners considered the test data more similar to wideband data when LB-ABE was used (PVIII).

5. Discussion

5.1 Theoretical implications

The imputation approaches studied in this work were evaluated and compared as noise compensation methods in LVCSR. The experiments indicated that various imputation approaches can improve transcription accuracies under environmental noise. The experiments also indicated that transcription accuracies improve (1) when temporal dependencies are used in imputation and (2) when the uncertainties related to imputation are introduced in acoustic model likelihood calculation as observation uncertainties.

5.1.1 Model comparison

The experiments reported in PI and PIII–PIV compared sparse imputation and model-based imputation methods. The experiments indicated that sparse imputation is more sensitive to mask estimation errors than model-based approaches. However, even when imputation methods were evaluated on estimated masks, sparse imputation outperformed model-based imputation approaches in most noise conditions. A previous comparison between sparse imputation and cluster-based imputation resulted in the same conclusion that sparse imputation is more sensitive to mask estimation errors but improves transcription accuracies more than cluster-based imputation [125]. Another comparison indicated that sparse imputation was better than or comparable to window-based cluster-based imputation when the methods were evaluated with oracle masks, while window-based cluster-based imputation was better when estimated masks were used [162].

5.1.2 Temporal dependencies

The experiments conducted in PIII and PIV evaluated temporal dependencies in model-based imputation. The experiments compared NSSM-based imputation and window-based cluster-based imputation to cluster-based imputation approaches that do not take into account temporal dependencies between incomplete data vectors in consecutive time frames. The comparisons were conducted on clean speech data mixed with babble and impulsive noise. The impulsive-noise condition was used because we wanted to evaluate temporal dependencies on incomplete speech data where the unrecorded values are clustered in time, as opposed to the babble noise condition where the ratio between recorded and unrecorded values is quite constant in consecutive time frames.

The experiments indicate that temporal dependencies can improve imputation-based ASR. NSSM-based imputation (PIII) improved transcription accuracies compared to frame-based cluster-based imputation in the impulsive-noise condition and resulted in comparable performance in the babble-noise condition. In contrast, the window-based cluster-based imputation method evaluated in PIV improved transcription accuracies compared to frame-based cluster-based imputation in both conditions. We note that the window-based approaches are expected to reduce reconstruction errors because (1) imputation can utilise temporal dependencies between consecutive observation vectors and (2) summation over several window-based estimates reduces occasional estimation errors. Hence we cannot conclude, based on the window-based cluster-based imputation experiments, that dependencies between consecutive observations improve transcription accuracies in the babble-noise condition.

The window-based cluster-based imputation experiments conducted on babble-noise-corrupted speech data indicate that the letter error rates increase when window width $T > 5$ (PIV). We assume this is because the increase in window width and data dimension reduces similarities between observations and increases the expected distance between an observation and the posterior predictive distribution mean. The window-based observation vectors are also expected to include components that do not improve prediction accuracies when dependencies between the unrecorded variables and other recorded variables are modelled. The experiments conducted in PI–PIV indicate that sparse imputation can utilise more time context than window-based cluster-based imputation, but even

sparse imputation errors increase when window width $T > 20$. This is because the observation vectors cannot be associated with a sparse representation, as reported in PI. The reason observation vectors cannot be associated with a sparse representation is the increase in data dimension which increases and equalises distances between an observation and basis vectors.

5.1.3 Estimation uncertainties

While temporal dependencies are expected to make reconstructions more accurate, uncertainties aim to ensure that standard errors or class probabilities calculated based on reconstructed data are unbiased. The experiments conducted in this work introduced imputation uncertainties into acoustic model likelihood calculation and acoustic model adaptation as observation uncertainties. This means that the uncertainties associated with the reconstructed observations were converted into uncertainties in the acoustic model domain and added to the acoustic model variance as proposed in [14, 15]. The uncertainties associated with sparse imputation estimates were estimated based on heuristic measures (PII) whereas uncertainties associated with bounded conditional mean imputation estimates were calculated as posterior predictive distribution variances (PV–PVI).

The experiments conducted with heuristic uncertainties in PI and with posterior variances in PV–PVI indicate that uncertainties improve transcription accuracies in both conditions. Furthermore, sparse imputation and uncertainties calculated based on the heuristic measures proposed in PII were evaluated on another dataset in [163] and uncertainties calculated based on related measures have been used in sparse separation [164, 165]. The conversion between heuristic measures and observation uncertainties was learned based on parallel clean and noise-corrupted speech data, whereas the uncertainties used in PV–PVI were determined based on variance calculation rules. The uncertainties were not optimised to test conditions because we wanted to minimise the increase that uncertainties introduce in computational cost.

The experiments conducted in PVI used uncertainties in CMLLR parameter estimation. CMLLR parameters are estimated based on first and second order statistics calculated based on the observations [160]. When observation uncertainties are used, the statistics are calculated based on the observations and acoustic model parameters. This indicates

the connection between CMLLR parameter estimation based on uncertain observations (PVI) and previous work on approaches that calculate speaker-based transformation parameters based on interpolated statistics [166, 167]. The approaches discussed in previous works [166, 167] assume that the available observations do not cover the complete dataset that is needed to estimate speaker-dependent linear transformation parameters. Hence parameter estimation based on the available observations is modelled as parameter estimation based on incomplete data. In contrast, we assume in PVI that the available observations would have covered the complete dataset unless environmental noise had masked certain components. CMLLR parameters were estimated based on uncertain observations as proposed in [93, 94].

5.2 Practical implications

The imputation and spectral envelope extension approaches studied in this work can be used in ASR and ABE. ASR applications include, for example, automatic broadcast news transcription and spoken content retrieval. ABE can be used in mobile terminals to improve user experience when the speech transmission line includes a narrowband connection or a narrowband terminal [161].

The experiments discussed in this work indicated that imputation can improve transcription transcription accuracies in LVCSR under unseen environmental noise. Experiments conducted with imputation and acoustic model adaptation also indicated that the reconstructed observations can be used in parameter estimation when estimation uncertainties are taken into account. This means that imputation could be used in acoustic model parameter estimation to estimate a clean-speech model based on mixed sources that include narrowband telephone speech and noise-corrupted data. Since data collection over a telephone is economical and effective, parameter estimation based on mixed sources can reduce data collection cost. For more discussion on imputation and acoustic model parameter estimation based on mixed-bandwidth data, see previous works [168, 169, 170, 171].

5.3 Limitations

The conclusions presented in the previous sections are based on experimental evaluation. We evaluated imputation as a noise compensation method in ASR and conducted ABE experiments on mobile telephone speech. The current section discusses limitations in the experimental evaluation and conclusions related to comparison between imputation approaches.

The experiments reported in PI–PIV indicated that the outcomes in comparisons between imputation approaches depend on the noise condition and mask estimation method. The conclusions presented in this work could hence be invalid in conditions that were not represented in the experiments. The experiments evaluated imputation approaches on real and simulated noise-corrupted speech data. The main comparisons between imputation methods were conducted on the real data recorded in public and car environments, while the simulated clean speech and noise mixtures were used to test uncertainties and temporal dependencies in controlled conditions. Evaluation based on simulated data alone is not recommended since there can be deviation between outcomes observed on simulated speech and noise mixtures and outcomes observed on real noise-corrupted speech data [172]. The environmental noise conditions represented in the evaluation data used in the main comparisons include various indoors and outdoors locations but are not comprehensive. Moreover, the mask estimation method used in PI–PVI was based on comparison between the noise-corrupted observations and a noise estimate, and the imputation approaches were not evaluated with alternative mask estimation methods.

While imputation was evaluated as a noise compensation method in LVCSR, imputation was not compared to other noise compensation methods and reconstructed features were not evaluated in deep neural network (DNN) based acoustic model likelihood calculation [72, 73, 74]. DNN-based approaches have improved transcription accuracies and noise robustness compared to GMM-HMM, and DNN-HMM has become the most common acoustic model. Front-end imputation methods can be used in DNN-HMM-based systems without modification, but the approaches that most improve transcription accuracies are not expected to be the same that most improved transcription accuracies in the current work where GMM-HMM-based acoustic likelihood calculation was used. DNN-based

approaches can learn noise-robust representations based on multicondition data [173] and noise compensation can be used to extend rather than substitute noise-corrupted observations [174].

5.4 Recommendations for future research

The experiments conducted in this work extended sparse imputation and bounded conditional mean imputation with uncertainties. ASR experiments indicated that the uncertainties improve transcription accuracies when introduced in acoustic model likelihood calculation as observation uncertainties. However, when the observation uncertainties were calculated based on sparse imputation uncertainties, the conversion between heuristic uncertainties in compressed spectral domain and observation uncertainties in acoustic-model domain relied on a dataset with parallel clean speech and noise-corrupted data (PII). Hence our recommendations include that future studies on sparse imputation and uncertainties evaluate the alternative approach that is based on frame-based uncertainties [135]. Frame-based uncertainties $0 \leq \alpha(n) \leq 1$ are used to post-process the state-conditioned likelihoods calculated based on reconstructed data, which means that imputation uncertainties do not need to be mapped to the acoustic-model domain. Instead, frame-based uncertainties can be calculated based on estimation uncertainties in a heuristic manner so that parallel data is not needed.

The experiments conducted in PVI used observation uncertainties in acoustic model likelihood calculation and speaker-based acoustic model adaptation. The experiments indicate that adaptation with CMLLR improves transcription accuracies more when uncertainties are used. We assume this is because uncertainties improve CMLLR parameter estimation, but we note that the evaluation was based on transcription accuracies rather than comparison between CMLLR parameters estimated based on reconstructed data and complete clean speech data. Since the parameters were estimated based on the evaluation data, transcription accuracies do not indicate whether adaptation captures speaker variation or environmental noise. Hence our recommendations include that future studies evaluate CMLLR and observation uncertainties so that noise robustness is needed in parameter estimation. For example, CMLLR parameters estimated based on reconstructed data recorded in certain environmental condition can be evaluated on clean speech data or

data recorded in another environmental condition [175, 176]. Alternatives include evaluation in another application. Examples of applications where the parameters must capture speaker variation and must not capture environmental noise include speaker verification [177, 178]. We believe that the parameters which are estimated based on expected clean speech statistics calculated based on uncertain observations and acoustic model parameters as proposed PVI are too conservative, which means that alternative approaches to parameter estimation based on uncertain data need to be studied, and we believe that the additional evaluations can contribute in this aim.

6. Conclusions

The current work discussed missing-data imputation in ASR and spectral envelope extension in ABE. The connection between the ASR and ABE experiments is that the imputation and envelope extension approaches evaluated in PI–PVIII model speech data in a compressed spectral domain and estimate unobserved values.

ASR experiments were conducted to evaluate additive-noise compensation with front-end imputation methods. The experiments compared sparse imputation and model-based imputation methods in noise-robust LVCSR. The aim was to find imputation methods that most improve transcription accuracies. The main research directions related to temporal dependencies and observation uncertainties. Temporal dependencies were proposed to improve reconstruction accuracies and uncertainties were proposed to improve transcription accuracies when acoustic likelihood calculation operates on reconstructed features. Related to imputation uncertainties, a novel bounded conditional mean imputation method was proposed.

The experiments reported in PI–PVI indicate that the best imputation method depends on the noise condition and missing-data mask. For example, temporal dependencies improve transcription accuracies when speech is corrupted with impulsive noise. However, the experiments also indicate that window-based imputation and observation uncertainties improve transcription accuracies in various conditions and improve both sparse imputation and model-based imputation approaches. Furthermore, missing-data imputation is compatible with speaker-based acoustic model adaptation when observation uncertainties are used.

The bandwidth extension experiments (PVII–PVIII) used a GMM-based spectral envelope extension method. The wideband envelope parameters that are not observed in narrowband telephone speech were estimated

based on narrowband envelope parameters and statistical dependencies learned based on parallel wideband and narrowband data. The envelope extension method was used in a highband extension system and a lowband extension system. The experiments indicated that bandwidth extension with the proposed system reduces the perceived difference between narrowband and wideband speech.

References

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, 2001.

[2] B. Raj, M. L. Seltzer, and R. M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43(4):275–296, 2004.

[3] T. Kohonen. The "neural" phonetic typewriter. *IEEE Computer*, 21(3): 11–22, 1988.

[4] M. Kurimo. *Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models*. PhD thesis, Helsinki University of Technology, 1997.

[5] V. Siivola. *Language models for automatic speech recognition: construction and complexity control*. PhD thesis, Helsinki University of Technology, 2007.

[6] T. Hirsimäki. *Advances in unlimited-vocabulary speech recognition for morphologically rich languages*. PhD thesis, Helsinki University of Technology, 2009.

[7] J. Pylkkönen. *Towards Efficient and Robust Automatic Speech Recognition: Decoding Techniques and Discriminative Training*. PhD thesis, Aalto University, 2013.

[8] V. T. Turunen. *Morph-Based Speech Retrieval: Indexing Methods and Evaluations of Unsupervised Morphological Analysis*. PhD thesis, Aalto University, 2012.

[9] S. Enarvi. *Finnish Language Speech Recognition For Dental Health Care*. Licentiate thesis, Aalto University, 2013.

[10] S. Keronen. *Approaching human performance in noise robust automatic speech recognition*. Licentiate thesis, Aalto University, 2014.

[11] H. Kallasjoki. *Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech*. PhD thesis, Aalto University, 2016.

[12] H. Pulakka. *Development and evaluation of artificial bandwidth extension methods for narrowband telephone speech*. PhD thesis, Aalto University, 2013.

[13] L. Laaksonen. *Artificial bandwidth extension of narrowband speech - enhanced speech quality and intelligibility in mobile devices*. PhD thesis, Aalto University, 2013.

[14] J. A. Arrowood and M. A. Clements. Using observation uncertainty in HMM decoding. In *Proc. ICSLP*, pages 1561–1564, 2002.

[15] L. Deng, J. Droppo, and A. Acero. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE Trans. Speech and Audio Processing*, 13(3):412–421, 2005.

[16] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541, 2006.

[17] T. Hirsimäki, J. Pylkkönen, and M. Kurimo. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 17(4):724–732, 2009.

[18] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling. SPEECON - speech databases for consumer devices: Database specification and validation. In *Proc. LREC*, pages 329–333, 2002.

[19] J. F. Gemmeke and B. Cranen. Using sparse representations for missing data imputation in noise robust speech recognition. In *Proc. EUSIPCO*, 2008.

[20] J. Gemmeke and B. Cranen. Missing data imputation using compressive sensing techniques for connected digit recognition. In *Proc. DSP*, 2009.

[21] T. Raiko, M. Tornio, A. Honkela, and J. Karhunen. State inference in variational Bayesian nonlinear state-space models. In *Proc. ICA*, pages 222–229, 2006.

[22] H. Yamamoto, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura. Parameter sharing in mixture of factor analyzers for speaker identification. *IEICE Trans. Inf. & Syst.*, E88–D(3):418–424, 2005.

[23] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen, and P. Alku. Evaluation of an artificial speech bandwidth extension method in three languages. *IEEE Trans. Audio, Speech, and Language Processing*, 16(6):1124–1137, 2008.

[24] H. Pulakka and P. Alku. Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum. *IEEE Trans. Audio, Speech, and Language Processing*, 19(7): 2170–2183, 2011.

[25] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley, 2nd edition, 2002.

[26] J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

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

[28] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *J. American Statistical Association*, 82(398):528–540, 1987.

[29] J. L. Schafer. *Analysis of incomplete multivariate data*. CRC Press, 1997.

[30] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[31] R. K. Pearson. The problem of disguised missing data. *ACM SIGKDD Explorations Newsletter*, 8(1):83–92, 2006.

[32] R. J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *J. American Statistical Association*, 90(431):1112–1121, 1995.

[33] J. W. Graham, B. J. Taylor, A. E. Olchowski, and P. E. Cumsille. Planned missing data designs in psychological research. *Psychological Methods*, 11 (4):323–343, 2006.

[34] N. V. Chawla and G. Karakoulas. Learning from labeled and unlabeled data: An empirical study across techniques and domain. *J. Artificial Intelligence Research*, 23:331–366, 2005.

[35] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[36] T. W. Anderson. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. American Statistical Association*, 52(278):200–203, 1957.

[37] D. B. Rubin. Characterizing the estimation of parameters in incomplete-data problems. *J. American Statistical Association*, 69(346):467–474, 1974.

[38] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[39] R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *J. Americal Statistical Association*, 88(421):125–134, 1993.

[40] M. Wu and R. J. Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175–188, 1988.

[41] D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *Ann. Statistics*, 19(4):2244–2253, 1991.

[42] J. W. Graham and S. I. Donaldson. Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *J. Applied Psychology*, 78(1):119–128, 1993.

[43] G. Molenberghs, C. Beunckens, C. Sotto, and M. G. Kenward. Every missingness not at random model has a missingness at random counterpart with equal fit. *J. Royal Statistical Society B*, 70(2):371–388, 2008.

[44] B. M. Marlin, R. S. Zemel, S. Roweis, and M. Slaney. Collaborative filtering and the missing at random assumption. In *Proc. UAI*, pages 267–275, 2007.

[45] L. M. Collins, J. L. Schafer, and C. M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001.

[46] L. C. Lazzeroni, N. Schenker, and J. M. G. Taylor. Robustness of multiple-imputation techniques to model misspecification. In *Proc. SRMS*, pages 260–265, 1990.

[47] X. L. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

[48] R. Eckardt, C. Berger, C. Thiel, and C. Schmullius. Removal of optically thick clouds from multi-spectral satellite images using multi-frequency SAR data. *Remote Sensing*, 5(6):2973–3006, 2013.

[49] F. Thoemmes and N. Rose. A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research*, 49(5):443–459, 2014.

[50] W. C. Horrace. Some results on the multivariate truncated normal distribution. *J. Multivariate Analysis*, 94(1):209–221, 2005.

[51] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

[52] Z. Ghahramani. Hierarchical and nonlinear models [lecture slides]. `http://mlg.eng.cam.ac.uk/zoubin/course04/`, 2004.

[53] R. R. Andridge and R. J. A. Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

[54] D. R. Wilson and T. R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38:257–286, 2000.

[55] R. J. A. Little. Missing-data adjustments in large surveys. *J. Business & Economic Statistics*, 6(3):287–296, 1988.

[56] D. B. Rubin. *Multiple imputation for non-response in surveys*. Wiley, 1987.

[57] J. N. K. Rao and J. Shao. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4):811–822, 1992.

[58] J. Shao and R. R. Sitter. Bootstrap for imputed survey data. *J. American Statistical Association*, 91(435):1278–1288, 1996.

[59] J. L. Schafer and N. Schenker. Inference with imputed conditional means. *J. American Statistical Association*, 95(449):144–154, 2000.

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

[61] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.

[62] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoustics, Speech and Signal Processing*, 34(1):52–59, 1986.

[63] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. Speech and Audio Processing*, 7(3):272–281, 1999.

[64] D. O'Shaughnessy. Acoustic analysis for automatic speech recognition. *Proc. IEEE*, 101(5):1038–1053, 2013.

[65] S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, 13(5):45–57, 1996.

[66] M. Gales, S. Watanabe, and E. Fosler-Lussier. Structured discriminative models for speech recognition. *IEEE Signal Processing Magazine*, 29(6): 70–81, 2012.

[67] F. Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64(4):532–556, 1976.

[68] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[69] K. F. Lee. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Trans. Acoustics, Speech and Signal Processing*, 38(4):599–609, 1990.

[70] S. J. Young and P. C. Woodland. The use of state tying in continuous speech recognition. In *Proc. EUROSPEECH*, pages 2203–2206, 1993.

[71] M. Gales and S. Young. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.

[72] A. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE Trans. Audio, Speech, and Language Processing*, 20 (1):14–22, 2012.

[73] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pretrained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[74] G. Hinton, L. Deng, Y. Dong, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[75] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle. Template-based continuous speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 15(4):1377–1390, 2007.

[76] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram. Exemplar-based processing for speech recognition. *IEEE Signal Processing Magazine*, 29(6):98–113, 2012.

[77] H. Liao and M. J. F. Gales. Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In *Proc. ICASSP*, pages 389–392, 2007.

[78] O. Kalinli, M. L. Seltzer, J. Droppo, and A. Acero. Noise adaptive training for robust automatic speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 18(8):1889–1901, 2010.

[79] M. Creutz and K. Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech and Language Processing*, 4(1), 2007.

[80] V. Siivola and B. Pellom. Growing an n-gram language model. In *Proc. INTERSPEECH*, pages 1309–1312, 2005.

[81] J. Pylkkönen. An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition. In *Proc. HLT*, pages 167–172, 2005.

[82] X. L. Aubert. An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech and Language*, 16(1):89–114, 2002.

[83] J. Droppo, A. Acero, and L. Deng. Uncertainty decoding with SPLICE for noise robust speech recognition. In *Proc. ICASSP*, pages 57–60, 2002.

[84] H. Liao and M. J. F. Gales. Joint uncertainty decoding for noise robust speech recognition. In *Proc. INTERSPEECH*, pages 3129–3132, 2005.

[85] H. Xu, L. Rigazio, and D. Kryze. Vector Taylor series based joint uncertainty decoding. In *Proc. INTERSPEECH*, pages 1125–1128, 2006.

[86] H. Liao and M. J. F. Gales. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Communication*, 50(4):265–277, 2008.

[87] D. K. Kim and M. J. F. Gales. Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 19(2):315–325, 2011.

[88] H. Jiang, K. Hirose, and Q. Huo. Robust speech recognition based on a Bayesian prediction approach. *IEEE Trans. Speech and Audio Processing*, 7(4):426–440, 1999.

[89] H. Jiang and Q. Huo. A Bayesian prediction approach to robust speech recognition. *IEEE Trans. Speech and Audio Processing*, 8(2):200–204, 2000.

[90] A. C. Morris, J. Barker, and H. Bourland. From missing data to maybe useful data: soft data modelling for noise robust ASR. In *Proc. WISP*, pages 153–164, 2001.

[91] M. Kühne, R. Togneri, and S. Nordholm. A new evidence model for missing data speech recognition with applications in reverberant multi-source environments. *IEEE Trans. Audio, Speech, and Language Processing*, 19 (2):372–384, 2011.

[92] R. Maas, C. Huemmer, A. Sehr, and W. Kellermann. A Bayesian view on acoustic model-based techniques for robust speech recognition. *EURASIP J. Advances in Signal Processing*, 2015(103), 2015.

[93] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 17(3):423–435, 2009.

[94] A. Ozerov, M. Lagrange, and E. Vincent. Uncertainty-based learning of acoustic models from noisy data. *Computer Speech and Language*, 27(3): 874–894, 2013.

[95] V. Stouten, H. Van hamme, and P. Wambacq. Model-based feature enhancement with uncertainty decoding for noise robust ASR. *Speech Communication*, 48(11):1502–1514, 2005.

[96] R. F. Astudillo, D. Kolossa, P. Mandelartz, and R. Orglmeister. An uncertainty propagation approach to robust ASR using the ETSI advanced frontend. *IEEE J. Selected Topics in Signal Processing*, 4(5):824–833, 2010.

[97] M. Delcroix, T. Nakatani, and S. Watanabe. Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing. *IEEE Trans. Audio, Speech, and Language Processing*, 17 (2):324–334, 2009.

[98] D. T. Tran, E. Vincent, and D. Jouvet. Extension of uncertainty propagation to dynamic MFCCs for noise robust ASR. In *Proc. ICASSP*, pages 5507–5511, 2014.

[99] D. T. Tran, E. Vincent, and D. Jouvet. Discriminative uncertainty estimation for noise robust ASR. In *Proc. ICASSP*, pages 5038–5042, 2015.

[100] D. T. Tran, E. Vincent, and D. Jouvet. Nonparametric uncertainty estimation and propagation for noise robust ASR. *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 23(11):1835–1846, 2015.

[101] V. Stouten, H. Van hamme, and P. Wambacq. Accounting for the uncertainty of speech estimates in the context of model-based feature enhancement. In *Proc. INTERSPEECH*, pages 105–108, 2004.

[102] N. B. Yoma, F. R. McInnes, and M. A. Jack. Weighted Viterbi algorithm and state duration modelling for speech recognition in noise. In *Proc. ICASSP*, pages 709–712, 1998.

[103] G. A. Miller and J. C. R. Licklider. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.*, 22(2):167–173, 1950.

[104] R. M. Warren. Perceptual restoration of missing speech sounds. *Science*, 167:393–393, 1970.

[105] R. M. Warren, K. Riener Hainsworth, B. S. Brubaker, J. A. Bashford, Jr., and E. W. Healy. Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps. *Perception & Psychophysics*, 52(2):275–283, 1997.

[106] L. Josifovski, M. Cooke, P. Green, and A. Vizinho. State based imputation of missing data for robust speech recognition and speech enhancement. In *Proc. EUROSPEECH*, pages 2837–2840, 1999.

[107] M. Van Segbroeck and H. Van Hamme. Advances in missing-feature techniques for robust large-vocabulary continuous speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 19(1):123–137, 2011.

[108] T. Virtanen, R. Singh, and B. Raj, editors. *Techniques for noise robustness in automatic speech recognition*. Wiley, 2012.

[109] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 22(4):745–777, 2014.

[110] A. Nádas, D. Nahamoo, and M. A. Picheny. Speech recognition using noise-adaptive prototypes. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 37(10):1495–1503, 1989.

[111] A. Varga and R. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. ICASSP*, pages 845–848, 1990.

[112] J. R. Hershey, P. A. Olsen, and S. J. Rennie. Signal interaction and the devil function. In *Proc. INTERSPEECH*, pages 334–337, 2010.

[113] M. L. Seltzer, B. Raj, and R. M. Stern. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43(4):379–393, 2004.

[114] N. Ma, P. Green, J. Barker, and A. Coy. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Communication*, 49(12):874–891, 2007.

[115] C. Cerisara, S. Demange, and J. P. Haton. On noise masking for automatic missing data speech recognition: A survey and discussion. *Computer Speech and Language*, 21(3):443–457, 2007.

[116] M. Cooke. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 119(3):1562–1573, 2006.

[117] P. Renevey and A. Drygajlo. Detection of reliable features for speech recognition in noisy conditions using a statistical criterion. In *Proc. CRAC*, 2001.

[118] J. Barker, L. Josifovski, M. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proc. ICSLP*, pages 373–376, 2000.

[119] B. Raj and R. Singh. Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition. In *Proc. ASRU*, pages 65–70, 2005.

[120] J. P. Barker, M. P. Cooke, and D. P. W. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45(1):5–25, 2005.

[121] A. M. Reddy and B. Raj. Soft mask methods for single-channel speaker separation. *IEEE Trans. Audio, Speech, and Language Processing*, 15(6): 1766–1776, 2007.

[122] F. Faubel, J. McDonough, and D. Klakow. Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features. In *Proc. ICASSP*, pages 3869–3872, 2009.

[123] J. A. González, A. M. Peinado, N. Ma, A. M. Gómez, and J. Barker. MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 21(3):624–635, 2013.

[124] B. Borgström and A. Alwan. Utilizing compressibility in reconstructing spectrographic data. *IEEE Signal Processing Letters*, 16(5):398–401, 2009.

[125] J. F. Gemmeke, H. Van hamme, B. Cranen, and L. Boves. Compressive sensing for missing data imputation in noise robust speech recognition. *IEEE J. Selected Topics in Signal Processing*, 4(2):272–287, 2010.

[126] W. Kim and J. H. L. Hansen. Missing-feature reconstruction by leveraging temporal spectral correlation for robust speech recognition in background noise conditions. *IEEE Trans. Audio, Speech, and Language Processing*, 18(8):2111–2120, 2010.

[127] S. Parveen and P. D. Green. Speech recognition with missing data using recurrent neural nets. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, pages 1189–1195. MIT Press, 2002.

[128] B. J. Borgström and A. Alwan. HMM-based reconstruction of unreliable spectrographic data for noise robust speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 18(6):1612–1623, 2010.

[129] A. Narayanan, X. Zhao, D. L. Wang, and E. Fosler-Lussier. Robust speech recognition using multiple prior models for speech reconstruction. In *Proc. ICASSP*, pages 4800–4803, 2011.

[130] Q. F. Tan, P. G. Georgiou, and S. S. Narayanan. Enhanced sparse imputation techniques for a robust speech recognition front-end. *IEEE Trans. Audio, Speech, and Language Processing*, 19(8):2418–2429, November 2011.

[131] Q. F. Tan and S. S. Narayanan. Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 20(4): 1337–1346, 2012.

[132] M. M. Goodarzi and F. Almasganj. Model-based clustered sparse imputation for noise robust speech recognition. *Speech Communication*, 76: 218–229, 2016.

[133] S. Srinivasan and D. L. Wang. A supervised learning approach to uncertainty decoding for robust speech recognition. In *Proc. ICASSP*, pages 297–300, 2006.

[134] S. Srinivasan and D. L. Wang. Transforming binary uncertainties for robust speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 15(7):2130–2140, 2007.

[135] J. A. González, A. M. Peinado, A. M. Gómez, N. Ma, and J. Barker. Combining missing-data reconstruction and uncertainty decoding for robust speech recognition. In *Proc. ICASSP*, pages 4693–4696, 2012.

[136] V. Ion and R. Haeb-Umbach. Uncertainty decoding for distributed speech recognition over error-prone networks. *Speech Communication*, 48(11): 1435–1446, 2006.

[137] R. Fernandez Astudillo and D. Kolossa. Uncertainty propagation. In D. Kolossa and R. Haeb-Umbach, editors, *Robust Speech Recognition of Uncertain and Missing Data*, pages 35–64. Springer Verlag, 2011.

[138] J. Makhoul and M. Berouti. High-frequency regeneration in speech coding systems. In *Proc. ICASSP*, pages 428–431, 1979.

[139] B. Iser and G. Schmidt. Bandwidth extension of telephony speech. *EURASIP Newsletter*, 16(2):2–14, 2005.

[140] G. Miet, A. Gerrits, and J. C. Valiére. Low-band extension of telephone-band speech. In *Proc. ICASSP*, pages 1851–1854, 2000.

[141] A. McCree. A 14 kb/s wideband speech coder with a parametric highband model. In *Proc. ICASSP*, pages 1153–1156, 2000.

[142] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein. Statistical recovery of wideband speech from narrowband speech. *IEEE Trans. Speech and Audio Processing*, 2(4):544–548, 1994.

[143] K. Y. Park and H. S. Kim. Narrowband to wideband conversion of speech using GMM based transformation. In *Proc. ICASSP*, pages 1843–1846, 2000.

[144] H. Carl and U. Heute. Bandwidth enhancement of narrow-band speech signals. In *Proc. EUSIPCO*, pages 1178–1181, 1994.

[145] Y. Yoshida and M. Abe. An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping. In *Proc. ICSLP*, pages 1591–1594, 1994.

[146] J. Epps and W. H. Holmes. A new technique for wideband enhancement of coded narrowband speech. In *Proc. SCW*, pages 174–176, 1999.

[147] Y. Nakatoh, M. Tsushima, and T. Norimatsu. Generation of broadband speech from narrowband speech using piecewise linear mapping. In *Proc. EUROSPEECH*, pages 1643–1646, 1997.

[148] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter. Speech enhancement via frequency bandwidth extension using line spectral frequencies. In *Proc. ICASSP*, pages 665–668, 2001.

[149] P. Jax and P. Vary. On artificial bandwidth extension of telephone speech. *Signal Processing*, 83(8):1707–1719, 2003.

[150] G. Chen and V. Parsa. HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies. In *Proc. ICASSP*, pages 709–712, 2004.

[151] G. B. Song and P. Martynovich. A study of HMM-based bandwidth extension of speech signals. *Signal Processing*, 89(10):2036–2044, 2009.

[152] A. Uncini, F. Gobbi, and F. Piazza. Frequency recovery of narrow-band speech using adaptive spline neural networks. In *Proc. ICASSP*, pages 997–1000, 1999.

[153] J. Kontio, L. Laaksonen, and P. Alku. Neural network-based artificial bandwidth expansion of speech. *IEEE Trans. Audio, Speech, and Language Processing*, 15(3):873–881, 2007.

[154] P. Jax and P. Vary. Feature selection for improved bandwidth extension of speech signals. In *Proc. ICASSP*, pages 697–700, 2004.

[155] A. H. Nour-Eldin and P. Kabal. Memory-based approximation of the Gaussian mixture model framework for bandwidth extension of narrowband speech. In *Proc. INTERSPEECH*, pages 1185–1188, 2011.

[156] U. Kornagel. Techniques for artificial bandwidth extension of telephone speech. *Signal Processing*, 86(6):1296–1306, 2006.

[157] M. Nilsson and W. B. Kleijn. Avoiding over-estimation in bandwidth extension of telephony speech. In *Proc. ICASSP*, pages 869–872, 2001.

[158] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11): 2647–2692, 2002.

[159] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, 1996.

[160] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[161] L. Laaksonen, H. Pulakka, V. Myllylä, and P. Alku. Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal. *IEEE Trans. Consumer Electronics*, 55(2):780–787, 2009.

[162] S. Keronen, H. Kallasjoki, U. Remes, G. J. Brown, J. F. Gemmeke, and K. J. Palomäki. Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment. *Computer Speech and Language*, 27(3):798–819, 2013.

[163] H. Kallasjoki, S. Keronen, G. J. Brown, J. F. Gemmeke, U. Remes, and K. J. Palomäki. Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. In *Proc. CHIME*, 2011.

[164] H. Kallasjoki, U. Remes, J. F. Gemmeke, T. Virtanen, and K. J. Palomäki. Uncertainty measures for improving exemplar-based sparse separation. In *Proc. INTERSPEECH*, pages 469–472, 2011.

[165] H. Kallasjoki, J. F. Gemmeke, and K. J. Palomäki. Estimating uncertainty to improve examplar-based feature enhancement for noise robust speech recognition. *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 22(2):368–380, 2014.

[166] V. Digalakis and L. Neumeyer. Speaker adaptation using combined transformation and Bayesian methods. In *Proc. ICASSP*, pages 680–683, 1995.

[167] A. Gunawardana and W. Byrne. Discounted likelihood linear regression for rapid speaker adaptation. *Computer Speech and Language*, 15(1):15–38, 2001.

[168] M. L. Seltzer and A. Acero. Training wideband acoustic models using mixed-bandwidth training data via feature bandwidth extension. In *Proc. ICASSP*, pages 921–924, 2005.

[169] M. L. Seltzer and A. Acero. Training wideband acoustic models using mixed-bandwidth training data for speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 15(1):235–245, 2007.

[170] D. Macho. Narrowband to wideband feature expansion for robust multilingual ASR. In *Proc. INTERSPEECH*, pages 1118–1121, 2007.

[171] J. Li, D. Yu, J. T. Huang, and Y. Gong. Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM. In *Proc. SLT*, pages 131–136, 2012.

[172] T. Winkler. How realistic is artificially added noise? In *Proc. INTERSPEECH*, pages 2605–2608, 2011.

[173] M. L. Seltzer, D. Yu, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *Proc. ICASSP*, pages 7398–7402, 2013.

[174] T. Yoshioka and M. J. F. Gales. Environmentally robust ASR front-end for deep neural network acoustic models. *Computer Speech and Language*, 31 (1):65–86, 2015.

[175] M. L. Seltzer and A. Acero. Separating speaker and environmental variability using factored transforms. In *Proc. INTERSPEECH*, pages 1097–1100, 2011.

[176] Y. Wang and M. J. F. Gales. Speaker and noise factorization for robust speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 20(7):2149–2158, 2012.

[177] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman. MLLR transforms as features in speaker recognition. In *Proc. EUROSPEECH*, pages 2425–2428, 2005.

[178] M. Ferras, C. C. Leung, C. Barras, and J. L. Gauvain. Comparison of speaker adaptation methods as feature extraction for SVM-based speaker recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 18(6): 1366–1378, 2010.

Errata

Publication V

The normalised lower bound variable in Equations (8) and (9) is missing a minus sign.

Aalto-DD 149/2016

ISBN 978-952-60-6936-4 (printed)
ISBN 978-952-60-6937-1 (pdf)
ISSN-L 1799-4934
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