Contributions to speech processing and ambient sound analysis

Romain Serizel

UNIVERSITÉ DE LORRAINE

Habilitation à diriger des recherches

Romain Serizel

Mémoire présenté en vue de l'obtention d'une
Habilitation de l'Université de Lorraine
Mention Informatique

École doctorale : IAEM

Unité de recherche : Laboratoire Lorrain de Recherche en Informatique et ses Applications UMR 7503

Soutenue le 16 mars 2022
Thèse N°:

Contributions to speech processing and ambient sound analysis

JURY

Rapporteuse :	**Shoko ARAKI**, Chercheuse, NTT Communication Science Laboratories, Kyoto, Japon
Rapporteur :	**Geoffroy PEETERS**, Professeur, Télécom ParisTech, France
Rapporteur :	**Tuomas VIRTANEN**, Professeur, Université de Tampere, Finlande
Examinatrice :	**Marie-Odile BERGER**, Directrie de Recherche, Inria Nancy-Grand Est, Nancy, France, **présidente du jury**
Examinatrice :	**Mounya ELHILALI**, Professeure, Johns Hopkins University, Baltimore, United States
Examinateur :	**Jonathan LEROUX**, Directeur de rechercher, Mitsubishi Electric Research Laboratories, Cambridge, United States

Résumé

Nous sommes constamment entourés de sons que nous exploitons pour adapter nos actions aux situations auxquelles nous sommes confrontés. Certains sons comme la parole peuvent avoir une structure particulière à partir de laquelle nous pouvons déduire des informations, explicites ou non. C'est l'une des raisons pour lesquelles la parole est peut-être le moyen le plus intuitif de communiquer entre humains. Au cours de la décennie écoulée, des progrès significatifs ont été réalisés dans le domaine du traitement de la parole et du son et en particulier dans le domaine de l'apprentissage automatique appliqué au traitement de la parole et du son. Grâce à ces progrès, la parole est devenue un élément central de nombreux outils de communication à distance d'humain à humain ainsi que dans les systèmes de communication humain-machine. Ces solutions fonctionnent bien sur un signal de parole propre ou dans des conditions contrôlées. Cependant, dans les scénarios qui impliquent la présence de perturbations acoustiques telles que du bruit ou de la réverbération les performances peuvent avoir tendance à se dégrader gravement. Dans cette HDR, nous nous concentrons sur le traitement de la parole et de son environnement d'un point de vue audio. Les algorithmes proposés ici reposent sur une variété de solutions allant des approches basées sur le traitement du signal aux solutions orientées données à base de factorisation matricielle supervisée ou de réseaux de neurones profonds. Nous proposons des solutions à des problèmes allant de la reconnaissance vocale au rehaussement de la parole ou à l'analyse des sons ambiants. L'objectif est d'offrir un panorama des différents aspects qui pourraient être améliorer un algorithme de traitement de la parole fonctionnant dans un environnement réel. Nous commençons par décrire la reconnaissance automatique de la parole comme une application finale potentielle et analysons progressivement les limites et les solutions proposées aboutissant à l'analyse plus générale des sons ambiants.

Abstract

We are constantly surrounded by sounds that we continuously exploit to adapt our actions to situations we are facing. Some of the sounds like speech can have a particular structure from which we can infer some information, explicit or not. This is one reason why speech is possibly that is the most intuitive way to communicate between humans. Within the last decade, there has been significant progress in the domain of speech and audio processing and in particular in the domain of machine learning applied to speech and audio processing. Thanks to these progresses, speech has become a central element in many human to human distant communication tools as well as in human to machine communication systems. These solutions work pretty well on clean speech or under controlled condition. However, in scenarios that involve the presence of acoustic perturbation such as noise or reverberation systems performance tends to degrade severely.
In this thesis we focus on processing speech and its environments from an audio perspective. The algorithms proposed here are relying on a variety of solutions from signal processing based approaches to data-driven solutions based on supervised matrix factorization or deep neural networks. We propose solutions to problems ranging from speech recognition, to speech enhancement or ambient sound analysis. The target is to offer a panorama of the different aspects that could improve a speech processing algorithm working in a real environments. We start by describing automatic speech recognition as a potential end application and progressively unravel the limitations and the proposed solutions ending-up to the more general ambient sound analysis.

Table of contents

List of figures

List of tables

Acronyms

ASR *automatic speech recognition*
CI *cochlear implant*
CNN *convolutional neural networks*
CQT *constant-Q transform*
CRNN *stacked convolutional recurrent neural network*
DCASE *Detection and Classification of Acoustic Scenes and Events*
DCT *discrete cosine transform*
DESED *domestic environment sound event detection*
DNN *deep neural network*
EFR *eigen factor radial normalization*
EVD *eigenvalue decomposition*
FBSNR *foreground-to-background signal-to-noise ratio*
GEVD *generalized eigenvalue decomposition*
GMM *gaussian mixture model*
GPGPU *general purpose graphical processing units*
HMM *hidden Markov model*
LSTM *long-short term memory network*
MFCC *Mel-frequency cepstrum coefficients*
MIL *multiple instance learning*
MSE *mean squared error cireterion*
MWF *multichannel Wiener filter*
NIST *National Institute of Standards and Technology*
NMF *nonnegative matrix factorization*
NNLS *nonnegative least square*
NR *noise reduction*
PER *phone error rate*
PLDA *probabilistic linear discriminant analysis*
R1-MWF *rank-1 multichannel Wiener filter*
RBM *restricted Boltzmann machine*
RNN *recurrent neural network*
SD *speech distortion*
SDW-MWF *speech distortion weighted multichannel Wiener filter*
SED *sound event detection and classification*
SIW-SD *intelligibility weighted speech distortion*
SIW-SNR *speech intelligibility-weighted signal-to-noise-ratio*
SNR *signal-to-noise-ratio*
SP-MWF *spatial-prediction multichannel Wiener filter*

TDL *task-driven dictionary learning*
TDNMF *task-driven nonnegative matrix factorization*
TGNMF *task-driven group nonnegative matrix factorization*
UBM *universal background model*
VAD *voice activity detection*
VTLN *vocal tract length normalization*
WER *word error rate*

Notations

$\odot$ element-wise product (Hadamard product)
$\|.\|$ Euclidean norm
$\|.\|_p$ p-norm (note that for $p =$ this is the Euclidean norm, denoted $\|.\|$ for simplicity)
$\|.\|_F$ Frobenius norm
$.^H$ Hermitian transpose
$[.]_{i,j}$ coefficient on the i^{th} row and j^{th} column of a matrix
$\mathbf{A}$ M-dimensional steering vector from the speech source to the microphone
b the bias of the layer
G number of speakers
g speaker index
^{g} current speaker
$\mathcal{C}$ sound event classes ensemble
c sound event class
D separable divergence
$\mathbf{e}_1$ all-zero vector except for a one in the first position
E error signal
$\mathbb{E}\{.\}$ expectation
F is the number of frequency components
FP_c number of false positives for the sound event class c
FN_c false negative for the sound event class c
$\mathbf{h}$ output of a hidden layer
$\mathbf{H}$ NMF activations
J cost function
K number of components in the NMF decomposition
K_{RES} number of components in the residual bases
K_{SES} number of components in the session-dependent bases
K_{SPK} number of components in the speaker-dependent bases
l segment index
$\mathcal{L}_s$ classification loss
$\mathcal{L}_{class_s}$ classification loss on strong labels (sound event detection)
$\mathcal{L}_{class_w}$ classification loss on weak labels (audio tagging)
$\mathcal{L}_{cons_s}$ consistency loss on strong labels (sound event detection)
$\mathcal{L}_{cons_w}$ consistency loss on weak labels (audio tagging)
M number of microphones
n frame index
N the number of frames
$n_{\mathcal{C}}$ number of sound event classes

N_{Mat} total number of estimated $\mathbf{R}_{s_{r1}}$
NPD number of estimated $\mathbf{R}_{s_{r1}}$ that are not positive semi-definite
P_s power of the speech source signal
q a HMM state
$\mathbf{R}_x$ autocorrelation matrix for the signal x
$\tilde{\mathbf{R}}_x$ estimate of the autocorrelation matrix for the signal x
$\mathbf{R}_{s_{r1}}$ rank 1 approximation of the matrix $\mathbf{R}_s$
$\mathbf{R}_{\text{rem}}$ "remainder" matrix
S recording sessions
s session index
^{s} current session
TP_c number of true positives for the sound event class c
$\mathbf{W}$ NMF dictionary
$\mathcal{W}$ set of nonnegative dictionaries containing unit l_2-norm basis vectors
$\mathbf{w}$ MWF (the filter type is declared in subscript)
X is the observation
$x_m(\omega)$ signal recorded at microphone m
$x_{m,n}(\omega)$ noise component of the signal recorded at microphone m
$x_{m,s}(\omega)$ speech component of the signal recorded at microphone m
$\mathbf{x}$ compound vector gathering the microphone signals
y label
$\mathcal{Y}$ set of labels
$\mathbf{y}$ the vector of input to a layer
z output of a filter
η noise applied on the student model
η' noise applied on the teacher model
$\mathbf{\Phi}$ parameters of the classifier
λ exponential forgetting factor
λ_1, λ_2 are nonnegative regularization parameters
μ SDW-MWF trade-off parameter
ν regularization parameter on the classifier parameters
$\sigma_{i,j}$ cross-correlation term between channel i and j
σ_{n_i} noise power on the ith microphone
σ_{s_i} speech power on the ith microphone
$\mathbf{\Sigma}_n$ diagonal matrix with the generalized eigenvalues of the noise
$\mathbf{\Sigma}_s$ diagonal matrix with the generalized eigenvalues of the speech
θ_s weights of the student sound event detection model (strong predictions)
θ weights of the student audio tagging model (weak predictions)
θ'_s weights of the teacher sound event detection model (strong predictions)
θ' weights of the teacher audio tagging model (weak predictions)
ω angular frequency
$\mathbf{\Omega}$ the weights of a layer

Curriculum Vitæ

1 Research interests

- Digital signal processing for speech, audio for hearing aid and cochlear implants application, multichannel noise reduction, adaptive filtering
- Sound scene analysis, ambient sound recognition
- Machine learning applied to signal processing, automatic speech recognition, (deep) neural networks for acoustic modeling, speaker identification

2 Degrees and positions held

- **Since 2016:** Associate professor, Université de Lorraine (FR)
- **2014 – 2016:** Post-doctoral researcher, Télécom Paristech (FR)
- **2013 – 2014:** Post-doctoral researcher, Fondazione Bruno Kessler (IT)
- **2011 – 2012:** Post-doctoral researcher, KU Leuven (BE)
- **2006 – 2011:** PhD, KU Leuven (BE), *Integrated active noise control and noise reduction in hearing aids* (average PhD duration in the department: 5 years)
- **Oct – Dec 2009:** Guest Researcher, Aalborg University (DK)
- **2006:** MSc, Université Rennes 1/ENSSAT (FR)
- **2005:** MEng, ENSEM (FR)

3 Project responsibilities

- **Coordinator** at Université de Lorraine (FR) for the project DiSCogs: *Far-field speech enhancement with ad-hoc antennas.* (ANR JCJC project, French young researcher grant), 2018 – 2022
- **Researcher** at Université de Lorraine (FR) for the project ROBOVOX: *Speaker identification with moving robots.* (ANR PRCE project with two academic partners and one industrial partner), 2019 – 2023
- **Researcher** at Université de Lorraine (FR) for the project LEAUDS: *Learning to understand sounds.* (ANR PRCE project with two academic partners and one industrial partner), 2019 – 2023
- **Researcher** at Université de Lorraine (FR) for the project CPS4EU: *Cyber-physical systems for Europe.* (PSPC/ECSEL project), 2019 – 2022
- **Local coordinator, Task co-leader, Use case leader** at Télécom Paristech (FR) for the project LASIE: *Speaker identification and sound scenes analysis.* (FP7 project with 18 partners both academic and industrial) & 2014 – 2016

4 Scientific responsibilities and supervision

- 4 PhD students co-supervised (thesis defended), co-supervising 3 PhD students
- **General co-chair** of DCASE challenge since 2019
- **Coordinator** of task 4 for DCASE since 2018
- **Member** of DCASE steering group since 2019
- **Technical committee co-chair** for ICASSP 2021, 2022 and WASPAA 2021, area: Detection and Classification of Acoustic Scenes and Events
- **Associate editor** for IEEE/ACM Transactions on Audio, Speech and Language processing since 2021
- **Member** of IEEE acoustic and audio signal processing technical committee since 2019
- **Member** of the organization committee PFIA 2018
- **Member** of the commission for the scientific staff (COMIPERS) of the research center Inria Nancy - Grand Est, since 2018
- **Member** of the commission for the technological developpment (CDT) of the research center Inria Nancy - Grand Est, since 2017
- Special session at EUSIPCO 2015 and 2018
- **Reviewer** for international scientific journals (IEEE/ACM Transactions on Audio, Speech and Language processing, IEEE Journal of Selected Topics in Signal Processing, IEEE Transactions on Emerging Topics in Computational Intelligence, Signal Processing, EURASIP Journal on Audio, Speech and Music Processing, The Journal of the Acoustical Society of America...) and international scientific conferences (ICASSP, EUSIPCO, WASPAA. . .)
- **Project reviewer :** ANR, MITACS (Canada), FWO (Belgium) and NWO (Netherlands)

5 Software

- Beta NMF: Theano based GPGPU implementation of NMF with beta-divergence and multiplicative updates.
- Group NMF: Theano based GPGPU implementation of group-NMF with class and session similarity constraints. The NMF works with beta-divergence and multiplicative updates.
- Mini batch NMF: Theano based GPGPU implementation of NMF with beta-divergence and mini-batch multiplicative updates.
- Supervised (group) NMF: Python code to perform task-driven NMF and task-driven group NMF
- DCASE 2018 baseline: Sound event detection system, code in Python, International audience, baseline for DCASE 2018.
- DCASE 2019 baseline: Sound event detection system, code in Python, International audience, baseline for DCASE 2019.
- DCASE 2020 baseline: Sound event detection and separation system, code in

Python, International audience, baseline for DCASE 2020.

6 Datasets

- **DCASE 2018 – Task 4:** Dataset for sound event detection with heterogeneous level of annotation. Contribution: dataset design, data annotation and dataset distribution.
- **DESED dataset:** Flexible dataset for sound event detection with heterogeneous level of annotation.
  Contribution: dataset design, data annotation and dataset distribution. Downloaded 4500 times since its release

7 Dissemination to the society

- Participation to *Fête de la science* – science fair (2017, 2018) : separate the sound sources
- Participation to *Nuit des chercheurs – Researchers' night* (2019) : Sound source localization with a robot

8 Distinctions

- 2020-2024: Doctoral supervision and research bonus (PEDR), rank A (top 20%)

9 Publications

9.1 Journal articles

[J1]N. Furnon, R. Serizel, S. Essid and I. Illina. DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays. In *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2021
[J2] Guillaume Carbajal, Romain Serizel, Emmanuel Vincent, Eric Humbert. Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise. In *EEE/ACM Transactions on Audio, Speech and Language Processing*, 2020
[J3] Antoine Deleforge, Diego Di Carlo, Martin Strauss, Romain Serizel, Lucio Marcenaro. Audio-Based Search and Rescue with a Drone: Highlights from the IEEE Signal Processing Cup 2019 Student Competition. *IEEE Signal Processing Magazine*, 2019
[J4] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin. CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 2019, Special Issue on Acoustic Source Localization and Tracking in Dynamic Real-life Scenes

[J5] Ziteng Wang, Emmanuel Vincent, Romain Serizel, Yonghong Yan. Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments. *Computer Speech and Language*, Elsevier, 2018
[J6] Victor Bisot, Romain Serizel, Slim Essid, Gael Richard. Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017
[J7] Romain Serizel, Diego Giuliani. Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *In Natural Language Engineering*, Cambridge University Press (CUP), 2016
[J8] Romain Serizel, Marc Moonen, Bas van Dijk, Jan Wouters. Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants. I*n IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2014
[J9] Romain Serizel, Marc Moonen, Jan Wouters, Søren Holdt Jensen. A Speech Distortion Weighting Based Approach to Integrated Active Noise Control and Noise Reduction in Hearing Aids. *Signal Processing*, Elsevier, 2013
[J10] Romain Serizel, Marc Moonen, Jan Wouters, Søren Jensen. Binaural Integrated Active Noise Control and Noise Reduction in Hearing Aids. *In IEEE Transactions on Audio, Speech and Language Processing*, 2013
[J11] Romain Serizel, Marc Moonen, Jan Wouters, Søren Jensen. A Zone of Quiet Based Approach to Integrated Active Noise Control and Noise Reduction for Speech Enhancement in Hearing Aids. *IEEE Transactions on Audio, Speech and Language Processing*, 2012
[J12] Romain Serizel, Marc Moonen, Jan Wouters, Søren Holdt Jensen. Output SNR analysis of integrated active noise control and noise reduction in hearing aids under a single speech source scenario. *Signal Processing*, Elsevier, 2011
[J13] Romain Serizel, Marc Moonen, Jan Wouters, Søren Jensen. Integrated active noise control and noise reduction in hearing aids. *IEEE Transactions on Audio, Speech, and Language Processing*, 2009,
[J14] Daniel Menard, Romain Serizel, Romuald Rocher, Olivier Sentieys. Accuracy Constraint Determination in Fixed-Point System Design. *In EURASIP Journal on Embedded Systems*, SpringerOpen, 2008

9.2 Conference papers

[C1] Felix Gontier, Romain Serizel, Christophe Cerisara. Automated Audio Captioning by Fine-Tuning BART with AudioSet Tags. In Proc. *In Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021
[C2] Francesca Ronchini, Romain Serizel, Nicolas Turpault, Samuele Cornell. The impact of non-target events in synthetic soundscapes for sound event detection. *In Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021
[C3] Mohammad Mohammadamini, Driss Matrouf, Jean-Francois Bonastre, Romain Serizel, Sandipana Dowerah, Denis Jouvet. Compensate multiple dis-

tortions for speaker recognition systems. In Proc. *European Signal Processing Conference (EUSIPCO)*, Aug 2021

[C4] Nicolas Furnon, Romain Serizel, Irina Illina, Slim Essid. Distributed speech separation in spatially unconstrained microphone arrays. In Proc. *European Signal Processing Conference (EUSIPCO)*, Aug 2021

[C5] Nicolas Furnon, Romain Serizel, Irina Illina, Slim Essid. DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021

[C6] Scott Wisdom, Hakan Erdogan, Daniel Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, John Hershey, What's All the FUSS About Free Universal Sound Separation Data? *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021

[C7] Nicolas Turpault, Romain Serizel, Scott Wisdom, Hakan Erdogan, John Hershey, Eduardo Fonseca, Prem Seetharaman, Justin Salamon. Sound Event Detection and Separation: a Benchmark on Desed Synthetic Soundscapes. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021

[C8] Giacomo Ferroni, Nicolas Turpault, Juan Azcarreta, Francesco Tuveri, Romain Serizel, Çagdaş Bilen, Sacha Krstulović. Improving Sound Event Detection Metrics: Insights from DCASE 2020. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021

[C9] Michel Olvera, Emmanuel Vincent, Romain Serizel, Gilles Gasso. Foreground-Background Ambient Sound Scene Separation. In Proc. *European Signal Processing Conference (EUSIPCO)*, Jan 2021

[C10] Nicolas Turpault, Scott Wisdom, Hakan Erdogan, John Hershey, Romain Serizel, Eduardo Fonseca, Prem Seetharaman, Justin Salamon. Improving Sound Event Detection In Domestic Environments Using Sound Separation. *In Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov 2020

[C11] Nicolas Turpault, Romain Serizel. Training Sound Event Detection On A Heterogeneous Dataset. *In Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov 2020

[C12] Nicolas Turpault, Romain Serizel, Emmanuel Vincent. Limitations of weak labels for embedding and tagging. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2020

[C13] Romain Serizel, Nicolas Turpault, Ankit Shah, Justin Salamon. Sound event detection in synthetic domestic environments. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2020

[C14] Nicolas Furnon, Romain Serizel, Irina Illina, Slim Essid. DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2020

[C15] Lauréline Perotin, Alexandre Défossez, Emmanuel Vincent, Romain

Serizel, Alexandre Guérin. Regression versus classification for neural network based audio source localization. *In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2019

[C16] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. *In Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Oct 2019

[C17] Romain Serizel, Nicolas Turpault. Sound Event Detection from Partially Annotated Data: Trends and Challenges. *In Proc. IcETRAN conference*, Jun 2019

[C18] Nicolas Turpault, Romain Serizel, Emmanuel Vincent. Semi-supervised triplet loss based learning of ambient audio embeddings. *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019

[C19] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, Ankit Parag Shah. Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments. *In Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE),* Nov 2018

[C20] Laurréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin. CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector *In Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep 2018

[C21] Mathieu Fontaine, Fabian-Robert Stöter, Antoine Liutkus, Umut Simsekli, Romain Serizel, Roland Badeau. Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition. *In Proc. Latent Variable Analysis and Signal Separation (LVA/ICA)*, Jul 2018

[C22] Laurréline Perotin, Romain Serizel, Emmanuel Vincent, Alexandre Guérin. Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr 2018

[C23] Guillaume Carbajal, Romain Serizel, Emmanuel Vincent, Eric Humbert. Multiple-input neural network-based residual echo suppression. *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr 2018

[C24] Victor Bisot, Romain Serizel, Slim Essid, Gael Richard. Nonnegative Feature Learning Methods for Acoustic Scene Classification. *In Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, Nov 2017

[C25] Victor Bisot, Romain Serizel, Slim Essid, Gael Richard. Leveraging deep neural networks with nonnegative representations for improved environmental sound classification. *In Proc. IEEE International Workshop on Machine Learning for Signal Processing MLSP*, Sep 2017

[C26] Romain Serizel, Victor Bisot, Slim Essid, Gael Richard. Supervised Group Nonnegative Matrix Factorisation With Similarity Constraints And Applications To Speaker Identification. *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar 2017

[C27] Romain Serizel, Victor Bisot, Slim Essid, Gael Richard. Machine listening techniques as a complement to video image analysis in forensics. *In IEEE International Conference on Image Processing (ICIP)*, Sep 2016

[C28] Romain Serizel, Slim Essid, Gael Richard. Mini-batch stochastic approaches for accelerated multiplicative updates in nonnegative matrix factorisation with beta-divergence. *In IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep 2016

[C29] Romain Serizel, Slim Essid, Gael Richard. Group Non-Negative Matrix Factorisation With Speaker And Session Similarity Constraints For Speaker Identification. *In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar 2016

[C30] Victor Bisot, Romain Serizel, Slim Essid, Gael Richard. Acoustic scene classification with matrix factorization for unsupervised feature learning. *In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar 2016

[C31] Romain Serizel, Diego Giuliani. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. *In Proc. Spoken Language Technology Workshop (SLT)*, Dec 2014

[C32] Romain Serizel, Diego Giuliani. Deep neural network adaptation for children's and adults' speech recognition. *In Proc. Italian Computational Linguistics Conference (CLiC-it)*, Dec 2014

[C33] Romain Serizel, Marc Moonen, Bas Dijk, Jan Wouters. Rank-1 Approximation Based Multichannel Wiener Filtering Algorithms For Noise Reduction In Cochlear Implants. *In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013

[C34] Romain Serizel, Marc Moonen, Jan Wouters, Søren Holdt Jensen. Output SNR analysis of integrated active noise control and noise reduction in hearing aids under a single speech source scenario. *In Proc. European Signal Processing Conference (EUSIPCO)*, Aug 2010

[C35] Romain Serizel, Marc Moonen, Jan Wouters, Søren Jensen. A Zone of Quiet Based Approach to Integrated Active Noise Control and Noise Reduction in Hearing Aids. *In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2009

[C36] Romain Serizel, Marc Moonen, Jan Wouters, Søren Jensen. A Weighted Approach for Integrated Active Noise Control and Noise Reduction in Hearing Aids. *In Proc. European Signal Processing Conference (EUSIPCO)*, Aug 2009

[C37] Romain Serizel, Marc Moonen, Jan Wouters, Søren Jensen. Combined Active Noise Control and noise reduction in Hearing Aids. *In Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sep 2008

[C38] Daniel Menard, Romain Serizel, Romuald Rocher, Olivier Sentieys. Noise model for Accuracy Constraint Determination in Fixed-Point Systems. *In Proc. Workshop on Design and Architectures for Signal and Image Processing DASIP 2007*, Nov 2007

9.3 Book chapters

[B1] Romain Serizel, Victor Bisot, Slim Essid, Gael Richard. Acoustic Features for Environmental Sound Analysis. *In Tuomas Virtanen; Mark D. Plumbley; Dan Ellis. (Eds.) Computational Analysis of Sound Scenes and Events*, Springer International Publishing AG, pp.71-101, 2017

[B2] Slim Essid, Sanjeel Parekh, Ngoc Duong, Romain Serizel, Alexey Ozerov, Fabio Antonacci, Augusto Sarti. Multiview Approaches to Event Detection and Scene Analysis. *In Tuomas Virtanen; Mark D. Plumbley; Dan Ellis. (Eds.) Computational Analysis of Sound Scenes and Events*, Springer, pp.243-276, 2017

1 Introduction

We, humans, are constantly surrounded by sounds. These complex soundscapes are one of the most important source of information regarding what is happening around us. We rely continuously on these sound to adapt our actions to the situations we are facing (e.g., am I in a quiet or a noisy environment?), to react to events (e.g., I hear someone entering the room, I hear a baby crying...) or to detect dangers (e.g., I hear a car passing, a dog barking...), consciously or not. One great advantage of sound over vision is that we can perceive sounds, in low light conditions or even in total darkness (At night, I can hear that my house is quiet), at 360 degree (I can hear the car coming from behind) and even through obstacles to some extent (I can hear the baby crying in another room). Some of the sounds can have a particular structure from which we can infer information, explicit or not. Speech is one example of structured sound signals, music is another one. Speech is possibly the most intuitive way to communicate between humans. One of the reason for this is that speech is explicitly informative. In a regular day we have tens of speech based interactions with other persons: we listen to news on radio, watch people talking on TV...

Within the last decade, there have been significant progresses in the domain of speech and audio processing and in particular in the domain of machine learning applied to speech and audio processing. Thanks to these progresses, speech has become a central element in many human to human distant communication tools as well as in human to machine communication systems. One typical example is the rapid development of smart speakers during the past years. Current algorithms now allows for a speech recognition quality or speaker identification accuracy that makes the use of such systems acceptable by the general public whereas before they were often considered as annoyingly unreliable. These solutions work pretty well on clean speech or under controlled condition. This can be true when using for example a close-up microphone but it using this type of microphone would be too constraining for some general public applications that have to rely on distant microphones. This latter scenario usually involves the presence of acoustic perturbations such as noise, concurrent speakers or reverberation which tends to degrades systems performance severely. An additional source of degradation is related to the speaker himself/herself. Indeed, if most systems relying on spoken language are fairly reliable for English speaking adults, performance can vary drastically depending on the amount of data available in a particular language as well as for a certain speaker gender or age.

There are several possible solutions to the problem mentioned above. A first solution is to deal with this at the model level by designing speech processing algorithms that are robust to additive noise, reverberation, speaker variabilities... Such algorithms can be trained in acoustic conditions that match the test conditions when the amount of

data available is sufficient. System flexibility can then be achieved with multi-condition training (including training data matching several test conditions). When the amount of data is not sufficient we can rely on domain adaptation in order to adapt a general model to a specific test domain. A second solution is to deal with variabilities at the signal level. In the case of noise and reverberation, this approach would rely on speech enhancement algorithms, more precisely noise reduction and dereverberation algorithms applied as a pre-processing. One problem with these algorithms is that they tend to introduce artifacts on the speech while attenuating the perturbations. When several microphones are available, it is possible to exploit spatial information to extract the desired speech signal from on noisy signal while limiting the amount of artifacts introduced on the processed speech. This kind of setup is now widely used in most handsfree communication systems.

For years, research in audio was mainly focusing on speech, music and to a lesser extent on animal sounds. However, the soundscapes around us are containing a much wider variety of sounds usually referred to as ambient sounds. As human we do rely heavily on these ambient sounds to adapt our behavior consciously or (should I pay a particular attention because a car is passing by? My baby is crying and probably need me; I'm in a noisy place so I should speak louder and slower...). Inspired by these human behavior, research on the automatic analysis and classification of ambient sounds has been attracting a consistently growing attention during the past decade. This has been motivated among other aspects by the possible applications to context awareness for speech communication algorithms or robots, home assisted living or security to name a few.

In this thesis we focus on processing speech and its environments from an audio perspective. The algorithms proposed here are relying on a variety of solutions from signal processing based approaches to data-driven solutions based on supervised matrix factorization or deep neural networks. We propose solutions to problems ranging from speech recognition, to speech enhancement or ambient sound analysis. The target is to offer a panorama of the different aspects that could be involved in a speech processing algorithm working in a real environment. We start by describing automatic speech recognition as a potential end application and progressively unravel the limitations and proposed solution ending-up to the more general ambient sound analysis. More precisely, the thesis is organized as follows.

Chapter 2: We explore the problem of speech recognition with varying age classes and gender. Developmental changes in the voice production system result in variabilities than can be a major source of errors for automatic speech recognition systems. This is particularly true when considering speech from population for which large scale data collection could be cumbersome or inappropriate (e.g. children). We propose adaptation methods for neural networks based acoustic models that allows for automatic speech recognition in under-resourced conditions with an heterogeneous population of speakers.

Chapter 3: We explore the problem of speaker recognition in possibly noisy environments. As indicated in Chapter 2, speaker variation could be a source of errors in speech recognition systems. Knowing the speaker identity could greatly help reducing these

degradations. The target here is to address the problem when the signal presented to the speaker recognition algorithm is not clean speech but is corrupted by background noise. We propose a matrix factorization approach where specific dictionaries are learned for different speakers and noise conditions.

Chapter 4: We explore the problem of noise compensation in far-field speech communication system. Speech captured with distant microphones in real environments is distorted and attenuated during the propagation from the speech source to the microphones and often corrupted by additive background noise. We propose to address the latter problem using multichannel filtering algorithms, more precisely multichannel Wiener filters. In real environment the estimation of these filters can become unstable because of the sometimes unrealistic simplification assumptions made to allow for estimating the filters. We propose a filter estimation based on generalized eigenvalue decomposition that allows for a stable filter estimation even in challenging condition as well as improved speech enhancement performance.

Chapter 5 and 6: We explore the problem of sound event detection. All the aforementioned algorithms can depend heavily on the acoustic context including (type of surrounding noise, stationnarity, interfering event density, relative signal to noise level...). We propose to benchmark the state-of-the-art systems performance on a dataset designed for the DCASE challenge. We propose a detailed analysis of the limitations of the systems submitted to the task we organize for the DCASE challenge. We first focus on the problem of learning a sound event segmentation from training clips without temporal segmentation. Then we exploit the possibility offered by synthetic soundscapes generation to design dataset targeting specific scientific problems in order to highlight the systems improvements and remaining limitations on these problems.

Chapter 7: We present the conclusions of the thesis and of the past decade of work in research in different institutions in Europe. We then propose perspectives for the coming years. In particular, most of the approaches presented or analyzed here require a large amount of data and computational resources. This can have a significant impact on our environment while sometimes producing only limited benefits. We propose to focus on low footprint algorithms, targeting audio applications that can have a positive impact on the environment.

The work presented here spans over a decade and each chapter corresponds to a specific period of time. We decided to present the contributions in the context of the works done at that time and indicate the time period when the work was done. The work we have done since then in relation to each contribution presented here is summarized at the end of each corresponding chapters. Note also that the contribution are not order chronologically but organized as described above.

2 Speech recognition for children

Context: This work was done when I was a postdoctoral researcher at Fondazione Bruno Kessler (Trento, Italy) between January 2013 and September 2014 together with Diego Guiliani. The work presented here has been previously published in articles [Serizel and Giuliani, 2014a,b, 2017].

2.1 Introduction

Speaker-related acoustic variability is a major source of errors in automatic speech recognition. In this chapter we cope with age group differences, by considering the relevant case of children versus adults, as well as with male/female differences. Here a deep neural network (DNN) is used to deal with the acoustic variability induced by age and gender differences.

Developmental changes in speech production introduce age-dependent spectral and temporal variabilities in speech produced by children. Studies on morphology and development of the vocal tract [Fitch and Giedd, 1999] reveal that during the childhood there is a steady gradual lengthening of the vocal tract as the child grows while a concomitant decrease in formant frequencies occurs [Huber et al., 1999, Lee et al., 1999]. In particular, for females there is an essential gradual continuous growth of vocal tract through puberty into adulthood, while for males during puberty there is a disproportionate growth of the vocal tract, which lowers formant frequencies, together with an enlargement of the glottis, which lowers the pitch. After age 15, males show a substantial longer vocal tract and lower formant frequencies than females. As consequence, voices of children tend to be more similar to the voices of women than to those of men.

When an automatic speech recognition (ASR) system trained on adults' speech is employed to recognize children's speech, performance decreases drastically, especially for younger children [Claes et al., 1998, Das et al., 1998, Gerosa et al., 2007, 2009b, Giuliani and Gerosa, 2003, Li and Russell, 2001, Potamianos and Narayanan, 2003, Wilpon and Jacobsen, 1996]. A number of attempts have been reported in the literature to contrast this effect. Most of them try to compensate for spectral differences caused by differences in vocal tract length and shape by warping the frequency axis of the speech power spectrum of each test speaker or transforming acoustic models [Claes et al., 1998, Das et al., 1998, Potamianos and Narayanan, 2003]. However, to ensure good recognition performance, age-specific acoustic models trained on speech collected from children of the target age, or group of ages, is usually employed [Gerosa et al., 2007, Hagen et al., 2003, Nisimura et al., 2004, Wilpon and Jacobsen, 1996]. Typically much less training data are available for children than for adults. The use of adults' speech for reinforc-

ing the training data in the case of a lack of children's speech was investigated in the past [Steidl et al., 2003, Wilpon and Jacobsen, 1996]. However, in order to achieve a recognition performance improvement when training with a mixture of children's and adults' speech, speaker normalization and speaker adaptive training techniques are usually needed [Gerosa et al., 2009a].

How to cope with acoustic variability induced by gender differences has been studied for adult speakers in a number of papers. Assuming that there was enough training data, one approach consisted in the use of gender-dependent models that are either directly used in the recognition process itself [Woodland et al., 1994, Yochai and Morgan, 1992] or used as a better seed for speaker adaptation [Lee and Gauvain, 1993]. Alternatively, when training on speakers of both genders, speaker normalization and adaptation techniques were commonly employed to contrast acoustic inter-speaker variability Gales [1998], Lee and Rose [1996].

Since the surfacing of efficient pre-training algorithms during the past years [Bengio et al., 2007, Erhan et al., 2010, Hinton et al., 2006, Seide et al., 2011], DNN has proven to be an effective alternative to Gaussian mixture models (GMM) in hidden Markov models (HMM) based ASR [Bourlard and Morgan, 1994, Hinton et al., 2012] and really good performance has been obtained with hybrid DNN-HMM systems [Dahl et al., 2012, Mohamed et al., 2012].

Capitalizing on their good classification and generalization capabilities the DNN have been used widely in multi-domain and multi-languages tasks [Sivadas and Hermansky, 2004, Stolcke et al., 2006]. The main idea is usually to first exploit a task independent (multi-lingual/multi-domain) corpus and then to use a task specific corpus. These different corpora can be used to design new DNN architectures with application to task specific ASR [Pinto et al., 2009] or task independent ASR [Bell et al., 2013]. Another approach consists in using the different corpora at different stages of the DNN training. The task independent corpus is used only for the pre-training [Swietojanski et al., 2012] or for a general first training [Le et al., 2010, Thomas et al., 2013] and the task specific corpus is used for the final training/adaptation of the DNN. In under-resourced scenarios, approaches based on DNN [Imseng et al., 2013] have then shown to outperform approaches based on subspace GMM [Burget et al., 2010]. However, to our best knowledge, at the time of the study presented in this chapter (2013-2014), apart from the very recent work on the subject by Metallinou and Cheng [2014] DNN was scarcely used in the context of children's speech recognition.

Three target groups of speakers are considered in this chapter, that is children, adult males and adult females. There is only a limited amount of labeled data for such groups. We investigate two approaches for ASR in under-resourced conditions with an heterogeneous population of speakers.

The first approach investigated in this chapter extends the idea introduced by Yochai and Morgan [1992] to the DNN context. The DNN trained on speech data from all the three groups of speakers is adapted to the age/gender group specific corpora. First it is shown that training a DNN only from a group specific corpus is not effective when only a limited amount of labeled data is available. Then the method proposed by Thomas

et al. [2013] is adapted to the age/gender specific problem and used in a DNN-HMM architecture instead of a tandem architecture.
The second approach introduced in this chapter relies on vocal tract length normalization (VTLN). Seide et al. [2011] conducted an investigation by training a DNN on VTLN normalised acoustic features, it was found that in a large vocabulary adults' speech recognition task limited gain can be achieved with respect to using acoustic features without normalization. It was argued that, when a sufficient amount of training data is available, DNN are already able to learn, to some extent, internal representations that are invariant with respect to sources of variability such as the vocal tract length and shape. However, when only a limited amount of training data is available from a heterogeneous population of speakers, made of children and adults as in our case, the DNN might not be able reach strong generalization capabilities [Serizel and Giuliani, 2014a]. In such case, techniques like DNN adaptation [Le et al., 2010, Swietojanski et al., 2012, Thomas et al., 2013], speaker adaptation [Abdel-Hamid and Jiang, 2013, Liao, 2013] or VTLN [Eide and Gish, 1996, Lee and Rose, 1996, Wegmann et al., 1996] can help to improve the performance. Here we consider first the application of a conventional VTLN technique to normalize mel-frequency cepstrum coefficients (MFCC) as input features to a DNN-HMM system.
Shortly before we conducted work on ASR for children, it was shown that augmenting the inputs of a DNN with, e.g. an estimate of the background noise [Seltzer et al., 2013] or utterance i-vector [Senior and Lopez-Moreno, 2014], can improve the robustness and speaker independence of the DNN. We then propose to augment the MFCC inputs of the DNN with the posterior probabilities of the VTLN-warping factors to improve the robustness with respect to inter-speaker acoustic variations.
An approach to optimize jointly the DNN that extracts the posterior probabilities of the warping factors and the DNN-HMM is proposed here, combination of the different approaches is considered and the different systems performance are evaluated not only on phone recognition but also on word recognition.
The rest of the chapter is organized as follows, Section 2.2 briefly introduces DNN for acoustic modeling in ASR and present the approach based on DNN adaptation. Approaches based on VTLN are presented in Section 2.3. The experimental setup is described in Section 2.4 and experiments results are presented in Section 2.5. Finally, conclusions of the chapter are drawn in Section 2.6.

2.2 DNN adaptation

The DNNs used here are feedforward neural networks where the neurons are arranged in fully connected layers (so-called multi-layer perceptrons). The input layer processes the feature vectors (augmented with context) and the output layer provides (in the case of ASR) the posterior probability of the (sub)phonetic units. The DNN used here have

sigmoid activation functions in the hidden layers:

$$h = \mathbf{\Omega}.\mathbf{y} + b$$
$$\text{sigmoid}(h) = \frac{1}{1+e^{-h}},$$

with $\mathbf{y}$ the vector of input to the layer, $\mathbf{\Omega}$ the weights of the layer and b the bias of the layer.
The target of the DNN presented here is to estimate posteriors probabilities. Therefore, it is chosen to use softmax activation in the output layer, as the the output then sum up to one:

$$\text{softmax}(\mathbf{h}_j) = \frac{e^{h_j}}{\sum_i e^{h_i}}.$$

When used in a DNN-HMM context, the posterior probabilities are normalised by the prior probabilities of the state to obtain the state emission likelihood used by the HMM. Following Bayes rule:

$$p(X|q) \propto \frac{p(q|X)}{p(q)}.$$

Where X is the observation and q the HMM state.

2.2.1 Pre-training/training procedure

Training a DNN is a difficult tasks mainly because the optimization criterion involved is non convex. Training a randomly initialized DNN with back-propagation would converge to one of the many local minima involved in the optimization problem sometimes leading to poor performance [Erhan et al., 2010]. In recent works this limitation has been partly overcome by training on a large amount of data (1700 hours [Senior and Lopez-Moreno, 2014]). However, this solution does not apply when tackling ASR for under-resourced groups of population where the amount of training data is limited by definition. In such cases, pre-training is a mandatory step to efficiently train a DNN. The aim of pre-training is to initialize the DNN weights to a better starting point than randomly initialized DNN and avoid the back-propagation training to be stuck in a poor local minimum. Here generative training based on Restricted Boltzmann Machines (RBM) [Erhan et al., 2010, Hinton et al., 2006] is chosen. Once the DNN weights have been initialized with stacked RBM, the DNN is trained to convergence with back-propagation. More details about training and network parameters are presented in Sections 2.4.2.2 and 2.4.3.2.

2.2.2 Age/gender independent training

The general training procedure described above can be applied, by using all training data available, in an attempt to achieve a system with strong generalization capabilities. Estimating the DNN parameters on speech from all groups of speakers, that is children, adult males and adult females, may however, have some limitation due to the heterogeneity of the speech data that may negatively impact on the classification accuracy compared to group-specific DNN.

2.2.3 Age/gender adaptation

ASR systems provide their best recognition performances when the operating (or testing) conditions match the training conditions. To be effective, the general training procedure described above requires that a sufficient amount of labeled data is available. Therefore, when considering training for under-resourced population groups (such as children or males/females in particular domains of applications) it might be more effective to train first a DNN on all data available and then to adapt this DNN to a specific group of speakers. A similar approach has been proposed by Thomas et al. [2013] for the case of multilingual training. Here the language does not change and the targets of the DNN remain the same when going from age/gender independent training to group specific adaptation. The DNN trained on speech data from all groups of speakers can then be used directly as initialization to the adaptation procedure where the DNN is trained to convergence with back-propagation only on group specific speech corpora.
This adaptation approach, however, suffers from a lack of flexibility: a new DNN would have to be adapted to each new group of speakers.

2.3 VTLN based approaches

In this section, we propose to define a more general framework inspired by VTLN approaches to ASR to tackle the problem of inter-speaker acoustic variability due to vocal tract length (and shape) variations among speakers. Two different approaches are considered here. The first one is based on the conventional VTLN approach [Eide and Gish, 1996, Lee and Rose, 1996, Wegmann et al., 1996]. The resulting VTLN normalized acoustic features are used as input to the DNN during both training and test [Seide et al., 2011]. The second approach, proposed in this chapter, has two main characteristics: a) by using a dedicated DNN, for each speech frame the posterior probability of each warping factor is estimated and b) for each speech frame the vector of the estimated warping factor posterior probabilities is appended to the acoustic features vector, extended with context, to form an augmented acoustic features vector for the DNN-HMM system.

2.3.1 VTLN normalised features as input to the DNN

In the conventional frequency warping approach to speaker normalization [Eide and Gish, 1996, Lee and Rose, 1996, Wegmann et al., 1996], typical issues are the estimation of a proper frequency scaling factor for each speaker, or utterance, and the implementation of the frequency scaling during speech analysis. A well known method for estimating the scaling factor is based on a grid search over a discrete set of possible scaling factors by maximizing the likelihood of warped data given a current set of HMM-based acoustic models [Lee and Rose, 1996]. Frequency scaling is performed by warping the power spectrum during signal analysis or, for filter-bank based acoustic front-end, by changing the spacing and width of the filters while maintaining the spectrum unchanged [Lee and Rose, 1996]. In this work we adopted the latter approach considering a discrete set of VTLN factors. Details on the VTLN implementation are provided in Section 2.4.5.

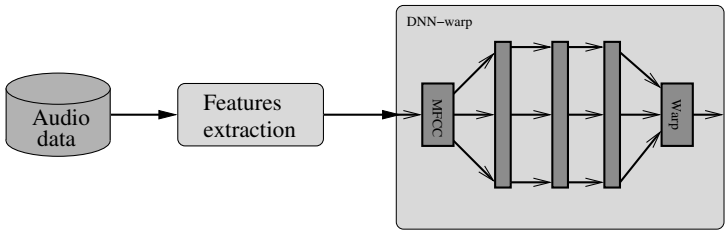


Figure 2.1: Training of the DNN-warp.

Similarly to the method proposed by Seide et al. [2011], the VTLN normalized acoustic features are used to form the input to the DNN-HMM system both during training and testing.

2.3.2 Posterior probabilities of VTLN warping factors as input to DNN

In this approach we propose to train a warping-factor aware DNN. We augment the acoustic features vector with the posterior probabilities of the VTLN warping factors (see also Figure 2.2). Similar approaches have been shown to improve the robustness to noise and speaker independence of the DNN [Seltzer et al., 2013, Senior and Lopez-Moreno, 2014].
We first propose to train a DNN that estimates the VTLN warping factors (Figure 2.1). This DNN will be referred to as DNN-warp. The VTLN procedure is first applied to generate a warping factor for each utterance in the training set. Each acoustic feature vector in the utterance is labeled with the utterance warping factor. Then, training acoustic feature vectors and corresponding warping factors are used to train a DNN-warp classifier. Each class of the DNN correspond to one of the discrete VTLN warping factors and the dimension of the DNN output corresponds to the number of discrete VTLN warping factors. The DNN learns to infer the VTLN warping factor from the acoustic feature vector or more precisely the posterior probability of each VTLN warping factors knowing the input acoustic feature vector.
During training and test of the DNN-HMM system, for each speech frame the warping factors posterior probabilities are estimated with the DNN-warp. These estimated posterior probabilities are appended to the acoustic feature vectors, extended with context, to form an augmented acoustic feature vectors. The augmented features vector is then normalized to zero mean and unit variance and used as input to the DNN-HMM (Figure 2.2).
This approach has the advantage to reduce considerably the complexity during decoding compared to the approach making use of VTLN normalized acoustic features that requires two decoding passes [Lee and Rose, 1996, Welling et al., 1999]. It also allows for a flexible estimation of the warping factors: they could either be updated on a frame to frame basis or averaged at utterance level (see also Section 2.5).

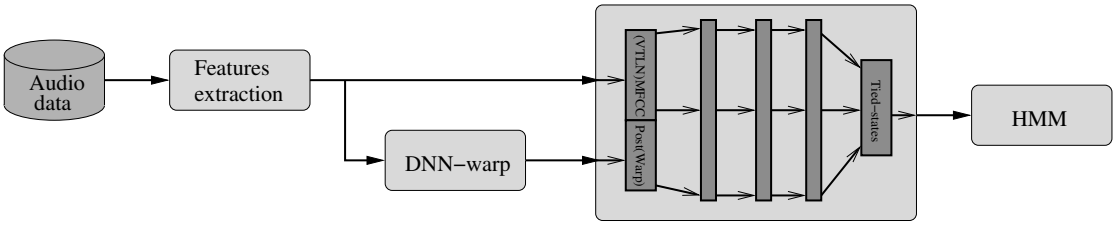


Figure 2.2: Training of the warping factor aware DNN-HMM.

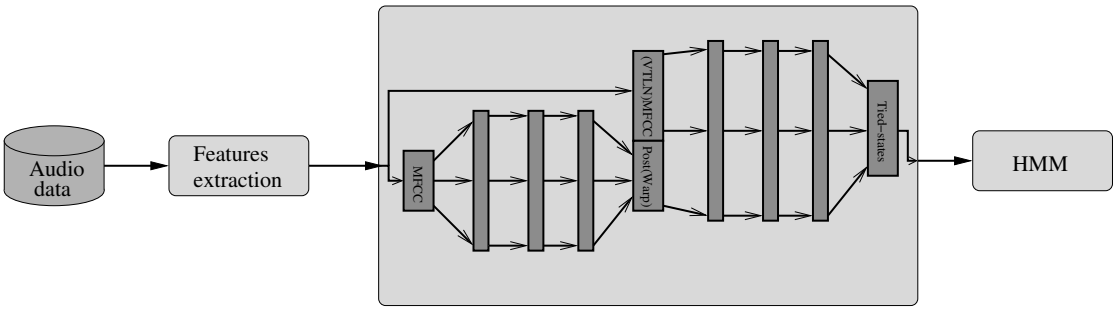


Figure 2.3: Joint optimisation of the DNN-warp and the DNN-HMM.

2.3.3 Joint optimisation

The ultimate goal here is not to estimate the VTLN warping factors but to perform robust speech recognition on heterogeneous corpora. To this end, the DNN-warp and the DNN-HMM can be optimized jointly (Figure 2.3). The procedure is the following: 1) first the DNN-warp is trained alone (Figure 2.1), 2) the posteriors of the warping factors on the training set are obtained with the DNN-warp, 3) these posteriors of the warping factors are used as input to the DNN-HMM together with the acoustic features to produce an augmented feature vector, 4) the DNN-HMM is trained (Figure 2.2), 5) the DNN-warp and the DNN-HMM are concatenated to obtained a deeper network that is fine-tuned with back-propagation on the training set (Figure 2.3). Details about joint optimization are presented in Section 2.4.6

2.4 Experimental setup

2.4.1 Speech corpora

For this study we relied on three Italian speech corpora that are described below: the ChildIt corpus consisting of children speech, the APASCI corpus and the IBN corpus consisting of adults' speech. All corpora were used for evaluation purposes, while the ChildIt and the APASCI provided similar amount of training data for children and adults, respectively. The IBN corpus contains approximately 5 times as much training data as ChildIt or APASCI for adult speech only (Table 2.1).

	Speech Corpus				
Subset	ChildIt	APASCI(f)	APASCI(m)	IBN(f)	IBN(m)
Training	7 h 15 min	2 h 40 min	2 h 40 min	23 h 00 min	25 h 00 min
Test	2 h 20 min	0 h 20 min	0 h 20 min	1 h 00 min	1 h 00 min

Table 2.1: Data distribution in the speech corpora. (f) and (m) denote speech from female and male speakers, respectively.

2.4.1.1 ChildIt

The ChildIt corpus [Gerosa et al., 2007, Giuliani and Gerosa, 2003] is an Italian, task-independent, speech corpus that consists of clean read speech from children aged from 7 to 13 years, with a mean age of 10 years. For each recording in the corpus a word-level transcription is available. The overall duration of audio recordings in the corpus is 10 h 24 min hours. Speech was collected from 171 children. The corpus was partitioned into: a training set consisting of data from 115 speakers for a total duration of 7 h 15 min; a development set consisting of data from 14 speakers, for a total duration of 0 h 49 min; a test set consisting of data from 42 speakers balanced with respect to age and gender for a total duration of of 2 h 20 min.

2.4.1.2 APASCI

The APASCI speech corpus [Angelini et al., 1994] is a task-independent, high quality, acoustic-phonetic Italian corpus. For each recording in the corpus a word-level transcription is available. APASCI was developed at ITC-irst (*Istituto per la ricerca scientifica e tecnologica*, Trento, Italy) and consists of speech data collected from 194 adult speakers for a total duration of 6h49m. The corpus was partitioned into: a training set consisting of data from 134 speakers for a total duration of 5 h 19 min; a development set consisting of data from 30 speakers balanced per gender, for a total duration of 0 h 39 min; a test set consisting of data from 30 speakers balanced per gender, for a total duration of 0 h 40 min.

2.4.1.3 IBN Corpus

The IBN corpus is composed of speech from several radio and television Italian news programs [Gerosa et al., 2009a]. It consists of adult speech only, with word-level transcriptions. The IBN corpus was partitioned into a training set, consisting of 52 h of speech, and a test set formed by 2 h of speech. During the experiments presented here 2h00m of male speech and 2 h of female speech are extracted from the training set to be used as development set during the DNN training. The resulting training set is then partitioned into 25 h of male speech and 23 h of female speech.

2.4.2 Phone recognition systems

The approaches proposed in this chapter have been first tested on small corpora (ChildIt + APASCI) for phone recognition in order to explore as many set-ups as possible in a limited amount of time. The reference phone-level transcriptions are obtained with Viterbi forced-alignment performed with our best GMM-HMM system at the moment of the experiments.

2.4.2.1 GMM-HMM

The acoustic features are 13 MFCC, including the zero order coefficient, computed on 20ms frames with 10ms overlap. First, second and third order time derivatives are computed after cepstral mean subtraction performed utterance by utterance. These features are arranged into a 52-dimensional vector that is projected onto a 39-dimensional feature space by applying a linear transformation estimated through Heteroscedastic Linear Discriminant Analysis (HLDA) [Kumar and Andreou, 1998].
Acoustic models are 3039 tied-state triphone HMM based on a set of 48 phonetic units derived from the SAMPA Italian alphabet. Each tied-state is modeled with a mixture of 8 Gaussian densities having a diagonal covariance matrix. In addition, "silence" is modeled with a Gaussian mixture model having 32 Gaussian densities.

2.4.2.2 DNN-HMM

The DNN uses the same 13 MFCC, including the zero order coefficient, computed on 20 ms frames with 10 ms overlap on which Hamming windowing is applied. The context spans on a 31 frames window. This 403 dimensional feature vector is then projected onto a 208 dimensional feature vector by applying principal component analysis (PCA) and normalized to zero mean and unit variance before being used as input to the DNN. The targets of the DNN are the 3039 tied-states obtained from the GMM-HMM training on the mixture of adults' and children's speech (ChildIt + APASCI). The DNN has 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can be summarized as follows: $208 \times 1500 \times 1500 \times 1500 \times 1500 \times 3039$.
The DNN are trained with the TNet software package [Veselỳ et al., 2010]. The DNN weights are initialized randomly and pre-trained with RBM. The first layer is pre-trained with a Gaussian-Bernouilli RBM trained during 10 iterations with a learning rate of 0.005. The following layers are pre-trained with a Bernouilli-Bernouilli RBM trained during 5 iterations with a learning rate of 0.05. Mini-batch size is 250.
For the back propagation training the learning rate is kept to 0.02 as long as the frame accuracy on the cross-validation set progresses by at least 0.5% between successive epochs. The learning rate is then halved at each epoch until the frame accuracy on the cross-validation set fails to improve by at least 0.1%. The mini-batch size is 512. In both pre-training and training, a first-order momentum of 0.5 is applied.
The DNN can be trained either on all speech data available (ChildIt + APASCI) or on group specific corpora (ChildIt, adult female speech in APASCI, adult male speech in APASCI).

2.4.2.3 Language model

A simple finite state network having just one state and a looped transition for each phone unit was employed. In this network uniform transition probabilities are associated to looped transitions. In computing recognition performance, in terms of PER, no distinction was made between single consonants and their geminate counterparts. In this way, the set of phonetic labels was reduced to 28 phone labels. For all corpora, reference phone transcriptions were derived from prompted texts by performing Viterbi decoding on a pronunciation network for each utterance. The pronunciation network of an utterance was built by concatenation of the phonetic transcriptions of the words in the prompted text. In doing this alternative word pronunciations were taken into account and an optional insertion of the silence model between words was allowed.

2.4.3 Word recognition systems

The approaches that performed best in phone recognition on the small corpora are validated in word recognition on a more realistic set-up (ChildIt + IBN) including a corpus of adult speech (IBN) that is larger than the corpus of children speech (ChildIt). In this setup, there is a bias toward adult speech that can correspond to the problem faced in real applications.

2.4.3.1 GMM-HMM

The GMM-HMM are similar to those used for the phone recognition except that they use more Gaussian densities to benefit from the extensive training data. Acoustic models are 5021 tied-state triphone HMM based on a set of 48 phonetic units derived from the SAMPA Italian alphabet. Each tied-state is modeled with a mixture of 32 Gaussian densities having a diagonal covariance matrix. In addition, "silence" is modeled with a Gaussian mixture model having 32 Gaussian densities.

2.4.3.2 DNN-HMM

The DNN are similar to those used for phone recognition except that they are trained on a different set of targets. The targets of the DNN are the 5021 tied-states obtained from the word recognition GMM-HMM training on the mixture of adults' and children's speech (ChildIt + IBN). The DNN has 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can be summarized as follows: $208 \times 1500 \times 1500 \times 1500 \times 1500 \times 5021$.

2.4.3.3 Language model

For word recognition, a 5-gram language model was trained on texts from the Italian news domain consisting in about 1.6 G words. Part of the textual data, consisting in about 1.0 G words, were acquired via web crawling of news domains. The recognition dictionary consists of the most frequent 250 K words.

2.4.4 Age/gender adapted DNN for DNN-HMM

One option is to adapt an already trained general DNN to group specific corpora. The data architectures are the same as described above. The initial DNN weights are the weights obtained with a pre-training/training procedure applied on all the training data available (ChildIt+APASCI for phone recognition and ChildIt + IBN for word recognition). The DNN is then trained with back propagation on a group specific corpora (ChildIt, adult female speech in APASCI and adult male speech in APASCI for phone recognition and ChilIt, adult female speech in IBN and adult male speech in IBN for word recognition). The training parameters are the same as during the general training (2.4.2.2 and 2.4.3.2, respectively) and the learning rate follows the same rule as above. The mini-batch size is 512 and a first-order momentum of 0.5 is applied.

2.4.5 VTLN

In this work we are considering a set of 25 warping factors evenly distributed, with step 0.02, in the range 0.76-1.24. During both training and test a grid search over the 25 warping factors was performed. The acoustic models for scaling factor selection, carried out on an utterance-by-utterance basis, were speaker-independent triphone HMM with 1 Gaussian per state and trained on un-warped children's and adults' speech [Gerosa et al., 2007, Welling et al., 1999].
The DNN-warp inputs are the MFCC with a 61 frame context window, DCT projected to a 208 dimensional features vector. The targets are the 25 warping factors. The DNN has 4 hidden layers, each of which contains 500 elements such that the DNN architecture can be summarized as follows: $208 \times 500 \times 500 \times 500 \times 500 \times 25$. The training procedure is the same as for the DNN acoustic model in the DNN-HMM.
The posterior probabilities obtained with the DNN-warp are concatenated with the 208-dimensional DCT projected acoustic features vector to produce a 233-dimensional features vector that is normalized to zero mean and unit variance before being used as input to the DNN. The new DNN acoustic model has 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can then be summarized as follows: $233 \times 1500 \times 1500 \times 1500 \times 1500 \times 3039$ for phone recognition ($233 \times 1500 \times 1500 \times 1500 \times 1500 \times 5021$ for word recognition).

2.4.6 Joint optimisation

The DNN-warp and DNN-HMM can be fine-tuned jointly with back-propagation. In such case, the starting learning rate is set to 0.0002 in the first 4 hidden layers (corresponding to the DNN-warp) and to 0.0001 in the last 4 hidden layers (corresponding to the DNN-HMM). The learning rate is chosen empirically as the highest value for which both training accuracy and cross-validation accuracy improve. Setting a different learning rate in the first 4 hidden layers and the last 4 hidden layers is done in a attempt to overcome the vanishing gradient effect in the 8 layers DNN obtained from the concatenation of the DNN-warp and the DNN-HMM. The learning rates are then adapted following the same

schedule as described above. The joint optimization is done with a modified version of the TNet software package [Veselỳ et al., 2010].

2.5 Experimental Results

Two sets of experiments are presented here. First the systems are tested extensively in terms of phone error rate (PER) on small corpora (ChildIt + APASCI), then the best performing systems are tested in terms of word error rate (WER) performance on a more realistic set-up including a larger adult speech corpus (ChildIt + IBN).

2.5.1 Phone recognition

The experiments presented here are designed to verify the validity of the following statements:

- The age/gender group specific training of the DNN does not necessarily lead to improved performance, specifically when only a small amount of data is available
- The age/gender group adaptation of a general DNN can help to design group specific systems, even when only a small amount of data is available
- VTLN can be beneficial to the DNN-HMM framework when targeting a heterogeneous speaker population with limited amount of training data
- Developing an "all-DNN" approach to VTLN for DNN-HMM framework, when targeting a heterogeneous speaker population, offers a credible alternative to the use of VTLN normalized acoustic features or to the use of age/gender group specific DNN
- Optimizing the DNN-warp and the DNN-HMM jointly can help to improve the performance in certain cases
- The different approaches introduced in this chapter can be complementary.

During the experiments the language model weight is tuned on the development set and used to decode the test set. Results were obtained with a phone loop language model and the PER was computed based on 28 phone labels. Variations in recognition performance were validated using the matched-pair sentence test [Gillick and Cox, 1989] to ascertain whether the observed results were inconsistent with the null hypothesis that the output of two systems were statistically identical. Considered significance levels were .05, .01 and .001.

2.5.1.1 Age/gender specific training for DNN-HMM

In this experiment, DNN are trained on group specific corpora (children's speech in ChildIt, adult female speech in APASCI and adult male speech in APASCI) and performance are compared with the DNN-HMM baseline introduced above where the DNN is trained on speech from all speaker groups. Recognition results are reported in Table 2.2, which includes results achieved with the DNN-HMM baseline in the row *Baseline*. In ChildIt there is about 7h of training data which is apparently sufficient to train an effective DNN and we can observe an improvement of 22% PER relative compared to the

	Evaluation Set		
Training Set	ChildIt	APASCI(f)	APASCI(m)
Baseline	15.56%	**10.91%**	**8.62%**
ChildIt	**12.76%**	29.59%	46.16%
APASCI(f)	34.23%	12.75%	31.21%
APASCI(m)	56.11%	30.81%	9.83%

Table 2.2: Phone error rate achieved with the DNN-HMM trained on age/gender groups specific data.

	Evaluation Set			
Adaptation Set	ChildIt	APASCI(f)	APASCI(m)	ChildIt + APASCI
Baseline	15.56%	10.91%	8.62%	14.32%
ChildIt	**12.43%**	16.93%	24.96%	–
APASCI(f)	21.91%	**9.65%**	17.01%	–
APASCI(m)	32.33%	16.99%	**7.61%**	–
Model selection	12.43%	9.65%	7.61%	11.59%

Table 2.3: Phone error rate achieved with the DNN-HMM trained on a mixture of adult and children's speech and adapted to specific age/gender groups.

baseline performance ($p < .001$). However, in adult data there is only about 2h40m of data for each gender. This is apparently not sufficient to train the DNN. In fact, the DNN-HMM system based on a DNN that is trained on gender specific data consistently degrades the PER. The degradation compared to the baseline performance is 14% PER relative on female speakers in APASCI ($p < .001$) and 12% PER relative on male speakers in APASCI ($p < .001$).

2.5.1.2 Age/gender adapted DNN-HMM

In this experiment the baseline model (trained on all corpora) is adapted to each group specific corpus. PER performance is presented in Table 2.3. The group adapted DNN-HMM consistently improve the PER compared to the DNN-HMM baseline. On children's speech the PER improvement compared to the baseline is 25% PER relative ($p < .001$). On adult female speakers in APASCI the age/gender adaptation improves the baseline performance by is 13% PER relative ($p < .001$). On adult male speakers the age/gender adaptation improves the baseline performance by 13% ($p < .05$).
From results in Table 2.3 it is also possible to note that the DNN-HMM system adapted to children's voices perform much better for adult female speakers than for adult male speakers. Symmetrically, the DNN-HMM system adapted to female voices perform better on children' speech than the system adapted to male voices. These results confirm that characteristics of children's voice is much more similar to those of adult female voices than those of adult male voices.
In the *Model selection* approach, we assumed that a perfect age/gender classifier exist

Model	**Evaluation Set**			
	ChildIt	APASCI(f)	APASCI(m)	ChildIt + APASCI
Baseline	15.56%	10.91%	8.62%	14.32%
VTLN-normalisation	12.80%	10.41%	**7.91%**	12.00%
Warp + MFCC	14.51%	10.48%	9.63%	13.46%
Warp-post + MFCC	14.10%	10.89%	8.34%	13.12%
Warp-post(utt)+ MFCC	13.43%	**9.66%**	8.06%	12.45%
Warp-post + MFCC(joint)	**12.52%**	11.23%	8.98%	**11.98%**

Table 2.4: PER achieved with VTLN approaches to DNN-HMM.

which allows us to know in which target group of speaker an incoming speech segment belongs. The recognition is then performed using the corresponding adapted model. On the evaluation set including all the target groups of speakers (ChildIt + APASCI) *Model selection* improves the baseline by 23% PER relative ($p < .05$).

2.5.1.3 VTLN based approaches

Table 2.4 presents the PER obtained with the DNN-HMM baseline, and the VTLN approaches:

- the VTLN applied to MFCC during training and test (row *VTLN-normalization*)
- the MFCC features vector augmented with the the warping factors obtained in a standard way (row *Warp + MFCC*)
- the MFCC features augmented with the posterior probabilities of the warping factors (row *Warp-post + MFCC*)
- the MFCC features augmented with the posterior probabilities of the warping factors averaged at utterance level (row *Warp-post (utt) + MFCC*)
- and the joint optimization of the DNN-warp and the DNN-HMM (row *Warp-post + MFCC (joint)*)

VTLN normalization allows for consistently obtaining a PER among the best for each group of speaker and is therefore the most robust approach presented up to here. The *Warp-post + MFCC (joint)* overall improvement is mainly due to the large improvement on the children evaluation set, 24% relative ($p < .001$) whereas it mildly degrades the performance on other groups of speakers. This is probably due to the fact that the training set is unbalanced towards children (7 h 15 min in ChildIt against 2 h 40 min for each adult group), therefore, performing the joint optimization biases the system in favor of children speech.

Using directly the warping factors obtained in a standard way consistently performs among the worst system and is outperformed by the systems using the MFCC augmented with the posterior probabilities of the warping factors. This seems to indicate that the ASR can benefit from the flexibility introduced by the posterior probabilities of the warping factors, in contrast with the hard decision in the standard warping factors

Model	Evaluation Set			
	ChildIt	APASCI(f)	APASCI(m)	ChildIt + APASCI
Baseline	15.56%	10.91%	8.62%	14.32%
Model selection	12.43%	9.65%	7.61%	11.59%
Warp-post + MFCC	14.10%	10.89%	8.34%	13.12%
Warp-post + MFCC (model selection)	11.71%	9.23%	7.28%	10.98%
VTLN-normalisation	12.80%	10.41%	7.91%	12.00%
VTLN (model selection)	**11.31%**	9.14%	**7.19%**	**10.61%**
Warp-post + VTLN (model selection)	11.34%	**9.04%**	7.32%	10.68%

Table 2.5: PER achieved with combination of approaches.

estimation. To perform at their best however, these estimation have to be aggregated by averaging at utterance level or adapted using joint-optimization. Note that both of these constraints were not compatibles with the framework used for these experiments.

2.5.1.4 Combination of approaches

To exploit the potential complementarity of the different approaches introduced until here we combine the different systems at features level.
Table 2.5 presents the PER obtained with the DNN-HMM baseline, VTLN approaches (rows *VTLN-normalization* and *Warp-post + MFCC*) and the combination of different approaches:

- the age/gender adaptation approach in combination with model selection (row *Model selection*)
- age/gender adaptation performed on a system trained with VTLN-normalized features (row *VTLN (model selection*)
- MFCC features vector augmented with the posterior probabilities of the warping factors (row *Warp-post + MFCC (model selection)*)
- VTLN-normalized features vector augmented with the posterior probabilities of the warping factors (row *Warp-post + VTLN (model selection)*)

Joint optimization is not applied at this stage as the unbalanced training corpus results in biased training and the corpora used here are too small to truncate them to produce a balanced heterogeneous corpus.
On the evaluation set including all the target groups of speakers (ChildIt + APASCI) the combination of approaches outperform all the individual approaches presented until here. When compared to the best system until now (*Model selection*), system combination allows for consistently improving the PER on every groups of speakers. The combination *Warp-post + MFCC (model selection)* represents the best single-pass system presented here.

	Evaluation Set			
Adaptation set	ChildIt	IBN(f)	IBN(m)	ChildIt+IBN
Baseline	12.83%	10.61%	11.02%	11.98%
Model selection	**10.89%**	**10.33%**	**10.99%**	**10.93%**
ChildIt + general model	**10.89%**	10.61%	11.02%	11.00%

Table 2.6: WER achieved with the DNN-HMM trained on a mixture of adult and children's speech and adapted to specific age/gender groups.

2.5.2 Word recognition

The experiments presented here are designed to verify that results obtained for phone recognition can be replicated in terms WER and on a more "realistic" set-up where the adult speech training corpus (IBN corpus) is larger than the children speech training corpus (ChildIt). During the experiments the language model weight is tuned on the development set and used to decode the test set. Variations in recognition performance were again validated using the matched-pair sentence test [Gillick and Cox, 1989].

2.5.2.1 Age/gender adapted DNN-HMM

Table 2.6 presents the WER obtained with a DNN-HMM baseline trained on the corpus composed of ChildIt and IBN (row *Baseline*). These performance are compared with the performance obtained with age/gender adaptation (row *Model selection*) and with the performance obtained with system performing model selection between age adapted system for children speaker and the general baseline for adult speaker (row *ChildIt + general model*).
On the evaluation set including all the target groups of speakers (ChildIt + INBC) the age-gender adaptation improves the performance of the baseline by 10% WER relative ($p < .001$). When targeting children speakers, the age-gender adaptation improves the performance of the baseline by 18% relative ($p < .001$). On the other hand, when targeting adult speakers, the age-gender adaptation does not significantly improve the WER compared to the baseline. This is due to the fact that the adult corpus is now considerably larger than for experiment on PER (52h00m for IBN against 5h19m for APASCI). This allows for achieving an effective training on the adult groups with the general corpus and the benefits from the age-gender adaptation are limited. Therefore for simplicity's sake, in the remainder of the chapter, the approach (row *ChildIt + general model*) is considered instead of age-gender adaptation for all groups of speakers (*Model selection*). Performance difference between *Model selection* and *ChildIt + general model* is not statistically significant.

2.5.2.2 VTLN based approaches and system combination

Table 2.7 presents the WER performance for:

- VTLN based approaches:

	Evaluation Set			
Model	ChildIt	IBN(f)	IBN(m)	ChildIt+IBN
Baseline	12.83%	10.61%	11.02%	11.98%
ChildIt + general model	10.89%	10.61%	11.02%	11.00%
Warp-post + MFCC	12.11%	10.52%	11.07%	11.57%
Warp-post + MFCC (joint)	11.81%	**10.49%**	**11.01%**	11.33%
Warp-post + MFCC (joint / ChildIt + general model)	11.06%	**10.49%**	**11.01%**	10.97%
VTLN-normalisation	12.21%	10.58%	11.25%	11.58%
Warp-post - VTLN (joint)	**10.83%**	**10.49%**	11.07%	**10.86%**
Warp-post - VTLN (joint / ChildIt + general model)	11.07%	**10.49%**	11.07%	10.96%

Table 2.7: WER achieved with several VTLN approaches to DNN-HMM.

  - VTLN applied to MFCC during training and testing (row *VTLN-normalization*)
  - MFCC features augmented with the posterior probabilities of the warping factors (row *Warp-post + MFCC*)
  - joint optimization of the DNN-warp and the DNN-HMM (row *Warp-post + MFCC (joint)*)
- system combination:
  - VTLN-normalized features vector augmented with the posterior probabilities of the warping factors and joint optimization (row *Warp-post + VTLN (joint)*)
  - age/gender adaptation for children speaker performed on a system working with the MFCC features vector augmented with the posterior probabilities of the warping factors with joint optimization (row *Warp-post + MFCC (joint/ChildIt + general model)*)
  - VTLN-normalized features vector augmented with the posterior probabilities of the warping factors with joint optimization (row *Warp-post + VTLN (joint/ChildIt + general model)*)

These approaches are compared to the baseline and to *ChildIt + general model*.
The approaches combining VTLN-normalized features and posterior probabilities aim at testing the complementary between VTLN-normalization that operates at utterances level and posterior probabilities that are obtained at frame level. While estimating VTLN factors on a longer time unit(utterance) should allow for a more accurate average estimation, the "true" warping factor might be fluctuating in time. Combining VTLN normalization at utterance level and posterior probabilities estimated at frame level should help overcoming this problem.
On the evaluation set including all the target groups of speakers (ChildIt + INBC) the VTLN based approaches (*Warp-post + MFCC* and *VTLN-normalization*) perform similarly. They improve the performance baseline ($p < .001$) but both the methods are outperformed by *ChildIt + general model* ($p < .001$). Performance difference between the VTLN based approaches, the baseline and *ChildIt + general model* on adult corpora

are in general not statistically significant.
During these experiment, the corpus was unbalanced towards adults (52 h for IBN against 7 h 15 min for ChildIt). Joint optimization is performed on a balanced training set in order to avoid introducing a bias in favor of the adults corpora. The balanced corpus is composed of 7 h of adult female and 7 h of adult male speech randomly selected from the IBN corpus. On the evaluation set composed of all target groups, joint optimization improves the *Warp-post + MFCC* performance ($p < .001$). The performance improvement in each speakers group is not statistically significant.
The combination of several approach improves the performance compared to both the baseline and the VTLN based approaches. Among the approaches proposed in the chapter, *ChildIt + general model* and *Warp-post + VTLN (joint)* perform equally well. However, their potential applications are different. Indeed, *ChildIt + general model* is the most simple approach but lacks flexibility and is difficult to generalize to new groups of speakers as a new DNN would have to be adapted to each new group of speakers. The VTLN based approach *Warp-post + VTLN (joint)* on the other end, does not rely on model adaptation/selection and is more general than *ChildIt + general model*. The drawback of this approach, however, is that it requires a two-pass decoding whereas *ChildIt + general model* operate in single-pass.

2.6 Conclusions

In this chapter we have investigated the use of the DNN-HMM approach to speech recognition targeting three groups of speakers, that is children, adult males and adult females. Two different kinds of approaches have been introduced here to cope with inter-speaker variability: approaches based on DNN adaptation and approaches relying on VTLN. The combination of the different approaches to take advantage of their complementarity has then been investigated.
The different approaches presented here have been tested extensively in terms of PER on small corpora first. Approaches based on VTLN have been shown to provide a significant improvement compared to the baseline (up to 19% relative) but were still outperformed by the DNN adaption approach (23% relative improvement compared to the baseline). System combination on the other hand effectively takes advantage of the complementarity of the different approaches introduced in this chapter and improves the baseline performance by up to 35% relative PER. Besides, system combination is shown to consistently outperform each approach used separately.
The best performing approaches have been validated in terms of WER on a more "realistic" set-up where the adult speech corpus (IBN) used for training is larger than the training children's speech corpus (ChildIt). DNN adaptation is then proved effective for the under-resourced target group (children) but not significantly on target group with sufficient training data (adults). The trend observed on PER is confirmed and approaches based on VTLN have been shown to provide a significant improvement compared to the baseline (5% to 6% relative) but were still outperformed by the DNN adaption approach (10% relative improvement compared to the baseline). System combination improves the

baseline performance by up to 11% WER relative. The two best performing approaches introduced here (*ChildIt + general model* and *Warp-post + VTLN (joint)*) can have different applications. *ChildIt + general model* is bas on age dependent model selection and is the most simple approach but lacks flexibility as it requires to train specific models beforehand. VTLN based approach *Warp-post + VTLN (joint)* is more general as it only relies on VTLN factors estimation but it requires a two-pass decoding.
The work done in this chapter highlights the importance of acoustic models dedicated to groups of speaker to get accurate speech recognition. When acoustic conditions are degraded the importance of specific models can become even larger. This is not the case in this chapter but it can occur frequently in hands-free communication or in human-machine interactions, which represent a large portion of the computer-based speech communication nowadays. When considering personal devices such as smartphones or smart-speaker, it could even be relevant to consider acoustic models that are specific to one person in which case, it might be useful to identify automatically who is talking to the device.

3 Speaker recognition in noisy environments

Context: This work was done when I was a postdoctoral researcher at Télécom ParisTech (Paris, France) between October 2014 and August 2016 together with Victor Bisot, Slim Essid and Gaël Richard.[1] The work presented here has been previously published in articles [Serizel et al., 2016a, 2017, 2016b].

3.1 Introduction

The main target of speaker identification is to assert whether or not the speaker of a test segment is known and if he/she is known, to find his/her identity. Applications of speaker identification are numerous, among which are speaker dependent automatic speech recognition and subject identification based on biometric information. The sentence pronounced by the subject can be unknown and the recordings can be of variable quality. The speaker identification then becomes a highly challenging problem.
Between 2011 and 2018, the i-vectors [Dehak et al., 2011] were the state-of-the-art approach for speaker identification [Greenberg et al., 2014]. A typical speaker identification system was composed of i-vector extraction, normalization [Bousquet et al., 2011, Garcia-Romero and Espy-Wilson, 2011] and classification with probabilistic linear discriminant analysis (PLDA) [Prince and Elder, 2007]. Research on the tandem i-vector/PLDA was focusing a lot of attention during this period and speaker identification systems had reached a high level of performance on databases such as those from the National Institute of Standards and Technology (NIST) [Greenberg et al., 2014, 2013].
On the other hand, studies have shown that approaches such as non-negative matrix factorization (NMF) [Lee and Seung, 1999] can be successfully applied to spectrogram factorization [Hurmalainen et al., 2015a,b, Saeidi et al., 2012] or to multimodal cofactorisation [Seichepine et al., 2014] to retrieve speaker identity. These results tend to indicate that the activations of NMF dictionary atoms can represent well the speaker identity [Saeidi et al., 2012]. Besides, exploiting group sparsity on the activations has then proven to improve further the performance of NMF-based approaches [Hurmalainen et al., 2015b]. NMF therefore offers a credible alternative to i-vectors that takes advantage of the intrinsic sparsity of speech [Hurmalainen et al., 2012, 2015b]. However, to the best of our knowledge, none of the NMF-based approaches proposed until 2014 took the recording sessions variability into account. Yet this is a crucial point in the success of i-vectors.

[1] Collaborators listed alphabetically, members of the host institution unless mentioned otherwise

This chapter proposes an approach to speaker identification that relies on group-NMF and that is inspired by the i-vector training procedure. Given data measured with several subjects, the key idea in group-NMF is to track inter-subject and intra-subject variations by constraining a set of common bases across subjects in the decomposition dictionaries. This has originally been applied to the analysis of electroencephalograms [Lee and Choi, 2009]. The approach presented here extends this idea and proposes to capture inter-speaker and inter-session variabilities by constraining a set of speaker-dependent bases across sessions and a set of session-dependent bases across speakers. This approach is inspired by the joint factor analysis [Kenny et al., 2007] and i-vectors as it takes both speaker variability and session variability into account. In this sense, it differs from previous approaches based on NMF [Hurmalainen et al., 2012, 2015a,b] that take only speaker variability into account. Besides, in these previous works similarity constraints were imposed on activations while in the approach proposed here the constraints are on the dictionaries.
The rest of the chapter is organized as follows the signal model, notations and the genreal NMF are described in Section 3.2. The group NMF approach is introduced in Section 3.3 and the task-driven NMF aproaches are presented in Section 3.4. The experimental setup is described in Section 3.5 and experiments results are presented in Section 3.6. A summary of the chapter and conclusions are provided in Section 3.7 and the other related works are presented in Section 3.8.

3.2 Problem statement

3.2.1 Notations

Consider the (nonnegative) time-frequency representation of a set of audio signals $\mathbf{X} \in \mathbb{R}_+^{F \times N}$ (this could be for example a mel-frequency spectrogram), where F is the number of frequency components and N the number of frames. $\mathbf{X}$ is composed of data collected during S recording sessions with speech segments originating from G speakers. In each session several speakers can be present and a particular speaker can be present in several sessions. Let $\mathcal{G}$ denote the set of speakers and $\mathcal{S}$ the set of sessions. The number of elements in an ensemble is denoted Card(.), such that $\text{Card}(\mathcal{G}) = G$ and $\text{Card}(\mathcal{S}) = S$. Let $\mathcal{G}^s$ denote the subset of speakers that appear in the session s ($\mathcal{G}^s \subset \mathcal{G}$) and $\mathcal{S}^g$ the subset of sessions in which the speaker g is active ($\mathcal{S}^g \subset \mathcal{S}$). In the remainder of this chapter, superscripts ^{g} and ^{s} will denote the current speaker and session, respectively.

3.2.2 NMF with Kullback-Leibler divergence

The goal of NMF [Lee and Seung, 1999] is to find a factorisation for $\mathbf{X}$ of the form:

$$\mathbf{X} \approx \mathbf{WH} \tag{3.1}$$

where $\mathbf{W} \in \mathbb{R}_+^{F \times K}$, $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ and K is the number of components in the decomposition. Given a separable divergence D, NMF model estimation can be formulated as the

following optimisation problem:

$$\min_{\mathbf{WH}} D(\mathbf{X}|\mathbf{WH}) \quad \text{s.t.} \quad \mathbf{W} \geq 0,\ \mathbf{H} \geq 0.$$

When considering audio signals, D is often chosen to be the Kullback-Leibler divergence (denoted D_{KL} here) [Kullback and Leibler, 1951] or the Itakura-Saito divergence [Itakura, 1975]. In most cases the NMF problem is solved using a two-block coordinate descent approach. Each of the factors $\mathbf{W}$ and $\mathbf{H}$ is optimised alternatively. The sub-problem in one factor can then be considered as a nonnegative least square problem (NNLS) [Gillis, 2014]. One of the approaches to solve these NNLS problems leads to the multiplicative update rules for the matrices $\mathbf{W}$ and $\mathbf{H}$, which can be expressed as follows for the D_{KL} [Févotte and Idier, 2011, Lee and Seung, 2000]:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \left[(\mathbf{WH})^{-1} \odot \mathbf{X}\right]}{\mathbf{W}^T \mathbf{1}} \tag{3.2}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left[(\mathbf{WH})^{-1} \odot \mathbf{X}\right] \mathbf{H}^T}{\mathbf{1}\mathbf{H}^T}; \tag{3.3}$$

where $\odot$ is the element-wise product (Hadamard product) and division and power are element-wise. $\mathbf{1}$ is a matrix of dimension $F \times N$ with all its coefficient equal to 1.

3.2.3 NMF for feature learning in speaker identification

In this chapter, NMF is used for feature learning in a speaker identification framework. First, the factorization is learned on a training set and the activations are used as input features to train a general classifier. The dictionaries $\mathbf{W}$ obtained on the training set are then used to extract features (activations) on the test set. These features are used as input to the general classifier to perform speaker identification.

3.3 Group NMF with speaker and session similarity

In the approach presented above, the feature learning step is totally unsupervised and does not account for speaker variability or session variability. The approach introduced here intends to take these variabilities into account. It derives from Group-NMF [Lee and Choi, 2009] and is inspired by exemplar-based approaches [Hurmalainen et al., 2015a,b]. The idea of a decomposition across speaker was originally used by Saeidi et al. [2012] but session variability was not considered.

3.3.1 NMF on speaker utterances for speaker identification

In order to better model speaker identity, we now consider the portion of $\mathbf{X}$ recorded in a session s in which only the speaker g is active. This is denoted by $\mathbf{X}^{(gs)}$, its length is $N^{(gs)}$ and it can be decomposed according to (3.1):

$$\mathbf{X}^{(gs)} \approx \mathbf{W}^{(gs)}\mathbf{H}^{(gs)} \quad \forall\ (g,s) \in \mathcal{G} \times \mathcal{S}^g$$

under nonnegative constraints.
We define a global cost function which is the sum of all local divergences:

$$J_{\text{global}} = \sum_{g=1}^{G} \sum_{s \in \mathcal{S}^g} D_{KL}(\mathbf{X}^{(gs)} | \mathbf{W}^{(gs)} \mathbf{H}^{(gs)}). \tag{3.4}$$

Each $\mathbf{X}^{(gs)}$ can be decomposed independently with standard multiplicative rules (3.2, 3.3). The bases learned on the training set are then concatenated to form a global basis. The latter basis is then used to produce features on test sets.

3.3.2 Class and session similarity constraints

In order to take the session and speaker variabilities into account we propose to further decompose the dictionaries $\mathbf{W}$ similarly as what was proposed by Lee and Choi [2009]. The matrix $\mathbf{W}^{(gs)}$ can indeed be arbitrarily decomposed as follows:

$$\mathbf{W}^{(gs)} = [\underbrace{\mathbf{W}^{(gs)}_{\text{SPK}}}_{\leftarrow K_{\text{SPK}} \rightarrow} | \underbrace{\mathbf{W}^{(gs)}_{\text{SES}}}_{\leftarrow K_{\text{SES}} \rightarrow} | \underbrace{\mathbf{W}^{(gs)}_{\text{RES}}}_{\leftarrow K_{\text{RES}} \rightarrow}]$$

with $K_{\text{SPK}} + K_{\text{SES}} + K_{\text{RES}} = K$ and where K_{SPK}, K_{SES} and K_{RES} are the number of components in the speaker-dependent bases, the session-dependent bases and the residual bases, respectively.

$$\mathbf{W}^{(gs)}_{\text{SPK}} \leftarrow \mathbf{W}^{(gs)}_{\text{SPK}} \odot \frac{\left[(\mathbf{W}^{(gs)}\mathbf{H}^{(gs)})^{-1} \odot \mathbf{X}^{(gs)}\right] {\mathbf{H}^{(gs)}_{\text{SPK}}}^T + \frac{\lambda_1}{2} \sum\limits_{\substack{s_1 \in \mathcal{S}_g \\ s_1 \neq s}} \mathbf{W}^{(gs_1)}_{\text{SPK}}}{\mathbf{1}{\mathbf{H}^{(gs)}_{\text{SPK}}}^T + \frac{\lambda_1}{2} (\text{Card}(\mathcal{S}_g) - 1) \mathbf{W}^{(gs)}_{\text{SPK}}} \tag{3.5}$$

$$\mathbf{W}^{(gs)}_{\text{SES}} \leftarrow \mathbf{W}^{(gs)}_{\text{SES}} \odot \frac{\left[(\mathbf{W}^{(gs)}\mathbf{H}^{(gs)})^{-1} \odot \mathbf{X}^{(gs)}\right] {\mathbf{H}^{(gs)}_{\text{SES}}}^T + \frac{\lambda_2}{2} \sum\limits_{\substack{g_1 \in \mathcal{G}_s \\ g_1 \neq g}} \mathbf{W}^{(g_1 s)}_{\text{SES}}}{\mathbf{1}{\mathbf{H}^{(gs)}_{\text{SES}}}^T + \frac{\lambda_2}{2} (\text{Card}(\mathcal{G}_s) - 1) \mathbf{W}^{(gs)}_{\text{SES}}} \tag{3.6}$$

The first target is to capture speaker variability. This is related to finding vectors for the speaker bases ($\mathbf{W}^{(gs)}_{\text{SPK}}$) for each speaker g that are as close as possible across all the sessions in which the speaker is present, leading to the constraint:

$$J_{\text{SPK}} = \frac{1}{2} \sum_{g=1}^{G} \sum_{s \in \mathcal{S}_g} \sum_{\substack{s_1 \in \mathcal{S}_g \\ s_1 \neq s}} \|\mathbf{W}^{(gs)}_{\text{SPK}} - \mathbf{W}^{(gs_1)}_{\text{SPK}}\|^2 < \alpha_1 \tag{3.7}$$

with $\|.\|^2$ the Euclidean distance and α_1 is the similarity constraint on speaker-dependent bases.

The second target is to capture session variability. This can be accounted for by finding vectors for the sessions bases ($\mathbf{W}_{\text{SES}}^{(gs)}$) for each session s that are as close as possible across all the speakers that speak in the session, leading to the constraint:

$$J_{\text{SES}} = \frac{1}{2}\sum_{s=1}^{S}\sum_{g\in\mathcal{G}_s}\sum_{\substack{g_1\in\mathcal{G}_s \\ g_1\neq g}} \|\mathbf{W}_{\text{SES}}^{(gs)} - \mathbf{W}_{\text{SES}}^{(g_1 s)}\|^2 < \alpha_2 \tag{3.8}$$

where α_2 is the similarity constraint on session-dependent bases.
The vectors composing the residual bases $\mathbf{W}_{\text{RES}}^{(gs)}$ are left unconstrained to represent characteristics that depend neither on the speaker nor on the session.
Minimizing the global divergence (3.4) subject to constraints (3.7) and (3.8) is equivalent to the following problem:

$$\min_{\mathbf{W},\mathbf{H}} J_{\text{global}} + \lambda_1 J_{\text{SPK}} + \lambda_2 J_{\text{SES}} \quad \text{s.t.} \quad \mathbf{W} \geq 0,\ \mathbf{H} \geq 0 \tag{3.9}$$

which in turn leads to the multiplicative update rules for the dictionaries $\mathbf{W}_{\text{SPK}}^{(gs)}$ and $\mathbf{W}_{\text{SES}}^{(gs)}$ that are given in equations (3.5) and (3.6), respectively. We obtained these update rules using the well known heuristic which consists in expressing the gradient of the cost function (3.9) as the difference between a positive contribution and a negative contribution [Lee and Seung, 2000]. The multiplicative update then has the form of a quotient of the negative contribution by the positive contribution. The update rules for $\mathbf{W}_{\text{RES}}^{(gs)}$ are similar to the standard rules:

$$\mathbf{W}_{\text{RES}}^{(gs)} \leftarrow \mathbf{W}_{\text{RES}}^{(gs)} \odot \frac{\left[(\mathbf{W}^{(gs)}\mathbf{H}^{(gs)})^{-1} \odot \mathbf{X}^{(gs)}\right]\mathbf{H}_{\text{RES}}^{(gs)\,T}}{\mathbf{1}\mathbf{H}_{\text{RES}}^{(gs)\,T}}.$$

Note that the update rules for the activations ($\mathbf{H}^{(gs)}$) are left unchanged.

3.4 Task-driven NMF based dictionary learning

Task driven dictionary learning [Mairal et al., 2012] can be applied with nonnegativity constraints to perform speech enhancement [Sprechmann et al., 2014] or to acoustic scene classification, where temporally integrated projections are classified with multinomial logistic regression [Bisot et al., 2017a]. In this section we extend the latter approach to the Group-NMF case.

3.4.1 Task-driven NMF

The general idea of nonnegative TDL or task-driven NMF (TNMF) is to unite the dictionary learning with NMF and the training of the classifier in a joint optimization problem [Bisot et al., 2017a, Sprechmann et al., 2014]. Influenced by the classifier, the basis vectors are encouraged to explain the discriminative information in the data while

keeping a low reconstruction cost. The TNMF model first considers the optimal projections $\mathbf{h}^\star(\mathbf{x}, \mathbf{W})$ of the data points $\mathbf{x}$ on the dictionary $\mathbf{W}$. $\mathbf{x}$ represents a single frame of the input observation $\mathbf{X}$. The projections $\mathbf{h}^\star(\mathbf{x}, \mathbf{W})$ are defined as solutions of the nonnegative elastic-net problem [Zou and Hastie, 2005], expressed as:

$$\mathbf{h}^\star(\mathbf{x}, \mathbf{W}) = \min_{\mathbf{h} \in \mathbb{R}_+^K} \frac{1}{2}\|\mathbf{X} - \mathbf{W}\mathbf{h}\|^2 + \lambda_1 \|\mathbf{h}\|_1 + \frac{\lambda_2}{2}\|\mathbf{h}\|^2; \tag{3.10}$$

where λ_1 and λ_2 are nonnegative regularization parameters. Given each data segment $\mathbf{X}^{(l)}$ of length N frames, associated with a label y in a fixed set of labels $\mathcal{Y}$, we want to classify the mean of the projections of the data points $\mathbf{x}^{(l)}$ belonging to the segment l, such that $\mathbf{X}^{(l)} = [\mathbf{x}_0^{(l)}, ..., \mathbf{x}_{M-1}^{(l)}]$. We define $\hat{\mathbf{h}}^{(l)}$ as the averaged projection of $\mathbf{X}^{(l)}$ on the dictionary, where $\hat{\mathbf{h}}^{(l)}(\mathbf{x}^{(l)}, \mathbf{W})) = \frac{1}{M}\sum_{m=0}^{M-1} \mathbf{h}^\star(\mathbf{x}_m^{(l)}, \mathbf{W})$. The corresponding classification loss (here using multinomial logistic regression) is defined as $\mathcal{L}_s(y, \mathbf{\Phi}, \hat{\mathbf{h}}^{(l)})$, where $\mathbf{\Phi}$ are the parameters of the classifier. The TNMF problem is then expressed as a joint minimization of the expected classification loss over $\mathbf{W}$ and $\mathbf{\Phi}$:

$$\min_{\mathbf{W} \in \mathcal{W}, \mathbf{\Phi} \in \mathcal{A}} f(\mathbf{W}, \mathbf{\Phi}) + \frac{\nu}{2}\|\mathbf{\Phi}\|^2, \tag{3.11}$$

with

$$f(\mathbf{W}, \mathbf{\Phi}) = \mathbb{E}_{y, \mathbf{X}^{(l)}}[l_s(y, \mathbf{\Phi}, \hat{\mathbf{h}}^{(l)}(\mathbf{x}^{(l)}, \mathbf{W}))]. \tag{3.12}$$

Here, $\mathcal{W}$ is defined as the set of nonnegative dictionaries containing unit l_2-norm basis vectors and ν is a regularization parameter on the classifier parameters, meant to prevent over-fitting. The problem in equation (3.12) is optimized with mini-batch stochastic gradient descent as described in Bisot et al. [2017a].

3.4.2 Task-driven Group-NMF

In task-driven Group-NMF (TGNMF) we propose to perform jointly the dictionary learning based on Group-NMF (Section 3.3) and the training of a multinomial logistic regression. The dictionary $\mathbf{W}$ is the concatenation of all the sub-dictionaries $\mathbf{W}^{(cs)}$ and the optimal projections $\mathbf{h}^\star(\mathbf{x}, \mathbf{W})$ are the solutions of (3.10).
Including the similarity constraints (3.7) and (3.8), the TGNMF is expressed as the minimization of the following problem:

$$\min_{\mathbf{W} \in \mathcal{W}, \mathbf{\Phi} \in \mathcal{A}} f(\mathbf{W}, \mathbf{\Phi}) + \frac{\nu}{2}\|\mathbf{\Phi}\|^2 + \mu_1 J_{\text{SPK}} + \mu_2 J_{\text{SES}}, \tag{3.13}$$

with $f(\mathbf{W}, \mathbf{\Phi})$ as defined above. The problem is again optimized with mini-batch stochastic gradient descent. However, as opposed to the previous algorithm, for each data point $\mathbf{x}$ belonging to a particular $\mathbf{X}^{(gs)}$, only the corresponding sub-dictionaries ($\mathbf{W}^{(gs)}$) are updated, whereas the other dictionaries are left unchanged in order to match the Group-NMF adaptation scheme [Serizel et al., 2016b].

Duration	< 1 min	1 min – 5 min	> 5 min
Number of speakers	25	26	44

Table 3.1: Speakers distribution according to the amount of available training data.

3.5 Experimental setup and corpus

3.5.1 Corpus

The approaches presented here are tested on a subset of the ESTER corpus [Gravier et al., 2004], a radio broadcast corpus. Only speakers with at least 10 sec of training data are selected from ESTER to compose the subset corpus. Speaker utterances are split in 10 sec segments in order to obtain enough segments to train the back-end classifier. The amount of training data is limited to 6 min per speaker. When there is more than 6 min of speech for a speaker, 10 sec segments are selected randomly to compose a 6 min subset. The resulting corpus is composed of 6 h 11 min of training data and 3 h 40 min of test data both distributed among 95 speakers. The training data is extracted from the original ESTER training set and the test data is extracted from the original ESTER development set. This way, there is no overlapping session between the training set and the test set. The amount of training data per speaker ranges from 10 sec to 6 min (Table 3.1).

3.5.2 i-vector baseline

A baseline i-vector-based system is trained with the LIUM speaker diarisation toolkit [Rouvier et al., 2013]. The acoustic features are computed with YAAFE [Mathieu et al., 2010]. They are 20 mel-frequency cepstral coefficients (MFCC) [Davis and Mermelstein, 1980], including the energy coefficient. They are computed on 32 ms frames with 16 ms overlap. The MFCC are augmented with their first and second derivatives to form a 60-dimensional feature vector. A universal background model (UBM) with 256 Gaussian components per acoustic feature is trained on the full training set described above and the dimension of the total variability space is set to 100. The parameter values are in the range of the values commonly found in the literature for datasets of similar size. Eigen factor radial normalisation (EFR) is applied on i-vectors before classification [Bousquet et al., 2011].

3.5.3 NMF-based feature learning

NMF-based systems are trained on general purpose graphical processing units (GPGPU) with an in-house software.[2] exploiting the Theano toolbox [Bastien et al., 2012] The acoustic features are 132 constant-Q transform coefficients (CQT) [Brown, 1991] computed on 16 ms frames with YAAFE [Mathieu et al., 2010]. To cope with the well-known problem of non-uniqueness of the NMF solution, NMF and Group-NMF are initialised

[2]Source code is available at https://github.com/rserizel/groupNMF

randomly 6 times and trained independently for 100 iterations. In each case, the factorization with the lowest cost function value at the end of the training is selected to extract features. After preliminary tests, the number of components for the NMF has been set to $K = 100$. The number of components for each data portion of the Group-NMF is set to $K = 8$ ($K_{\text{SPK}} = 4$, $K_{\text{SES}} = 2, K_{\text{RES}} = 2$). Only speaker-related bases and session-related bases are kept to project the data at runtime. There are 236 unique (speaker, session) couples, so the dimension of the feature vectors extracted with the Group-NMF is $K = 1416$. The weights μ_1 and μ_2 are scaled such that, respectively, for $\mu_1 = 1$ the contributions from (3.4) and (3.7) to (3.9) are equivalent, and for $\mu_2 = 1$ the contributions from (3.4) and (3.8) to (3.9) are equivalent. The features extracted with NMF are scaled to unit variance before classification. In the remainder of this chapter, Group-NMF applied without similarity constraints ($\mu_1 = 0$ and $\mu_2 = 0$) is denoted Group-NMF$_0$. Similarly, Group-NMF with similarity constraints ($\mu_1 = 0.4$ and $\mu_2 = 0.15$) is denoted Group-NMF$_c$.

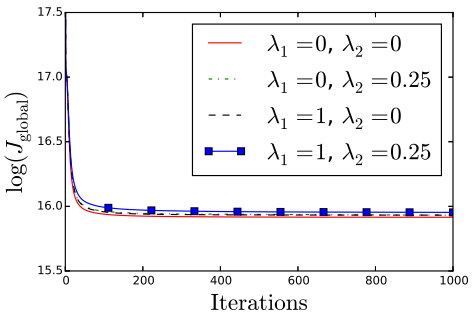
3.5.4 Task-driven approaches

TNMF and TGNMF are applied to fine-tune the dictionaries obtained with the unsupervised NMF and Group-NMF described above.[3] The projections on the dictionary (corresponding to equation (3.10)) are computed using the *lasso* function from the *spams* toolbox [Mairal et al., 2010]. The classifier is updated using one iteration of the scikit-learn [Pedregosa et al., 2011] implementation of the multinomial logistic regression with the L-BFGS solver. The model is trained over $I = 5$ full passes over the data (epochs). When the initial dictionary is obtained with standard NMF ($K = 100$), the initial gradient update step is 0.0005 and the parameters for the elastic net problem are $\lambda_1 = 0.001$ and $\lambda_2 = 0.001$. When the initial dictionary is obtained with Group-NMF ($K = 1416$), the initial gradient update step is 0.0001 and the parameter for the elastic net problem are $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$. The decaying of the gradient steps over iterations follows the same heuristic as suggested in [Mairal et al., 2012]. The hyper parameters are obtained after performing a grid search over several reasonable values. After 5 epochs, the dictionaries are kept fixed and the classifier alone is trained for at most 50 epochs. In the remainder of this chapter, TGNMF applied without similarity constraints ($\mu_1 = 0$ and $\mu_2 = 0$) is denoted TGNMF$_0$. Similarly, TGNMF with similarity constraints ($\mu_1 = 0.0001$ and $\mu_2 = 0.0001$) is denoted TGNMF$_g$.

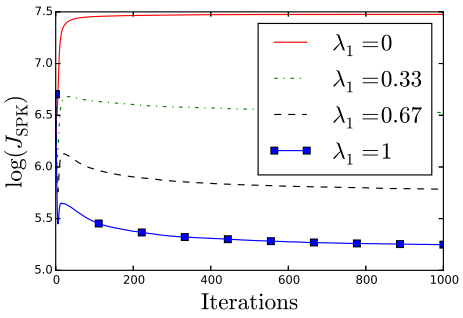
3.5.5 Multinomial logistic regression

Normalised i-vectors and feature vectors extracted with NMF and Group-NMF are classified with a multinomial logistic regression performed with the scikit-learn toolkit [Pedregosa et al., 2011]. The logistic regression is preferred to PLDA as the latter is known to perform quite poorly when the number of samples becomes small compared to the feature dimensionality, which is the case here.

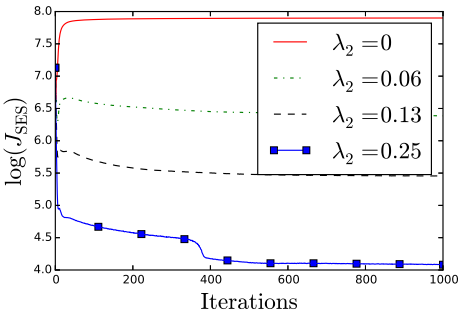
[3] Source code is available at https://github.com/rserizel/TGNMF



(a) KL-divergence



(b) Intra-speaker distance across sessions($\lambda_2 = 0$)



(c) Intra-session distance across speakers ($\lambda_1 = 0$)

Figure 3.1: Convergence of the different criteria depending on the weights λ_1 and λ_2

3.5.6 Performance evaluation

In order to mitigate the effect of the imbalance between speakers in the test set, the classification performance is measured with weighted F1-score [Rijsbergen, 1979] where the F1-score is computed for each class (here the classes are the speakers to identify) separately and weighted by the number of utterances in the class. Variations in identification performance are validated using the McNemar test [McNemar, 1947] with significance level .05. In the remainder of the chapter, unless stated otherwise explicitly, when a performance change is mentioned it is statistically significant.

3.6 Results and discussion

3.6.1 Group NMF

The first important test is to control that the constraints imposed on the speaker bases and the session bases do not degrade the stability of the NMF algorithm. Indeed, convergence can quickly become problematic when imposing constraints on NMF. The KL-

	λ_2			
λ_1	0	0.06	0.12	0.25
0	77.8%	76.5%	76.0%	76.7%
0.33	75.6%	80.2%	78.9%	79.7%
0.67	74.1%	77.3%	77.4%	75.1%
1	78.8%	76.0%	75.7%	79.1%

Table 3.2: Weighted F1-scores obtained for different values of λ_1 and λ_2.

	Features			
			Group-NMF	
	i-vector	NMF	$\lambda_1 = 0$	$\lambda_1 = 0.33$
			$\lambda_2 = 0$	$\lambda_2 = 0.06$
F1-score	76.1%	70.7%	77.8%	**80.2%**

Table 3.3: Weighted F1-scores obtained for a classification with multinomial logistic regression.

divergence still varies uniformly even with constraints on the cost function (3.9) (Figure 3.1 (a)). Yet the constraints are effective at reducing the distance between the speaker bases (Figure 3.1 (b)) and between the sessions bases (Figure 3.1 (c)).
In a second experiment the proposed approach is tested for different value of the weight applied to the constraints. Weighted F1-score performance is presented in Table 3.2. A few trends appear on this table. Firstly it seems clear that imposing constraints on the speaker bases and the session bases does have an impact on the performance of the speaker identification. Secondly, it appears that there is a trade-off between the weight λ_1 and λ_2. Indeed, for a fixed λ_1, the performance reaches a maximum for a particular value of λ_2. Increasing λ_2 beyond this value results in a performance degradation.
In a final experiment, the systems described above and the i-vector baseline are compared on the subset of ESTER described in Section 3.5 (Table 3.3). The Group-NMF has been tested for different values of the weight applied to the constraints and two different configurations have been selected. The first configuration is fully unconstrained ($\lambda_1 = 0$ and $\lambda_2 = 0$) and both constraints are active in the second configuration ($\lambda_1 = 0.33$ and $\lambda_2 = 0.06$). The first remark is that all systems perform reasonably well even if standard NMF is clearly behind the other approaches ($p < .001$). The unconstrained Group-NMF and the i-vector approach perform similarly (the difference is not statistically significant). Imposing constraints on both the speaker bases and the session bases improves significantly the performance compared to the i-vector approach and the unconstrained Group-NMF ($p < .01$ in both cases).

3.6.2 Task-Driven NMF

F1-score performance obtained with the different approaches described above is presented in Table 3.4. Each column corresponds to a different initialization method (NMF, Group-NMF$_0$ and Group-NMF$_g$). The first row (labeled **unsupervised**) presents the

reference performance for each initialization method, where the feature learning model and the classifier are learned independently. For the sake of simplicity, in the remainder of the chapter these methods are referred to as unsupervised, as opposed to supervised methods (TNMF and TGNMF), even though some level of supervision is necessary for Group-NMF. The second row (labeled **TNMF**) presents the performance obtained when applying TNMF in a similar way as in Bisot et al. [2017a], initialized with the dictionaries obtained with NMF and Group-NMF. The last rows present the performance obtained when applying TGNMF$_0$ and TGNMF$_g$, initialized with the dictionaries obtained with Group-NMF.
Two main tendencies can be observed from the results in Table 3.4. First, on small dictionaries (NMF with $K = 100$), TNMF allows for a large improvement compared to unsupervised methods and good performance. Secondly, TGNMF can sometimes provide large improvement reducing the performance difference between systems using initializations with Group-NMF$_0$ and Group-NMF$_g$.

Unsupervised reference methods

The performance obtained with unsupervised methods tends to confirm previous findings where NMF (75.6%) is behind other systems and where the Group-NMF (81.7% with Group-NMF$_g$ or 80.7% Group-NMF$_0$) is better than the baseline i-vector system (76.1%). These systems also improve the performance compared to previous experiments from the authors with Group-NMF with generalized Kullback-Leibler divergence [Kullback and Leibler, 1951] applied on Mel-spectrums coefficients [Serizel et al., 2016b].

TNMF

Applying TNMF in a similar way as in Bisot et al. [2017a], initialized on the dictionaries learned with standard NMF allows for a large performance improvement (from 75.6% to 79.9%), whereas TNMF initialized with concatenated dictionaries obtained with Group-NMF leads to improvements that are not statistically significant. This could be due to the fact that the dictionaries are then too large and that one of the advantages of TNMF is that it is the most efficient when considering dictionaries smaller than those used with unsupervised methods.

TGNMF$_0$

Group-NMF$_0$ allows for focusing on learning some sub-dictionaries related to portions of the data originating from a specific speaker or session. This already proved effective on the unsupervised methods. This observation is confirmed when applying TGNMF$_0$ on dictionaries obtained with Group-NMF$_0$. TGNMF$_0$ then allows for a F1-score increase from 80.7% to 81.7%. The system obtains similar performance as the best reference system (Group-NMF$_g$), without exploiting the similarity constraints. The gain is less important when applying TGNMF$_0$ initialized with Group-NMF$_g$ where the annotations were already exploited to some extent.

	Initialization			
Features	i-vector	NMF	Group-NMF$_0$	Group-NMF$_g$
Unsupervised	76.1%	75.6%	80.7%	81.7%
TNMF	–	79.9%	81.1%	81.9%
TGNMF$_0$	–	–	81.7%	82.1%
TGNMF$_c$	–	–	82.0%	**82.2%**

Table 3.4: Weighted F1-scores for speaker classification ($K = 100$ for NMF and $K = 1446$ for Group-NMF). Each column corresponds to a different initialization method and each row corresponds to the method applied after the initialization (for the first row no processing is done after the initialization). The subscripts $_0$ and $_g$ correspond to method without and with constraints, respectively (see also 3.5.3, 3.5.4 and 3.5.6 for more detailed explanations).

TGNMF$_g$

Imposing similarity constraints during TGNMF helps improving the performance further, up to 82.2% when initialized with dictionaries obtained with Group-NMF$_g$. This is our best performance to date on this corpus. However, this is not significantly better than performance obtained with other TGNMF systems. This tends to indicate that both methods (Group-NMF and TGNMF) are to some extent redundant in the way to exploit the information from the annotations to structure the dictionaries and that we maybe reached a saturation point for these methods applied to speaker identification on rather small corpora such as the subset of ESTER.

3.7 Conclusions

This chapter introduced a new feature learning approach for speaker identification that is based on NMF. Works on exemplar based speaker identification have shown that dictionary atoms in an NMF system can represent well speaker identity. Capitalizing on this statement, we proposed an approach based on group-NMF that is inspired by the state-of-the-art i-vector approach and tries to capture both speaker variability and session variability. The central idea is to impose similarity constraints on speaker-dependent bases and session-dependent bases in the decomposition dictionaries. The proposed approach has proven to be competitive with i-vectors on a small corpus.

An alternative approach to model based compensation of the noise is to use a noise reduction front-end before the speaker identification back-end. This approach can be particularly useful when addressing far-field speaker recognition problems with devices that are equipped of several microphones. In this case the so-called multichannel noise reduction algorithms applied in front-end can exploit the spatial properties of the acoustic scene and in particular the fact that two different sound sources are usually located at two different places in space. This approach is investigated within the ANR project Robovox (see also 3.8).

3.8 Other related works

While at Télécom ParisTech, I collaborated with Victor Bisot a PhD student at Télécom ParisTech supervised by Gaël Richard (Full professor at Télécom ParisTech, Paris, France) and Slim Essid (Full professor at Télécom ParisTech, Paris, France). This work is methodologically close to the work presented in this chapter. Victor Bisot proposed the original task-driven NMF framework for acoustic scene classification. It proved competitive with DNN based approaches [Bisot et al., 2016, 2017a,c] during DCASE 2016 challenge. This approach was later adapted to sound event detection [Bisot et al., 2017b]. Note that this work is also close, from an application point of view to the work presented in Chapters 5 and 6.

Since my arrival at Université de Lorraine in September 2016 I have also continued working on speaker identification and verification. I have been collaborating with Md Sahidullah (Inria starting researcher, Nancy, France) and Emmanuel Vincent (Inria senior researcher scientist, Nancy, France) on some work on speaker verification on short speech segments. One problem when performing speaker identification is the phonetic unbalance across the possible test utterance and also the possible mismatch with the phonetic content in the utterances used for enrollment [Poddar et al., 2017]. One of the idea that we explored was to train a speaker embedding networks that would disentangle the phonetic contribution from the speaker contribution in each speaker utterance. In order to do so, we investigated architectures with several downstream branches targeting for example phoneme recognition, utterance reconstruction besides the standard speaker classification branch use in x-vectors [Snyder et al., 2018]. The cost from each separate branch was either combined in an adversarial mode [Lample et al., 2017] or in an multi-task mode. In the former we tried to obtain an model that would explicitly discard the phonetic unbalance while the latter was aiming at taking all the aspects of the speech utterances into account separately during the training phase. Both approaches showed to be effective when a limited amount of training data was available but the benefits vanished with large training corpora. Part of this work has been used in a cross-institution collaborative submission to the Short-duration speaker verification challenge [Sahidullah et al., 2021].

Since October 2019, I am co-director of Sandipana Dowerah's PhD thesis with Denis Jouvet (Inria senior researcher scientist, Nancy, France). The PhD takes place within the ANR PRCE project Robovox involving the laboratory of computer science in Avignon (Laboratoire d'informatique d'Avignon – LIA) and the company A.I. Mergence in Paris. The goal of the project is to perform speaker verification with a mobile robot in challenging conditions (high reverberation or low SNR). Sandipana Dowerah is working on the impact of using a speech enhancement front-end for far-field speaker verification in noisy environments. She first investigated the degradation caused by reverberation, additive noise or both on speaker recognition. She proposed to use a multichannel front-end based on signal processing and DNN based masks to compensate from these degradation and compared the performance obtained to that obtained with a state-of-the-art pure DNN multichannel speech enhancement approach, FaSNet [Luo et al., 2019]. The first conclu-

sion was that when applying a pre-processing on the speech signals, it was important to perform the speaker verification enrollment on conditions that are matching the operating conditions (in order to account for the distortions introduced by the pre-processing algorithm). When the enrollment and operating conditions are matching, then a multichannel speech enhancement pre-processing can improve the speaker verification performance in challenging conditions. The second conclusion was that even if the pure DNN approach was generally outperforming the approach relying partly on signal processing in terms of speech enhancement metrics, the pure-DNN multichannel speech enhancement approach was systematically outperformed on the downstream speaker verification task. Finally, even-though the pre-processing was trained on synthetic data (dry sources convoluted with room impulse responses), using the multichannel pre-processing exhibited performance improvement on real recorded signals from the VOiCES dataset [Nandwana et al., 2019]. This work is at the interface of the work presented in this chapter and in Chapter 4.

4 Multichannel speech enhancement

Context: This work was done when I was a postdoctoral researcher at KU Leuven (Leuven, Belgium) between July 2011 and December 2012 together with Marc Moonen, Bas Van Dijk (Cochlear Ltd. Belgium) and Jan Wouters (UZ Leuven, Belgium).[1] The work presented here has been previously published in articles [Serizel et al., 2013, 2014].

4.1 Introduction

A major challenge in cochlear implant design is to improve the speech understanding in noise for cochlear implant recipients [Hu and Loizou, 2008]. To this end, having an efficient front-end noise reduction (NR) is important. Therefore, several NR algorithms have been developed and tested with cochlear implant recipients [Hamacher et al., 1997, Van Hoesel and Clark, 1995, Weiss, 1993]. Commercial cochlear implants usually include multiple microphones and allow for multichannel adaptive NR algorithms, such as the BEAM in the Cochlear Freedom device, which have been shown to greatly improve speech understanding for cochlear implant recipients [Spriet et al., 2007].
In general, cochlear implant recipients need a 10dB to 25dB higher signal-to-noise-ratio (SNR) than normal hearing subjects to achieve a similar speech understanding performance [Wouters and Berghe, 2001] but they can tolerate a much higher speech distortion (SD). This motivates the use of more aggressive noise reduction (NR) strategies. The speech distortion weighted multichannel Wiener filter (SDW-MWF) has been developed to allow for tuning multichannel Wiener filter (MWF)-based NR and perform a more aggressive NR by allowing for more SD [Doclo et al., 2007, Ephraim and Van Trees, 1995, Ngo et al., 2009, Spriet et al., 2004]. In the case of a single speech source the SDW-MWF performance can sometimes be improved if the filters are reformulated based on the assumption that the frequency-domain autocorrelation matrix of the speech signal is a rank-1 matrix, leading to the so-called spatial-prediction MWF (SP-MWF) [Benesty et al., 2008, Cornelis et al., 2010] and the rank-1 MWF (R1-MWF) [Souden et al., 2009]. In this paper, the difference is investigated between the original SDW-MWF and these two rank-1 approximation based NR filters when the rank of autocorrelation matrix of the speech signal is actually greater than one.
All these NR algorithms rely on the estimation of the autocorrelation matrix of the speech signal, which is based on a rank-1 approximation with a so-called first column decomposition, as well as on the assumption that the (unknown) speech signal and the noise are uncorrelated and that these signals are locally stationary. In low input SNR scenarios, if these assumptions are violated, the autocorrelation matrix of the speech

[1] Collaborators listed alphabetically, members of the host institution unless mentioned otherwise

signal can be wrongly estimated and become non positive semi-definite. The SDW-MWF as well as the rank-1 approximation based filters can then deliver unpredictable NR performance. This chapter proposes a solution to this problem that is to select an alternative rank-1 approximation based on an eigenvalue decomposition (EVD) [Serizel et al., 2013], or a generalized eigenvalue decomposition (GEVD) [Dendrinos et al., 1991, Doclo and Moonen, 2002, Jensen et al., 1995], of the autocorrelation matrix of the speech signal.
These alternative NR filters are demonstrated to deliver a better NR performance especially in low input SNR scenarios and are especially useful in cochlear implants, where more SD and hence a more aggressive NR can be tolerated. The GEVD based NR filter is also extended to a rank-R approximation based filter, in which the rank reduction is shown to be equivalent to tuning the NR to be more aggressive. The rank-1 approximation based filter then indeed represents the extreme case with the most aggressive NR. A performance comparison is provided between the original SDW-MWF, the EVD based NR filter and the GEVD based NR filter applied on both bilateral and binaural set-ups [Doclo et al., 2006, Hamacher, 2002].
The signal model and the SDW-MWF are described in Section 4.2. The so-called first column decomposition and how this provides an interpretation of the SDW-MWF versus the SP-MWF and the R1-MWF is described in Section 4.3. The EVD based NR filter is introduced in Section 4.4. The GEVD based NR filter is presented in Section 4.5 and is extended to a rank-R approximation based filter in Section 4.6. The performance of the original SDW-MWF, the EVD based NR filter and the GEVD based NR filter are compared in Section 4.7. A summary of the chapter and conclusions are provided in Section 4.8 and the other related works are presented in Section 4.9.

4.2 Backgound and problem statement

4.2.1 Signal model

Let M be the number of microphones (channels). Let us consider the time-frequency domain representation of the signal x recorded by the microphones. For each frame n, the frequency domain representation $x_m(\omega, n)$ of the input signal for microphone m has a speech component $x_{m,s}(\omega, n)$ and an additive noise component $x_{m,n}(\omega, n)$, i.e.:

$$x_m(\omega, n) = x_{m,s}(\omega, n) + x_{m,n}(\omega, n) \; m \in \{1...M\}, \tag{4.1}$$

where $\omega = 2\pi f$ is the frequency-domain variable. For conciseness, (ω, n) will be omitted in all subsequent equations. Subscripts s and n will also be used to denote the "speech" and "noise" component of other quantities. Signal model (4.1) holds for so-called "speech plus noise periods". There are also "noise only periods" (i.e., speech pauses), during which only a noise component is observed.
In practice, in order to distinguish between "speech plus noise periods" and "noise only periods" it is necessary to use a voice activity detector (VAD). The performance of the VAD can affect the performance of the NR. For the time being, a perfect VAD is assumed.

The compound vector gathering all microphone signals is:

$$\mathbf{x} = [x_1 \dots x_M]^T. \tag{4.2}$$

The autocorrelation matrix of the microphone signals in "speech plus noise periods", and of the speech component and the noise component of the microphone signals are given by:

$$\mathbf{R}_x = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} \tag{4.3}$$

$$\mathbf{R}_s = \mathbb{E}\{\mathbf{x}_s\mathbf{x}_s^H\} \tag{4.4}$$

$$\mathbf{R}_n = \mathbb{E}\{\mathbf{x}_n\mathbf{x}_n^H\}, \tag{4.5}$$

where ^{H} denotes the Hermitian transpose and $\mathbb{E}\{.\}$ is the expectation. $\mathbf{R}_n$ can be estimated during "noise only periods" and $\mathbf{R}_x$ can be estimated during "speech plus noise periods". If the speech and noise signals are assumed to be uncorrelated and if the noise signal is stationary, $\mathbf{R}_s$ can be estimated by using:

$$\mathbf{R}_s = \mathbf{R}_x - \mathbf{R}_n. \tag{4.6}$$

In practice, the autocorrelation matrices are estimated recursively in time. The estimate of the autocorrelation matrix of the microphone signals is updated during "speech plus noise periods", using:

$$\tilde{\mathbf{R}}_x = \lambda\tilde{\mathbf{R}}_x + (1-\lambda)\mathbf{x}\mathbf{x}^H, \tag{4.7}$$

where $\lambda \in [0, 1]$ is an exponential forgetting factor that depends on the number of past frames to be taken into account (here the forgetting time is about 1 sec). This clearly exceeds the spectral stationarity of speech signals (around 20 ms) but not necessarily the spatial stationarity of the sources.
The estimate of the autocorrelation matrix of the noise component of the microphone signals is updated similarly during "noise only periods", using:

$$\tilde{\mathbf{R}}_n = \lambda\tilde{\mathbf{R}}_n + (1-\lambda)\mathbf{x}\mathbf{x}^H \tag{4.8}$$

$$= \lambda\tilde{\mathbf{R}}_n + (1-\lambda)\mathbf{x}_n\mathbf{x}_n^H. \tag{4.9}$$

The estimate of the autocorrelation matrix of the speech component of the microphone signals is then given by:

$$\tilde{\mathbf{R}}_s = \tilde{\mathbf{R}}_x - \tilde{\mathbf{R}}_n. \tag{4.10}$$

It is noted that in the sequel, NR filters are specified as functions of $\mathbf{R}_x$, $\mathbf{R}_n$, and $\mathbf{R}_s$, whereas in practice these matrices are replaced by their estimated versions $\tilde{\mathbf{R}}_x$, $\tilde{\mathbf{R}}_n$, and $\tilde{\mathbf{R}}_s$ (or modifications thereof).

4.2.2 MWF-based Noise Reduction

An MWF $\mathbf{w} = [w_1 \dots w_M]^T$ will be designed and applied to the microphone signals, which minimizes a Mean Squared Error (MSE) criterion:

$$J_{\mathrm{MWF}} = \mathbb{E}\{\|E\|^2\}, \tag{4.11}$$

where E is an error signal to be defined next, depending on the scheme applied. The filter output signal z is defined as:

$$z = \mathbf{w}^H \mathbf{x}. \tag{4.12}$$

The desired signal for the MWF is arbitrarily chosen to be the (unknown) speech component of the first microphone signal ($m = 1$). This can be written as:

$$d_{\text{MWF}} = \mathbf{e}_1^H \mathbf{x}_s, \tag{4.13}$$

where $\mathbf{e}_1$ is an all-zero vector except for a one in the first position.
The MWF aims to minimize the squared distance between the filtered microphone signal 4.12 and the desired signal (4.13). The corresponding MSE criterion is:

$$J_{\text{MWF}} = \mathbb{E}\{\|\mathbf{w}^H \mathbf{x} - \mathbf{e}_1^H \mathbf{x}_s\|^2\}. \tag{4.14}$$

The MWF solution is given as:

$$\mathbf{w}_{\text{MWF}} = (\mathbf{R}_s + \mathbf{R}_n)^{-1} \mathbf{R}_s \mathbf{e}_1. \tag{4.15}$$

The SDW-MWF has been proposed to provide an explicit trade-off between the NR and the SD [Doclo et al., 2007, Ephraim and Van Trees, 1995, Ngo et al., 2009, Spriet et al., 2004]. Changing the optimization problem to a constrained optimization problem, the MSE criterion effectively becomes:

$$J_{\text{SDW-MWF}} = \mathbb{E}\{\|\mathbf{w}^H \mathbf{x}_s - \mathbf{e}_1^H \mathbf{x}_s\|^2\} + \mu \mathbb{E}\{\|\mathbf{w}^H \mathbf{x}_n\|^2\}, \tag{4.16}$$

where μ is a trade-off parameter. The SDW-MWF solution is then given as:

$$\mathbf{w}_{\text{SDW-MWF}} = (\mathbf{R}_s + \mu \mathbf{R}_n)^{-1} \mathbf{R}_s \mathbf{e}_1. \tag{4.17}$$

In a single speech source scenario, the autocorrelation matrix of the speech component of the microphone signals $\mathbf{R}_s$ is often assumed to be a rank-1 matrix and can then be rewritten as:

$$\mathbf{R}_s = P_s \mathbf{A} \mathbf{A}^H, \tag{4.18}$$

where P_s is the power of the speech source signal and $\mathbf{A}$ is the M-dimensional steering vector, containing the acoustic transfer functions from the speech source position to the microphones (including the microphone characteristics).
Based on this rank-1 assumption it is possible to derive the so-called SP-MWF [Benesty et al., 2008, Cornelis et al., 2010]:

$$\mathbf{w}_{\text{SP-MWF}} = \mathbf{R}_n^{-1} \mathbf{R}_s \mathbf{e}_1 \frac{\mathbf{e}_1^H \mathbf{R}_s \mathbf{e}_1}{\mu \mathbf{e}_1^H \mathbf{R}_s \mathbf{e}_1 + \text{Tr}\{\mathbf{R}_n^{-1} \mathbf{R}_s \mathbf{e}_1 \mathbf{e}_1^H \mathbf{R}_s\}} \tag{4.19}$$

and the R1-MWF [Souden et al., 2009]:

$$\mathbf{w}_{\text{R1-MWF}} = \mathbf{R}_n^{-1} \mathbf{R}_s \mathbf{e}_1 \frac{1}{\mu + \text{Tr}\{\mathbf{R}_n^{-1} \mathbf{R}_s\}}. \tag{4.20}$$

The filters (4.17), (4.19) and (4.20) are fully equivalent if $rank(\mathbf{R}_s) = 1$. In practice, however, (4.18) may not hold, *i.e.*, $rank(\mathbf{R}_s) > 1$ even for a single speech source scenario and then (4.17), (4.19) and (4.20) are different filters.

4.3 First column decomposition

When $rank(\mathbf{R}_s) > 1$ the matrix $\mathbf{R}_s$ can be decomposed as:

$$\mathbf{R}_s = \mathbf{R}_{s_{r1}} + \mathbf{R}_{\text{rem}}, \tag{4.21}$$

where $\mathbf{R}_{s_{r1}}$ is a rank-1 approximation of $\mathbf{R}_s$ and $\mathbf{R}_{\text{rem}}$ is a "remainder" matrix. The decomposition is not unique and so several choices for $\mathbf{R}_{s_{r1}}$ can be considered.
The most obvious choice for $\mathbf{R}_{s_{r1}}$ is a rank-1 extension of the first column and row of $\mathbf{R}_s$, i.e.:

$$\mathbf{R}_s = \underbrace{\mathbf{d}\mathbf{d}^H\sigma_{1,1}}_{\mathbf{R}_{s_{r1}}} + \underbrace{\left[\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \hline 0 & x & \cdots & x \\ \vdots & \vdots & & \vdots \\ 0 & x & \cdots & x \end{array}\right]}_{\mathbf{R}_z}, \tag{4.22}$$

where

$$\sigma_{i,j} = [\mathbf{R}_s]_{i,j} \tag{4.23}$$

$$\mathbf{d} = [1 \ \frac{\sigma_{2,1}}{\sigma_{1,1}} \dots \frac{\sigma_{1,N}}{\sigma_{1,1}}]^T \tag{4.24}$$

and $\sigma_{1,1}$is the speech power in microphone 1. This decomposition will be referred to as the "first column decomposition". It allows to pinpoint the differences between the filters (4.17), (4.19) and (4.20) whenever $rank(\mathbf{R}_s) > 1$. This decomposition has also been exploited in [Benesty et al., 2012].
It is noted that:

$$\mathbf{R}_s\mathbf{e}_1 = \mathbf{R}_{s_{r1}}\mathbf{e}_1 + \underbrace{\mathbf{R}_{\text{rem}}\mathbf{e}_1}_{=\mathbf{0}}, \tag{4.25}$$

which means that the (rightmost) "desired signal part" $\mathbf{R}_s\mathbf{e}_1$ in (4.17), (4.19) and (4.20) can be (obviously) replaced by the "rank-1 approximation desired signal part" $\mathbf{R}_{s_{r1}}\mathbf{e}_1$. The difference between the filters (4.17), (4.19) and (4.20) then effectively depends on how $\mathbf{R}_{\text{rem}}$ is treated, as will be explained next. Note that when $rank(\mathbf{R}_s) = 1$, then $\mathbf{R}_{\text{rem}} = 0$ and so it is again seen that the filters are fully equivalent.

4.3.1 SDW-MWF

Plugging (4.22) into the SDW-MWF formula (4.17) leads to:

$$\mathbf{w}_{\text{SDW}-\text{MWF}} = (\mathbf{R}_{s_{r1}} + \mu(\mathbf{R}_n + \frac{1}{\mu}\mathbf{R}_{\text{rem}}))^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.26}$$

This means that in the SDW-MWF (4.17) $\mathbf{R}_s$ can be replaced by $\mathbf{R}_{s_{r1}}$ and then the remainder matrix $\mathbf{R}_{\text{rem}}$ is effectively treated as noise (up to a scaling with $\frac{1}{\mu}$).

To avoid the scaling with $\frac{1}{\mu}$ an alternative approach is to start from the MSE criterion (4.11). Plugging (4.22) into (4.11), merging $\mathbf{R}_{\text{rem}}$ with the noise and (only then) introducing the trade-off factor μ, leads to:

$$J^{\diamond}_{\text{SDW}-\text{MWF}} = \mathbf{w}^H\mathbf{R}_{s_{r1}}\mathbf{w} - \mathbf{w}^H\mathbf{R}_{s_{r1}}\mathbf{e}_1 - \mathbf{e}_1^H\mathbf{R}_{s_{r1}}\mathbf{w} + \mathbf{e}_1^H\mathbf{R}_{s_{r1}}\mathbf{e}_1 + \mu\left(\mathbf{w}^H\mathbf{R}_z\mathbf{w} + \mathbf{w}^H\mathbf{R}_n\mathbf{w}\right), \tag{4.27}$$

where the superscript $^{\diamond}$ is used to denote the alternative formulation where the trade-off factor μ is introduced after $\mathbf{R}_{\text{n}}$ and $\mathbf{R}_{\text{rem}}$ are merged.
The filter minimizing (4.27) is then:

$$\mathbf{w}^{\diamond}_{\text{SDW}-\text{MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\left(\mathbf{R}_n + \mathbf{R}_z\right)\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.28}$$

This again means that in the SDW-MWF (4.17) $\mathbf{R}_s$ is replaced by $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_{\text{rem}}$ is effectively returned to noise, i.e, $\mathbf{R}_n$ is replaced by $\mathbf{R}_n + \mathbf{R}_{\text{rem}}$. The initial speech plus noise decomposition $\mathbf{R}_x = \mathbf{R}_n + \mathbf{R}_s$ is then effectively reshuffled into $\mathbf{R}_x = \mathbf{R}_{s_{r1}} + (\mathbf{R}_n + \mathbf{R}_{\text{rem}})$.
It is seen that (4.26) and (4.28) only differ in the weighting applied to $\mathbf{R}_{\text{rem}}$. While (4.26) is fully equivalent to (4.17), (4.28) adopts a weighting that is intuitively more appealing if $\mathbf{R}_{\text{rem}}$ is considered to be a noise contribution. However, $\mathbf{R}_{\text{rem}}$ can come not only from noise estimate leaking into the speech estimate but also from various factors such as VAD errors, over/underestimation of the noise during speech periods, correlation between speech and noise... Therefore, it is unclear which of noise weighting strategies (4.26) and (4.28) is the more appropriate. For $\mu = 1$, filters (4.26) and (4.28) are equivalent.

4.3.2 SP-MWF

Plugging (4.22) into the SP-MWF formula (4.19) leads to:

$$\mathbf{w}_{\text{SP}-\text{MWF}} = \mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1\frac{1}{\mu + \text{Tr}\left\{\mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}}\right\}} \tag{4.29}$$

and

$$\mathbf{w}_{\text{SP}-\text{MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\mathbf{R}_n\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.30}$$

This means that the SP-MWF effectively corresponds to the SDW-MWF (4.17) where $\mathbf{R}_s$ is replaced by $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is simply ignored.

4.3.3 R1-MWF

Plugging (4.22) into the R1-MWF formula (4.20) leads to:

$$\mathbf{w}_{\text{R1}-\text{MWF}} = \mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1\frac{1}{\mu + \text{Tr}\left\{\mathbf{R}_n^{-1}\left(\mathbf{R}_{s_{r1}} + \mathbf{R}_{\text{rem}}\right)\right\}} \tag{4.31}$$

and

$$\mathbf{w}_{\text{R1}-\text{MWF}} = \left(\mathbf{R}_{s_{r1}} + \bar{\mu}\mathbf{R}_n\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1, \tag{4.32}$$

where

$$\bar{\mu} = \mu + \mathrm{Tr}\left\{\mathbf{R}_n^{-1}\mathbf{R}_{\mathrm{rem}}\right\} \neq \mu. \tag{4.33}$$

By comparing (4.32) with (4.28) and (4.30), it is seen that the R1-MWF represents an intermediate approach between the SDW-MWF and the SP-MWF. Indeed, in the R1-MWF the remainder matrix $\mathbf{R}_z$ is ignored in the spatial filter ($\mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1$) as it is also the case for the SP-MWF filter (see (4.30) and (4.31)). The remainder matrix $\mathbf{R}_z$ changes the trade-off parameter from μ to $\bar{\mu}$ which effectively changes the spectral postfilter in (4.30) and (4.31). If $\mathbf{R}_z$ is positive semi-definite, $\mu > \bar{\mu}$ which corresponds to putting a higher weight on the noise. This is similar to $\mathbf{R}_z$ being treated as noise in the SDW-MWF case.

4.3.4 Speech autocorrelation matrix estimation

In low input SNR scenarios it is observed that:

$$\tilde{\mathbf{R}}_x \approx \tilde{\mathbf{R}}_n \tag{4.34}$$

and then the estimated $\tilde{\mathbf{R}}_s = \tilde{\mathbf{R}}_x - \tilde{\mathbf{R}}_n$ can loose its positive semi-definiteness, especially so if the noise is non-stationary. This is problematic and has been observed to lead to unpredictable NR performance. The first column decomposition in particular suffers from this estimation problem where the estimated speech power in microphone 1

$$\tilde{\sigma}_{1,1} \triangleq \left[\tilde{\mathbf{R}}_s\right]_{1,1} = \left[\tilde{\mathbf{R}}_x\right]_{1,1} - \left[\tilde{\mathbf{R}}_n\right]_{1,1} \tag{4.35}$$

can become negative (which is meaningless) so that $\mathbf{R}_{s_{r1}}$ is negative semi-definite and hence the desired signal is ill-defined. This explains why the first column decomposition based filters often provide poor NR performance in low input SNR scenarios. In addition, if $\mathbf{R}_z$ is non positive definite, then the $\bar{\mu}$ in the R1-MWF (4.32) may be spuriously decreased instead of increase compared to the μ in the SP-MWF (4.30).

4.4 EVD Based NR Filters

An alternative to the first column decomposition based rank-1 approximation is a rank-1 approximation based on an EVD of $\mathbf{R}_s$, as also introduced in Serizel et al. [2013]:

$$\mathbf{R}_s = \underbrace{\mathbf{d}_{\max}\mathbf{d}_{\max}^H\lambda_{\max}}_{\mathbf{R}_{s_{r1}}} + \mathbf{R}_z, \tag{4.36}$$

where $\lambda_{\max}$ is $\mathbf{R}_s$'s (real-valued) largest eigenvalue, $\mathbf{d}_{\max}$ is the corresponding normalized eigenvector and $\mathbf{R}_z$ is again a remainder matrix. When $\mathrm{rank}(\mathbf{R}_s) = 1$, then $\mathbf{R}_z = 0$ and $\mathbf{R}_{s_{r1}}$ is the same as in the first column decomposition. When $\mathrm{rank}(\mathbf{R}_s) > 1$, then the rank-1 estimated part $\mathbf{R}_{s_{r1}}$ is positive semi-definite if the dominant eigenvalue of $\mathbf{R}_s$ is positive (which is more likely than the first diagonal element $\tilde{\sigma}_{1,1}$ of $\mathbf{R}_s$ being positive as needed in the first column decomposition approach).

It is noted that:

$$\mathbf{R}_s\mathbf{f}_1 = \mathbf{R}_{s_{r1}}\mathbf{f}_1 + \underbrace{\mathbf{R}_z\mathbf{f}_1}_{=0} = \mathbf{R}_{s_{r1}}\mathbf{e}_1, \tag{4.37}$$

to be compared to (4.25), where

$$\mathbf{f}_1 = \mathbf{d}_{\max}\mathbf{d}_{\max}(1)^*, \tag{4.38}$$

with $\mathbf{d}_{\max}(1)$ is the first element of $\mathbf{d}_{\max}$.
An analysis similar to the analysis for the first column decomposition in Section 4.3 can then be done where $\mathbf{R}_s$ is replaced by the rank-1 approximation $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is either treated as noise or ignored. Equivalently, one can start from a modified MSE criterion where, compared to (4.11), the (arbitrary) $\mathbf{e}_1$ is replaced by $\mathbf{f}_1$:

$$J_{\text{EVD-SDW-MWF}} = \mathbb{E}\left\{\|\mathbf{w}^H\mathbf{x}^s - \mathbf{f}_1^H\mathbf{x}^s\|^2\right\} + \mu\mathbb{E}\left\{\|\mathbf{w}^H\mathbf{x}^n\|^2\right\}. \tag{4.39}$$

Replacing the desired signal $\mathbf{e}_1^H\mathbf{x}_s$ by $\mathbf{f}_1^H\mathbf{x}_s$ is equivalent to replacing $\mathbf{R}_s$ by the EVD based $\mathbf{R}_{s_{r1}}$ as demonstrated by (4.37).

4.4.1 EVD-SDW-MWF

The filter minimizing (4.39) is given as:

$$\mathbf{w}_{\text{EVD-SDW-MWF}} = \left(\mathbf{R}_s + \mu\mathbf{R}_n\right)^{-1}\mathbf{R}_s\mathbf{f}_1. \tag{4.40}$$

Plugging (4.36) and (4.37) into (4.40) leads to:

$$\mathbf{w}_{\text{EVD-SDW-MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\left(\mathbf{R}_n + \frac{1}{\mu}\mathbf{R}_z\right)\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.41}$$

This means that in the SDW-MWF (4.17) $\mathbf{R}_s$ is replaced by the EVD based $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is effectively treated as noise (up to a scaling with $\frac{1}{\mu}$).
To avoid the scaling with $\frac{1}{\mu}$, the same alternative derivation as for (4.28) can be applied leading to:

$$\mathbf{w}^{\diamond}_{\text{EVD-SDW-MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\left(\mathbf{R}_n + \mathbf{R}_{\text{rem}}\right)\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.42}$$

This means that in the SDW-MWF (4.17) the desired signal vector $\mathbf{R}_s$ is replaced by the EVD based $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is effectively returned to noise.

4.4.2 EVD-SP-MWF

Based on the MSE criterion (4.39) it is also possible to derive the SP-MWF:

$$\mathbf{w}_{\text{EVD-SP-MWF}} = \mathbf{R}_n^{-1}\mathbf{R}_s\mathbf{f}_1\frac{\mathbf{f}_1^H\mathbf{R}_s\mathbf{f}_1}{\mu\mathbf{f}_1^H\mathbf{R}_s\mathbf{f}_1 + \text{Tr}\left\{\mathbf{R}_n^{-1}\mathbf{R}_s\mathbf{f}_1\mathbf{f}_1^H\mathbf{R}_s\right\}}. \tag{4.43}$$

Plugging (4.37) into the EVD-SP-MWF formula (4.43) leads to:

$$\mathbf{w}_{\text{EVD-SP-MWF}} = \mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}}\mathbf{f}_1\frac{1}{\mu + \text{Tr}\left\{\mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}}\right\}}. \tag{4.44}$$

and

$$\mathbf{w}_{\text{EVD-SP-MWF}} = (\mathbf{R}_{s_{r1}} + \mu\mathbf{R}_n)^{-1}\,\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.45}$$

This means that in the SDW-MWF (4.17) $\mathbf{R}_s$ is replaced by the EVD based $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is simply ignored. The EVD-R1-MWF derivation is omitted for conciseness.

4.4.3 A matrix approximation based derivation of EVD-SDW-MWF and EVD-SP-MWF

From a given $\mathbf{R}_x$ and $\mathbf{R}_n$ the autocorrelation matrix of the speech component can be computed as $\mathbf{R}_s = \mathbf{R}_x - \mathbf{R}_n$ and these matrices can be plugged in the SDW-MWF formula (4.17). It has been mentioned in Section 4.3.4 that this may result in poor NR performance, in particular in low input SNR scenarios, where then the estimated $\mathbf{R}_s$ is oftentimes indefinite rather than positive semi-definite. To avoid this, an alternative approach can be followed where first a better autocorrelation matrix of the speech component is computed (call it $\mathbf{R}_{s_{r1}}$) together with a better autocorrelation matrix of the noise component (call it $\mathbf{R}_{n_{r1}}$). To compute the $\{\mathbf{R}_{s_{r1}},\mathbf{R}_{n_{r1}}\}$, a matrix approximation problem is formulated, specifying that $\mathbf{R}_{n_{r1}}$ should provide a good approximation to the given $\mathbf{R}_n$, while $(\mathbf{R}_{n_{r1}} + \mathbf{R}_{s_{r1}})$ should provide a good approximation to the given $\mathbf{R}_x$. In addition, "a priori knowledge" is incorporated, namely that $\mathbf{R}_{s_{r1}}$ should be a rank-1 matrix. The so obtained $\{\mathbf{R}_{s_{r1}}, \mathbf{R}_{n_{r1}}\}$ can then be used in the SDW-MWF formula (4.17). It is demonstrated in this section that this approach indeed leads to the EVD-SDW-MWF and EVD-SP-MWF, and so provides an alternative interpretation of these filters.
It is noted that the rank-1 condition for the autocorrelation matrix of the speech component is generalized to a rank-K condition in Section 4.6. The rank condition is then also seen to be a crucial ingredient, where in the extreme case of K = M (i.e., effectively no rank condition) the solution to the matrix approximation problem is merely $\{\mathbf{R}_x - \mathbf{R}_n, \mathbf{R}_n\}$, i.e., the autocorrelation matrices remain unchanged.
The $\{\mathbf{R}_{s_{r1}}, \mathbf{R}_{n_{r1}}\}$ should minimize the following criterion:

$$J_{\text{r1}} = \alpha\|\mathbf{R}_x - (\mathbf{R}_{n_{r1}} + \mathbf{R}_{s_{r1}})\|_F^2 + (1-\alpha)\|\mathbf{R}_n - \mathbf{R}_{n_{r1}}\|_F^2, \tag{4.46}$$

with $\|\cdot\|_F$ the Frobenius norm. Here, $\mathbf{R}_{n_{r1}}$ and $\mathbf{R}_{s_{r1}}$ are positive semi-definite matrices and $\mathbf{R}_{s_{r1}}$ is a rank-1 matrix. The two approximations may be given a different weight, i.e., α and $(1-\alpha)$, where α is a constant $(0 < \alpha < 1)$. In the case of estimated autocorrelation matrices, for instance, it may make sense to give a smaller weight to the approximation of the noise autocorrelation matrix (i.e. $\alpha > 0.5$), as this is estimated in older (hence possibly more outdated) "noise only" frames whenever a noise reduction is computed in a "speech plus noise" frame.
It is easy to check that when an optimal $\mathbf{R}_{s_{r1}}$ is given, the optimal solution for $\mathbf{R}_{n_{r1}}$ is:

$$\mathbf{R}_{n_{r1}} = \alpha\,(\mathbf{R}_x - \mathbf{R}_{s_{r1}}) + (1-\alpha)\mathbf{R}_n, \tag{4.47}$$

with the positive semi-definiteness of $\mathbf{R}_{n_{r1}}$ yet to be checked. As $\mathbf{R}_n$ is positive semi-definite by construction, it remains to check if $\mathbf{R}_x$ - $\mathbf{R}_{s_{r1}}$ is positive semi-definite (see below).

The $\mathbf{R}_{n_{r1}}$ can then be eliminated from the optimization problem by plugging (4.47) into (4.46). Therefore, after some simple manipulation, $\mathbf{R}_{s_{r1}}$ should minimize the following criterion:

$$J_{\text{sr}} = \alpha(1-\alpha)\|\mathbf{R}_x - \mathbf{R}_n - \mathbf{R}_{s_{r1}}\|_F^2. \tag{4.48}$$

The optimal solution is then known to be:

$$\mathbf{R}_{s_{r1}} = \mathbf{d}_{\max}\mathbf{d}_{\max}^H \max\left(\lambda_{\max}, 0\right), \tag{4.49}$$

as defined in (4.36) (assuming $\lambda_{\max}$ is non-negative). For this $\mathbf{R}_{s_{r1}}$, the matrix $\mathbf{R}_x$ - $\mathbf{R}_{s_{r1}}$ is indeed seen to be positive semi-definite, as required.
Once $\mathbf{R}_{s_{r1}}$ is defined according to (4.49), $\mathbf{R}_{n_{r1}}$ is computed based on (4.47). Two extreme cases can then be considered, as follows:

- If $\alpha \to 1$, which means that $\mathbf{R}_{n_{r1}} + \mathbf{R}_{s_{r1}}$ is to give the best possible approximation to $\mathbf{R}_x$ (first term in the original optimization function (4.46)), then $\mathbf{R}_{n_{r1}} = \mathbf{R}_x - \mathbf{R}_{s_{r1}} = \mathbf{R}_n + \mathbf{R}_z$ with $\mathbf{R}_z$ defined in (4.37). By replacing $\{\mathbf{R}_s, \mathbf{R}_n\}$ by this $\{\mathbf{R}_{s_{r1}}, \mathbf{R}_{n_{r1}}\}$ in formula (4.17), the EVD-SDW-MWF formula (4.42) is obtained.
- If $\alpha \to 0$, which means that $\mathbf{R}_{n_{r1}}$ is to give the best possible approximation to $\mathbf{R}_n$ (second term in the original optimization function (4.46)), then $\mathbf{R}_{n_{r1}} = \mathbf{R}_n$. By replacing $\{\mathbf{R}_s, \mathbf{R}_n\}$ by this $\{\mathbf{R}_{s_{r1}}, \mathbf{R}_{n_{r1}}\}$ in formula (4.17), the EVD-SP-MWF formula (4.45) is obtained.

4.5 GEVD based NR filters

A second alternative to the first column decomposition based rank-1 approximation is a rank-1 approximation based on the GEVD [Dendrinos et al., 1991, Doclo and Moonen, 2002, Jensen et al., 1995] of the matrix pencil $\{\mathbf{R}_x, \mathbf{R}_n\}$:

$$\begin{aligned} \mathbf{R}_n &= \mathbf{Q}\mathbf{\Sigma}_n\mathbf{Q}^H \\ \mathbf{R}_x &= \mathbf{Q}\mathbf{\Sigma}_x\mathbf{Q}^H \\ \Rightarrow \mathbf{R}_n^{-1}\mathbf{R}_x &= \mathbf{Q}^{-H}\left(\mathbf{\Sigma}_n^{-1}\mathbf{\Sigma}_x\right)\mathbf{Q}^H = \mathbf{Q}^{-H}\mathbf{\Sigma}\mathbf{Q}^H, \end{aligned} \tag{4.50}$$

where $\mathbf{Q}$ is an invertible matrix, the columns of which are normalized and define the generalized eigenvectors. $\mathbf{\Sigma}_x$, $\mathbf{\Sigma}_n$ and $\mathbf{\Sigma}$ are real-valued diagonal matrices with $\mathbf{\Sigma}_x = diag\{\sigma_{x_1} \dots \sigma_{x_M}\}$, $\mathbf{\Sigma}_n = diag\{\sigma_{n_1} \dots \sigma_{n_M}\}$ and $\mathbf{\Sigma} = diag\{\frac{\sigma_{x_1}}{\sigma_{n_1}} \dots \frac{\sigma_{x_M}}{\sigma_{n_M}}\}$ (with ordering $\frac{\sigma_{x_1}}{\sigma_{n_1}} \geq \frac{\sigma_{x_2}}{\sigma_{n_2}} \geq \dots \geq \frac{\sigma_{x_M}}{\sigma_{n_M}}\}$) defining the generalized eigenvalues.
The $\mathbf{R}_s$ is then obtained as:

$$\mathbf{R}_s = \mathbf{R}_x - \mathbf{R}_n = \mathbf{Q}(\underbrace{\mathbf{\Sigma}_x - \mathbf{\Sigma}_n}_{\mathbf{\Sigma}_s})\mathbf{Q}^H, \tag{4.51}$$

where $\text{SNR}_i = \frac{\sigma_{s_i}}{\sigma_{n_i}} = \frac{\sigma_{x_i}}{\sigma_{n_i}} - 1$ is the SNR in the i^{th} "mode".
The rank-1 approximation is then based on the decomposition:

$$\mathbf{R}_s = \underbrace{\mathbf{q}_1\mathbf{q}_1^H\sigma_{s_1}}_{\mathbf{R}_{s_1}} + \mathbf{R}_z, \tag{4.52}$$

where $\mathbf{q}_1$ is the first column of the matrix $\mathbf{Q}$, which corresponds to the highest SNR mode and $\mathbf{R}_z$ is again a remainder matrix. The decomposition can then be summarized as follows:

$$\mathbf{R}_s = \mathbf{Q}\,\mathrm{diag}\left\{\sigma_{s_1}, \sigma_{s_2}, \ldots, \sigma_{s_M}\right\}\mathbf{Q}^H \tag{4.53}$$

$$\mathbf{R}_{s_{r1}} = \mathbf{Q}\,\mathrm{diag}\left\{\sigma_{s_1}, 0, \ldots, 0\right\}\mathbf{Q}^H \tag{4.54}$$

$$\mathbf{R}_{\mathrm{rem}} = \mathbf{Q}\,\mathrm{diag}\left\{0, \sigma_{s_2}, \ldots, \sigma_{s_M}\right\}\mathbf{Q}^H. \tag{4.55}$$

When $rank(\mathbf{R}_s) = 1$, then $\mathbf{R}_z = 0$ and $\mathbf{R}_{s_{r1}}$ is the same as in the first column decomposition. When $rank(\mathbf{R}_s) > 1$, then the estimated rank-1 approximation $\mathbf{R}_{s_{r1}}$ is positive semi-definite if the estimated $\tilde{\sigma_{s_1}} = \tilde{\sigma_{x_1}} - \tilde{\sigma_{n_1}}$ is positive (which is again more likely than the first diagonal element $\tilde{\sigma}_{1,1}$ of the matrix $\mathbf{R}_s$ being positive as needed in the first column decomposition approach).
It is noted that:

$$\mathbf{R}_s\mathbf{t}_1 = \mathbf{R}_{s_{r1}}\mathbf{t}_1 + \underbrace{\mathbf{R}_{\mathrm{rem}}\mathbf{t}_1}_{=0} = \mathbf{R}_{s_{r1}}\mathbf{e}_1, \tag{4.56}$$

to be compared to (4.25), where

$$\mathbf{t}_1 = \mathbf{Q}^{-H}\mathbf{e}_1\mathbf{q}_1(1)^*, \tag{4.57}$$

with $\mathbf{q}_1(1)$ is the first element of $\mathbf{q}_1$.
An analysis similar to the analysis for the first column decomposition in Section 4.3 and the EVD based decomposition in Section 4.4 can then be done where $\mathbf{R}_s$ is replaced by the rank-1 approximation $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is either treated as noise or ignored. Equivalently, one can start from a modified MSE criterion where, compared to (4.11), the (arbitrary) $\mathbf{e}_1$ is replaced by $\mathbf{t}_1$:

$$J_{\mathrm{GEVD-SDW-MWF}} = \mathbb{E}\left\{\|\mathbf{w}^H\mathbf{X}_s - \mathbf{t}_1^H\mathbf{X}_s\|^2\right\} \tag{4.58}$$

$$+ \mu\mathbb{E}\left\{\|\mathbf{w}^H\mathbf{X}_n\|^2\right\}. \tag{4.59}$$

Replacing the desired signal $\mathbf{e}_1\mathbf{X}_s$ by $\mathbf{t}_1\mathbf{X}_s$ is indeed equivalent to replacing $\mathbf{R}_s$ by the GEVD based $\mathbf{R}_{s_r}$ as demonstrated by (4.56).

4.5.1 GEVD-SDW-MWF

The filter minimizing (4.58) is given as:

$$\mathbf{w}_{\mathrm{GEVD-SDW-MWF}} = \left(\mathbf{R}_s + \mu\mathbf{R}_n\right)^{-1}\mathbf{R}_s\mathbf{t}_1. \tag{4.60}$$

Plugging (4.52) and (4.56) into (4.60) leads to:

$$\mathbf{w}_{\mathrm{GEVD-SDW-MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\left(\mathbf{R}_n + \frac{1}{\mu}\mathbf{R}_z\right)\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.61}$$

To avoid the scaling with $\frac{1}{\mu}$, the same alternative derivation as for (4.28) can be applied leading to:

$$\mathbf{W}^{\circ}_{\text{GEVD-SDW-MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\left(\mathbf{R}_n + \mathbf{R}_z\right)\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.62}$$

This means that in the SDW-MWF (4.17) the desired signal vector $\mathbf{R}_s$ is replaced by the GEVD based $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is treated as noise.
Plugging (4.53),(4.51) and (4.55) into the GEVD-SDW-MWF formula (4.61) also leads to:

$$\mathbf{w}_{\text{GEVD-SDW-MWF}} = \mathbf{Q}^{-H}\left[\begin{array}{c|c} \frac{\frac{\sigma_{s_1}}{\sigma_{n_1}}}{\mu+\frac{\sigma_{s_1}}{\sigma_{n_1}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array}\right]\mathbf{Q}^{H}\mathbf{e_1}. \tag{4.63}$$

Note that (4.63) is still true if (4.61) is replaced by (4.62).
From (4.53) and (4.51) it appears that (4.63) can be reformulated as follows:

$$\mathbf{w}_{\text{GEVD-SDW-MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\mathbf{R}_n\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{l}_1. \tag{4.64}$$

By comparing (4.64) to (4.61) it is seen that the remainder matrix $\mathbf{R}_z$ actually has no influence on the GEVD-SDW-MWF (see also Section 4.5.3).

4.5.2 GEVD-SP-MWF

Based on the MSE criterion (4.58) it is also possible to derive the SP-MWF:

$$\mathbf{W}_{\text{GEVD-SP-MWF}} = \mathbf{R}_n^{-1}\mathbf{R}_s\mathbf{t}_1\frac{\mathbf{t}_1^H\mathbf{R}_s\mathbf{t}_1}{\mu\mathbf{t}_1^H\mathbf{R}_s\mathbf{t}_1 + \text{Tr}\left\{\mathbf{R}_n^{-1}\mathbf{R}_s\mathbf{t}_1\mathbf{f}_1^H\mathbf{R}_s\right\}}. \tag{4.65}$$

Plugging (4.52) into the GEVD-SP-MWF formula (4.65) leads to:

$$\mathbf{w}_{\text{GEVD-SP-MWF}} = \mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}\mathbf{t}}\mathbf{t}\frac{1}{\mu + \text{Tr}\left\{\mathbf{R}_n^{-1}\mathbf{R}_{s_{r1}}\right\}} \tag{4.66}$$

and

$$\mathbf{w}_{\text{GEVD-SP-MWF}} = \left(\mathbf{R}_{s_{r1}} + \mu\mathbf{R}_n\right)^{-1}\mathbf{R}_{s_{r1}}\mathbf{e}_1. \tag{4.67}$$

This means that in the SDW-MWF (4.17) $\mathbf{R}_s$ is replaced by the GEVD based $\mathbf{R}_{s_{r1}}$ and the remainder matrix $\mathbf{R}_z$ is simply ignored. From equations (4.64) and (4.67) it appears that the GEVD-SDW-MWF and the GEVD-SP-MWF are fully equivalent.

$$\mathbf{w}_{\text{GEVD-SDW-MWF}} = \mathbf{w}_{\text{GEVD-SP-MWF}}. \tag{4.68}$$

The good news here is that the question as to whether $\mathbf{R}_z$ should be either treated as noise (GEVDSDW-MWF) or ignored (GEVD-SP-MWF) becomes void, as the corresponding solutions are indeed the same.

4.5.3 A matrix approximation based derivation of GEVD-SDW-MWF and GEVD-SP-MWF

In matrix approximation problem (4.46), rather than using an unweighted Frobenius norm, where absolute (squared) approximation errors are summed, it may be more appropriate to consider relative approximation errors, where larger errors are tolerated in places where there is a lot of noise. This is standardly done by including a noise prewhitening operation. From the GEVD (4.53) it follows that:

$$\mathbf{R}_n = \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)\left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^H. \tag{4.69}$$

The noise prewhitening is then done by premultiplying each vector with $\left(\mathbf{Q}\Sigma_n^{1/2}\right)^{-1}$. Each autocorrelation matrix is premultiplied with $\left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-1}$ and postmultiplied with $\left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-H}$ (so that for instance $\mathbf{R}_n$ is prewhitened into $\mathbf{I}$).
The criterion (4.46) is then replaced by:

$$\begin{aligned} J_{\text{pw-r1}} =& \alpha \left\| \left(\mathbf{Q}\Sigma_n^{\frac{1}{2}}\right)^{-1} \left[\mathbf{R}_x - (\mathbf{R}_{n_{r1}} + \mathbf{R}_{s_{r1}})\right] \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-H} \right\|_F^2 \\ &+ (1-\alpha) \left\| \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-1} \left[\mathbf{R}_n - \mathbf{R}_{n_{r1}}\right] \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-H} \right\|_F^2, \end{aligned} \tag{4.70}$$

where now $\mathbf{R}_{s_{r1}}$ and $\mathbf{R}_{n_{r1}}$ are sought such that after the prewhitening the Frobenius norms are minimal. It can be verified that the prewhitening does not change (4.47), i.e., when an optimal $\mathbf{R}_{s_{r1}}$ is given, the optimal solution for $\mathbf{R}_{n_{r1}}$ is still given by (4.47).
The $\mathbf{R}_{n_{r1}}$ can then again be eliminated from the optimization problem by plugging (4.47) into (4.70). Therefore, after some simple manipulation, $\mathbf{R}_{s_{r1}}$ should minimize the following:

$$J_{\text{s}_{\text{pw-r1}}} = \alpha(1-\alpha) \| \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-1} \left[\mathbf{R}_x - \mathbf{R}_n - \mathbf{R}_{s_{r1}}\right] \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-H} \|_F^2 \tag{4.71}$$

$$= \alpha(1-\alpha) \| \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\Sigma}_s - \mathbf{I} - \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-1} \mathbf{R}_{s_{r1}} \left(\mathbf{Q}\boldsymbol{\Sigma}_n^{\frac{1}{2}}\right)^{-H} \|_F^2. \tag{4.72}$$

The optimal solution is then shown to be:

$$\mathbf{R}_{s_{r1}} = \mathbf{q}_{\max}\mathbf{q}_{\max}^H \max\left(\sigma_{x_1} - \sigma_{n_1}, 0\right) \tag{4.73}$$

$$= \mathbf{Q}\,\text{diag}\{\max\left(\sigma_{x_1} - \sigma_{n_1}, 0\right), 0, \ldots, 0\}. \tag{4.74}$$

Assuming $\sigma_{x_1} - \sigma_{n_1}$ is non-negative, this corresponds to (4.56). Once $\mathbf{R}_{s_{r1}}$ is defined according to (4.73), $\mathbf{R}_{n_{r1}}$ is computed based on (4.47), leading to:

$$\mathbf{R}_{n_{r1}} = \mathbf{Q}\,\text{diag}\left\{\sigma_{n_1}, \alpha\sigma_{x_2} + (1-\alpha)\sigma_{n_2}, \ldots, \alpha\sigma_{x_M} + (1-\alpha)\sigma_{n_M}\right\}\mathbf{Q}^H. \tag{4.75}$$

Again, two extreme cases can be considered, as follows:

- If $\alpha \to 1$ then $\mathbf{R}_{n_{r1}} = \mathbf{R}_x - \mathbf{R}_{s_{r1}} = \mathbf{R}_n + \mathbf{R}_z$ with $\mathbf{R}_z$ defined in (4.52). By replacing $\{\mathbf{R}_s, \mathbf{R}_n\}$ by this $\{\mathbf{R}_{s_{r1}}, \mathbf{R}_{n_{r1}}\}$ in formula (4.17), the EVD-SDW-MWF formula (4.64) is obtained.
- If $\alpha \to 0$ then $\mathbf{R}_{n_{r1}} = \mathbf{R}_n$. By replacing $\{\mathbf{R}_s, \mathbf{R}_n\}$ by this $\{\mathbf{R}_{s_{r1}}, \mathbf{R}_{n_{r1}}\}$ in formula (4.17), the EVDSP-MWF formula (4.67) is obtained.

It is reiterated that the GEVD-SDW-MWF and GEVD-SP-MWF are found to be fully equivalent (formula (4.68)), so that in this case, the selection of a good α, remarkably, becomes irrelevant.

4.6 Rank-R Approximation GEVD Based NR Filters

The GEVD based rank-1 approximation in (4.52) can be seen as an extreme case of a more general rank-R approximation, which then leads to more general rank-R approximation based NR filters.
Plugging (4.53) and (4.51) into the SDW-MWF formula (4.17) leads to:

$$\mathbf{w}_{\text{SDW-MWF}} = \mathbf{Q}^{-H} \left(\text{diag} \left\{ \frac{\frac{\sigma_{s_i}}{\sigma_{n_i}}}{\mu + \frac{\sigma_{s_i}}{\sigma_{n_i}}} \right\} \right) \mathbf{Q}^H \mathbf{e}_1. \tag{4.76}$$

Considering the gains in the diagonal matrix in (4.76):

$$1 \geq \frac{\frac{\sigma_{s_1}}{\sigma_{n_1}}}{\mu + \frac{\sigma_{s_1}}{\sigma_{n_1}}} \geq \frac{\frac{\sigma_{s_2}}{\sigma_{n_2}}}{\mu + \frac{\sigma_{s_2}}{\sigma_{n_2}}} \geq \cdots \geq \frac{\frac{\sigma_{s_M}}{\sigma_{n_M}}}{\mu + \frac{\sigma_{s_M}}{\sigma_{n_M}}} \geq 0. \tag{4.77}$$

It has been demonstrated that cochlear implant recipients can tolerate a much higher speech distortion than normal hearing subjects. This means that the noise reduction can be tuned to be more aggressive, which in the SDW-MWF corresponds to increasing the trade-off parameter μ. Following (4.77), by increasing μ, a relatively larger weight is given to the modes with the highest SNR. The modes with the lowest SNR are eventually set to 0.
This can also be pursued more explicitly by setting the $M - R$ components with the lowest SNR to 0, leading to a rank-R approximation based NR filter:

$$\mathbf{w}_{\text{GEVD-R}} = \mathbf{Q}^{-H} \left[\begin{array}{ccc|c} \frac{\frac{\sigma_{s_1}}{\sigma_{n_1}}}{\mu + \frac{\sigma_{s_1}}{\sigma_{n_1}}} & & 0 & \mathbf{0} \\ & \ddots & & \\ 0 & & \frac{\frac{\sigma_{s_R}}{\sigma_{n_R}}}{\mu + \frac{\sigma_{s_R}}{\sigma_{n_1}}} & \mathbf{0} \\ \hline \mathbf{0} & & \mathbf{0} & \mathbf{0} \end{array} \right] \mathbf{Q}^H \mathbf{e_1}. \tag{4.78}$$

This NR filter is then equivalent to (4.76) where the trade-off parameter μ is mode-dependent. In the $M - R$ modes with the lowest SNR the trade-off parameter $\mu = \infty$

whereas in the R modes with the highest SNR μ is set to a real value. This approach then again corresponds to tuning the SDW-MWF to perform a more agressive NR, which indeed makes sense for cochlear implant recipients. Note that for $R = 1$ (rank-1 approximation case) only the mode with the highest SNR is not set to 0 and then (4.78) reduces to (4.63), i.e.,

$$\mathbf{w}_{\text{GEVD}-1} = \mathbf{w}_{\text{GEVD}-\text{SDW}-\text{MWF}} = \mathbf{w}_{\text{GEVD}-\text{SP}-\text{MWF}}. \tag{4.79}$$

For $R = M$, none of the modes is set to 0 and hence

$$\mathbf{w}_{\text{GEVD}-\text{M}} = \mathbf{w}_{\text{SDW}-\text{MWF}}. \tag{4.80}$$

The rank-R approximation is effectively based on the decomposition:

$$\mathbf{R}_s = \mathbf{R}_{s_{rR}} + \mathbf{R}_z, \tag{4.81}$$

where $\mathbf{R}_{s_{rR}}$ is a rank-R approximation of $\mathbf{R}_s$. For $R > 1$ the matrices can be expressed as follows:

$$\mathbf{R}_s = \mathbf{Q}\,\text{diag}\left\{\sigma_{s_1}, \sigma_{s_2}, \ldots, \sigma_{s_M}\right\}\mathbf{Q}^H \tag{4.82}$$

$$\mathbf{R}_{s_{rR}} = \mathbf{Q}\,\text{diag}\left\{\sigma_{s_1}, \ldots, \sigma_{s_R}, 0, \ldots, 0\right\}\mathbf{Q}^H \tag{4.83}$$

$$\mathbf{R}_{\text{rem}} = \mathbf{Q}\,\text{diag}\left\{0, \ldots, 0, \sigma_{s_{(R+1)}}, \ldots, \sigma_{s_M}\right\}\mathbf{Q}^H. \tag{4.84}$$

The rank-R approximation can be further decomposed into a sum of rank-1 terms:

$$\mathbf{R}_{s_{rR}} = \sum_{i=1}^{R} \underbrace{\mathbf{q}_i \mathbf{q}_i^H \sigma_{s_i}}_{\mathbf{R}_{s_{Ri}}}, \tag{4.85}$$

where $\mathbf{q}_i$ i^{th} column of the matrix $\mathbf{Q}$, which corresponds to the i^{th} mode, leading to:

$$\mathbf{R}_{s_{Ri}} = \mathbf{Q}\,\text{diag}\left\{0, \ldots, \sigma_{s_i}, \ldots, 0\right\}\mathbf{Q}^H. \tag{4.86}$$

It is then noted that:

$$\mathbf{R}_s \sum_{i=1}^{R} \mathbf{t}_i = \sum_{i=1}^{R} \left(\mathbf{R}_{s_{ri}} \mathbf{t}_i + \left(\mathbf{R}_z + \sum_{\substack{j=1 \\ j \neq i}}^{R} \mathbf{R}_{s_{ri}} \right) \mathbf{t}_i \right) = \mathbf{R}_{s_{rR}} \mathbf{e}_1, \tag{4.87}$$

to be compared to (4.25), where

$$\mathbf{t}_i = \mathbf{Q}^{-H} \mathbf{e}_i \mathbf{q}_i(1)^*, \tag{4.88}$$

with $\mathbf{q}_i(1)$ is the first element of $\mathbf{q}_i$ and $\mathbf{e}_i$ an all-zero vector except for a one in the i^{th} position.

An analysis similar to the analysis for the first column decomposition in Section 4.3, the EVD based decomposition in Section 4.4 and the GEVD based decomposition in

Section 4.5 can then be done where $\mathbf{R}_s$ is replaced by the rank-R approximation $\mathbf{R}_{s_{rR}}$ and the remainder matrix $\mathbf{R}_{\text{rem}}$ is either treated as noise or ignored. Equivalently, one can start from a modified MSE criterion where, compared to (4.11), the (arbitrary) $\mathbf{e}_1$ is replaced by $\mathbf{t}_{rR}$:

$$J_{\text{GEVD-R}} = \mathbb{E}\left\{\|\mathbf{W}^H\mathbf{X}^s - \mathbf{t}_{rR}^H\mathbf{X}^s\|^2\right\} + \mu\mathbb{E}\left\{\|\mathbf{W}^H\mathbf{X}^n\|^2\right\}, \tag{4.89}$$

where:

$$\mathbf{t}_{rR} = \sum_{i=1}^{R}\mathbf{t}_i. \tag{4.90}$$

Replacing the desired signal $\mathbf{e}_1\mathbf{X}_s$ by $\mathbf{t}_{rR}\mathbf{X}_s$ is indeed equivalent to replacing $\mathbf{R}_s$ by the GEVD based $\mathbf{R}_{s_{rR}}$ as demonstrated by (4.87). Note that $\mathbf{t}_{rM} = \mathbf{e}_1$ and so

$$J_{\text{GEVD-M}} = J_{\text{SDW-MWF}}, \tag{4.91}$$

again leading to (4.80).
As in the rank-1 approximation case (Section 4.5), it can easily be shown that the $\mathbf{R}_z$ can be either treated as noise or ignored as the corresponding NR filters are both equal to $\mathbf{w}_{\text{GEVD-R}}$ as given in (4.78).

4.7 Experimental Results

4.7.1 Experimental setup

The simulations were run on acoustic path measurements obtained in a reverberant room (RT_{60} = 0.61 s [Van den Bogaert et al., 2008, 2009]) with a CORTEX MK2 manikin equipped with two Cochlear SP15 behind-the-ear devices. Each device has two omnidirectional microphones. The manikin head is used so that the head shadow effects are taken into account. The sound sources (FOSTEX 6301B loudspeakers) were positioned at 1 meter from the center of the head directed towards the artificial head. The system was calibrated with a microphone placed at the position of the center of the head. The input SNR is then the SNR at the center of the head.
In each experiment, the speech signal was composed of five consecutive sentences from the English Hearing-In-Noise Test (HINT) database [Nilsson et al., 1994] concatenated with five second silence periods between each sentence. The noise was the multitalker babble signal from Auditec. Three spatial scenarios were considered, two single noise source scenarios (S0N45 and S90N270) and one scenario with multiple noise sources (S0N90-180-270) where the speech source (S) and the noise source(s) (N) are located at a specified angle. Not that 90°corresponds to a source facing the right ear while 270°corresponds to a source facing the left ear. When multiple noise sources are present, different time shifted versions of the multitalker babble signal were used to ensure uncorrelated noise sources. In each scenario, signals with input SNR varying from -15 dB to 5 dB are presented to the left and right devices. The microphone signals are then filtered by several NR algorithms and the performance is compared.

All the signals were sampled at 20480Hz. The filter lengths and DFT size were set to $N = 128$ and the frame overlap was set to half of the DFT size ($L = 64$). When mentioned, the so-called input SNR is the SNR at the center of the head (excluding the HRTF effects).

4.7.2 Performance measures

An intelligibility weighted speech distortion (SIW-SD) measure is used defined as

$$\text{SIW} - \text{SD} = \sum_i I_i \text{SD}_i, \tag{4.92}$$

where I_i is the band importance function defined in [29] and SD_i the average SD (in dB) in the i-th one third octave band,

$$\text{SD}_i = \frac{1}{\left(2^{1/6} - 2^{-1/6}\right) f_i^c} \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} \|10 \log_{10} G^s(f)\| \text{f}, \tag{4.93}$$

with center frequencies f_i^c and $G^s(f)$ is given by:

$$G^s(f) = \frac{P_{X_s}(f)}{P_{Z_s}(f)}, \tag{4.94}$$

where $P_{X_s}(f)$ and $P_{Z_s}(f)$ are the power, for the frequency f, of the speech component of the input signal X_s and the speech component of the signal processed by one of the approaches described above Z_s, respectively.
The speech intelligibility-weighted SNR (SIW-SNR) [Greenberg et al., 1993] is used here to compute the SIW-SNR improvement which is defined as

$$\Delta\text{SIW} - \text{SNR} = \sum_i I_i \left(\text{SNR}_{i,\text{out}} - \text{SNR}_{i,\text{front}}\right), \tag{4.95}$$

where $\text{SNR}_{i,\text{out}}$ and $\text{SNR}_{i,\text{front}}$ represent the output SNR (at the considered ear) of the NR filter and the SNR of the signal in the front microphone (at the considered ear) of the i^{th} band, respectively.
The percentage of estimated $\mathbf{R}_{s_{r1}}$'s that are not positive semi-definite (%NPD) is defined as follows.

$$\%\text{NPD} = \frac{NPD}{N_{\text{Mat}}} * 100 \tag{4.96}$$

where NPD is the number of estimated $\mathbf{R}_{s_{r1}}$ that are not positive semi-definite and N_{Mat} is the total number of estimated $\mathbf{R}_{s_{r1}}$.

4.7.3 Algorithms tested

For each of the three spatial scenarios the performance of the SDW-MWF and the GEVD-MWF are compared. For each algorithm (SDW-MWF and GEVD-SDW-MWF), three cases are then considered: a bilateral (2+0) system, a binaural (2+1) system where only the signal from the front microphone of the contralateral device is used (this is referred to as "front") and a binaural (2+2) system where both microphone signals from the contra-lateral device are used (this is referred to as "binaural").

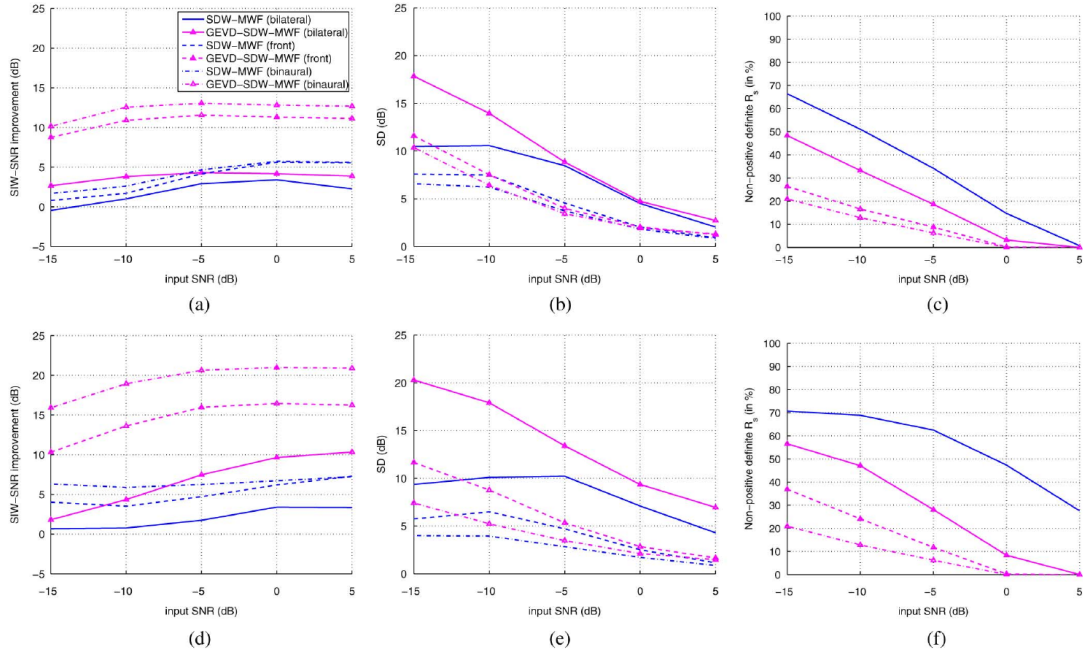


Figure 4.1: Performance for the S0N45 scenario, comparison between SDW-MWF and GEVD-SDW-MWF. (a) SIW-SNR performance at the left ear. (b) SIW-SD performance at the left ear. (c) %NPD for the left ear. (d) SIW-SNR performance at the right ear. (e) SIW-SD performance at the right ear. (f) %NPD for the right ear

4.7.4 Speech source at 0°, single noise source at 45°(S0N45)

In the first spatial scenario, the speech source is located at 0°and the noise source at 45°. The aim of this scenario is to investigate to which extent an MWF-based NR can benefit from the the GEVD based approach when the speech source and the noise source are closely spaced.

Figure 4.1(c) presents the %NPD, at the left ear, as a function of the input SNR for bilateral, front and binaural SDW-MWF and GEVD-SDW-MWF. For the SDW-MWF's the first diagonal element of $\mathbf{R}_{s_{r1}}$, i.e., $[\mathbf{R}_{s_{r1}}]_{1,1}$ is the same in all three cases, therefore, bilateral, front and binaural return the same %NPD that can be as high as 65% at -15 dB input SNR. For the GEVD-SDW-MWF's on the other hand, the positive semi-definiteness of $\mathbf{R}_{s_{r1}}$ depends on and each additional channel can help to improve this. Therefore, whereas the bilateral GEVD-SDW-MWF already decreases the %NPD and the binaural GEVD-SDW-MWF allow to further decrease the %NPD to 20%.

Figures 4.1(a) and 4.1(b) present the SIW-SNR improvement and the SIW-SD introduced at the left ear, respectively. At low input SNR, the bilateral approahes barely give any SIW-SNR improvement while it is still introducing about 10 dB SD. The binaural approaches allows for improving the SIW-SNR while introducing lower SIW-SD than the bilateral appraohes. It is important to notice that in this scenario, the left ear is the so-

called best ear (i.e., the ear with the highest input SNR) and an improvement of the SIW-SNR of a few dB at the best ear can already improve comfort and speech understanding tremendously. The GEVD-SDW-MWF provides an SIW-SNR improvement that is higher than the improvement for the SDW-MWF but at the cost of a higher SD.
Figure 4.1(f) presents the %NPD, at the right ear, as a function of the input SNR for bilateral, front and binaural SDW-MWF and GEVD-SDW-MWF. The SDW-MWF returns a %NPD that can be as high as 70% at -15 dB input SNR whereas the GEVD-SDW-MWF can decrease this percentage down to about 20%. In this scenario, as the right ear in the worst ear, the $\mathbf{R}_{s_{r1}}$ can benefit from the higher SNR of the signal from the contra-lateral device. This is especially the case for the binaural GEVD-SDW-MWF that is delivering a %NPD as low as 20% at -15 dB SNR, which is the same figure as for the best ear (see also Figure 4.1(c)).
Figures 4.1(d) and 4.1(e) present the SIW-SNR improvement and the SIW-SD introduced at the right ear, respectively. At low input SNR, the binaural GEVD-SDW-MWF provides an SIW-SNR improvement that is higher than the improvement for the corresponding SDW-MWF, at a cost of a higher SIW-SD. For input SNR higher than -10 dB, however, the GEVD-SDW-MWF and the SDW-MWF are introducing a similar SD. In this scenario, as the right ear in the worst ear, the NR can benefit from the higher SNR of the signals from the contra-lateral device which is especially the case for the GEVD-SDW-MWF.
The next two experiments support the claim that the GEVD-SDW-MWF allows to increase the SIW-SNR while introducing only a controlled SD (Figure 4.2). In the first experiment, the trade-off parameter μ in the GEVD-SDW-MWF is set such that the same amount of SIW-SD is introduced as with the corresponding SDW-MWF with a $\mu = 1$. Figures 4.2(a) and 4.2(b) present the SIW-SNR improvement at the left and right ear, respectively, for the SDW-MWF and the GEVD-SDW-MWF. In all cases the SIW-SNR performance of the GEVD-SDW-MWF improves the SIW-SNR by up to 10 dB compared to the SDW-MWF.
In the second experiment, the trade-off parameter μ in the SDW-MWF is set such that the SDW-MWF delivers the same SIW-SNR improvement as the corresponding GEVD-SDW-MWF with $\mu = 1$. Figures 4.2(c) and 4.2(d) present the SIW-SD introduced by the SDW-MWF and the GEVD-SDW-MWF at the left and right ear, respectively. In order to deliver a similar SIW-SNR performance, the SDW-MWF has to introduce 5 dB to 10 dB more SIW-SD than the corresponding GEVD-SDW-MWF at low input SNR.

4.7.5 Speech source at 90°, single noise source at 270°(S90N270)

In the second spatial scenario, the speech source is located at 90°(facing the left ear) and the noise source at 270°(facing the right ear). The aim of this scenario is to investigate to which extent the GEVD based NR can improve the SIW-SNR performance at the best ear and the worst ear (Figure 4.3).
Figure 4.3(c) presents the %NPD, at the left ear, for the SDW-MWF and GEVD-SDW-MWF. The GEVD-SDW-MWF can decrease the %NPD to about 20% in the binaural case. In this scenario, as the left ear in the worst ear, the $\mathbf{R}_{s_{r1}}$ can benefit from the

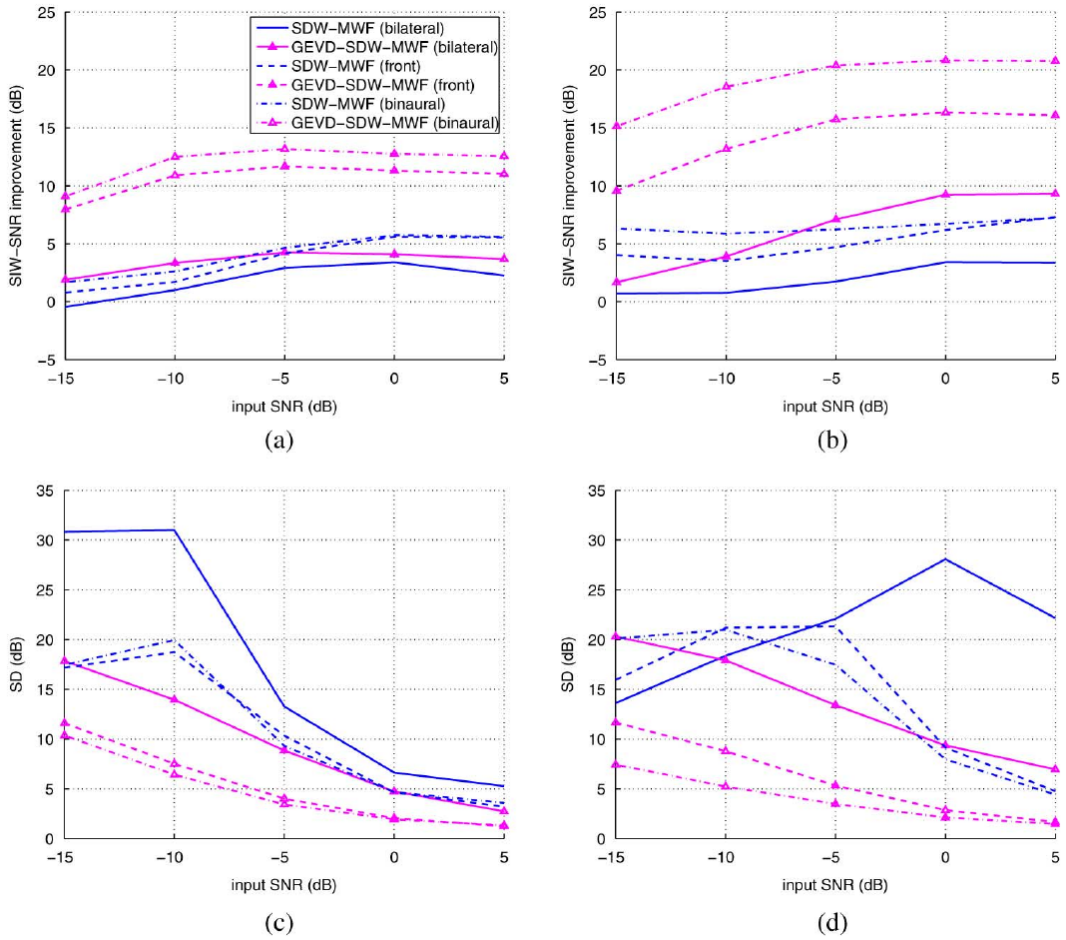


Figure 4.2: Performance for the S0N45 scenario, comparison between SDW-MWF and GEVD-SDW-MWF with equal SIW-SD (a) and (b) and with equal SIW-SNR improvement (c) and (d). (a) SIW-SNR performance at the left ear. (b) SIW-SNR performance at the right ear. (c) SIW-SD performance at the left ear. (d) SIW-SD performance at the right ear.

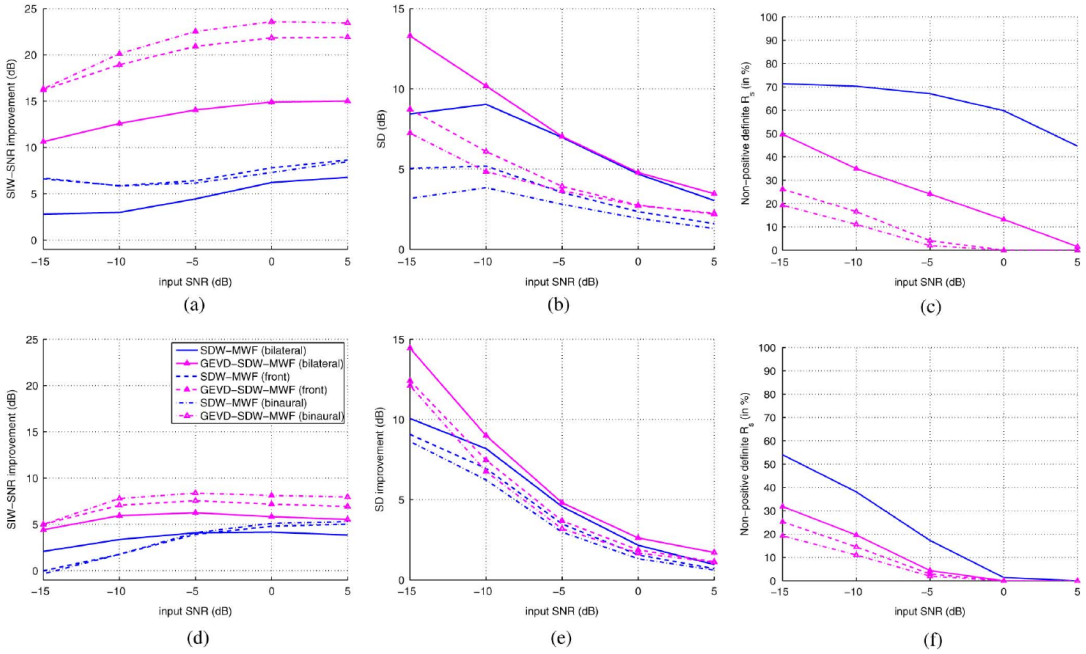


Figure 4.3: Performance for the S90N270 scenario, comparison between SDW-MWF and GEVD-SDW-MWF. (a) SIW-SNR performance at the left ear. (b) SIW-SD performance at the left ear. (c) %NPD for the left ear. (d) SIW-SNR performance at the right ear. (e) SIW-SD performance at the right ear. (f) %NPD for the right ear

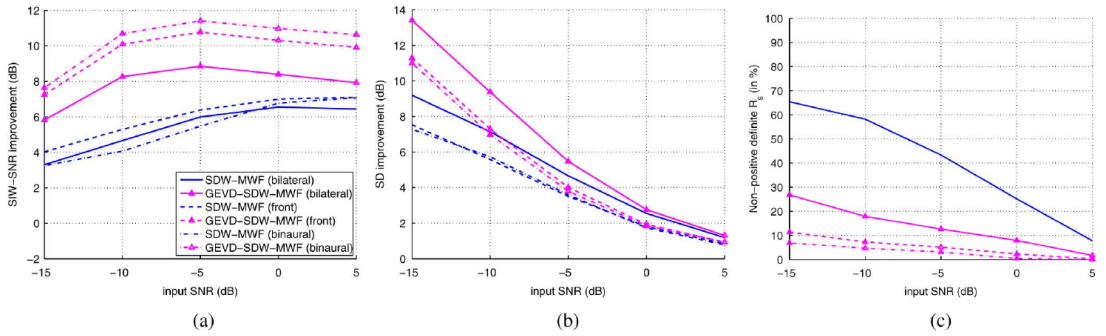


Figure 4.4: Performance for the S0N90-180-270 scenario, comparison between SDW-MWF and GEVD-SDW-MWF. (a) SIW-SNR performance at the right ear. (b) SIW-SD performance at the right ear. (c) %NPD for the right ear.

higher SNR of the signals from the contra-lateral device. This is especially the case for the binaural approaches that are delivering the same %NPD as the for the best ear (see also Figure 4.3(f)).
For both ears, the binaural SDW-MWF allow to improve the SIW-SNR while introducing lower SIW-SD than the bilateral SDW-MWF. The GEVD-SDW-MWF provides an SIW-SNR improvement higher than the improvement for the corresponding SDW-MWF, at the cost of higher SIW-SD at low input SNR.
Figures 4.3(d) and 4.3(e) present the SIW-SNR improvement and the SIW-SD introduced at the right ear (considered as the best ear in this spatial scenario). At low input SNR, the front and the binaural SDW-MWF suffer from the low input SNR of the signals from the contra-lateral device and deliver a lower SIW-SNR than the bilateral SDW-MWF. The GEVD-SDW-MWF delivers an SIW-SNR higher than the improvement for the corresponding SDW-MWF. It is important to note that, at low input SNR, the front and the binaural GEVD-SDW-MWF still deliver a better SIW-SNR improvement than the bilateral GEVD-SDW-MWF and are therefore less affected by the low SNR of the signals from the contra-lateral device than the corresponding SDW-MWF.

4.7.6 Speech source at 0°, multiple noise sources (S0N90-180-270)

In the third spatial scenario, the speech source is located at 0°and the uncorrelated noise sources at 90°, 180°and 270°. The aim of this scenario is to investigate to which extent the GEVD based approach can improve the robustness in multiple noise sources scenarios. The scenario is spatially symmetrical so the NR should perform similarly in both ears. Therefore, only the results for the right ear are presented here (Figure 4.4).
Figure 4.4(c) presents the %NPD, as a function of the input SNR for the SDW-MWF and the GEVDSDW-MWF. The SDW-MWF returns a %NPD that can be up to 65% at -15 dB SNR. The GEVD-SDW-MWF can decrease this percentage to less than 10%.
Figures 4.4(a) and 4.4(b) present the SIW-SNR improvement and the SIW-SD introduced at the right ear, respectively. As the scenario is symmetrical, the input SNR is similar

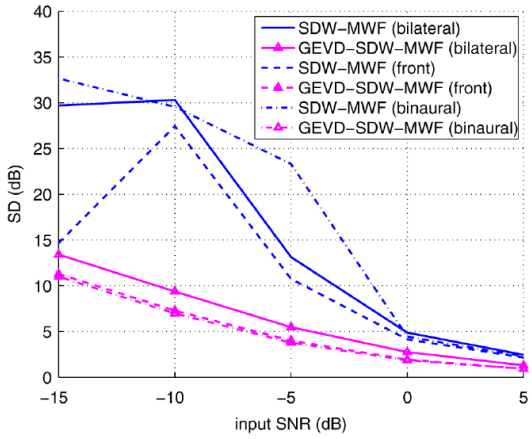


Figure 4.5: SIW-SD performance at the right ear, comparison between SDW-MWF and GEVD-SDW-MWF with equal SIW-SNR improvement.

at both ears and the front and the binaural NR cannot benefit from the higher input SNR at the contra-lateral device. There is no clear benefit either from the increased number of channels in the case of the SDW-MWF. The GEVD-SDW-MWF provides an SIW-SNR improvement 3 dB to 4 dB higher than the improvement for the corresponding SDW-MWF.
When the trade-off parameter μ is set so that the GEVD-SDW-MWF introduces the same amount of SIW-SD as the corresponding SDW-MWF (with $\mu = 1$), the GEVD-SDW-MWF performance is just slightly reduced and still better than the SDW-MWF performance. Figure 4.5 presents the SIW-SD introduced by the SDW-MWF and the GEVD-SDW-MWF at the right ear when the trade-off parameter μ is set so that the SDW-MWF delivers the same SIW-SNR as the corresponding GEVD-SDW-MWF (with $\mu = 1$). The SDW-MWF then introduces up to 20 dB more SIW-SD than the GEVD-SDW-MWF.

4.8 Conclusions

In this chapter we first analyze the difference between the SDW-MWF, the R1-MWF and the SP-MWF (which are equivalent when the autocorrelation matrix of the speech signal is a rank-1 matrix) when the rank of autocorrelation matrix of the speech signal is effectively greater than one. In this case, it is possible to decompose the autocorrelation matrix of the speech signal into the sum of a rank-1 approximation and a remainder matrix. The SDW-MWF, the R1-MWF and the SP-MWF then differ in the way this remainder matrix is treated.
At low input SNR, due to noise non-stationarity, the estimated autocorrelation matrix of the speech signal may not be positive semi-definite. To tackle this problem, an EVD based rank-1 approximation approach to SDW-MWF and to SP-MWF has been introduced. It

is then again possible to decompose the autocorrelation matrix of the speech signal into the sum of a rank-1 approximation and a remainder matrix and the difference between the EVD based SDW-MWF and SP-MWF again depends in the way the remainder matrix is treated. It has been demonstrated that the EVD-SDW-MWF provides an improved SIW-NSR performance.
A GEVD based rank-1 approximation approach to SDW-MWF and to SP-MWF has finally been proposed. The rank-1 approximation based SDW-MWF and SP-MWF have then been shown to be fully equivalent even when the rank of the autocorrelation matrix of the speech signal is greater than one. As it effectively selects the mode with the highest SNR this approach has been shown to allow for a more reliable estimation of the autocorrelation matrix of the speech signal than both the original SDW-MWF an SP-MWF approaches and the EVD based approaches, fully taking advantage of the high input SNR at best ear in the case of a binaural system.
The GEVD-SDW-MWF has been shown to deliver a better SIW-SNR than the corresponding SDW-MWF while introducing the same SD. Similarly, it has been shown that if the SDW-MWF was to be set to deliver similar SIW-SNR as the corresponding GEVD-SDW-MWF, it introduces a large amount of SD. Finally, the rank-1 approximation based GEVD-SDW-MWF has been generalized to a rank-R approximation based approach (GEVD-R), which encompasses the GEVD-SDW-MWF (GEVD-1) and the SDW-MWF (GEVD-M) as extreme cases.
In this chapter a perfect VAD is used and the benefits of the presented algorithms might be limited by the need of a VAD at SNR ranging from -15 dB to 5 dB. In recent work we have done since my arrival at Université de Lorraine, we replace the perfect VAD used here by time-frequency masks estimated with a DNN or to estimate directly the individual signals spectra (see also Section 4.9).
Additionally the type of noises dealt with is limited here but it can have an impact on the noise reduction performance, especially so in data-driven approaches if the acoustic environment faced at test time differs severely from the environments seen at training. This is one of the motivations for exploring algorithms that can automatically characterize acoustic scenes in details.

4.9 Other related works

Since my arrival in Nancy, I have collaborated with Ziteng Wang, a visiting PhD student from the University of Chinese Academy of Sciences (Beijing, China), supervised by Emmanuel Vincent during his stay. Ziteng Wang proposed a benchmark of several multichannel noise reduction techniques for far-field speech recognition [Wang et al., 2018b]. Each filter was relying on the same DNN-based time frequency mask estimation to compute the filters parameters but the multichannel filtering approach itself changed. The study showed that a filter that would remove more noise (but at the cost of higher speech distortions) can actually have a beneficial impact on the back-end ASR performance. This is the case with the proposed version of the GEVD-MWF tuned to output a constant noise power in each frequencies.

From October 2016 to October 2019, I was co-director of Laureline Perotin's PhD together with Emmanuel Vincent and Alexandre Guerin (Scientist researcher at Orange, Rennes, France). Laureline Perotin's work during her PhD first focused on exploring DNN-based multichannel noise reduction from high-order ambisonics recordings [Perotin et al., 2018b]. The proposed approach was based on a delay-and-sum ambisonic beamformer followed by a mask estimation with a long-short term memory network (LSTM) and a GEVD-MWF. The proposed approach was evaluated in terms in word error rate with a fixed automatic speech recognition back-end. The study showed that the signals estimated by the ambisonic beamformer can greatly improve the mask estimation and the multichannel filtering performance. The approach was shown to match performance obtained with filter estimated from oracle masks in scenarios with two concurrent speakers. The second part of Laureline Perotin's thesis was focused on sound source localization from high-order ambisonic recordings [Perotin et al., 2019a, 2018a, 2019b]. The approache based on convolutionnal recurrent neural networks applied on the active and reactive intensity of the first order ambisonic signals was shown to generalized to unseen recorded signals even if it was trained on simulated signals. An analysis of the role played by specific time-frequency points of the input with layerwise relevance propagation [Montavon et al., 2018] was proposed. When only one sound source was present, the network relies mainly on sound onsets for localization, similarly to what was observed with humans [Litovsky et al., 1999]. Finally an analysis of the importance of the type of cost function of the network (classification vs. regression) was proposed. When using a classification cost, two alternatives to the one-hot encoding were proposed. They were based on a Gibbs distribution derived from the squared angular distance between the predicted position and the ground truth. The corresponding probabilities were used either as soft targets or as weight to the cross-entropy in order to take the angular distance into account during the network training.

From October 2018 to December 2021, I have been supervising Nicolas Furnon's PhD. The PhD was co-directed by Irina Ilina (Univeristé de Lorraine, Nancy, France) and Slim Essid (Télécom ParisTech, Paris, France) and took place within the framework of the ANR JCJC (young researcher grant) DiSCogs that I obtained in 2018. The main topic of this project was far-field speech enhancement with ad-hoc microphone arrays. Nicolas Furnon proposed a multi-node DNN based mask estimation integrated within the distributed node-specific MWF proposed by Bertrand and Moonen [2010]. Each node operates a first local filtering step. The signals obtained during this step are then exchanged between nodes. These local signals are used by the mask estimation networks and as additional channels to compute the MWF filter used during a second filtering step. The approach proposed allows for efficiently exploiting the diversity of the information provided by each node during the mask estimation [Furnon et al., 2021b, 2020]. The proposed approach has been shown to perform on par with pure neural network approaches such as FaSNet [Luo et al., 2019] while being less demanding computationally and offering more flexibility thanks to the MWF framework. Extensions of the algorithm have been proposed using attention to enforce robustness to missing nodes [Furnon et al., 2021a] or with application to speech separation in a meeting setup [Furnon et al., 2021c].

In October 2020, Louis Delebecque joined the DiSCogs project as a research engineer. He is working together with Nicolas Furnon on validating the algorithms developed during Nicolas Furnon's PhD on real scenarios. This work involve two complementary steps. Louis is recording audio data with hearing aids simulators in order to validate the algorithms on a binaural hearing aid setup (where each hearing aid is considered as a node). The second step is the analysis of the computational needs of the algorithm that led to the conclusion that the network part is more demanding than the MWF part at run-time. The work then focused on reducing the network computational needs while preserving the overall performance of the filtering algorithm. This work led to a reduction by a factor 5 in computational needs while maintaining the performance of the original system.
From March 2017 to April 2020, I was co-director of Guillaume Carbajal's PhD together with Emmanuel Vincent and Eric Humbert (Invoxia SAS, Paris, France). During his PhD, Guillaume Carbajal proposed a DNN based solution for jointly compensating for noise, echo and reverberation. In a first stage, an echo reduction approach was propose that relied on a DNN possibly exploiting different inputs [Carbajal et al., 2018]. The use of the far-end signal was shown to be particularly beneficial to the network. It was also shown that using a phase sensitive mask [Erdogan et al., 2015] as training criterion could help improving the performance further. This approach was then generalized to a context were the algorithm also has to compensate for additive noise and reverberation [Carbajal et al., 2019, 2020]. The solution proposed is a sequence of filters optimized iteratively within an expectation-maximization framework were a DNN jointly estimates the spectra of all the signals of interest. The approach was shown to outperform previous cascaded approaches [Togami and Kawaguchi, 2014] while maintaining the performance when only some of the perturbations are present in the input signal.
From April 2019 to August 2019, I co-supervised Michel Olvera's master internship together with Emmanuel Vincent. Michel Olvera explored the application of DNN-based speech separation algorithms to separate foreground and background sound events in complex sound scene [Olvera et al., 2021]. This research work is at the interface between the work presented in this chapter and the work presented in Chapters 5 and 6.

5 Sound event detection with weakly labeled data

Context: This work was done as an associate professor at Université de Lorraine (Nancy, France) since September 2016 together with Hamid Egbal-Zadeh (Johannes Kepler University, Linz, Austria), Ankit Shah (Carnegie Mellon University, Pittsburgh, United States) and Nicolas Turpault.[1] The work presented here has been previously published in articles [Serizel and Turpault, 2019, Serizel et al., 2018].

5.1 Introduction

We are constantly surrounded by sounds and we rely heavily on these sounds to obtain important information about what is happening around us. Ambient sound analysis aims at automatically extracting information from these sounds. It encompasses disciplines such as sound scene classification (in which context does this happen?) or sound event detection and classification (SED) (what happens during this recording?) [Virtanen et al., 2018]. This area of research has been attracting a continuously growing attention during the past years as it can have a great impact in many applications in noise monitoring in smart cities [Bello et al., 2018a,b], surveillance [Radhakrishnan et al., 2005], urban planning [Bello et al., 2018b], multimedia information retrieval [Jin et al., 2012, Wold et al., 1996]; and domestic applications such as smart homes, health monitoring systems and home security solutions [Debes et al., 2016, Serizel et al., 2018, Zigel et al., 2009] to name a few.
Since 2018 I am participating to the organization of a task in Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge. In 2018 we organize DCASE challenge task 4 (large-scale weakly labeled semi-supervised sound event detection in domestic environments) that focused on SED with time boundaries in domestic applications [Serizel et al., 2018]. The systems submitted had to detect when a sound event occurred in an audio clip and what was the class of the event (as opposed to audio tagging where only the presence of a sound event is important regardless of when it happened). We proposed to investigate the scenario where a large scale corpus is available but only a small amount of the data is labeled. Task 4 corpus was derived from the Audioset corpus [Gemmeke et al., 2017] targeting classes of sound events related to domestic applications. The labels are provided at clip level (an event is present or not within a sound clip) but without the time boundaries (weak labels, that can also be referred to as tags) in order to decrease the annotation time. These constraints indeed correspond

[1] Collaborators listed alphabetically, members of the host institution unless mentioned otherwise

to constraints faced in many real applications where the budget allocated to annotating is limited.
In order to fully exploit this dataset, the submitted systems had to tackle two different problems. The first problem is related to the exploitation of the unlabeled part of the dataset either in unsupervised approaches [Jansen et al., 2018, Salamon and Bello, 2015b] or together with the labeled subset in semi-supervised approaches [Elizalde et al., 2017, Komatsu et al., 2016, Zhang and Schuller, 2012]. The second problem was related to the detection of the time boundaries and how to train a system that can detect these boundaries from weakly labeled data [Kumar and Raj, 2016, 2017]. The evaluation metric chosen was selected because it was penalizing these boundary estimation errors heavily. The goal was to encourage participants to focus on the time localization aspect.
Through a detailed overview of the systems submitted to DCASE 2018 task 4 we propose an overview of some advances in SED with partially annotated data.[2] We will first briefly describe task 4 and the related audio corpus in Section 5.2. Systems performance over all classes will be presented and analyzed in Section 5.3. We will present a class-wise analyze in Section 5.4 and discuss the impact of the metric chosen in Section 5.5. Section 5.6 will draw the conclusions of the chapter and present some perspectives for SED.

5.2 DCASE 2018 task 4

5.2.1 Audio dataset

The task relies on a subset of Audioset that focuses on 10 classes of sound events [Serizel et al., 2018]. Audioset consists in 10-second audio clips extracted from youtube videos [Gemmeke et al., 2017]. The development set provided for task 4 is split into a training set and a test set.

5.2.1.1 Training set

In order to reflect what could possibly happen in a real-world scenario, we provide three different splits of training data in task 4 training set: a labeled training set, an unlabeled in domain training set and an unlabeled out of domain training set (clips that do not contain any of the target classes):
Labeled training set: contains 1,578 audio clips (2,244 class occurrences) for which weak labels provided in Audioset have been verified and corrected by human annotators. One-third of the audio clips in this set contain at least two different classes of sound events.
Unlabeled in domain training set: contains 14,412 audio clips. The audio clips are selected such that the distribution per class of sound event (based on Audioset labels) is close to the distribution in the labeled set.
Unlabeled out of domain training set: is composed of 39,999 audio clips extracted

[2] Additional result plots and analysis can be be found at https://turpaultn.github.io/dcase2018-results/

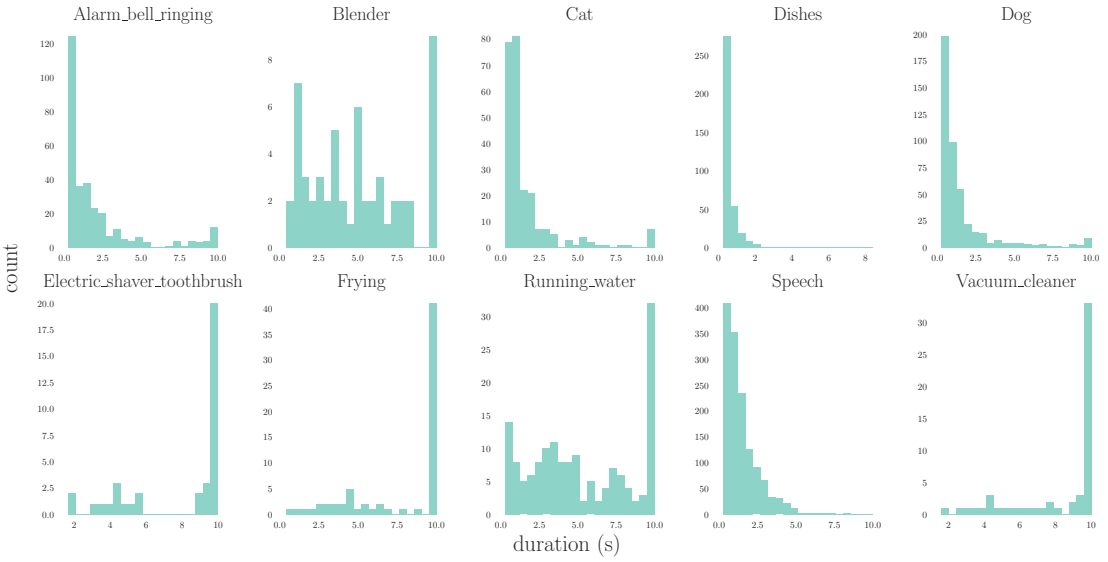


Figure 5.1: Duration distribution by class of sound events on the evaluation set.

from classes of sound events that are not considered in the task (according to unverified Audioset labels).

5.2.2 Test set

The test set is designed such that the distribution in term of clips per class of sound event is similar to that of the weakly labeled training set. The test set contains 288 audio clips (906 events). The test set is annotated with strong labels, with time boundaries (obtained from human annotators).

5.2.3 Evaluation set

The evaluation set contains 880 audio clips (3,187 events). The process to select the audio clips was similar to the process applied to select audio clips in the training set and the test set, in order to obtain a set with comparable classes distribution (see also Table 5.1). Labels with time boundaries are obtained from human annotators.
The duration distribution for each sound event class is presented on Figure 5.1. One of the focus of this task is the development of approaches that can provide fine time-level segmentation while learning on weakly labeled data. The observation of the event duration distribution confirms that in order to perform well it is essential to design approaches that are efficient at detecting both short events and events that have a longer duration.

	Subset	
Class	Test	Eval
Alarm/bell/ringing	112	306
Blender	40	56
Cat	97	243
Dishes	122	370
Dog	127	450
Electric shaver/toothbrush	28	37
Frying	24	67
Running water	76	154
Speech	261	1401
Vacuum cleaner	36	56
Total	906	3187

Table 5.1: Number of sound events per class in the test set and the evaluation set.

5.2.4 Task description

The task consists of detecting sound events within web videos using weakly labeled training data. The detection within a 10-seconds clip should be performed with start and end timestamps.

5.2.4.1 Task evaluation

Submissions were evaluated with event-based measures for which the system output is compared to the reference labels event by event [Mesaros et al., 2016] (see also Figure 5.2). The correspondence between sound event boundaries are estimated with a 200 ms tolerance collar on onsets and a tolerance collar on offsets that is the maximum of 200 ms and 20 % of the duration of the sound event. The collars are defined to account for the potential inaccuracy during the manual labeling process. For long event, the offset can be less clearly defined than the onset (an an be interpreted differently by different annotators). This motivates the use of a collar on the offsets that take into account the sound event duration.

- True positives are the occurrences when a sound event present in the system output corresponds to a sound event in the reference annotations.
- False positives are obtained when a sound event is present in the system output but not in the reference annotations (or not within the tolerance collars on the onset or the offset).
- False negatives are obtained when a sound event is present in the reference annotations but not in the system output (or not within the tolerance collars).

Submissions were ranked according to the event-based F-score. The F-score was first computed class-wise over the whole evaluation set:

$$F_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c}, \tag{5.1}$$

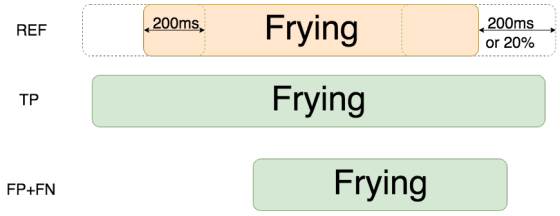


Figure 5.2: Event-based F-score.

where TP_c, FP_c and FN_c are the number of true positives, false positives and false negative for sound event class c over the whole evaluation set, respectively.
The final score is the F-score average over sound event classes regardless of the number of sound events per class (macro-average):

$$F_{\text{macro}} = \frac{\sum_{c \in \mathcal{C}} F_c}{n_{\mathcal{C}}}, \tag{5.2}$$

where $\mathcal{C}$ is the sound event classes ensemble and $n_{\mathcal{C}}$ the number of sound event classes.

5.3 Analysis of the performance over all sound event classes

In this section we present and analyze performance of the submitted systems regardless of the sound event classes.

5.3.1 Task submissions and results overview

DCASE 2018 task 4 gathered 50 submissions from 16 different research teams involving 57 researchers overall. The official team ranking and some characteristics of the submitted systems are presented in Table 5.2. The best two submissions quite clearly stand out from other submissions. They also go beyond the rather standard approaches based convolutional neural networks (CNN) or stacked CNN and recurrent neural networks (RNN) also denoted as CRNN. The best system, submitted by JiaKai (**jiakai_psh**) [JiaKai, 2018], relies on a mean-teacher model that exploits unlabeled data to regularize the classifier learned on the weakly labeled data [Tarvainen and Valpola, 2017]. The system submitted by Liu et al. (**liu_ustc**) [Liu et al., 2018] that ranked second relies on an energy based sound event detection as a pre-processing to a capsule network [Sabour et al., 2017]. The output of the network is then post processed to ensure that silence between events and events themselves are longer than a minimum duration.
Other notable submissions include the system from Kothinti et al. (**kothinti_jhu**) [Kothinti et al., 2018] that relies on a sound event detection based on restricted Boltzmann machines (RBM) as a pre-processing. This solution performs well at detecting onsets but not so much for offset detection (see also Section 5.4.1). Dinkel et al. proposed a system (**dinkel_sjtu**) that uses Gaussian mixture models (GMM) and hidden Markov models (HMM) to perform sound event alignment [Dinkel et al., 2018]. Gaussian filtering is then

Rank	System	Classifier	Parameters	F (%)
1	**jiakai_psh** [JiaKai, 2018]	CRNN	1M	**32.4**
2	**liu_ustc** [Liu et al., 2018]	CRNN, Capsule-RNN	4M	**29.9**
3	**kong_surrey** [Kong et al., 2018]	VGGish 8 layer CNN	4M	**24.0**
4	**kothinti_jhu** [Kothinti et al., 2018]	CRNN, RBM, cRBM, PCA	1M	**22.4**
5	**harb_tug** [Harb and Pernkopf, 2018]	CRNN, VAT	497k	**21.6**
6	**koutini_jku** [Koutini et al., 2018]	CRNN	126k	**21.5**
7	**guo_thu** [Guo et al., 2018]	multi-scale CRNN	970k	**21.3**
8	**hou_bupt** [Hou and Li, 2018]	CRNN	1M	**21.1**
9	**lim_etri** [Lim et al., 2018]	CRNN	239k	**20.4**
10	**avdeeva_itmo** [Avdeeva and Agafonov, 2018]	CRNN, CNN	200k	**20.1**
11	**wangjun_bupt** [Jun and Shengchen, 2018]	RNN	1M	**17.9**
12	**pellegrini_irit** [Cances et al., 2018]	CNN, CRNN with MIL	200k	**16.6**
13	**moon_yonsei** [Hyeongi et al., 2018]	RseNet, SENet	10M	**15.9**
14	**dinkel_sjtu** [Dinkel et al., 2018]	CRNN, HMM-GMM	126k	**13.4**
15	**wang_nudt** [Wang et al., 2018a]	CRNN	24M	**12.6**
	baseline [Serizel et al., 2018]	CRNN	126k	**10.8**
16	**raj_iit** [Raj et al., 2018]	CRNN	215k	**9.4**

Table 5.2: Team ranking and submitted systems characteristics.

used as post-processing. Pellegrini et al. proposed a system (**pellegrini_irit**) that relies on multiple instance learning (MIL) to exploit weakly labeled data [Cances et al., 2018]. Both these systems perform pretty decently on segmentation (see also Section 5.3.2) but they suffer from pretty poor sound event classification performance (see also Figure 5.10).

5.3.2 Segmentation

In this section, we focus on the segmentation performance. That is, the ability of the submitted systems to localize sound events in time without having to predict the class. To compute the F-score in practice, the 10 original classes are collapsed into a single class and we only evaluate whether or not the systems are able to detect that a sound event of interest is occruing (regardless of the class). Figures 5.3, 5.4 and 5.5 present the event-based F-score computed without taking the sound event class labels into account and for a tolerance collar of 200 ms, 1 s and 5 s, respectively. The fact that there is only little performance difference between the sound event detection performance (Table 5.2) and the segmentation performance tends to indicates that segmentation is possibly the main limiting factor in overall performance. This is actually confirmed by the rather high tagging performance of most systems presented on Figure 5.10.

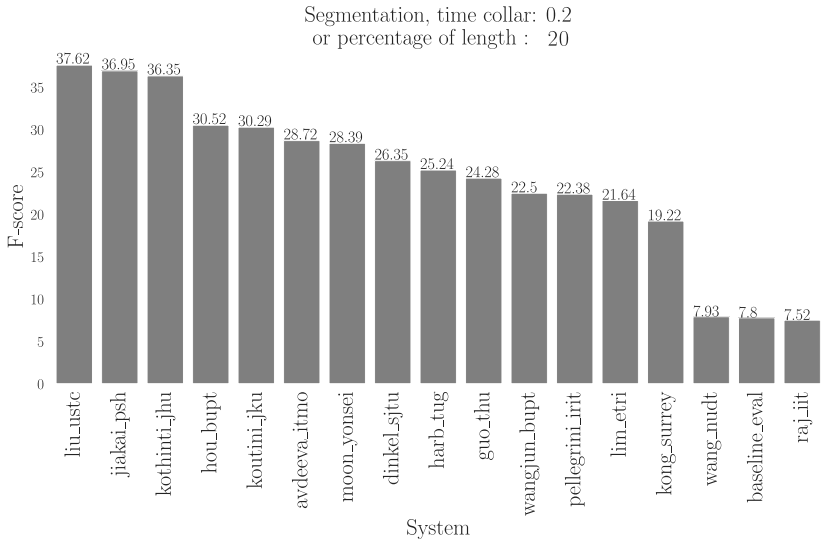


Figure 5.3: Segmentation performance (tolerance collar on onsets is 200 ms and tolerance collar on offsets is the maximum of 200 ms and 20 % of the event length).

Most of the systems are able to detect if an event occurred within a rather crude time area (see Figure 5.5 but are not able to properly segment the audio clips in terms of sound events (see Figure 5.3). The systems that performed best in terms of segmentation are the systems that actually implemented some sort of segmentation among which **liu_ustc** [Liu et al., 2018] and **kothinti_jhu** [Kothinti et al., 2018]. The winning system is ranked second in term of segmentation and owe its first overall rank to a much better classification than competing systems (see Figure 5.10).

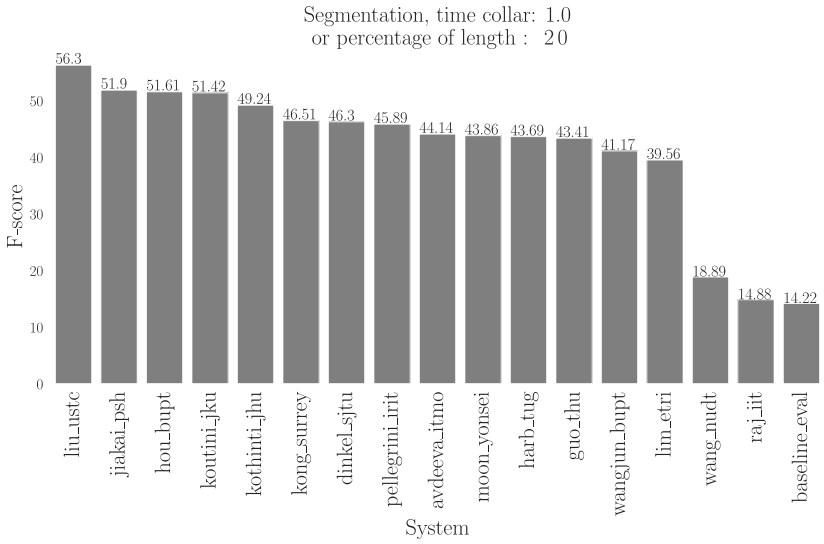


Figure 5.4: Segmentation performance (tolerance collar on onsets is 1 s and tolerance collar on offsets is the maximum of 1 s and 20 % of the event length).

5.3.3 Use of unlabeled data

One of the challenges proposed by DCASE 2018 task 4 was to exploit a large amount of unlabeled data. In the section we analyze the approaches proposed by participants. Most of the systems submitted used a pseudo-labeling approach where a first system trained on the labeled data is used to obtain labels for the unlabeled set (**liu_ustc**) [Liu et al., 2018], **hou_bupt** [Hou and Li, 2018]). Variations on this included setting a confidence threshold to decide to keep the label or not (**koutini_jku** [Koutini et al., 2018], **wang_nudt** [Wang et al., 2018a], **pellegrini_irit** [Cances et al., 2018], **harb_tug** [Harb and Pernkopf, 2018], **moon_yonsei** [Hyeongi et al., 2018]) and gradually introducing new audio clips with these pseudo labels (**wangjun_bupt** [Jun and Shengchen, 2018]). The winning system (**jiakai_psh** [JiaKai, 2018]) used the unlabeled data within a mean-teacher scheme [Tarvainen and Valpola, 2017]. It is composed of two models: a student model and a mean-teacher model whose weights are the exponential average of the student's weights. On labeled data, the student model weights are updated to optimize a classification cost on the sound event classes. Additionally, consistency costs are computed to compare the output of the student model and the mean-teacher model on both the labeled and the unlabeled data. Kothinti et al. (**kothinti_jhu** [Kothinti et al., 2018]) proposed to use both the weakly labeled and unlabeled in-domain data to train several RBM that are used to detect sound event boundaries.

5.3.4 Complexity

The complexity of the submitted systems (in terms of number of parameters) is presented in Table 5.2. The only system that used raw waveforms as input (**moon_yonsei** [Hyeongi

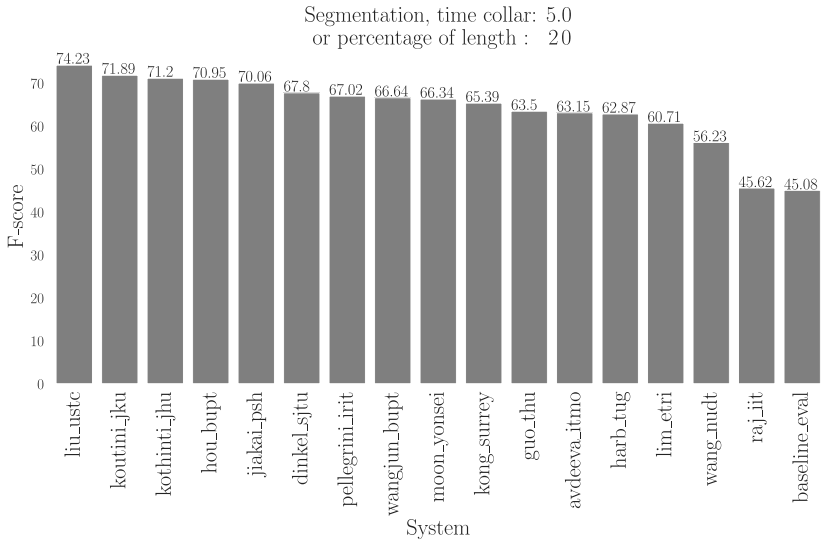


Figure 5.5: Segmentation performance (tolerance collar on onsets is 5 s and tolerance collar on offsets is the maximum of 5 s and 20 % of the event length).

et al., 2018]) is among the most complex systems yet it is not even among the top 10 systems. This tends to indicate that the dataset proposed for task 4 is too small to train SED systems using raw waveforms that are usually known to require a lot of training data. The most complex system (**wang_nudt** [Wang et al., 2018a]) is about 200 times more complex than the baseline in particular because it combines several complex models. However it performs only slightly better than the baseline. The winning system (**jiakai_psh** [JiaKai, 2018]) is about 10 times more complex than the baseline. The best performing system that has a number of parameters similar to that of the baseline (**koutini_jku** [Koutini et al., 2018]) improves the baseline F-score performance by more than 10 % absolute.

5.3.5 Duration of events

It has been shown above that the systems performance largely depends on the systems ability to properly segment the audio clips in terms of sound events. Figure 5.1 presents the duration distribution for each class of sound events on the evaluation set. From this distribution we can separate the sound events into two categories of events: short sound events ('Alarm/bell/ringing', 'Cat', 'Dishes', 'Dog' and 'Speech") and long sound events ('Blender', 'Electric shaver/toothbrush', 'Frying', 'Running water' and 'Vacuum cleaner').

Figure 5.6 presents the performance of the submitted systems on short sound events depending on their performance on long sound events. No system is clearly outperforming the others on both short and long sound events. This is confirmed when looking at the top performing systems on short sound events (Table 5.3) and on long sound events (Table 5.4). These rankings tend to show that the approaches proposed were either

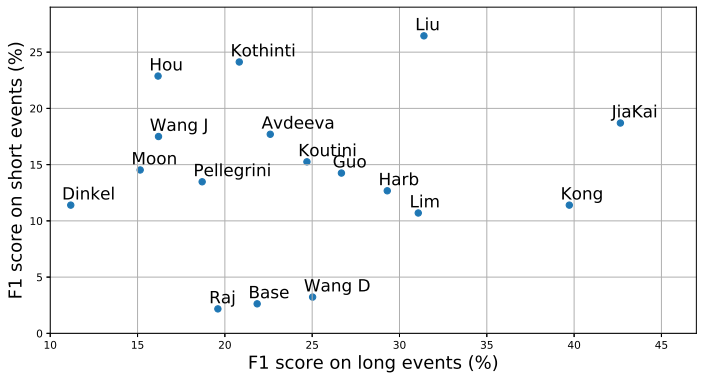


Figure 5.6: Systems performance on short sound events depending on their performance on long sound events.

	Sound event type			
System	**Short**	Long	All	Official rank
liu_ustc	**26.4**	31.4	29.9	2
kothinti_jhu	**24.1**	20.8	22.4	4
hou_bupt	**22.9**	16.2	21.1	8
jiakai_psh	**18.7**	42.6	32.4	1
avdeeva_itmo	**17.7**	22.6	20.1	10
baseline	**2.6**	21.8	10.8	

Table 5.3: Top 5 systems on short events ('Alarm/bell/ringing', 'Cat', 'Dishes', 'Dog' and 'Speech").

tailored to perform well on short sound events (top systems are also the systems that performed best in terms of segmentation, see also Figure 5.3) or on long sound events (top systems are also among the best systems in terms of tagging, see also Figure 5.10). However, in order to perform well on the SED task systems had to perform reasonably well on both short and long sound events. This is the case for the top two systems (**jiakai_psh** [JiaKai, 2018] and **liu_ustc**) [Liu et al., 2018]) that are in the top five both short sound events and long sound events.

5.4 Analysis of the class-wise performance

It have been shown above that systems performance can vary to a great extent depending on the sound events duration that is tightly related to the sound event class itself. Therefore, in this section we focus on the performance of the submitted systems depending on the sound event classes. Table 5.5 presents the class-wise event-based F-score for the 10 best performing submitted systems. The best system (**jiakai_psh** [JiaKai, 2018]) out-

System	Sound event type			Official rank
	Long	Short	All	
jiakai_psh	**42.6**	18.7	32.4	1
kong_surrey	**39.7**	11.4	24	3
liu_ustc	**31.4**	26.4	29.9	2
lim_etri	**31.1**	10.7	20.4	9
harb_tug	**29.3**	12.7	21.6	5
baseline	**21.8**	2.6	10.8	

Table 5.4: Top 5 systems on long events ('Blender', 'Electric shaver/toothbrush', 'Frying', 'Running water' and 'Vacuum cleaner').

System	Sound event class									
	Alar.	Blen.	Cat	Dish.	Dog	Shav.	Fry.	Wat.	Sp.	Vac.
jiakai_psh	**49.9**	**38.2**	3.6	3.2	18.1	**48.7**	35.4	**31.2**	**46.8**	48.3
liu_ustc	46.0	27.1	20.3	13.0	**26.5**	37.6	10.9	23.9	43.1	50.0
kong_surrey	24.5	18.9	7.8	7.7	5.6	46.4	**43.6**	15.2	19.9	50.0
kothinti_jhu	36.7	22.0	**20.5**	12.8	**26.5**	24.3	0.0	9.6	34.3	37.0
harb_tug	15.4	30.0	8.1	17.5	9.7	21.0	34.7	17.3	31.1	31.5
koutini_jku	30.0	16.4	13.1	9.5	8.4	23.5	18.1	12.6	42.9	40.8
guo_thu	35.3	31.8	7.8	4.0	9.9	17.4	32.7	18.3	31.0	24.8
hou_bupt	41.4	16.4	6.4	**23.5**	20.2	9.8	6.2	14.0	40.6	32.3
lim_etri	11.6	21.6	7.9	5.9	17.4	27.8	14.9	15.5	21.0	**60.0**
avdeeva_itmo	33.3	15.2	14.9	6.3	16.3	15.8	24.6	13.3	27.2	34.8
baseline	4.8	12.7	2.9	0.4	2.4	20.0	24.5	10.1	0.1	30.2

Table 5.5: Class-wise event-based F-score for the top 10 submitted systems.

performs other systems on five sound event classes upon ten (mainly long sound events). However, it performs rather poorly on some of the remaining sound event classes (mainly short sound events). On the other hand, the second best system (**liu_ustc** [Liu et al., 2018]) outperforms other systems on a single sound event class ('Dog') but is generally not too far from the best performance on several other sound event class. This explains why it can still compare with the winning system in terms of overall performance.

In general 'Speech' and 'Alarm bell ringing' seem to be the easiest sound event classes to detect and classify. This could be explained by the fact that sound events from these classes are not too short (with a median duration of 1.17 s and 0.57 s, respectively), occurs many times in the training set (in 550 clips and 205 clips, respectively) and generally have rather clear onsets and offsets (see also Section 5.4.1). There is a clear separation between 'Cat', 'Dishes' and 'Dog' and other sound event classes. The former seems more difficult to detect and classify than the latter. This can be due to the fact that sound events in these classes are short and present a large acoustic variability. Interestingly, the submitted systems that perform best on these sound event classes are not necessarily

among the top three systems. For example **hou_bupt** [Hou and Li, 2018] obtains the best performance on 'Dishes' and clearly outperforms other submissions with 23.5 % F-score. However, it ranked eighth overall (but was among the top five systems on short sound events, see also Table 5.3). The best system on 'Cat' (by a rather large margin) with 25.3 % F-score is **pellegrini_irit** [Cances et al., 2018] that relies on MIL and that is not even in the top 10 in terms of overall performance.

5.4.1 Performance on onset and offset detection

For some sound event classes that slowly decay the time location of offsets can be difficult to locate (and the concept of offset itself can even become ambiguous in reverberant scenarios). Therefore, we now evaluate the detection of onsets and offsets separately. In the plots presented in this section, sound events are classified from the shortest (on the left) to the longest (on the right) according to their median duration. Additionally, for the sake of clarity, only the systems among the top four in overall performance are presented here. Systems are presented in decaying overall onset or offset detection performance (the best system is on the left side).

5.4.1.1 Onset

Figure 5.7 presents F-score for onset detection for varying tolerance collars (in seconds). Performance generally increases when the tolerance collar is increased. For small tolerance collars, **liu_ustc** [Liu et al., 2018] performs best which confirms previous analysis about the relatively good segmentation of their system. When the tolerance collar is larger than 1 s **jiakai_psh** [JiaKai, 2018] outperforms other system which also confirm that the proposed segmentation is a bit too coarse.

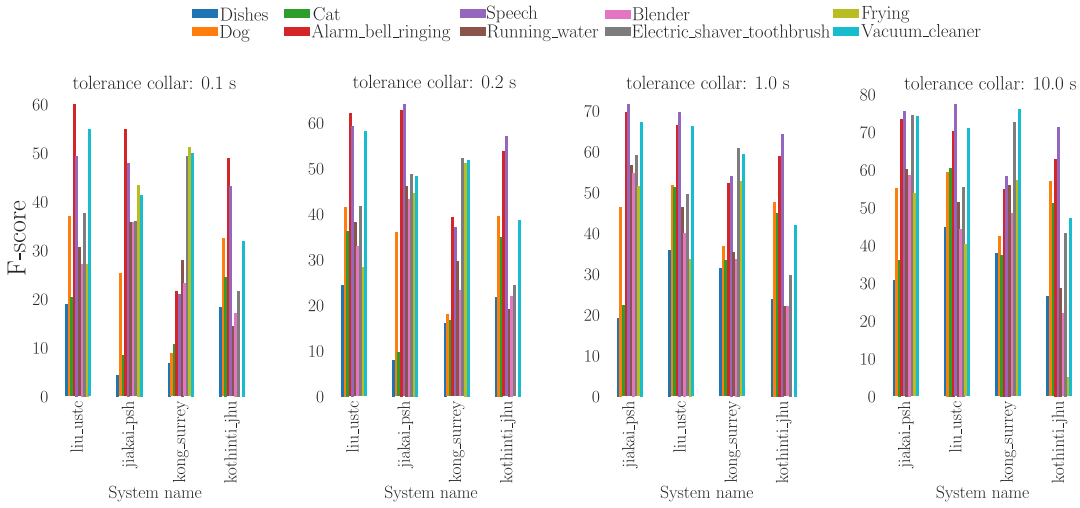


Figure 5.7: Event-based F-score for onset detection with absolute tolerance collars.

The remaining errors for a 10 sec tolerance collar indicate that the systems were not able

to predict how many onsets for the specific sound event class occurred within the audio clip. In most cases this could also corresponds to the case where the sound event was not detected at all (see also Figure 5.8).
When looking at particular sound event classes, in general systems exhibit good onset detection performance for 'Speech' and 'Alarm bell ringing'. As mentioned above, this can be due to the fact that these sound events occur frequently in the training set but it can also be related to the fact that the sound events from these classes indeed have rather clear onsets that appear to be easier to detect. On the other hand, sound event classes as 'Cat' and 'Dishes' seem to be difficult to detect. For the former it is probably due to the fact that the onsets are not always clear as for the latter it is most generally related to sound events that are simply missed by the systems because they are too short. For the remaining sound event classes, the performance varies a lot from one system to another and seems to be affected by the segmentation strategy implemented.

5.4.1.2 Offset

Figure 5.8 presents F-score for offset detection for varying tolerance collars (in seconds). When comparing with Figure 5.7 it appears that offsets are indeed more difficult to detect. The high F-score for some sound event classes such as ('Electric shaver/toothbrush', 'Frying' or 'Vacuum cleaner') is mainly due to the fact that many of the sound events in these classes do not have an offset within the audio clips and therefore the offset to be detected is simply the final boundary of the audio clip.

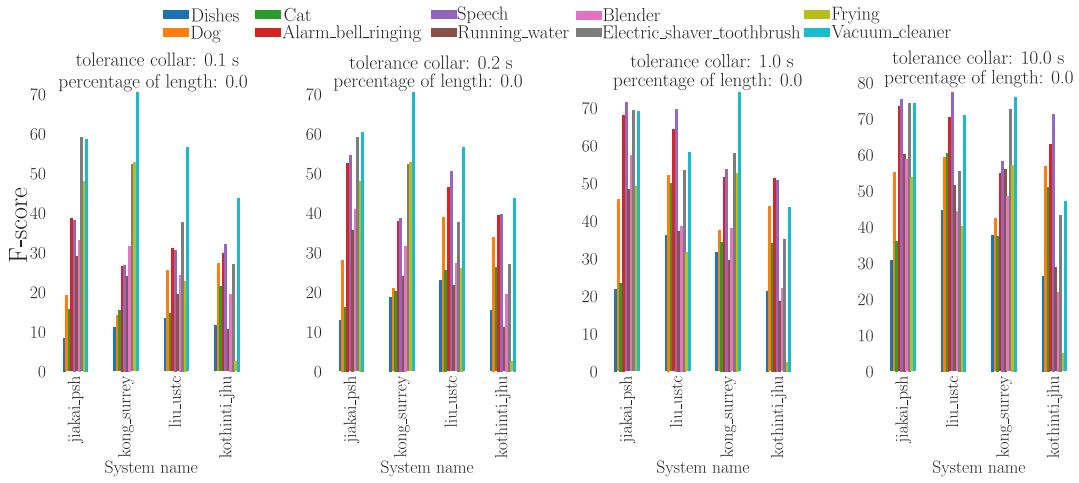


Figure 5.8: Event-based F-score for offset detection with absolute tolerance collars.

It is generally admitted that penalizing offset detection based on an absolute time tolerance collar is not a reasonable choice specially for long sound events. In particular because this type of tolerance collar might be affecting long sound events (with longer decay) much more than short (possibly percussive) sound events. Therefore, the metric retained for DCASE 2018 task 4 include both an absolute time tolerance collar and a

tolerance collar that was computed as a percentage of the sound event duration (the maximum of these two values was retained). With this approach, the absolute time tolerance collar usually applies to short sound events while the tolerance collar relative to event length applies to longer sound events.
Figure 5.9 presents F-score for offset detection for varying tolerance collars (in percent of the sound event duration). Note that the absolute time tolerance collar is kept to 0.1 s here in order to avoid unreasonably small tolerance collars for short sound events. As expected, this kind of tolerance collar has less effect than absolute time tolerance collar on offset detection of short sound events such as 'Cat', 'Dishes' or 'Dog' but can affect greatly the offset detection performance on long sound events such as 'Running water' or 'Blender'.

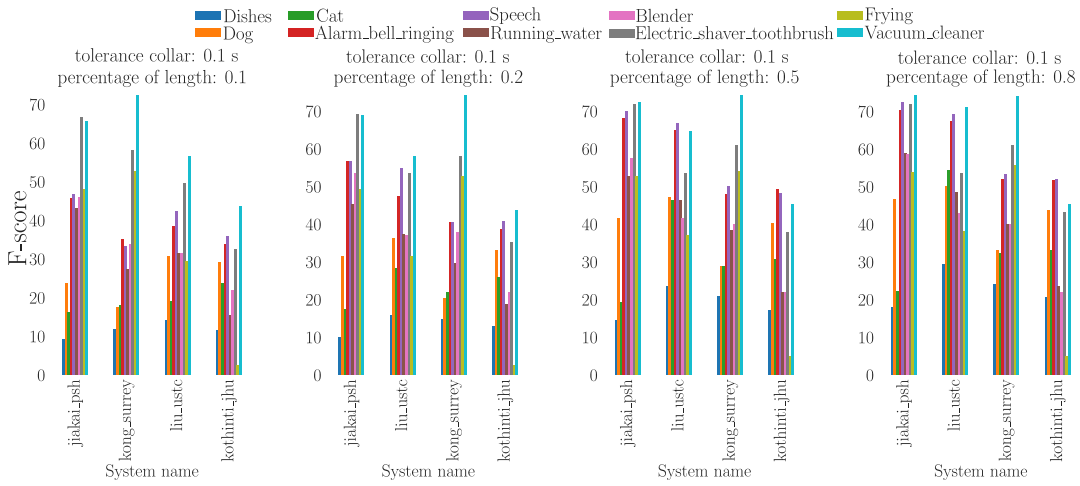


Figure 5.9: Event-based F-score for offset detection with tolerance collars relative to event duration.

jiakai_psh [JiaKai, 2018] outperforms the other submitted systems (even those which had demonstrated a better segmentation performance until now) including with low tolerance collars. When looking at particular sound event classes, in general the submitted systems exhibit good offset detection performance for 'Speech' and 'Alarm bell ringing' even if in this case offsets are usually not as well defined as onsets were.

5.5 Impact of the metric

For DCASE 2018, the F-score was computed in an event-based fashion in order to put on strong focus on the sound event segmentation. Class-wise performance was averaged in order to discard the effects of the sound event classes imbalance (5.2). In this section, we study the impact of these choices on the performance evaluation of the submitted systems.

5.5.1 F-score computation relatively to events or segments

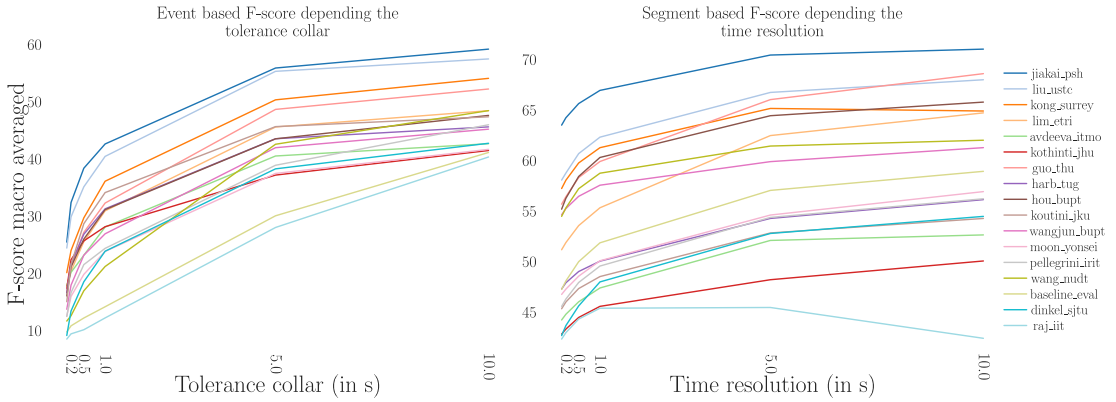


Figure 5.10: Comparison between event-based and segmented-based F-scores for various submitted systems depending on the tolerance collar and time resolution, respectively.

As opposed to event-based metrics, segment-based metrics are computed by comparing the system outputs and the reference on short segments extracted from the original audio clip. The sound event classes are then considered to be active or not on the full segment. The final metric is computed on all the segments [Mesaros et al., 2016]. This approach reports if a system is able to detect if a sound event class is active with a specific time resolution (the segment length) and can prove more robust than event-based metrics to phenomena such as short pauses between consecutive sound events. Figure 5.10 presents a comparison between the event-based F-scores (on the left) and the segment-based F-scores (on the right) for varying tolerance collars and time resolutions, respectively.
As expected, segmented-based metrics are more permissive to errors in the detection of the sound event boundaries. Indeed the reported segment-based F-scores (from 40 % to 70 % depending on the time resolution) are much higher than their event-based counterpart (from 5 % to 60 % depending on the tolerance collar). Additionally, the segment-based F-score seems to be favoring systems that are good at tagging while event-based F-score favors systems that have good segmentation performance. This is particularly clear for systems like **hou_bupt** [Hou and Li, 2018], **guo_thu** [Guo et al., 2018] and the task baseline [Serizel et al., 2018] which perform much better in terms of segment-based F-score and for **kothinti_jhu** [Kothinti et al., 2018] that performs much better in terms of event-based F-score.
When the time resolution for the segment-based F-scores is 10 s the reported performance is actually that of a tagging task. The tagging ranking is then rather different than the general ranking (see also Table 6.2) and the ranking for segmentation (see also Figure 5.3). This emphasizes once again that none of the submitted systems is actually outperforming others in both segmentation and tagging but that in order to perform well on the task, systems had to perform at least decently on both. This is the case for **jiakai_psh** [JiaKai, 2018] and **liu_ustc** [Liu et al., 2018] that clearly stand out in the final ranking.
As the choice of the metric is tightly related to the targeted application, some approaches

can be better suited when you need to know exactly when a sound event from a specific class did occur (in which case you might select a system that performs well in terms event-based F-score) some other approaches can be suited to monitor the activity within a time period (approximately when was each sound event class active, depending on the time resolution, in which case we might select systems that perform well in terms segment-based F-score)

5.5.2 Micro average

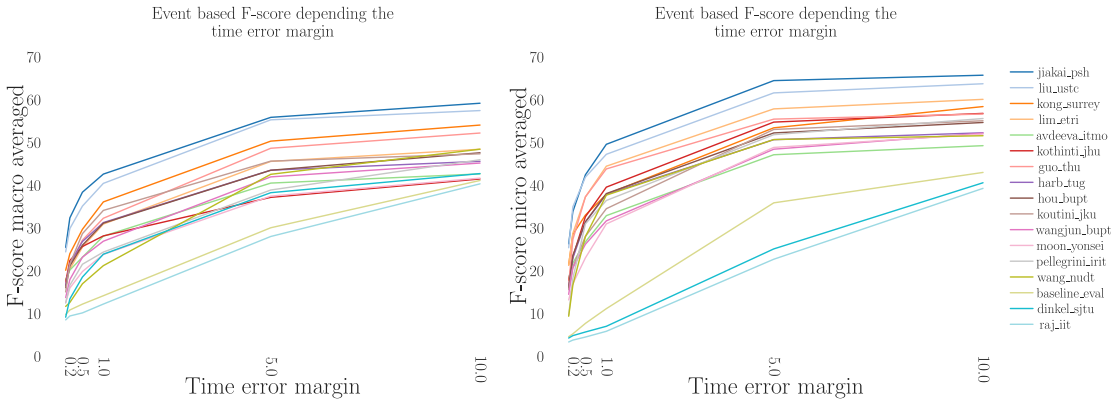


Figure 5.11: Event-based F-score for various submitted systems depending on the class averaging method.

While macro-averaging (used in task 4) computes the final F-score as the average across sound event classes (regardless of the number of events for each class), micro-averaging computes the final F-score as the average of each single decision. It therefore gives more importance to sound event classes that occur more frequently (see also Table 5.1 for the distribution). For example, 'Speech' events will account for almost half of the performance when using micro-averaged F-score.

Figure 5.11 presents event-based F-score depending on the averaging method. We can observe a clear score increase between macro-averaged and micro-average F-score for the systems that performed well the most frequent sound event classes ('Alarm bell ringing', 'Dishes', 'Dog' or 'Speech') such as **lim_etri** [Lim et al., 2018]. On the other hand the systems that were able to perform well on less frequent sound event classes ('Electric shaver/toothbrush', 'Frying'...) but not on frequent sound event classes can see their performance decreased between macro-averaged and micro-averaged F-score as this is the case for **kong_surrey** [Kong et al., 2018]. The top two systems (**jiakai_psh** [JiaKai, 2018] and **liu_ustc**) [Liu et al., 2018]) were performing reasonably well on the most frequent sound event classes and therefore still outperform other systems in terms of micro-averaged F-score.

The choice of the metric is related to the targeted application. If you want to detect mainly the sound event classes that occur the most frequently and that missing rare sound event classes is not really a problem then you should select approaches that perform well

in terms of micro-averaged F-score. On the contrary if detecting rare sound event classes is important then approaches that perform well in terms of macro-averaged F-score seem better suited.

5.6 Conclusion

In this chapter we proposed an overview of some of advances and challenges in sound event detection with systems trained on partially annotated data through the analysis of the results of DCASE 2018 challenge task 4. The chapter focused on the scientific aspects highlighted by the task: exploiting both unlabeled and weakly labeled data to train a system that provides not only the event class but also the event time boundaries. It has been shown that both the segmentation and the classification ability play an important role in the final performance. However whereas the tagging performance (related to the classification ability) is generally rather good for many systems, only few systems did implement an explicit segmentation strategy. This aspect actually remains quite challenging as training a system to detect sound events and predict their time localization from weakly labeled data is far from trivial. Therefore, one question investigated in the following DCASE challenge task 4 iterations is to analyze if strongly labeled data that is generated synthetically can help solving this issue. This latter topic is also the focus of Chapter 6

5.6.1 Other related works

From January 2017 to May 2021, I was co-director of Nicolas Turpault's PhD together with Emmanuel Vincent. In a first stage Nicolas Turpault investigated the use of triplet based approaches to learn representation from partially annotated dataset [Turpault et al., 2019b]. The triplet could be drawn differently depending on whether the considered anchor clip was annotated or not. In this study, the annotations available were limited to weak annotations. The proposed triplet based approach exhibited strong limitations in particular because of the impact of weak labeling that could lead to learning inaccurate representations [Turpault et al., 2020a]. In particular, portions of an audio clip could be considered as positive sample in a triplet (because a sound event class was annotated as active at the clip level) while the portion itself actually did not contain the target event and should be considered as a negative sample in the triplet. In order to understand clearly the impact of the weak labels during the learning process, Nicolas Turpault then proposed a detailed study on the subject [Turpault et al., 2021a]. A specific dataset was designed synthetically in order to isolate each particular aspects of the problem. It was shown in particular that the aggregation function used to convert frame level prediction to clip level prediction is crucial when dealing with weak labels that are actually true only for a portion of a clip. The impact of the temporal granularity of the annotations at training time was also analyzed. It was shown that having a coarser time granularity at training that the target test granularity is not necessarily a problem if the mismatch is not too important. This work is related to the work presented in Chapter 6.

Since 2018, I have been actively involved in the DCASE community in particular trough the organization of a sound event detection task during the DCASE challenge (task 4). The task as now been running for 4 consecutive iterations attracting up to 20 team submissions (for a total of about 80 researchers participating worldwide). The dataset designed specifically for the task, DESED ([Turpault et al., 2019a], Chapter 6), have been downloaded more than 4000 times.
In 2018, we proposed a sound event detection task based on semi-supervised learning [Serizel et al., 2018]. One challenge to be addressed within the task is to explore the possibility to exploit a large amount of unbalanced and unlabeled data together with a small weakly annotated data to train a sound event detection system that should estimate not only the event class but also the start and stop time instants for the events. Based on the submission to the challenge, we proposed a detailed performance analysis highlighting the remaining challenges in the task [Serizel and Turpault, 2019]. This latter study motivated the introduction of strongly labeled synthetic clips in the task in 2019 [Turpault et al., 2019a]. The next Chapter describes an analysis of SED systems performance based on these synthetic soundscapes.

6 Sound event detection with synthetic soundscapes

Context: This work was done as an associate professor at Université de Lorraine (Nancy, France) since September 2016 together with Justin Salamon (Adobe research, United states) Ankit Shah (Carnegie Mellon University, United states) and Nicolas Turpault.[1] The work presented here has been previously published in articles [Serizel et al., 2020, Turpault et al., 2019a].

6.1 Introduction

As seen in Chapter 5, using weakly labeled data at training can have an impact on the ability of systems to detect sound event classes [Turpault et al., 2021a] and to segment sound event in time. One cheap alternative to manually annotate data with strong annotation is to generate synthetic soundscapes. In this chapter, we generate strongly annotated synthetic soundscapes using the Scaper library [Salamon et al., 2017]. Given a set of user-specified background and foreground sound event recordings, Scaper automatically generates soundscapes containing random mixtures of the provided events sampled from user-defined distributions. These distributions are defined via a sound event specification including properties such as event duration, onset time, signal-to-noise ratio (SNR) with respect to the background and data augmentation (pitch shifting and time stretching). This allows us to generate multiple different soundscape instantiations from the same specification, which is chosen based on our general requirements for the soundscapes. Since generating such strongly labeled synthetic data is feasible on a large scale, in DCASE 2019 task 4, we provided a strongly labeled synthetic dataset in order to explore whether it can help improve SED models.
One problem is that the evaluation on complex recorded soundscapes does not allow to disentangle the several challenges faced in SED in real environments. Capitalizing on the possibility to have a full control on the properties of the soundscapes generated with Scaper [Salamon et al., 2017], we also exploit synthetic soundscapes at evaluation time. In this chapter we benchmark SED submissions to DCASE 2019 task 4 on synthetic soundscapes designed to investigate several SED challenges such as foreground event to background ratio or time localization of the sound event within a clip.
The chapter is organized as follows: Section 6.2 provides a brief overview of the task definition. Section 6.3 describes how the development and evaluation datasets were created. Section 6.4 describes the baseline system. Section 6.5 describes the evaluation

[1] Collaborators listed alphabetically, members of the host institution unless mentioned otherwise

procedure for DCASE 2019 Task 4 and gives an overview of the systems submitted to the challenge for this task. The robustness to noise degradation and segmentation are presented in Sections 6.6 and 6.7, respectively. Finally, conclusions from the challenge are provided in section 6.8.

6.2 Task description

This chapter focuses on the same 10 classes of sound events as in previous chapter. Systems are expected to produce strongly labeled output (i.e. detect sound events with a start time, end time, and sound class label). Multiple events can be present in each audio recording, including overlapping events. However, unlike in Chapter 5, in this chapter we also provide an additional training set with strongly annotated synthetic soundscapes. This opens the door to exploring scientific questions around the informativeness of real (but weakly labeled) data versus strongly-labeled synthetic data, whether the two data sources are complementary or not, and how to best leverage these datasets to optimize system performance.

6.3 DESED dataset

The Domestic Environment Sound Event Detection (DESED) development dataset is composed of 10-sec audio clips recorded in a domestic environment or synthesized to simulate a domestic environment. The real recordings are taken from AudioSet [Gemmeke et al., 2017]. The dataset is divided in three subsets:

- A training subset composed of real recordings similar to the dataset used in Chapter 5 and synthetic soundscapes generated using Scaper (see also Table 6.1).
- A validation subset composed of real recordings with strong label which is the combination of the validation and evaluation sets used in Chapter 5.
- An evaluation subset composed of real recordings with strong labels and synthetic soundscapes with strong labels. The synthetic subsets are designed to isolated specific challenges faced in real-world SED and analyze the behaviour of the submissions regarding these challenges (see 6.3.3).

6.3.1 Synthetic soundscape generation procedure

The subset of synthetic soundscapes is comprised of 10 second audio clips generated with Scaper [Salamon et al., 2017], a python library for soundscape synthesis and augmentation. Scaper operates by taking a set of foreground sounds and a set of background sounds and automatically sequencing them into random soundscapes sampled from a user-specified distribution controlling the number and type of sound events, their duration, signal-to-noise ratio, and several other key characteristics.

	Unique isolated sound events	
Class	Development set	Evaluation set
Alarm/bell/ringing	190	63
Blender	98	27
Cat	88	26
Dishes	109	34
Dog	136	43
Electric shaver/toothbrush	56	17
Frying	64	17
Running water	68	20
Speech	128	47
Vacuum cleaner	74	20
Total	1011	314

Table 6.1: Class-wise statistics for unique isolated sound events in the DESED dataset.

6.3.2 DESED development dataset

The development dataset is composed of the training subset and the validation subset. The foreground events used to generate the synthetic soundscapes are obtained from the Freesound [Fonseca et al., 2017, Font et al., 2013]. Each sound event clip was verified by a human to ensure that the sound quality and the event-to-background ratio were sufficient to be used as an isolated sound event. We also controlled for whether the sound event onset and offset were present in the clip. Each selected clip was then segmented when needed to remove silences or mild background noise periods before and after the sound event and between sound events when the file contained multiple occurrences of the sound event class. The number of unique isolated sound events per class used to generate the subset of synthetic soundscapes is presented in Table 6.1.
The background textures are obtained from the SINS dataset (from the activity class "other") [Dekkers et al., 2017]. This particular activity class was selected because it contains a low amount of sound events from our 10 target foreground sound event classes. However, there is no guarantee that these sound event classes are completely absent from the background clips. A total of 2060 unique background clips are used to generate the synthetic subset.
Scaper scripts are designed such that the distribution of sound events per class, the number of sound events per clip (depending on the class) and the sound event class co-occurrence are similar to that of the validation set which is composed of real recordings. The synthetic soundscapes are annotated with strong labels that are automatically generated by Scaper [Salamon et al., 2017].

6.3.3 DESED evaluation dataset

The evaluation set is composed of two subsets: a subset with real recordings and a subset with synthetic soundscapes.

6.3.3.1 Real recordings

The first subset contains 1,013 audio clips and is used for ranking purposes. It is comprised of audio clips extracted from 692 YouTube and 321 Vimeo videos under creative common licenses. Each clip is annotated in terms of sound event classes and time boundaries by a human and annotations are verified by a second annotator.

6.3.3.2 Synthetic soundscapes

The DESED synthetic soundscapes evaluation set is comprised of 10 second audio clips generated with Scaper [Salamon et al., 2017]. This set is used for analysis purposes and its design is motivated by the analysis of DCASE 2019 task 4 results [Serizel and Turpault, 2019]. In particular, most submissions performed poorly in terms of event segmentation. One of the goals of this subset is to facilitate studies on the extent to which including strongly labeled data in the training set helps improve and refine the segmentation output.
The foreground events are obtained from the Freesound [Fonseca et al., 2017, Font et al., 2013]. The selection and sound event clip processing is the same as for the development set. The number of unique isolated sound events per class used to generate the subset of synthetic soundscapes is presented in Table 6.1.
Background sounds are extracted from YouTube videos under a Creative Common license and from the Freesound subset of the MUSAN dataset [Snyder et al., 2015]. These recordings were selected because they contains a low amount of sound events from our 10 target foreground sound event classes. However, there is no guarantee that these sound event classes are completely absent from the background clips.
DESED synthetic soundscapes evaluation set is further divided into several subsets (described below) for a total of 12,139 audio clips synthesized from 314 isolated events. The synthetic soundscapes are annotated with strong labels that are automatically generated by Scaper [Salamon et al., 2017].

Varying foreground-to-background SNR

A subset of 754 soundscapes is generated with Scaper scripts are designed such that the distribution of sound events per class, the number of sound events per clip (depending on the class) and the sound event class co-occurrence are similar to that of the validation set which is composed of real recordings. The foreground event SNR parameter was randomly drawn between 6 dB and 30 dB. Four versions of this subset are generated varying the value of the background SNR parameter (resulting in different foreground-to-background SNR (FBNSR)):

- 0 dB (the FBSNR is between 6 dB and 30 dB);
- 6 dB (the FBSNR is between 0 dB and 24 dB);
- 15 dB (the FBSNR is between -9 dB and 15 dB);
- 30 dB (the FBSNR is between -24 dB and 0 dB).

In the remainder of the chapter, these subsets will be referred to as **synth_30dB**, **synth_24dB**, **synth_15dB** and **synth_0dB**, respectively. These subsets are designed

to study the impact of the FBSNR of the SED systems performance. Related results are discussed in Section 6.6.2.

Audio degradation

Six alternative versions of the subset **synth_30dB** are generated introducing artificial degradation with the Audio Degradation Toolbox [Mauch and Ewert, 2013]. The signal degradations are generated to simulate degradation faced in real environments. The following degradations are used (with default parameters) :

- **smartPhonePlayback**: Apply the response of a Google Nexus One loudspeaker and add pink noise at 40 dB SNR
- **smartPhoneRecording**: Apply the response of a Google Nexus One front microphone, apply dynamic range compression with a threshold of -35 dB and a slope of 0.5, apply clipping on 30% of the samples and add pink noise at 35 dB SNR
- **unit_applyClippingAlternative**: 10% of the samples are clipped
- **unit_applyDynamicRangeCompression**: The threshold is -40 dB and the slope is 0.9
- **unit_applyLowpassFilter**: The cut-off frequency is 800 Hz
- **unit_applyHighpassFilter**: The cut-off frequency is 1 kHz

These subsets will be referred to as **phone_play**, **phone_record**, **clipping**, **compression**, **lowpass** and **highpass**, in the remainder of the chapter. This subset is designed to study the robustness of the SED to audio degradation. Related results are discussed in Section 6.6.1.

Varying onset time

A subset of 750 soundscapes is generated with uniform sound event onset distribution and only one event per soundscape. The parameters are set such that the FBSNR is between 6 dB and 24 dB. Three variants of this subset are generated with the same isolated events, only shifted in time. In the first version, all sound events have an onset located between 250 ms and 750 ms, in the second version the sound event onsets are located between 4.75 s and 5.25 s and in the last version the sound event onsets are located between 9.25 s and 9.75 s. In the remainder of the chapter, these subsets will be referred to as **500ms**, **5500ms** and **9500MS**, respectively. This subset is designed to study of the SED segmentation to the event location in time. In particular, we wanted to verify if SED system were not learning a bias in term of time localization depending on the event length. Related results are discussed in Section 6.7.

Long sound events vs. short sound events

A subset with 522 soundscapes is generated where the background is selected from one of the five long sound event classes (Blender, Electric shaver/toothbrush, Frying, Running water and Vacuum cleaner). The foreground sound events are selected from the five short sound event classes (Alarm/bell/ringing, Cat, Dishes, Dog and Speech). Three variants of this subset are generated with the same sound event scripts and varying values of the background SNR parameter. In a first subet the resulting FBSNR is 0 dB, the FBSNR

is 15 dB is the second and 30 dB in the last subset. In the remainder of the chapter, these subsets will be referred to as **ls_0dB**, **ls_15dB** and **30_dB**, respectively. This subset is designed to study of the impact of a sound event being in the background or the foreground on SED performance [Salamon and Bello, 2015a]. Related results are discussed in Section 6.6.2.

6.4 Baseline

The baseline system is inspired by the winning system from DCASE 2018 Task 4 by JiaKai [2018].[2] It uses a mean-teacher model which is a combination of two models: a student model and a teacher model (both have the same architecture). The implementation of the mean-teacher model is based on the work of Tarvainen and Valpola [Tarvainen and Valpola, 2017]. The student model is the final model used at inference time, while the teacher model is aimed at helping the student model during training and its weights are an exponential moving average of the student model's weights. A depiction of the baseline model is provided in Figure 6.1.

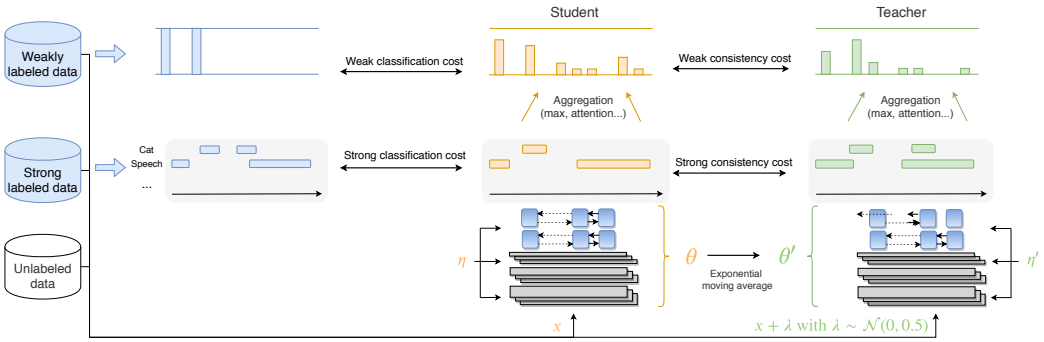


Figure 6.1: Mean-teacher model. η and η' represent noise applied to the different models (in this case dropout).

The models are a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN) followed by an aggregation layer (in our case an attention layer). The output of the RNN gives strong predictions while the output of the aggregation layer gives the weak predictions (the weights of the model are denoted θ).
The student model is trained on the synthetic and weakly labeled data. The loss (binary cross-entropy) is computed at the frame level for the strongly labeled synthetic data and at the clip level for the weakly labeled data. The teacher model is not trained, rather, its weights are a moving average of the student model (at each epoch). During training, the teacher model receives the same input as the student model but with added Gaussian noise, and helps train the student model via a consistency loss (mean-squared error) for both strong (frame-level) and weak predictions. Every batch contains a combination of

[2] Open source code available at: https://github.com/turpaultn/DCASE2019_task4/tree/public/baseline

unlabeled, weakly and strongly labeled samples.
This results in four cost components: a weak classification cost ($\mathcal{L}_{class_w}(\theta)$), strong classification cost ($\mathcal{L}_{class_s}(\theta)$), a weak consistency cost ($\mathcal{L}_{cons_w}$) and strong consistency cost ($\mathcal{L}_{cons_s}$). These costs are combined as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_{class_w}(\theta) + \sigma(\lambda)\mathcal{L}_{cons_w}(\theta) + \mathcal{L}_{class_s}(\theta_s) + \sigma(\lambda)\mathcal{L}_{cons_s}(\theta_s) \tag{6.1}$$

		Real recordings					Synthetic
		Event-based				Seg.-based	Event-based
Rank	Classifier	**Eval**	Youtube	Vimeo	Valid	Eval	synth_0dB
1	**Lin, ICT**	**42.7%**	47.7%	29.4%	45.3%	64.8%	47.6%
2	**Delphin-Poulat, OL**	**42.1%**	45.8%	33.3%	43.6%	71.4%	59.8%
3	**Shi, FRDC**	**42.0%**	46.1%	31.5%	42.5%	69.8%	53.2%
4	**Cances, IRIT**	**39.7%**	43.0%	30.9%	39.9%	64.7%	50.8%
5	**Yan, USTC**	**36.2%**	38.8%	28.7%	42.6%	65.2%	41.8%
6	**Lim, ETRI**	**34.4%**	38.6%	23.7%	40.9%	66.4%	42.5%
7	**Kiyokawa, NEC**	**32.4%**	36.2%	23.8%	36.1%	65.3%	42.3%
8	**Chan, NU**	**31.0%**	34.7%	21.6%	30.4%	58.2%	46.7%
9	**Zhang, UESTC**	**30.8%**	34.5%	21.1%	35.6%	60.9%	49.2%
10	**Kothinti, JHU**	**30.7%**	33.2%	23.8%	34.6%	53.1%	35.6%
11	**Wang B., NWPU**	**27.8%**	30.1%	21.7%	31.9%	61.6%	32.9%
12	**Lee, KNU**	**26.7%**	28.1%	22.9%	31.6%	50.2%	33.0%
	Baseline 2019	**25.8%**	29.0%	18.1%	23.7%	53.7%	40.6%
13	**Agnone, PDL**	**25.0%**	27.1%	20.0%	59.6%	60.4%	46.7%
14	**Rakowski, SRPOL**	**24.2%**	26.2%	19.2%	24.3%	63.4%	29.7%
15	**Kong, SURREY**	**22.3%**	24.1%	17.0%	21.3%	59.4%	23.6%
16	**Mishima, NEC**	**19.8%**	21.8%	15.0%	24.7%	58.7%	33.0%
17	**Wang D., NUDT**	**17.5%**	19.2%	13.3%	22.4%	63.0%	14.0%
18	**Yang, YSU**	**6.7%**	7.6%	4.6%	19.4%	26.3%	7.5%

Table 6.2: F-score performance on the evaluation sets

6.5 Submission evaluation

DCASE 2019 Task 4 obtained 57 submissions from 18 different teams involving 60 researchers overall.

6.5.1 Evaluation metrics

Submissions were evaluated according to an event-based F-score with a 200 ms collar on the onsets and a collar on the offsets that is the greater of 200 ms and 20% of the sound event's length. The overall F-score is the unweighted average of the class-wise F-scores

(macro-average). In addition, we provide the segment-based F-score on 1 s segments as a secondary measure. The metrics are computed using the sed_eval library [Mesaros et al., 2016].

6.5.2 System performance

The official team ranking (the best system from each team based on the performance on the evaluation set) along with some characteristics of the submitted systems is presented in Table 6.2. Submissions are ranked according to the event-based F-score computed over the real recordings in the evaluation set. For a more detailed comparison, we also provide the event-based F-score on the YouTube and Vimeo subsets and the segment-based F-score over all real recordings. The event-based F-score on the validation set is reported for the sake of comparison with previous year's results (75% of the 2019 validation set is comprised of the 2018 evaluation set). Performance on synthetic recordings is not taken into account in the ranking, but the event-based F-score on **synth_0dB** is presented here as well.

Twelve teams outperform the baseline. The best systems [Delphin-Poulat and Plapous, 2019, Lin and Wang, 2019, Shi, 2019] outperform the baseline by 16% points and the best system from 2018 by over 10 % points. While the ranking on the YouTube subset is similar to the official ranking, their rankings based on the Vimeo and synthetic subsets are notably different. Performance on the Vimeo set is in general considerably lower than on the YouTube set and **synth_0dB**. The fact that no data from Vimeo was used during training (unlike data from YouTube and synthetic data) suggests that the submitted systems struggle to generalize to an entirely unseen set of recording conditions.

All three top-performing teams used a semi-supervised mean-teacher model [Tarvainen and Valpola, 2017]. Lin et al. [Lin and Wang, 2019] focused on the importance of semi-supervised learning with a guided learning setup [Lin et al., 2019] and on how synthetic data can help when used together with a sufficient amount of real data. Delphin-Poulat et al. [Delphin-Poulat and Plapous, 2019] focused on data augmentation and Shi [Shi, 2019] focused on a specific type of data augmentation where both audio files and their labels are mixed. Cances et al. [Cances et al., 2019] proposed a multi-task learning setup where audio tagging (producing weak predictions) and the sound event localization in time (strong predictions) are treated as two separate subtasks [Caruana, 1997]. The latter was also the least complex of the top-performing systems.

Most of the top-performing systems also demonstrate the importance of employing class-dependent post-processing [Cances et al., 2019, Delphin-Poulat and Plapous, 2019, Lin and Wang, 2019], which improves performance significantly compared to e.g. using a fixed median filtering approach. This highlights the benefits of applying dedicated segmentation post-processing [Cances et al., 2019, Kothinti et al., 2019].

6.6 Robustness to noise and degradations

In this section we focus on the impact of the signal degradation on the SED and the êrformance with respect to the FBSNR. Only the F-score for the top performing system (on **synth_24dB**) for each submission is presented here. We restrain the analysis to the 10 top-performing systems.

6.6.1 Simulated degradations

	Degradation						
System	synth_24dB	ph_play	ph_rec.	clip.	compr.	high.	low.
Agnone, PDL	39.1%	15.4%	9.2%	14.6%	29.6%	8.5%	0.9%
Cances, IRIT	47.1%	25.7%	35.8%	42.6%	44.3%	19.2%	1.2%
Chan, NU	41.2%	25.9%	17.5%	22.8%	33.4%	19.3%	1.2%
Delphin-Poulat, OL	53.6%	32.9%	23.7%	29.5%	48.2%	23.3%	4.8%
Kiyokawa, NEC	36.8%	33.9%	21.9%	35.6%	40.2%	22.1%	4.2%
Kothinti, JHU	47.7%	30.7%	33.2%	23.8%	34.6%	53.1%	35.6%
Lim, ETRI	38.9%	26.9%	30.3%	39.7%	48.1%	15.4%	0.7%
Lin, ICT	43.7%	22.4%	9.3%	19.8%	35.3%	17.6%	0.5%
Shi, FRDC	46.4%	35.0%	36.4%	48.3%	54.1%	17.4%	4.0%
Yan, USTC	36.5%	22.1%	21.5%	18.3%	32.7%	16.6%	1.0%
Zhang, UESTC	43.7%	21.8%	15.3%	24.6%	41.4%	14.1%	1.7%

Table 6.3: F-score performance on the degraded synthetic soundscapes

The F-score obtained on the degraded subsets (see also Section 6.6.1) is presented in Table 6.3. The performance on **synth_24dB** subset are presented here for comparison purpose. The system are ordered alphabetically.
Some systems seem to have somehow over-fitted the synthetic soundscapes subset of the training set and see their performance decreased for most of the degradations. Otherwise, the trend is similar for most of the systems. The proposed systems seem to be rather robust to smartphone related degradations and compression which can be related to the fact that they have been trained on audio data extracted for Youtube and that has most probably been recorded with smartphones. On the other hand, all systems seem to be very sensitive to low-pass and high-pass filtering which tends to indicates that either having a wide-band signal is important to identify the different sound event classes proposed in DESED or the systems are sensitive to mismatch between training and evaluation. This latter hypothesis could verified by training systems on clips with only low/high frequency content.

6.6.2 Foreground-to-background Signal-to-noise ratio

Figure 6.2 presents the F-score performance for the 10 top-performing systems mentioned above under varying FBSNR (see also Section 6.3.3.2). The trend for all systems is sim-

ilar, so no submission really stands out in terms of robustness to noise. Interestingly, on **synth_15dB** where FBSNR should be distributed almost evenly around 0 dB, F-score performance are still acceptable for most systems and remain in the range of what was obtained on real recording clips [Turpault et al., 2019a]. Unsurprisingly, on **synth_0dB**, the FBSNR is always negative and the performance for all systems collapses.

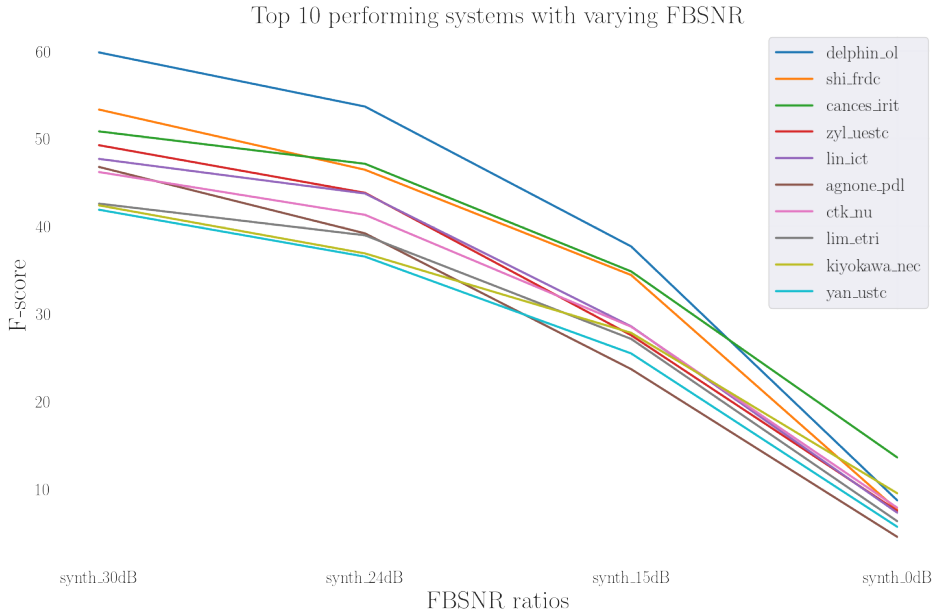


Figure 6.2: SED performance of various submitted systems depending on the FBSNR.

We then propose to analyze the systems performance when the background is actually one event from the long sound event classes and the foreground sound events are selected in the short sound event classes (see Section 6.3.3.2). Figure 6.3 presents the F-score performance for the 3 top performing systems (on this particular task) together with the performance averaged over all systems.
In all cases, when the FBSNR is low, all systems consistently obtain better performance on long sound event classes. Whereas when the FBSNR is high, all systems obtain better performance on short sound event classes. When the FBSNR is 0 dB most of the systems perform similarly on short sound event classes and long sound event classes. This tends to show that the bias toward long event classes observed in DCASE 2018 [Serizel and Turpault, 2019, Serizel et al., 2018] was less important in 2019. This is somehow confirmed by the performance on short or long sound event classes that are within the same range in the most favorable cases (0 dB FBSNR for the long sound event classes, 30 dB for the short sound event classes). However, the system proposed by Lin and Wang [2019] does not follow this trend and performs almost always better on long events. This could be due to some post-filtering that is introducing a bias toward long sound event classes (see also Figures 6.4 and 6.5).

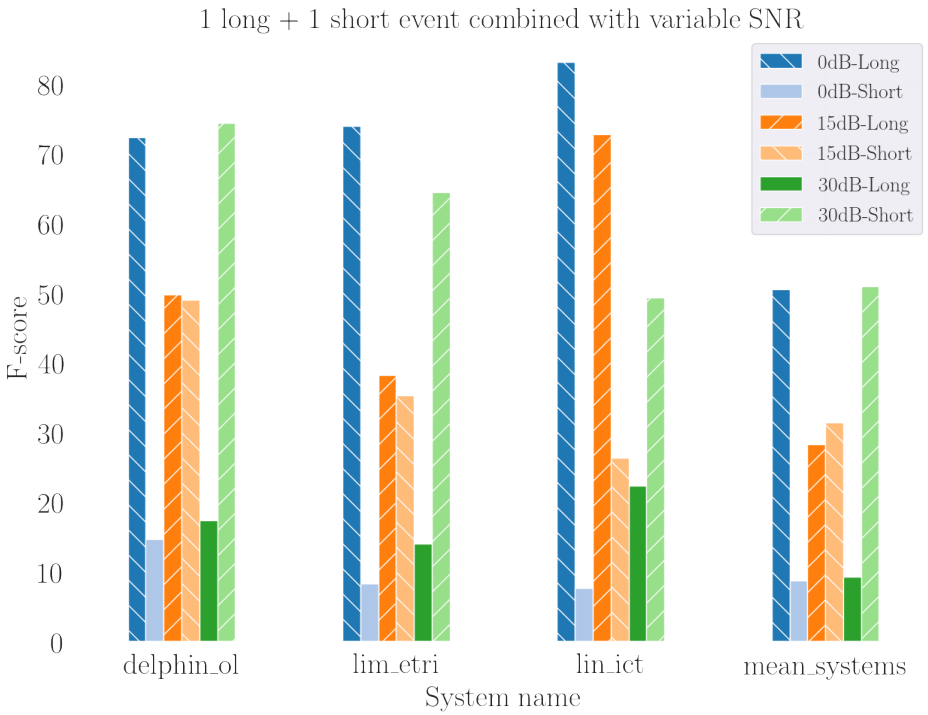


Figure 6.3: SED performance depending on the FBSNR when the soundscape is composed of a long event and a short event.

6.7 Segmentation

In this section we focus on the analysis of the systems performance in term of segmentation. In particular, we consider the scenario described in Section 6.3.3.2 in which three versions of a sound clip are generated with the same background and the same sound event starting either at the beginning, in the middle or at the end of the sound clip. The F-score performance is presented for the 3 top performing systems (on this particular task) together with the performance of the baseline system [Turpault et al., 2019a].
Figures 6.4 and Figure 6.5 present the F-score in term of segmentation (the event is properly localized in time regardless of its class), onset and offset detection (also regardless of the sound event class). Figure 6.4 presents the performance for the subset of short sound event classes and Figure 6.5 presents the performance for the subset of long sound event classes.
For short sound event classes, the F-score performance is more or less the same wherever the sound event is located within the segment. For the long sound event classes, the sound event position within the clip seems to have a large impact as performance dramatically decreases when the sound event is located towards the end of the audio clip. This could have arguably be due to the fact that in the training set, long sound events have onsets and offsets mostly located in the beginning of the audio clips. However, as shown on Figure 6.6, the onset distribution over time for long sound event classes is close to

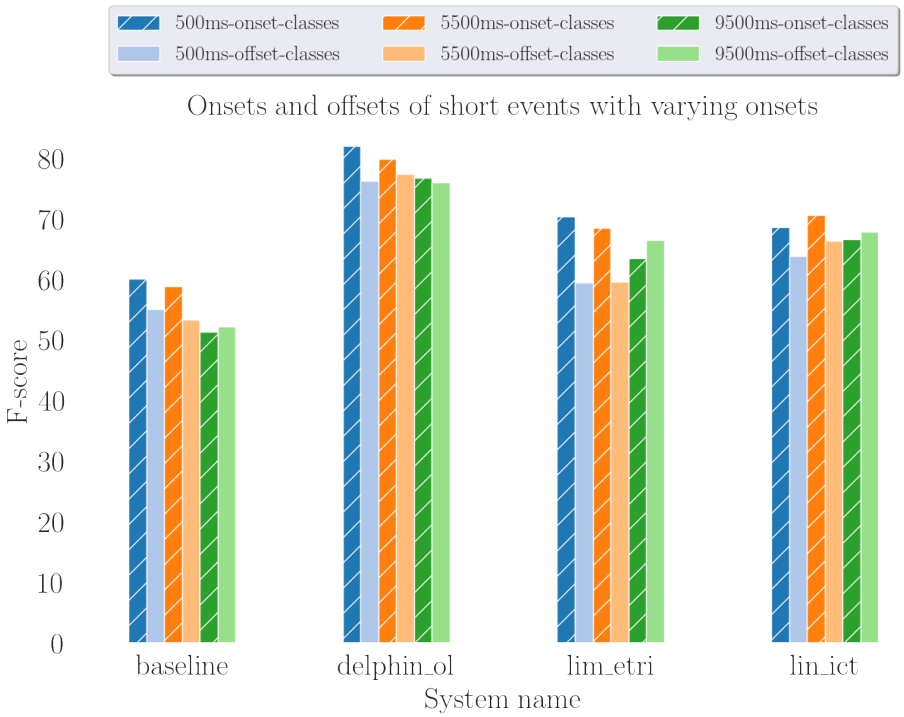


Figure 6.4: Segmentation performance depending on the event localization in time (performance for the short sound event classes).

uniform. Whereas the offsets of long sound event classes are often located toward the end of the sound clip. Therefore, if any bias was introduced by the training it should probably have led to a better offset detection for long sound event classes toward the end of the sound clips. This is somehow confirmed by the fact that all systems are able to detect quite accurately the offsets of long sound event classes when the sound event onset is located toward the middle of the sound clip. However, in this case the sound event offsets are located toward the end of the sound clip in most of the time (see also Figures 6.7 and 6.8).

One alternative explanation is that the proposed systems are simply not able to detect a long sound event class toward the end of the sound clip, not because of a bias on the classifier model but because of a bias in the post-processing. For example, median filtering with variable length that are used in most of the submissions would make it unlikely to detect a long sound event class at the end of the sound clip. This hypothesis tends to be confirmed by the fact that the baseline that is using fixed length median filtering as post-processing performs similarly wherever the sound event is located.

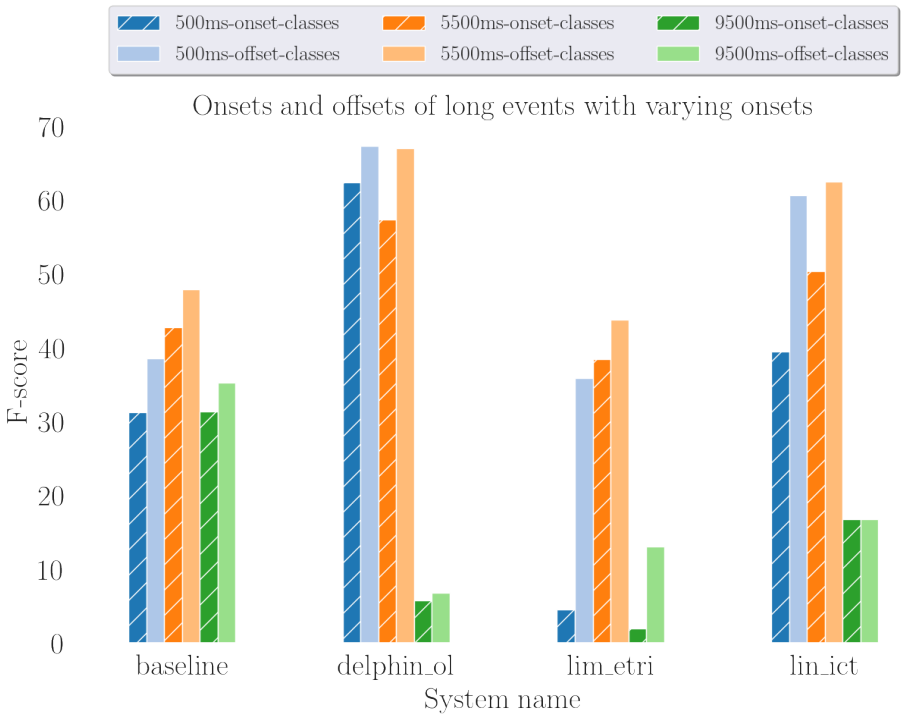


Figure 6.5: Segmentation performance depending on the event localization in time (performance for the long sound event classes).

6.8 Conclusion

This chapter presented DCASE 2019 Task 4 and the DESED dataset, which focus on SED in domestic environments. The goal of the task is to exploit a small dataset of weakly labeled sound clips together with a larger unlabeled dataset to perform SED. An additional training dataset composed of synthetic soundscapes with strong labels is provided to explore the gains achievable with simulated data. The best submissions from this year outperform last year's winning submission by over 10 % points, representing a notable advancement. We presented the performance on different subsets of the evaluation set. The performance tends to show that the proposed systems in general did not overfit the soundscape dataset and still perform well on the real dataset. However, the performance from all systems degrades in the unseen Vimeo subset which indicates a lack of ability to generalize. The progress are encouraging but the remaining limitations indicate possible directions for follow-up to this task. Evaluation on the Vimeo subset, suggests there is still a significant challenge in generalizing to unseen recording conditions.
We also presented an analysis of the performance of submissions to DCASE 2019 task 4 that were evaluated on a subset composed of synthetic soundscapes. The analysis shows that training SED on sound clips extracted from internet video makes the systems somehow robust towards degradation related to recording and playing sound on a smartphone.

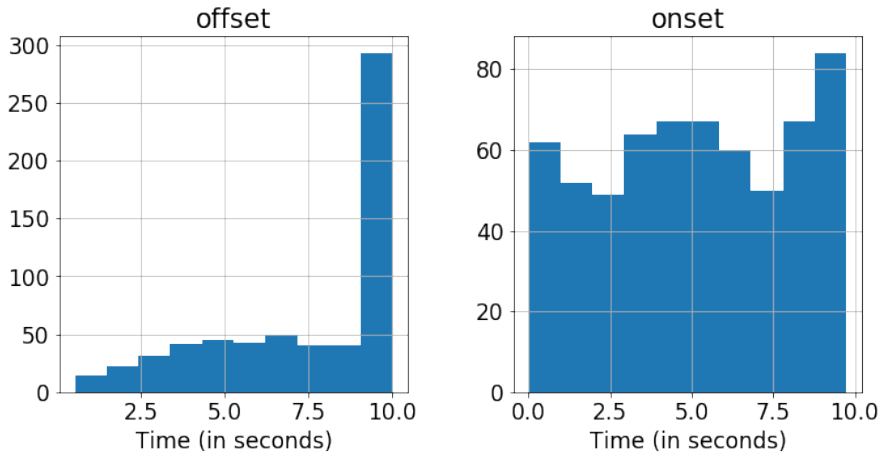


Figure 6.6: Time distribution of the onsets and the offsets in the synthetic soundscapes subset of DESED training set for the long sound event classes.

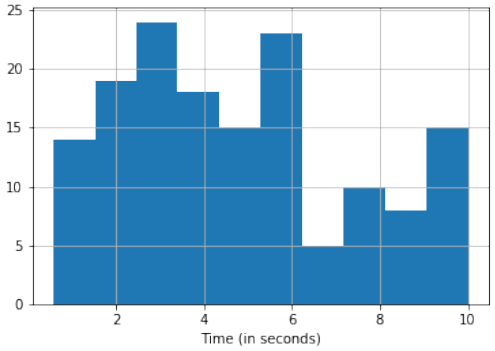


Figure 6.7: Time distribution of the offsets for long event classes in the synthetic soundscapes subsets **500ms** of DESED Evaluation set.

Additionally, we emphasize that even though performance has drastically improved since DCASE 2018 task 4 (See also Chapter 5), SED systems still rely largely on biased data (in particular for segmentation) that would probably prevent from generalizing to real case conditions.

6.9 Other related works

In 2019, we proposed an extension of the DCASE challenge task 4 as designed in 2018 where the training set is augmented with an additional set composed of soundscapes synthesized from isolated sound events of interest combined with additional noise. One advantage of this approach is that it allows for easily creating strongly labeled soundscapes and also to design specific soundscapes that can allow for targeting specific problems faced in real scenarios [Serizel et al., 2020, Turpault et al., 2021a]. The introduction of

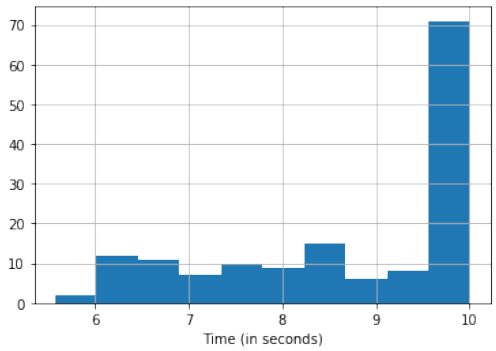


Figure 6.8: Time distribution of the offsets for long event classes in the synthetic soundscapes subsets **5500ms** of DESED Evaluation set.

these synthetic soundscapes allowed for improving the sound event detection performance but also raised the issue of how to exploit an heterogeneous training dataset [Turpault and Serizel, 2020].

In 2020 and 2021, we proposed to investigate the use of sound separation as a pre-processing to sound event detection, in particular in order to mitigate the problem of overlapping sound events [Turpault et al., 2020b, Wisdom et al., 2021]. We carefully designed synthetic evaluation sets targeting specific problems and submitted them to the participants for evaluation. This allowed us to point the impact of the sound separation processing depending on the sound to background noise ratio but also regarding the clips length, sound event density or the sound event position within the soundscape [Turpault et al., 2021b]. This work can be related to some extent to the work described in Chapter 4 and in particular to Michel Olvera's work mentioned at the end of the chapter [Olvera et al., 2021]. Additionally, based on the challenge submissions, we also proposed an analysis of the impact of the metric chosen when comparing different systems [Ferroni et al., 2021].

Until 2020 the synthetic soundscapes only included target sound events and background texture. Real recorded soundscapes also contain not target sound events that can affect the SED performance. Sound events occurrences and co-occurrences distribution were computed from a strongly labeled subset of Audioset [Hershey et al., 2021]. The impact of non target sound events was analyzed both at training and test showing the importance of having match conditions in particular regarding the target to non-target ratio [Ronchini et al., 2021].

Since October 2019, I am co-director of Mohammad Abdollahi PhD thesis with Alain Rakotomamonjy (Senior Researcher scientist at AI Criteo Lab, on secondment from Université de Rouen). The PhD takes place within the ANR PRCE project LAEUDS involving the Université de Rouen and the company Netatmo in Paris. Mohammad Abdollahi's work focuses on learning ambient sound analysis systems with limited supervision. Part of the work focused on exploiting heterogeneous dataset including isolated sound events (as DESED). The isolated sound events were used to learn class-wise representations

that could in turn be used for unsupervised segmentation and labeling of the unlabeled data. Another aspect of the work focused on model adaptation through task specific self-supervised learning. In particular, when considering a specific downstream task, Mohammad Abdollahi studied the benefits of using a large but generic corpus compared to a smaller task specific corpus.

Since December 2020, Francesca Ronchini joined the CPS4EU project as a research engineer. CPS4EU is an ECSEL project that aims at proposing solution for cyber-physical systems. Within this project our work focus on sound event detection in urban scenarios. In particular, the work addresses the problem of vehicles detection and classification. The target is to leverage the existing datasets [Mesaros et al., 2017, 2018, Zinemanas et al., 2019] to design a custom heterogeneous dataset that would allow for training a vehicle detection and classification system that generalizes to several operating conditions. The approach considered to handle the heterogeneous dataset is similar to that of the baseline of DCASE task 4 [Turpault et al., 2019a].

On a topic related to ambient sound, since February 2021, I am co-supervising Félix Gontier in his post-doctoral researches together with Christophe Cerisara (CNRS Researcher scientist, Nancy, France). Félix Gontier is working on automatic audio captioning within the ANR PRCE project LEAUDS. The work focuses on incorporating knowledge from pre-trained audio tagging models and natural language processing models within an audio captioning solution adapted to a specific corpus [Gontier et al., 2021]. Félix Gontier also investigated the generalization ability of the models across datasets depending on the audio mismatch between datasets as well as the lexical mismatch.

Since January 2021 I am co-director of François Effa's PhD thesis together with Jean-Pierre Arz (Researcher scientist at INRS in Nancy, a French research institute focusing on safety at work) and Nicolas Grimault (CNRS Researcher scientist in Lyon Neuroscience Research Center, CRNL). François Effa's work focuses on alarm detection in noisy environments. The target is to design models that can predict if an alarm would be detected given a certain alarm to background noise ratio and predict what is the alarm's salience. Current models rely mainly on psycho-acoustic and signal detection theory which can result in a few bottlenecks (for example in terms of scenarios considered). The goal here is to replace part of the current models by data-driven blocs.

7 Conclusions and Future Work

7.1 Conclusions

In this thesis, I presented an overview of the research work I have been doing with multiple collaborators during the past 10 years. The main goal here was to consider speech from the more global point of view of complex audio scenes usually faced in real-life scenarios that are currently one of the main applications of speech processing nowadays.
We first focused on addressing the variabilties related to the speaker itself. In particular we propose several approaches for automatic speech recognition with under-resourced groups of speaker such as children. Indeed, the generalization of deep learning based approaches that require a large amount of training data tends to penalize groups of speaker for which a low amount of training data is available. We proposed approaches that incorporated knowledge obtained from vocal tract length normalization to adapt models trained on well resourced population (such as male adults) to children.
At a finer granularity than addressing groups of speaker, adapting the speech recognition model to a specific person could greatly reduce the variability the system has to cope with and can be relevant in scenarios where a limited number of speakers need to interact with the system. In this case, it is also important to be able to identify the speaker in conditions that are possibly noisy when dealing with far-field speech. We proposed an approach based on supervised non-negative matrix factorization for speaker identification. The algorithm learns separate dictionaries related to the speaker identity, the acoustic conditions and the speech content within the considered audio recordings. This allows for disentangling the aspects related to the speaker identity from other acoustic variabilities therefore enforcing a more robust speaker identification in noisy conditions.
An alternative to compensating for variabilities such as noise at the model level is to apply a dedicated pre-processing such as speech enhancement in the case of noisy speech. We proposed a multichannel speech enhancement approach based on generalized eigenvalue decomposition and multichannel Wiener filter. The generalized eigenvalue decomposition of the correlation matrices of the input signals allows for projecting the signals on a subspace were it is easy to select the input channel with the highest signal-to-noise ratio. Based on this approach it is possible to derive a filter that is more robust to noise non-stationarity and that can better adjust the trade-off between noise reduction and signal distortion. Additionally, the eigenvalue decomposition can have a positive impact in binaural and ad-hoc setups where the input signal-to-noise ratio can vary importantly from one device to the other.
Our speech production and perception can be directly affected by the sounds around us and more generally so are our behavior and our actions. This is also true for speech

processing algorithms that can benefit from knowledge on the acoustic context. This is typically the type of problems addressed in ambient sound analysis. We proposed to work in particular on the fine-grained task of sound event detection where algorithms should detect not only the sound event class but also its time boundaries. In order to alleviate the problem of the costly and tedious annotations phase to obtain strongly labeled data, we proposed to exploit heterogeneous training datasets that contains unlabeled and weakly labeled recorded data, and strongly labeled synthetic data. Within the framework of a task we organized for the DCASE challenge, we proposed a detail analysis of the limitations of the state-of-the-art systems in particular regarding the exploitation of weakly labeled data and the segmentation capabilities of the systems.
From a methodological point of view, the work presented here covers approaches ranging from pure signal processing to data-driven approaches such as dictionary learning, matrix factorization or deep learning approaches. One common aspect of the work presented here is that a particular attention has been paid in keeping approaches that are centered on the signal of interest in order to avoid black box systems and to re-use previous knowledge during the design of the models or together with the models. One other related aspect that has emerged during the past years is the emphasis we put on understanding the problem itself in order to be able to propose reasonable solutions instead of trying to improve the performance at all cost using generic solution mildly adapted to the specificity of the problem. An example of this is the way we have been organizing task 4 on sound event detection for the past 4 years. We tried to go beyond the simple competition and provide each year a detailed analysis of the submissions and their limitations but also of the limitations of the metrics that were in turn benchmarked on the submissions. This analysis also motivated the evolution of the task across years in order to get a finer understanding of the problem itself and how to solve it.

7.2 Perspectives

7.2.1 Context

Over the past decades, a consensus has appeared in the scientific community on global climate change and the role we are playing in this phenomenon. The consequences of climate change are now a reality for most of us with temperature records being broken each year, recurrent droughts, storms, flooding and wildfires hitting several regions of the globe more and more frequently. A symptomatic example is the polar ice floe reaching its second lowest area ever recorded recently. One cause for this global climate change is the level of greenhouse gas emissions among which CO2 accounts for a large part. Even with the Paris Climate Agreement signed in 2015 and targeting drastic reduction of CO2 emissions by 2025, CO2 emissions have steadily increased in the past years according to the International Energy Agency [IEA, 2020].
Digital technologies have often been presented as a way to maintain a certain lifestyle while reducing CO2 [Initiative, 2020]. However, looking at the trend over the past few years and the predictions for the years to come we can hardly conclude on this. Indeed,

CO2 emissions generated by digital technologies have almost doubled since 2010 to represent almost 4% of the global CO2 emissions in 2020 and most of the scenarios predict an acceleration of the emissions in the years to come [Andrae and Edler, 2015, Arshad et al., 2017, Ferreboeuf et al., 2019]. Machine learning has often been presented as one part of the solution to the climate change problem for example by allowing for optimal energy production and supply, accelerating research in some domains such as smart city and traffic management, to name a few [Rolnick et al., 2019]. However, because of the increasingly complex models used in machine learning and the large amount of data needed to train these models, machine learning based solutions can have a dramatic impact in terms of CO2 emissions. For example, training one single model can cause as much CO2 emissions as the full life cycle of 5 cars or 300 plane trips from New York City to San Francisco [Strubell et al., 2019] and this is not even considering the latest outrageously complex models [Brown et al., 2020]. Even if a few hundred experiments are sometimes needed to train a working model, the cost of the training phase represents only 10% to 20% of the total CO2 emissions of the related machine learning usage (the rest lying in the inference phase). Yet, the cost of training machine learning models has increased exponentially, exceeding the progress in GPU energy efficiency [Biewald, 2019], and this is symptomatic of the trends in machine learning during the past years.

As humans, we constantly rely on the sounds around us to get information about our environment (birds singing, a car passing passing, the constant hum from a highway nearby...) and to get feedback about our actions (the noise of a door closing, the bips from an ATM keyboard...). Machine listening is a domain at the interface of audio signal processing and machine learning. The ultimate target is to design algorithms that can analyze and interpret sounds as a human would do. In a broad sense, machine listening encompasses all applications based on speech, music and/or ambient sounds. Among these, speech processing can find applications in hearing aids or, when combined with ambient sound processing, in home assisted living solutions. When used in applications such as noise pollution control, traffic management or monitoring of the human impact on wildlife, machine listening could also be an incredibly useful tool to raise awareness about our impact on the environment and guide us towards solutions that are less damageable. This is a first but mandatory step into acting at a local scale to mitigate the global climate change problem. However, machine listening relies heavily on machine learning techniques and it is no exception to their inherent problems [Parcollet and Ravanelli, 2021]. Models tend to be more and more complex [Battenberg et al., 2017, Défossez et al., 2019] and there is a tendency to use an increasing amount of data just because it is available, without questioning its relevance [Gemmeke et al., 2017].

In the case of speech signals, several solutions which have recently emerged in the literature operate on signals sampled at 8 kHz (i.e., the quality of voice over phone in the late 90s) because they're so complex that they become untractable on wideband speech without a large computing cluster at hand [Kolbæk et al., 2017, Xu et al., 2018]. The motivation of working on very large models that mainly apply on speech at a quality none would want to listen to is questionable. One of the current trends is often that, within a constant budget, researchers adapt the signal they are working on such that

it can accommodate with the complex models they are designing. I believe we should approach the problem the other way round: if there is any sacrifices to be made (and there are) we should adapt the models to the intrinsic properties of the physical signals we are working with.
The trends mentioned above are partly driven by the constant need to outperform the previous state-of-the-art system even by a small margin. This results in new models surfacing every few months that provide sometimes only a marginal performance improvement at the cost of increased model complexity. This way of working is similar to the planned obsolescence of digital devices: what is the cost we are ready to pay for a small increase in performance? Would these performance improvements be even perceived by end-users? Should the research in our domain be driven only by performance, with the eyes focused on one number, or should we value more systems that can maintain or improve performance albeit on a constrained budget? We cannot afford to continue like this and simply wait for the benefits of machine listening, and machine learning in general, to overcome their environmental cost. I am convinced that it is our duty, as researchers in computer science, to change the current dynamic and propose solutions for sustainable digital sciences. In order to progress towards more environmentally responsible machine listening algorithms, I propose a research program articulated around the three following axes:

- develop machine listening algorithms for applications which can help raise awareness about our impact on the environment and guide us towards more sustainable solutions.
- move towards environmentally responsible machine listening by proposing general methodologies and software tools to reduce model complexity and learn from smaller data and by more systematically assessing model energy efficiency in addition to accuracy.
- propose tools to better integrate machine listening algorithms and the hardware used in order to improve the energy efficiency further.

7.2.2 State-of-the-art

Speech and music processing has decades of history but machine listening applied to ambient sounds is a research field that emerged more recently motivated by potential applications to home automation [Debes et al., 2016, Serizel et al., 2018], home assisted living [Navarro et al., 2018] or security [Radhakrishnan et al., 2005]. In these scenarios, the advantages of machine listening compared to computer vision are that it can operate at 360°, from a distance up to a few tens of meters and that it is robust to low light conditions and visual occlusion. Machine listening applied to ambient sounds poses additional challenges compared to speech and music processing, in particular because the signals to be analyzed are not necessarily structured.
Machine listening has been applied to domains where it can have an impact on energy consumption, such has detecting occupation patterns in smart houses [Vuegen et al., 2015], or in domains where it can raise awareness about our impact on the environment, such as monitoring city and traffic noise [Bello et al., 2018b, Gontier et al., 2019]

and its impact of the remaining wildlife in cities [Fairbrass et al., 2019]. However these remain punctual experiments, as the large-scale deployment of such solutions would require energy-efficient algorithms that can work on long segments of audio, under varying acoustic conditions while preserving privacy. Most of these requirements are not met by current algorithms. Additionally, most of the current approaches are applied for analysis purposes only but they could also play an important role if included in control and design strategies [Rashidi and Cook, 2009, Silva et al., 2018].

Another domain where machine listening can have a large impact is the analysis of sounds of the nature. The analysis of sounds of the nature can help monitor species population and behavior [Lostanlen et al., 2018] and study the impact of humans on wildlife [Fairbrass et al., 2019]. Work in that direction can have a direct impact on the United Nations' Sustainable Development Goal 15. [1] Researchers and enthusiasts have been collecting sounds of the nature for many years [Ranft, 2004] but until recently most of the analysis remained manual or based on simple algorithms [Gillespie et al., 2008]. Progress has been achieved recently with methods relying on machine learning [Cramer et al., 2020, Joly et al., 2019]. However, most of the algorithms developed remain complex and the large-scale deployment of wildlife monitoring algorithms would also require energy-efficient algorithms that can run on embedded devices [Stowell and Sueur, 2020].

Over the past years, concerns about the environmental footprint of machine learning have increased. While model size is probably the most obvious cause for the large footprint of machine learning algorithms (both at training and test time), other factors can negatively affect it such as the amount of training data used and the number of experiments needed to obtain one single working model [Schwartz et al., 2020].

One solution to reduce model complexity is to train a complex model and then compress it, for example by transferring the large model to a simpler model using knowledge distillation [Cerutti et al., 2020]. Networks can also be simplified by enforcing sparsity [Louizos et al., 2017] or by pruning the neurons, weights, or channels that are least relevant [He et al., 2017]. These approaches still require a rather heavy training procedure but are efficient at reducing the energy consumption at test time. Another approach consists in factorizing the network weights in a lower dimensional space [Sun et al., 2020] or factorizing a complex task into several less complex tasks that require simpler models [Morfi and Stowell, 2018]. It is also possible to exploit knowledge about the task to be solved in order to limit the degrees of freedom of the overall system and reduce the complexity of the part that has to be learned [Aydore et al., 2019]. This has proven to be efficient in low-resource scenarios. Finally, model complexity can be further reduced by quantizing the weights [Hubara et al., 2017], sometimes up to binary weights [Qin et al., 2020]. However, the latter approaches sometime come at the cost of large performance degradation.

The other factors that can impact a system's environmental footprint are the amount of data used at training time and the number of experiments required to obtain a working system. Regarding the former aspect, there has been some work on selecting only the data that is most relevant [Kamthe and Deisenroth, 2018] or designing approaches that

[1] https://www.un.org/sustainabledevelopment/biodiversity/

show little degradation with less data and therefore allow for reducing the amount of data [Schwartz et al., 2018]. Highly complex models usually depend on several hyper-parameters to be tuned properly in order to work efficiently. Adjusting the values for these hyper-parameters requires additional experiments and the impact of these experiments is rarely taken into account when reporting system complexity [Dodge et al., 2019, Strubell et al., 2019]. There exist strategies to optimize these hyper-parameters in ways that are more efficient than grid search [Dodge et al., 2017], or strategies to stop the training procedure early if the set of hyper-parameters turns out to be suboptimal [Li et al., 2017]. Most of these approaches have been largely overlooked until now in the domain of ambient and nature sound processing. The few studies that have proposed approaches with reduced complexity were targeting the embedded implementation of a complex model simplified with knowledge distillation [Cerutti et al., 2020] or task factorization for low-resource tasks [Morfi and Stowell, 2018]. The former approach can still require a large amount of data and many experiments to train the original model. The latter approach is really tailored for one specific task but might not generalize to other setups.
Most of the approaches described are tested on generic hardware which limits the potential reduction of energy consumption. Solutions have been proposed to adapt the hardware to the quantization of the network weights [Lee et al., 2018, Pagliari et al., 2018], to exploit sparsity in networks to speed up the processing [Zhang et al., 2016], or to adjust the processor voltage depending on the computational load [Zhang et al., 2018]. Hardware manufacturers are also proposing solutions that can exploit these aspects but they are mainly limited to high-end products (at least for experimentation platforms) that are used to train already extremely complex models that are computationally expensive.[2] High performance computing shares some aspects with machine learning in terms of computational requirements and there has been some work on energy efficiency that can benefit the machine learning field [Borghesi et al., 2019, Ozer et al., 2019]. However, cross-community studies are rare, the proposed solutions are generally evaluated on simple examples that are much simpler than current models [Kang and Youn, 2019] and, to the best of my knowledge, none of these studies has targeted audio related tasks.

7.2.3 Future work

As sound can provide us with detailed information about our environment (traffic, construction work, wildlife...), I propose to work on machine listening algorithms that can both be useful in the current context of global climate change while being designed under the constraint of energy efficiency in order to make sure that their benefits overcome their cost. As the design of these approaches requires solving some more fundamental and practical issues, on the longer term, I propose to extend the work done here to other domains relying on machine learning, for example through collaboration with researchers in the domain of high performance computing, in order to achieve truly energy-efficient, application-driven machine learning algorithms. In order to address these challenges, I propose to work on three complementary research axes: one application-driven axis

[2] https://www.nvidia.com/en-us/data-center/a100/

on machine listening approaches for the environment, one more fundamental axis on energy efficient approaches for environmentally responsible machine listening and one exploratory axis on the integration between hardware and machine listening algorithms driven by application needs.

Machine listening for the environment Machine listening can play a major role in the climate change crisis by allowing for a fine-grained quantification of human impact on the environment. This is a crucial step to raise awareness about climate change and its consequences. Machine listening can also play an important role in evaluating the impact of city planning solutions and the deployment of new energy production installations on wildlife.

I propose to work on machine listening applications with a potentially beneficial impact on the environment. I propose to extend our current work on sound event detection in domestic environments [Serizel et al., 2018] to monitor occupancy and activity patterns that can be used to adjust household energy consumption [Rashidi and Cook, 2009]. At a city scale, I propose to apply machine listening to monitor traffic, people's daily trips and noise pollution in order to provide a diagnosis of the transportation patterns in a city that can be used for transportation planning at city scale or to regulate public lighting for pedestrians [Silva et al., 2018].

Human activity can have a dramatic impact on wildlife and biodiversity. Working together with bio-acousticians and biologists I propose to apply machine listening approaches to monitor the impact of human activity and road traffic in cities on wildlife. Reducing our carbon emissions will also require different (renewable) sources of energy such as solar energy or hydroelectric stations that can have an impact on wildlife too. Machine listening can help verify that this impact remains limited.

Public corpora exist in both domains that will allow for conducting medium-scale experiments. Further studies will require data collection in consultation with biologists, bio-acousticians, urbanists, acousticians and local authorities. Collecting sounds in environments visited by persons (like city street) or even just processing these recordings poses the question of respecting the privacy of the persons that are in the area of action of the recording devices. One way to solve this problem is to ensure that any data collected or used is be trimmed in order to remove any piece of information that could allow to identify people or their whereabouts.

Environmentally responsible machine listening All the applications described above require systems that can run on long time scales, possibly on embedded devices to avoid the cost of full-band audio transmission. Both of these aspects call for machine listening models that are simpler than current state-of-the-art models but still perform on par with (or close to) them.

I propose to explore approaches to simplify the neural networks used in machine listening. Most general approaches exploit intrinsic properties of the models to simplify them. In addition, I propose to exploit knowledge about the task and signal specificities to further reduce model complexity. For example, Aydore et al. [2019] rely on the fact that

the signals they are processing could be considered as images but not as translation-invariant images to impose regularizations that allow for model simplification. This could be extended to time-frequency representations of audio signals. Efficient signal representation Gontier et al. [2019], tasks and model factorization Morfi and Stowell [2018], Sun et al. [2020] have already been applied to audio to some extent but I propose to study their combination more extensively, adapting the constraints to the specific machine listening task to be solved and the energy budget allowed.
In addition to task-motivated reduction of model complexity, I propose to reduce the complexity of the datasets used for training with a two-fold approach. On the one hand, I propose to adapt active learning approaches to select only the data points that are relevant and actually improve the model performance [Kamthe and Deisenroth, 2018]. On the other hand, in order to maximize the efficiency of these approaches, I propose to design models that are more robust to a lower amount of training data [Schwartz et al., 2018]. I propose to explore strategies alternating between training data reduction and model reduction to obtain the most efficient solutions for the machine listening task at hand.
Finally, in order to quantify the progress made with the approaches proposed above, I propose to systematically report model energy efficiency instead of the sole accuracy [Schwartz et al., 2020]. I am actively involved in the animation of the machine listening community through the coordination of the flagship DCASE (Detection and Classification of Acoustic Scenes and Events) challenge series, the organization of evaluation tasks within the DCASE challenge every year, and active membership in the DCASE steering committee. I will exploit this role to propose to the community to systematically report efficiency of the machine listening algorithms developed. I will propose evaluation tasks related to the problem of environmentally responsible machine listening including investigations around low-complexity models, data efficiency and model re-usability to foster studies on this topic.

Integration of machine listening algorithms with hardware The improvements in terms of efficiency of the approaches proposed above are limited to some extent by the potential inadequacy with the hardware used. If reducing the model complexity just results in under-using a computer or a workstation that needs to be powered anyway during the training phase, then the gains in term of energy efficiency are limited. There are some solutions for efficient hardware design and control but they are mainly designed by researchers in fields related to fundamental computer science and are hardly adopted by machine listening researchers. High-level solutions (such as TensorFlow light) on the other hand usually include only limited capabilities when targeting low-consumption hardware. In order to foster a wider adoption of energy-efficient frameworks by researchers in machine listening there is a need for some pioneering work that demonstrates its feasibility and utility.
I am currently in contact with CEA List who has been developing a framework called N2D2 for fast and efficient neural network implementation. The framework allows for automatically computing the computational cost and exporting the models to several

hardware types. I will further collaborate with CEA List to adapt popular machine listening algorithm designs to the N2D2 framework. I will implement and distribute a software library that will propose efficient implementation of popular algorithms and a dedicated software overlay to ease the implementation of energy efficient versions of machine listening algorithms based on this framework. The aim is to demonstrate the efficiency of such tools on targeted machine listening applications and to provide easy access for machine listening practitioners to frameworks that allow energy efficient implementation and computational cost monitoring.

Bibliography

Abdel-Hamid, O. and Jiang, H. (2013). Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition. In *Proc. Annual Conference of the International Speech Communication Association*, pages 1248–1252.

Andrae, A. S. and Edler, T. (2015). On global electricity usage of communication technology: trends to 2030. *Challenges*, 6(1):117–157.

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., and Omologo, M. (1994). Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proc. International Conference on Spoken Language Processing*, pages 1391–1394.

Arshad, R., Zahoor, S., Shah, M. A., Wahid, A., and Yu, H. (2017). Green iot: An investigation on energy saving practices for 2020 and beyond. *IEEE Access*, 5:15667–15681.

Avdeeva, A. and Agafonov, I. (2018). Sound event detection using weakly labeled dataset with convolutional recurrent neural network. Technical report, DCASE2018 Challenge.

Aydore, S., Thirion, B., and Varoquaux, G. (2019). Feature grouping as a stochastic regularizer for high-dimensional structured data. In *Proc. International Conference on Machine Learning*, pages 385–394.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. (2012). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning Conference on Neural Information Processing Systems 2012 Workshop.

Battenberg, E., Chen, J., Child, R., Coates, A., Li, Y. G. Y., Liu, H., Satheesh, S., Sriram, A., and Zhu, Z. (2017). Exploring neural transducers for end-to-end speech recognition. In *Proc. Automatic Speech Recognition and Understanding Workshop*, pages 206–213.

Bell, P., Swietojanski, P., and Renals, S. (2013). Multi-level adaptive networks in tandem and hybrid ASR systems. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 6975–6979.

Bello, J. P., Mydlarz, C., and Salamon, J. (2018a). Sound analysis in smart cities. In *Computational Analysis of Sound Scenes and Events*, pages 373–397. Springer.

Bello, J. P., Silva, C., Nov, O., DuBois, R. L., Arora, A., Salamon, J., Mydlarz, C., and Doraiswamy, H. (In press, 2018b). SONYC: A system for the monitoring, analysis and mitigation of urban noise pollution. *Communications of the ACM*.

Benesty, J., Chen, J., and Huang, Y. (2008). Noncausal (frequency-domain) optimal filters. *Microphone array signal processing*, pages 115–137.

Benesty, J., Chen, J., Huang, Y., and Gaensler, T. (2012). Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction. *The Journal of the Acoustical Society of America*, 132(1):452–464.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al. (2007). Greedy layer-wise training of deep networks. *Proc. Advances in Neural Information Processing Systems*, 19:153.

Bertrand, A. and Moonen, M. (2010). Distributed adaptive node-specific signal estimation in fully connected sensor networks—part i: Sequential node updating. *IEEE Transactions on Signal Processing*, 58(10):5277–5291.

Biewald, L. (2019). Deep learning and carbon emissions, towards data science.

Bisot, V., Serizel, R., Essid, S., and Richard, G. (2016). Acoustic scene classification with matrix factorization for unsupervised feature learning. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6445–6449.

Bisot, V., Serizel, R., Essid, S., and Richard, G. (2017a). Feature learning with matrix factorization applied to acoustic scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1216–1229.

Bisot, V., Serizel, R., Essid, S., and Richard, G. (2017b). Leveraging deep neural networks with nonnegative representations for improved environmental sound classification. In *Proc. International Workshop on Machine Learning for Signal Processing*.

Bisot, V., Serizel, R., Essid, S., and Richard, G. (2017c). Nonnegative Feature Learning Methods for Acoustic Scene Classification. In *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events*.

Borghesi, A., Bartolini, A., Milano, M., and Benini, L. (2019). Pricing schemes for energy-efficient hpc systems: Design and exploration. *The International Journal of High Performance Computing Applications*, 33(4):716–734.

Bourlard, H. A. and Morgan, N. (1994). *Connectionist speech recognition: a hybrid approach*, volume 247. Springer.

Bousquet, P.-M., Matrouf, D., and Bonastre, J.-F. (2011). Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition. In *Proc. Annual Conference of the International Speech Communication Association*, pages 485–488.

Brown, J. (1991). Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Povey, D., Rastrow, A., Rose, R., and Thomas, S. (2010). Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 4334–4337.

Cances, l., Pellegrini, T., and Guyot, P. (2018). Sound event detection from weak annotations: Weighted gru versus multi-instance learning. Technical report, DCASE2018 Challenge.

Cances, L., Pellegrini, T., and Guyot, P. (2019). Multi task learning and post processing optimization for sound event detection. Technical report, DCASE2019 Challenge.

Carbajal, G., Serizel, R., Vincent, E., and Humbert, E. (2018). Multiple-input neural network-based residual echo suppression. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.

Carbajal, G., Serizel, R., Vincent, E., and Humbert, E. (2019). Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise: Supporting Document. Research Report RR-9303, INRIA Nancy ; Invoxia SAS.

Carbajal, G., Serizel, R., Vincent, E., and Humbert, E. (2020). Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise. *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Cerutti, G., Prasad, R., Brutti, A., and Farella, E. (2020). Compact recurrent neural networks for acoustic event detection on low-energy low-complexity platforms. *IEEE Journal of Selected Topics in Signal Processing*, 14(4):654–664.

Claes, T., Dologlou, I., ten Bosch, L., and Compernolle, D. V. (1998). A Novel Feature Transformation for Vocal Tract Length Normalisation in Automatic Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 6(6):549–557.

Cornelis, B., Moonen, M., and Wouters, J. (2010). Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1368–1381.

Cramer, J., Lostanlen, V., Farnsworth, A., Salamon, J., and Bello, J. P. (2020). Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 901–905.

Dahl, G., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42.

Das, S., Nix, D., and Picheny, M. (1998). Improvements in Children's Speech Recognition Performance. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

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

Debes, C., Merentitis, A., Sukhanov, S., Niessen, M., Frangiadakis, N., and Bauer, A. (2016). Monitoring activities of daily living in smart homes: Understanding human behavior. *IEEE Signal Processing Magazine*, 33(2):81–94.

Défossez, A., Usunier, N., Bottou, L., and Bach, F. (2019). Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Dekkers, G., Lauwereins, S., Thoen, B., Adhana, M. W., Brouckxon, H., van Waterschoot, T., Vanrumste, B., Verhelst, M., and Karsmakers, P. (2017). The SINS database for detection of daily activities in a home environment using an acoustic sensor network. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*, pages 32–36.

Delphin-Poulat, L. and Plapous, C. (2019). Mean teacher with data augmentation for dcase 2019 task 4. Technical report, DCASE2019 Challenge.

Dendrinos, M., Bakamidis, S., and Carayannis, G. (1991). Speech enhancement from noise: A regenerative approach. *Speech Communication*, 10(1):45–57.

Dinkel, H., Qiand, Y., and Yu, K. (2018). A hybrid asr model approach on weakly labeled scene classification. Technical report, DCASE2018 Challenge.

Doclo, S., Klasen, T. J., Van den Bogaert, T., Wouters, J., and Moonen, M. (2006). Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions. In *Proc. International Workshop on Acoustic Echo and Noise Control*, pages 1–4.

Doclo, S. and Moonen, M. (2002). Gsvd-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on signal processing*, 50(9):2230–2244.

Doclo, S., Spriet, A., Wouters, J., and Moonen, M. (2007). Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction. *Speech Communication*, 49(7-8):636–656.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. (2019). Show your work: Improved reporting of experimental results. *arXiv preprint arXiv:1909.03004*.

Dodge, J., Jamieson, K., and Smith, N. A. (2017). Open loop hyperparameter optimization and determinantal point processes. *arXiv preprint arXiv:1706.01566*.

Eide, E. and Gish, H. (1996). A Parametric Approach to Vocal Tract Lenght Normalization. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 346–349.

Elizalde, B., Shah, A., Dalmia, S., Lee, M. H., Badlani, R., Kumar, A., Raj, B., and Lane, I. (2017). An approach for self-training audio event detectors using web data. In *Proc. European Signal Processing Conference*, pages 1863–1867.

Ephraim, Y. and Van Trees, H. L. (1995). A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing*, 3(4):251–266.

Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 708–712.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660.

Fairbrass, A. J., Firman, M., Williams, C., Brostow, G. J., Titheridge, H., and Jones, K. E. (2019). Citynet—deep learning tools for urban ecoacoustic assessment. *Methods in ecology and evolution*, 10(2):186–197.

Ferreboeuf, H., Berthoud, F., Bihouix, P., Fabre, P., Kaplan, D., Lefèvre, L., and Ducass, A. (2019). Lean ict, towards digital sobriety. Technical report, The Shift Project.

Ferroni, G., Turpault, N., Azcarreta, J., Tuveri, F., Serizel, R., Bilen, Ç., and Krstulović, S. (2021). Improving sound event detection metrics: insights from dcase 2020. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 631–635.

Févotte, C. and Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456.

Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3):1511–1522.

Fonseca, E., Pons, J., Favory, X., Font, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., and Serra, X. (2017). Freesound datasets: a platform for the creation of open audio datasets. In *Proc. International Society for Music Information Retrieval Conference*, pages 486–493.

Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *Proc. International Conference on Multimedia*, pages 411–412.

Furnon, N., Serizel, R., Essid, S., and Illina, I. (2021a). Attention-based distributed speech enhancement for unconstrained microphone arrays with varying number of nodes. In *Proc. European Signal Processing Conference*.

Furnon, N., Serizel, R., Essid, S., and Illina, I. (2021b). Dnn-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2310–2323.

Furnon, N., Serizel, R., Illina, I., and Essid, S. (2020). DNN-Based Distributed Multichannel Mask Estimation for Speech Enhancement in Microphone Arrays. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Furnon, N., Serizel, R., Illina, I., and Essid, S. (2021c). Distributed speech separation in spatially unconstrained microphone arrays. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

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

Garcia-Romero, D. and Espy-Wilson, C. Y. (2011). Analysis of I-vector length normalization in speaker recognition systems. In *Proc. Annual Conference of the International Speech Communication Association*, pages 249–252.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.

Gerosa, M., Giuliani, D., and Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49(10–11):847 – 860.

Gerosa, M., Giuliani, D., and Brugnara, F. (2009a). Towards age-independent acoustic modeling. *Speech Communication*, 51(6):499 – 509.

Gerosa, M., Giuliani, D., Narayanan, S., and Potamianos, A. (2009b). A Review of ASR Technologies for Children's Speech. In *Proc. Workshop on Child, Computer and Interaction*, pages 7:1–7:8.

Gillespie, D., Mellinger, D., Gordon, J., Mclaren, D., Redmond, P., McHugh, R., Trinder, P., Deng, X., and Thode, A. (2008). Pamguard: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans. *Journal of the Acoustical Society of America*, 30(5):54–62.

Gillick, L. and Cox, S. (1989). Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages I–532–535.

Gillis, N. (2014). The why and how of nonnegative matrix factorization. In J.A.K. Suykens, M. S. and Argyriou, A., editors, *Regularization, Optimization, Kernels, and Support Vector Machines*, Machine Learning and Pattern Recognition Series, pages 257 – 291. Chapman & Hall/CRC.

Giuliani, D. and Gerosa, M. (2003). Investigating Recognition of Children Speech. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 137–140.

Gontier, F., Lavandier, C., Aumond, P., Lagrange, M., and Petiot, J.-F. (2019). Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques. *Acta Acustica united with Acustica*, 105(6):1053–1066.

Gontier, F., Serizel, R., and Cerisara, C. (2021). Automated audio captioning by fine-tuning bart with audioset tags. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*, pages 170–174.

Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., Mc Tait, K., and Choukri, K. (2004). ESTER, une campagne d'évaluation des systemes d'indexation automatique d'émissions radiophoniques en français. In *Proc. Journées d'Etude sur la Parole.*

Greenberg, C. S., Bansé, D., Doddington, G. R., Garcia-Romero, D., Godfrey, J. J., Kinnunen, T., Martin, A. F., McCree, A., Przybocki, M., and Reynolds, D. A. (2014). The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge. In *Proc. Odyssey*, pages 224–230.

Greenberg, C. S., Stanford, V. M., Martin, A. F., Yadagiri, M., Doddington, G. R., Godfrey, J. J., Hernandez-Cordero, J., and Meade, F. (2013). The 2012 NIST Speaker Recognition Evaluation. In *Proc. Annual Conference of the International Speech Communication Association*, pages 1971–1975.

Greenberg, J., Peterson, P., and Zurek, P. (1993). Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *The Journal of the Acoustical Society of America*, 94(5):3009–3010.

Guo, Y., Xu, M., Wu, J., Wang, Y., and Hoashi, K. (2018). Multi-scale convolutional recurrent neural network with ensemble method for weakly labeled sound event detection. Technical report, DCASE2018 Challenge.

Hagen, A., Pellom, B., and Cole, R. (2003). Children's Speech Recognition with Application to Interactive Books and Tutors. In *Proc. Automatic Speech Recognition and Understanding Workshop*.

Hamacher, V. (2002). Comparison of advanced monaural and binaural noise reduction algorithms for hearing aids. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 4.

Hamacher, V., Doering, W., Mauer, G., Fleischmann, H., and Hennecke, J. (1997). Evaluation of noise reduction systems for cochlear implant users in different acoustic environment. *The American journal of otology*, 18(6 Suppl):S46–9.

Harb, R. and Pernkopf, F. (2018). Sound event detection using weakly labeled semi-supervised data with gcrnns, vat and self-adaptive label refinement. Technical report, DCASE2018 Challenge.

He, Y., Zhang, X., and Sun, J. (2017). Channel pruning for accelerating very deep neural networks. In *Proc. International onference on computer vision*, pages 1389–1397.

Hershey, S., Ellis, D. P., Fonseca, E., Jansen, A., Liu, C., Moore, R. C., and Plakal, M. (2021). The benefit of temporally-strong labels in audio event classification. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 366–370.

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

Hinton, G., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Hou, Y. and Li, S. (2018). Semi-supervised sound event detection with convolutional recurrent neural network using weakly labelled data. Technical report, DCASE2018 Challenge.

Hu, Y. and Loizou, P. C. (2008). A new sound coding strategy for suppressing noise in cochlear implants. *The Journal of the Acoustical Society of America*, 124(1):498–509.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. (2017). Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research*, 18(1):6869–6898.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., and Johnson, K. (1999). Formants of children women and men: The effect of vocal intensity variation. *The Journal of the Acoustical Society of America*, 106(3):1532–1542.

Hurmalainen, A., Saeidi, R., and Virtanen, T. (2012). Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition. In *Proc. Annual Conference of the International Speech Communication Association*, pages 2–5.

Hurmalainen, A., Saeidi, R., and Virtanen, T. (2015a). Noise Robust Speaker Recognition with Convolutive Sparse Coding. In *Proc. Annual Conference of the International Speech Communication Association.*

Hurmalainen, A., Saeidi, R., and Virtanen, T. (2015b). Similarity induced group sparsity for non-negative matrix factorisation. In *Proc. International Conference on Acoustic, Speech and Signal Processing*, pages 4425–4429.

Hyeongi, M., Joon, B., Bum-Jun, K., Shin-hyuk, J., Youngho, J., Young-cheol, P., and Sung-wook, P. (2018). End-to-end crnn architectures for weakly supervised sound event detection. Technical report, DCASE2018 Challenge.

IEA (2020). Co2 emissions from fuel combustion: Overview. Technical report, IEA.

Imseng, D., Motlicek, P., Garner, P., and Bourlard, H. (2013). Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition. In *Proc. Automatic Speech Recognition and Understanding Workshop*, pages 332–337.

Initiative, I. S. I. (2020). A smart and sustainable world through ict. Technical report, IEEE Sustainable ICT.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):67–72.

Jansen, A., Plakal, M., Pandya, R., Ellis, D., Hershey, S., Liu, J., Moore, C., and Saurous, R. A. (2018). Unsupervised learning of semantic audio representations. In *Proc. International Conference on Acoustics, Speech, and Signal Processing.*

Jensen, S. H., Hansen, P. C., Hansen, S. D., and Sorensen, J. A. (1995). Reduction of broad-band noise in speech by truncated qsvd. *IEEE Transactions on Speech and Audio Processing*, 3(6):439–448.

JiaKai, L. (2018). Mean teacher convolution system for dcase 2018 task 4. Technical report, DCASE2018 Challenge.

Jin, Q., Schulam, P., Rawat, S., Burger, S., Ding, D., and Metze, F. (2012). Event-based video retrieval using audio. In *Proc. Annual Conference of the International Speech Communication Association.*

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.-P., Lombardo, J.-C., Planqué, R., Palazzo, S., and Müller, H. (2019). Biodiversity information retrieval through large scale content-based identification: a long-term evaluation. In *Information Retrieval Evaluation in a Changing World*, pages 389–413. Springer.

Jun, W. and Shengchen, L. (2018). Self-attention mechanism based system for dcase2018 challenge task1 and task4. Technical report, DCASE2018 Challenge.

Kamthe, S. and Deisenroth, M. (2018). Data-efficient reinforcement learning with probabilistic model predictive control. In *Proc. International conference on artificial intelligence and statistics*, pages 1701–1710.

Kang, D.-K. and Youn, C.-H. (2019). Hierarchical power control in heterogeneous hpc clusters for deep learning processing. In *Proc. International Conference on Artificial Intelligence*, pages 473–478.

Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15:1435–1447.

Kolbæk, M., Yu, D., Tan, Z.-H., and Jensen, J. (2017). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913.

Komatsu, T., Toizumi, T., Kondo, R., and Senda, Y. (2016). Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop)*, pages 45–49.

Kong, Q., Turab, I., Yong, X., Wang, W., and Plumbley, M. D. (2018). DCASE 2018 challenge baseline with convolutional neural networks. Technical report, DCASE2018 Challenge.

Kothinti, S., Imoto, K., Chakrabarty, D., Gregory, S., Watanabe, S., and Elhilali, M. (2018). Joint acoustic and class inference for weakly supervised sound event detection. Technical report, DCASE2018 Challenge.

Kothinti, S., Sell, G., Watanabe, S., and Elhilali, M. (2019). Integrated bottom-up and top-down inference for sound event detection. Technical report, DCASE2019 Challenge.

Koutini, K., Eghbal-zadeh, H., and Widmer, G. (2018). Iterative knowledge distillation in r-cnns for weakly-labeled semi-supervised sound event detection. Technical report, DCASE2018 Challenge.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 2(1):79–86.

Kumar, A. and Raj, B. (2016). Audio event detection using weakly labeled data. In *Proc. International onference on Multimedia*.

Kumar, A. and Raj, B. (2017). Audio event and scene recognition: A unified approach using strongly and weakly labeled data. In *Proc. International Joint Conference on Neural Networks*, pages 3475–3482.

Kumar, N. and Andreou, A. G. (1998). Heteroscedastic discriminant analysis and reduced rank {HMMs} for improved speech recognition. *Speech Communication*, 26(4):283 – 297.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. (2017). Fader networks: Manipulating images by sliding attributes. *arXiv preprint arXiv:1706.00409*.

Le, V.-B., Lamel, L., and Gauvain, J. (2010). Multi-style ML features for BN transcription. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 4866–4869.

Lee, C.-H. and Gauvain, J.-L. (1993). Speaker adaptation based on map estimation of hmm parameters. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 558–561.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Proc. Conference on Neural Information Processing Systems*, pages 556–562.

Lee, H. and Choi, S. (2009). Group nonnegative matrix factorization for EEG classification. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 320–327.

Lee, J., Kim, C., Kang, S., Shin, D., Kim, S., and Yoo, H.-J. (2018). Unpu: An energy-efficient deep neural network accelerator with fully variable weight bit precision. *IEEE Journal of Solid-State Circuits*, 54(1):173–185.

Lee, L. and Rose, R. C. (1996). Speaker Normalization Using Efficient Frequency Warping Procedure. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 353–356.

Lee, S., Potamianos, A., and Narayanan, S. (1999). Acoustic of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Li, L., Jamieson, K. G., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2017). Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *Proc International Conference on Learning Representations*.

Li, Q. and Russell, M. (2001). Why is Automatic Recognition of Children's Speech Difficult? In *Proc. European Conference on Speech Communication and Technology*.

Liao, H. (2013). Speaker adaptation of context dependent deep neural networks. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 7947–7951.

Lim, W., Suh, S., and Jeong, Y. (2018). Weakly labeled semi-supervised sound event detection using crnn with inception module. Technical report, DCASE2018 Challenge.

Lin, L. and Wang, X. (2019). Guided learning convolution system for dcase 2019 task 4. Technical report, DCASE2019 Challenge.

Lin, L., Wang, X., Liu, H., and Qian, Y. (2019). What you need is a more professional teacher. *arXiv preprint arXiv:1906.02517*.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (1999). The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654.

Liu, Y. L., Yan, J., Song, Y., and Du, J. (2018). Ustc-nelslip system for dcase 2018 challenge task 4. Technical report, DCASE2018 Challenge.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., and Bello, J. P. (2018). Birdvox-full-night: A dataset and benchmark for avian flight call detection. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 266–270.

Louizos, C., Welling, M., and Kingma, D. P. (2017). Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*.

Luo, Y., Ceolini, E., Han, C., Liu, S.-C., and Mesgarani, N. (2019). Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing. In *Proc. Automatic Speech Recognition and Understanding Workshop*, pages 260–267.

Mairal, J., Bach, F., and Ponce, J. (2012). Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60.

Mathieu, B., Essid, S., Fillon, T., Prado, J., and Richard, G. (2010). Yaafe, an easy to use and efficient audio feature extraction software. In *Proc. International Society for Music Information Retrieval*, pages 441–446.

Mauch, M. and Ewert, S. (2013). The audio degradation toolbox and its application to robustness evaluation. In *Proc. International Society for Music Information Retrieval Conference*, pages 83–88.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., and Virtanen, T. (2017). DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*.

Mesaros, A., Heittola, T., and Virtanen, T. (2016). Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162.

Mesaros, A., Heittola, T., and Virtanen, T. (2018). A multi-device dataset for urban acoustic scene classification. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*.

Metallinou, A. and Cheng, J. (2014). Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners. In *Proc. Annual Conference of the International Speech Communication Association*, pages 1468–1472.

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

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Morfi, V. and Stowell, D. (2018). Deep learning for audio event detection and tagging on low-resource datasets. *Applied Sciences*, 8(8):1397.

Nandwana, M. K., van Hout, J., Richey, C., McLaren, M., Barrios, M., and Lawson, A. (2019). The voices from a distance challenge 2019. In *Proc. Annual Conference of the International Speech Communication Association*.

Navarro, J., Vidaña-Vila, E., Alsina-Pagès, R. M., and Hervás, M. (2018). Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios. *Sensors*, 18(8):2492.

Ngo, K., Spriet, A., Moonen, M., Wouters, J., and Jensen, S. H. (2009). Incorporating the conditional speech presence probability in multi-channel wiener filter based noise reduction in hearing aids. *EURASIP Journal on Advances in Signal Processing*, 2009:1–11.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2):1085–1099.

Nisimura, R., Lee, A., Saruwatari, H., and Shikano, K. (2004). Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Olvera, M., Vincent, E., Serizel, R., and Gasso, G. (2021). Foreground-Background Ambient Sound Scene Separation. In *Proc. European Signal Processing Conference*.

Ozer, G., Garg, S., Poerwawinata, G., Davoudi, N., LRZ, M. D. T., Maiterth, M., LRZ, J. W., and LRZ, A. N. (2019). Energy-efficient runtime in hpc systems with machine learning. *Technical University of Munich, Data Innovation Lab*.

Pagliari, D. J., Macii, E., and Poncino, M. (2018). Dynamic bit-width reconfiguration for energy-efficient deep learning hardware. In *Proc. International Symposium on Low Power Electronics and Design*, pages 1–6.

Parcollet, T. and Ravanelli, M. (2021). The energy and carbon footprint of training end-to-end speech recognizers.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duche, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perotin, L., Défossez, A., Vincent, E., Serizel, R., and Guérin, A. (2019a). Regression versus classification for neural network based audio source localization. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*.

Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2018a). CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector. In *Proc. International Workshop on Acoustic Signal Enhancement*.

Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2018b). Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2019b). CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):22–33.

Pinto, J., Magimai-Doss, and Bourlard, H. (2009). MLP based hierarchical system for task adaptation in ASR. In *Proc. Automatic Speech Recognition and Understanding Workshop*, pages 365–370.

Poddar, A., Sahidullah, M., and Saha, G. (2017). Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*, 7(2):91–101.

Potamianos, A. and Narayanan, S. (2003). Robust Recognition of Children's Speech. *IEEE Transasctions on Speech and Audio Processing*, 11(6):603–615.

Prince, S. J. D. and Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Proc. International conference on Computer Vision*, pages 1–8.

Qin, H., Gong, R., Liu, X., Bai, X., Song, J., and Sebe, N. (2020). Binary neural networks: A survey. *Pattern Recognition*, 105:107281.

Radhakrishnan, R., Divakaran, A., and Smaragdis, A. (2005). Audio analysis for surveillance applications. In *Proc. Workshop on Applications of SignalProcessing toAudio and Acoustics*, pages 158–161.

Raj, R., Waldekar, S., and Saha, G. (2018). Large-scale weakly labelled semi-supervised cqt based sound event detection in domestic environments. Technical report, DCASE2018 Challenge.

Ranft, R. (2004). Natural sound archives: past, present and future. *Anais da Academia Brasileira de Ciências*, 76:456–460.

Rashidi, P. and Cook, D. J. (2009). Keeping the resident in the loop: Adapting the smart home to the user. *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans*, 39(5):949–959.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al. (2019). Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*.

Ronchini, F., Serizel, R., Turpault, N., and Cornell, S. (2021). The impact of non-target events in synthetic soundscapes for sound event detection. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*.

Rouvier, M., Dupuy, G., Gay, P., and Khoury, E. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In *Proc. Annual Conference of the International Speech Communication Association*.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. In *Proc. Conference on Neural Information Processing Systems*.

Saeidi, R., Hurmalainen, a., Virtanen, T., and A, V. L. D. (2012). Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification. In *Proc. Odyssey*.

Sahidullah, M., Kumar Sarkar, A., Vestman, V., Liu, X., Serizel, R., Kinnunen, T., Tan, Z.-H., and Vincent, E. (2021). UIAI System for Short-Duration Speaker Verification Challenge 2020. In *Proc. Spoken Language Technology Workshop*.

Salamon, J. and Bello, J. P. (2015a). Feature learning with deep scattering for urban sound analysis. In *Proc. European Signal Processing Conference*, pages 724–728.

Salamon, J. and Bello, J. P. (2015b). Unsupervised feature learning for urban sound classification. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 171–175.

Salamon, J., MacConnell, D., Cartwright, M., Li, P., and Bello, J. P. (2017). Scaper: A library for soundscape synthesis and augmentation. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 344–348.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Communications of the ACM*, 63(12):54–63.

Schwartz, R., Thomson, S., and Smith, N. A. (2018). Sopa: Bridging cnns, rnns, and weighted finite-state machines. *arXiv preprint arXiv:1805.06061*.

Seichepine, N., Essid, S., Fevotte, C., and Cappe, O. (2014). Soft nonnegative matrix co-factorization. *IEEE Transactions on Signal Processing*, 62(22):5940–5949.

Seide, F., Li, G., Chen, X., and Yu, D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Proc. Automatic Speech Recognition and Understanding Workshop*, pages 24–29.

Seltzer, M., Yu, D., and Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Senior, A. and Lopez-Moreno, I. (2014). Improving DNN speaker independence with I-vector inputs. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Serizel, R., Bisot, V., Essid, S., and Richard, G. (2016a). Machine listening techniques as a complement to video image analysis in forensics. In *Proc. International Conference on Image Processing*, pages 5470–5474.

Serizel, R., Bisot, V., Essid, S., and Richard, G. (2017). Supervised group nonnegative matrix factorisation with similarity constraints and applications to speaker identification. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 36–40.

Serizel, R., Essid, S., and Richard, G. (2016b). Group non-negative matrix factorisationwith speaker and session similarity constraints for speaker identification. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Serizel, R. and Giuliani, D. (2014a). Deep neural network adaptation for children's and adults' speech recognition. In *Proc. Italian Computational Linguistics Conference*.

Serizel, R. and Giuliani, D. (2014b). Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition. In *Proc. Spoken Language Technology Workshop*, pages 135–140.

Serizel, R. and Giuliani, D. (2017). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering*, 23(3):325–350.

Serizel, R., Moonen, M., Van Dijk, B., and Wouters, J. (2013). Rank-1 approximation based multichannel wiener filtering algorithms for noise reduction in cochlear implants. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 8634–8638.

Serizel, R., Moonen, M., Van Dijk, B., and Wouters, J. (2014). Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):785–799.

Serizel, R. and Turpault, N. (2019). Sound event detection from partially annotated data: Trends and challenges. In *Proc. IcETRAN conference*.

Serizel, R., Turpault, N., Eghbal-Zadeh, H., and Shah, A. P. (2018). Large-scale weakly labeled semi-supervised sound event detection in domestic environments. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*, pages 19–23.

Serizel, R., Turpault, N., Shah, A., and Salamon, J. (2020). Sound event detection in synthetic domestic environments. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 86–90.

Shi, Z. (2019). Hodgepodge: Sound event detection based on ensemble of semi-supervised learning methods. Technical report, DCASE2019 Challenge.

Silva, M., Leal, V., Oliveira, V., and Horta, I. M. (2018). A scenario-based approach for assessing the energy performance of urban development pathways. *Sustainable cities and society*, 40:372–382.

Sivadas, S. and Hermansky, H. (2004). On use of task independent training data in tandem feature extraction. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–541–4.

Snyder, D., Chen, G., and Povey, D. (2015). MUSAN: A Music, Speech, and Noise Corpus. arXiv:1510.08484v1.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 5329–5333.

Souden, M., Benesty, J., and Affes, S. (2009). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on audio, speech, and language processing*, 18(2):260–276.

Sprechmann, P., Bronstein, A. M., and Sapiro, G. (2014). Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement. In *Proc. Hands-free Speech Communication and Microphone Arrays*, pages 11–15.

Spriet, A., Moonen, M., and Wouters, J. (2004). Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction. *Signal Processing*, 84(12):2367–2387.

Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., Van Dijk, B., Van Wieringen, A., and Wouters, J. (2007). Speech understanding in background noise with the two-microphone adaptive beamformer beam™ in the nucleus freedom™ cochlear implant system. *Ear and hearing*, 28(1):62–72.

Steidl, S., Stemmer, G., Hacker, C., Nöth, E., and Niemann, H. (2003). Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer. In *Pattern Recognition, 25th DAGM Symposium*, pages 600–607.

Stolcke, A., Grezl, F., Hwang, M.-Y., Lei, X., Morgan, N., and Vergyri, D. (2006). Cross-Domain and Cross-Language Portability of Acoustic Features Estimated by Multilayer Perceptrons. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 321–334.

Stowell, D. and Sueur, J. (2020). Ecoacoustics: acoustic sensing for biodiversity monitoring at scale.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Sun, X., Gao, Z.-F., Lu, Z.-Y., Li, J., and Yan, Y. (2020). A model compression method with matrix product operators for speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2837–2847.

Swietojanski, P., Ghoshal, A., and Renals, S. (2012). Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proc. Spoken Language Technology Workshop*, pages 246–251.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Conference on Neural Information Processing Systems*, page 10.

Thomas, S., Seltzer, M., Church, K., and Hermansky, H. (2013). Deep neural network features and semi-supervised training for low resource speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 6704–6708.

Togami, M. and Kawaguchi, Y. (2014). Simultaneous optimization of acoustic echo reduction, speech dereverberation, and noise reduction against mutual interference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(11):1612–1623.

Turpault, N. and Serizel, R. (2020). Training Sound Event Detection On A Heterogeneous Dataset. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*.

Turpault, N., Serizel, R., Salamon, J., and Shah, A. P. (2019a). Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*, pages 253–257.

Turpault, N., Serizel, R., and Vincent, E. (2019b). Semi-supervised triplet loss based learning of ambient audio embeddings. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Turpault, N., Serizel, R., and Vincent, E. (2020a). Limitations of weak labels for embedding and tagging. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*.

Turpault, N., Serizel, R., and Vincent, E. (2021a). Analysis of weak labels for sound event tagging. working paper or preprint.

Turpault, N., Serizel, R., Wisdom, S., Erdogan, H., Hershey, J. R., Fonseca, E., Seetharaman, P., and Salamon, J. (2021b). Sound event detection and separation: a benchmark on desed synthetic soundscapes. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 840–844.

Turpault, N., Wisdom, S., Erdogan, H., Hershey, J. R., Serizel, R., Fonseca, E., Seetharaman, P., and Salamon, J. (2020b). Improving Sound Event Detection In Domestic Environments Using Sound Separation. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*.

Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M. (2008). The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids. *The Journal of the Acoustical Society of America*, 124(1):484–497.

Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M. (2009). Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids. *The Journal of the Acoustical Society of America*, 125(1):360–371.

Van Hoesel, R. and Clark, G. M. (1995). Evaluation of a portable two-microphone adaptive beamforming speech processor with cochlear implant patients. *The Journal of the Acoustical Society of America*, 97(4):2498–2503.

Veselỳ, K., Burget, L., and Grézl, F. (2010). Parallel training of neural networks for speech recognition. In *Text, Speech and Dialogue*, pages 439–446.

Virtanen, T., Plumbley, M. D., and Ellis, D. (2018). *Computational analysis of sound scenes and events*. Springer.

Vuegen, L., Van Den Broeck, B., Karsmakers, P., Van Hamme, H., and Vanrumste, B. (2015). Monitoring activities of daily living using wireless acoustic sensor networks in clean and noisy conditions. In *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4966–4969.

Wang, D., Xu, K., Zhu, B., Zhang, L., Peng, Y., and Wang, H. (2018a). A crnn-based system with mixup technique for large-scale weakly labeled sound event detection. Technical report, DCASE2018 Challenge.

Wang, Z., Vincent, E., Serizel, R., and Yan, Y. (2018b). Rank-1 Constrained Multichannel Wiener Filter for Speech Recognition in Noisy Environments. *Computer Speech and Language*, 49:37–51.

Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B. (1996). Speaker Normalisation on Conversational Telephone Speech. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages I–339–341.

Weiss, M. R. (1993). Effects of noise and noise reduction processing on the operation. *Journal of rehabilitation research and development*, 30(1):117.

Welling, L., Kanthak, S., and Ney, H. (1999). Improved Methods for Vocal Tract Normalization. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 761–764.

Wilpon, J. G. and Jacobsen, C. N. (1996). A Study of Speech Recognition for Children and Elderly. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 349–352.

Wisdom, S., Erdogan, H., Ellis, D. P., Serizel, R., Turpault, N., Fonseca, E., Salamon, J., Seetharaman, P., and Hershey, J. R. (2021). What's all the fuss about free universal sound separation data? In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 186–190.

Wold, E., Blum, T., Keislar, D., and Wheaten, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE multimedia*, 3(3):27–36.

Woodland, P., Odell, J. J., Valtchev, V., and Young, S. J. (1994). Large vocabulary continuous speech recognition using htk. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 125–128.

Wouters, J. and Berghe, J. V. (2001). Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system. *Ear and hearing*, 22(5):420–430.

Xu, C., Rao, W., Xiao, X., Chng, E. S., and Li, H. (2018). Single channel speech separation with constrained utterance level permutation invariant training using grid lstm. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 6–10.

Yochai, K. and Morgan, N. (1992). GDNN: a gender-dependent neural network for continuous speech recognition. In *Proc. Iternational Joint Conference on Neural Networks*, volume 2, pages 332–337.

Zhang, J., Rangineni, K., Ghodsi, Z., and Garg, S. (2018). Thundervolt: enabling aggressive voltage underscaling and timing error resilience for energy efficient deep learning accelerators. In *Proc. Design Automation Conference*, pages 1–6.

Zhang, S., Du, Z., Zhang, L., Lan, H., Liu, S., Li, L., Guo, Q., Chen, T., and Chen, Y. (2016). Cambricon-x: An accelerator for sparse neural networks. In *Proc. International Symposium on Microarchitecture*, pages 1–12.

Zhang, Z. and Schuller, B. (2012). Semi-supervised learning helps in sound event classification. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 333–336.

Zigel, Y., Litvak, D., and Gannot, I. (2009). A method for automatic fall detection of elderly people using floor vibrations and sound—proof of concept on human mimicking doll falls. *IEEE Transactions on Biomedical Engineering*, 56(12):2858–2867.

Zinemanas, P., Cancela, P., and Rocamora, M. (2019). Mavd: A dataset for sound event detection in urban environments. In *Proc. Detection and Classification of Acoustic Scenes and Events Workshop*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.