NOVEL METHODS FOR INCORPORATING PRIOR KNOWLEDGE FOR AUTOMATIC SPEECH ASSESSMENT

Thèse n. 8793 2021
présenté le 08 Juillet 2021
à la Faculté des Sciences et Techniques de l'Ingénieur
laboratoire de l'IDIAP
programme doctoral en Génie Électrique
École polytechnique fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Subrahmanya Pavankumar Dubagunta

acceptée sur proposition du jury:

Dr J.-M. Vesin, président du jury
Prof H. Bourlard, Dr M. Magimai Doss, directeurs de thèse
Dr M. Cernak, rapporteur
Dr H. Christensen, rapporteur
Prof J.-Ph. Thiran, rapporteur

Lausanne, EPFL, 2021

I do not know what I may appear to the world; but to myself I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.

- Isaac Newton.

నేను ప్రపంచానికి ఎలా కనబడతానో తెలియదు, కానీ నాకు మాత్రం ఒక సముద్రతీరాన ఆడుకుంటున్న బాలుడిలాగా ఉన్నాను; నేను నునుపైన గులకరాళ్ళను, అందమైన గవ్వలను కనుగొని ఏరుకుంటుండగా మహాసముద్రం వంటి సత్యరాశి నా ముందు కనుగొనబడక నిలిచి ఉన్నది.

– ఐసాక్ న్యూటన్.

Acknowledgements

I express my sincere gratitude to my advisor Dr. Mathew Magimai Doss, without whom this thesis and the knowledge I gained in these years would not have been possible. I thank Dr. Rob van Son for several months of collaboration and guidance that bore fruit into an entire chapter in the thesis. I thank Prof. Hervé Bourlard, Dr. Phil Garner and Dr. Petr Motlicek for their valuable feedback and suggestions in several meetings over the years that shaped parts of the thesis. I sincerely thank Alexandra Gherghina for hosting my internship at Google and for all her support, co-hosts Félix de Chaumont Quitry and Prof. Marco Tagliasacchi for their invaluable guidance, and mentor Nick Moukhine for the advice throughout the internship. I thank Dr. Aravind Ganapathiraju for all the encouragement and support he gave when I wanted to pursue a PhD. I will always be indebted to my former advisor, Prof. S. Umesh, for giving me a solid foundation in the field and in conducting research in general.

During my stay in Switzerland, several people made a positive impact on me, some in a very few and short interactions, and some over several conversations that made lasting friendships: Vinayak Abrol, Yacine Aghzaf, Alejandro Gómez Alanís, Hazel Anjith, Niccolò Antonello, Deepak Baby, Murali Karthick Baskar, Sophie Bentin, Milos Cernak, Anupama Chingacham, Subhadeep Dey, Pranay Dighe, Christian Frank, Julian Fritsch, Mark Gallagher, Anjith George, Anshul Gupta, Harshit Gupta, Weipeng He, Enno Hermann, Pierre-Edouard Honnet, Parvaneh Janbakhshi, Neethu Mariam Joy, Hande Kabil, Angelos Katharapoulos, Banriskhem Khonglah, Abbas Khosravani, Ina Kodrasi, Ketan Kotwal, Neethu Kuruvilla, Nam Le, Srikanth Madikeri, Florian Mai, François Marelli, Angel Martinez-González, Viviana Mendoza, Amir Mohammadi, Manaswini Mohapatra, Zohreh Mostaani, Hannah Muckenhirn, Skanda Muralidhar, Srikanth Nallanthighal, Alexandre Nanchen, Mahesh Nandwana, Hannah Neuser, Aishani Parida, Shantipriya Parida, Amrutha Prasad, Ravi Shankar Prasad, Tilak Purohit, Dhananjay Ram, Marzieh Razavi, Eklavya Sarkar, Bastian Schnell, Jilt Sebastian, Ravi Shankar, Shivam Sharma, Suhan Shetty, Sargam Shukla, Prabhu Teja Sivaprasad, Sunit Sivasankaran, Suraj Srinivas, Ajay Srinivasamurthy, Neha Tarigopula, Alexandre Tasso, Sibo Tong, Sandrine Tornay, Mihajlo Velimirovic, Soumya Venugopal, Esaú Villatoro Tello, Bogdan Vlasenko, Apoorv Vyas, Quan Wang, Qingran Zhan and Juan Pablo Zuluaga.

I specifically thank Laura Coppey, Frank Formaz, Sylvie Meier, Louis-Marie Plumel and Nadine Rosseau for all the administrative support. I thank Hasler Foundation for funding my PhD through the project Flexible Linguistically guided Objective Speech aSsessment (FLOSS) and

Innosuisse through the project Conversation Member Match (CMM). I thank Netherlands Cancer Institute and *speak and lunch* for a fruitful collaboration.
I thank my parents Dubagunta Nageswara Rao and Tumu Vanaja, parents-in-law Dhulipalla Umamaheswara Rao and Dhulipalla Bhanu Vyjayanthi, grandparents-in-law Dr. Padala Purnachandra Rao and Padala Vijayalakshmi for their unconditional love and support. I thank my sister-in-law, cousins and cousins-in-law for all the regular video interactions that made us feel closer to the entire family. I thank my wife Dhulipala Niharika for riding the rollercoaster with me with all the patience. Finally I thank my son Dubagunta Krithik Nayan for spicing up this journey with countless moments of fun and joy.

Abstract

Speech signal conveys several kinds of information such as a message, speaker identity, emotional state of the speaker and social state of the speaker. Automatic speech assessment is a broad area that refers to using automatic methods to predict human judgements regarding different kinds of information conveyed in speech, such as intelligibility of the spoken message, dialect and fluency of the speaker. Unlike other speech technology areas, such as automatic speech recognition, text-to-speech synthesis and automatic speaker recognition, automatic speech assessment is an emerging direction of research. One of the challenges in this field is that there is no single method or framework that scales across diverse speech assessment tasks. Thus, this thesis takes a broader outlook and focuses on prior knowledge incorporation for diverse data-driven speech assessment problems.

First, we focus on the development of end-to-end acoustic modelling methods for non-verbal cue-based speech assessment. More precisely, we develop neural network-based methods that can integrate prior knowledge about speech production to learn to assess speech from raw waveform. We validate the developed methods through investigations on several speech assessment tasks, viz. dialect identification, depression detection and speech fluency rating prediction.

Second, we focus on advancing a recently proposed phone posterior feature-based intelligibility estimation technique. Specifically, to enhance phone posterior probability estimation, we propose two novel approaches to incorporate linguistic segment level knowledge during the training of neural networks through estimation of confidence measures. We validate the two proposed approaches through automatic speech recognition and dysarthric speech intelligibility assessment studies.

Finally, in the context of privacy preservation, we develop a signal processing-based speech pseudonymization approach that alters voice source information and vocal tract system information based on prior knowledge to obfuscate the speaker identity, while retaining intelligibility, i.e. the phones and words remain recognizable. We validate the proposed pseudonymization approach through listening experiments and automatic evaluations.

Key words: Automatic speech assessment, end-to-end modelling, raw speech modelling, convolutional neural networks, source-filter decomposition, zero frequency filtering, articulatory

features, segment-level training, voice privacy, speech intelligibility, dialect identification, fluency prediction, depression detection.

Résumé

Le signal vocal transmet plusieurs types d'information tels qu'un message, l'identité du locuteur, l'état émotionnel du locuteur et l'état social du locuteur. L'évaluation automatique de la parole est un vaste domaine qui fait référence à l'utilisation de méthodes automatiques pour prédire les jugements humains par rapport aux différents types d'informations véhiculés dans la parole, tels que l'intelligibilité du message parlé, le dialecte et la fluidité du locuteur. Contrairement à d'autres domaines de la technologie vocale tels que la reconnaissance automatique de la parole, la synthèse texte-parole et la reconnaissance automatique du locuteur, l'évaluation automatique de la parole est une nouvelle direction de recherche. Un des défis de ce domaine est qu'il n'y a pas de méthode ou de cadre unique qui s'adapte à diverses tâches d'évaluation de la parole. Ainsi, cette thèse adopte une perspective plus large et se concentre sur l'incorporation des connaissances antérieures pour divers problèmes d'évaluation de la parole basés sur les données.

Tout d'abord, nous nous concentrons sur le développement de méthodes de modélisation acoustique de bout-en-bout pour l'évaluation de la parole basée sur des indices non verbaux. Plus précisément, nous développons des méthodes basées sur les réseaux de neurones qui peuvent intégrer des connaissances préalables sur la production de la parole pour apprendre à évaluer la parole à partir d'une forme d'onde brute. Nous validons les méthodes développées au travers d'investigations portant sur plusieurs tâches d'évaluation de la parole, à savoir l'identification du dialecte, la détection de la dépression et la prédiction de l'évaluation de la fluidité vocale.

Deuxièmement, nous nous concentrons sur l'avancement d'une technique, récemment proposée, d'estimation de l'intelligibilité basée sur les probabilités postérieures de phones. Plus précisément, pour améliorer l'estimation de la probabilité postérieure du phone, nous proposons deux nouvelles approches qui incorporent la connaissance des segments linguistiques lors de l'entraînement des réseaux de neurones par l'estimation de mesures de confiance. Nous validons les deux approches proposées par des études de reconnaissance vocale automatique et d'évaluation de l'intelligibilité de la parole dysarthrique.

Enfin, dans le contexte de la préservation de la vie privée, nous développons une approche de pseudonymisation de la parole basée sur le traitement du signal qui modifie les informations de la source vocale et du conduit vocal en fonction de connaissances antérieures dans le but de masquer l'identité du locuteur tout en conservant l'intelligibilité, c'est-à-dire que les

phones et les mots restent reconnaissables. Nous validons l'approche de pseudonymisation proposée par des expériences d'écoute et des évaluations automatiques.

Mots clefs : Évaluation automatique de la parole, modélisation de bout-en-bout, modélisation de la parole brute, réseaux de neurones convolutifs, décomposition source-filtre, filtrage à fréquence zéro, caractéristiques articulatoires, formation au niveau du segment, confidentialité de la voix, intelligibilité de la parole, identification du dialecte, prédiction de la fluidité, détection de la dépression.

Contents

List of Figures

List of Tables

1 Introduction

Speech is the most common mode of communication among people. During speech communication, a human listener can assess several aspects from the speech. As illustrated in Fig. 1.1, apart from the spoken content and the speaker identity, the listener can identify the spoken language and dialect, fluency and proficiency of the speaker, social state such as being sleepy, drunk, personality traits such as likeability, emotional state such as being happy, sad or angry, mental state such as being depressed, pathological conditions such as dysarthria, etc. The term *speech assessment* refers to predicting human judgements regarding different kinds of information present in speech. Speech and speaker recognition technologies have been popular in the literature for several decades; however, with the advancement of speech technologies and emergence of more resources, several other assessment tasks have been gaining interest recently.

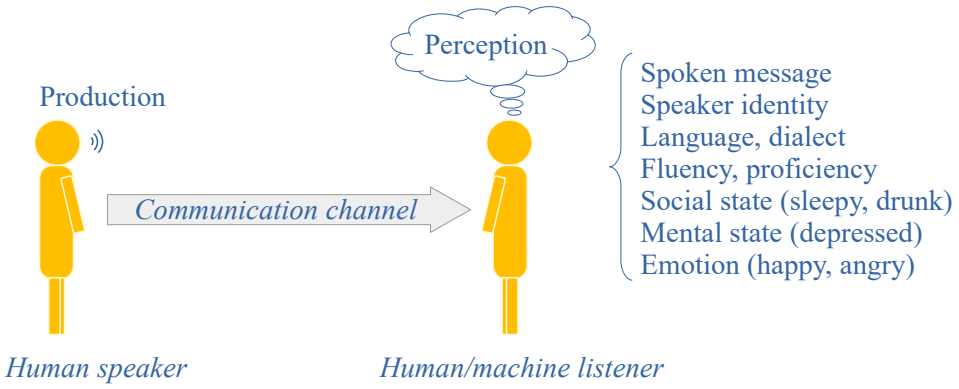


Figure 1.1: Speech assessment.

Speech assessment is essential in several human-to-machine interaction systems and human-to-human interaction systems facilitated by machines. For instance, speech received over a communication channel needs to be *intelligible*, i.e. the constituent sounds are discernible, and needs to maintain a good *quality*, i.e. contains less distortion and requires less listening

effort. Voice assistants and call centre automation systems can benefit from assessing the user's *emotion* or *satisfaction* and identify the speaker's *dialect* to provide a better interactive experience, in terms of both speech recognition and synthesis. Computer assisted language learning systems can assess the learners' *fluency* to provide tailored feedback. Car automation systems can alert the drivers of their *sleepiness*. Voice assistants and social robots can detect the presence of *pathological conditions* such as dysarthria or mental state such as *depression* for early identification and timely intervention.

Speech is conventionally assessed through subjective listening tests and human annotations. In the intelligibility test, a set of human listeners transcribe the provided speech in terms of words or syllables, which is then compared to the ground truth to compute an error rate, where a lower error implies better intelligibility. Several other assessment tasks involve human listeners annotating the utterances into two or more categories (e.g. dialect), or providing an opinion score on an ordinal scale (e.g. fluency, sleepiness). Subjective assessment requires time, funds and manual effort, relies on listener availability and lacks scalability. Opinion scores vary with factors such as the rater's preferences, nativity and demographics, and expertise in detecting pathological conditions (McHenry, 2011). Consequently, *subjective assessment is difficult to reproduce*, and there is a growing interest in machine-based automation, where the methods developed are evaluated in terms of accuracy, consistency and linearity against subjective annotations and scores (ITU-T Recommendation, 2012).

1.1 Motivation, objectives and contributions

This thesis deals with development of automatic speech assessment methods. This is an emerging area of research with increasing adoption into speech-based applications in day-to-day life. However, unlike other speech technologies such as automatic speech recognition and speech synthesis, automatic speech assessment is still an emerging area of research. There is no one single framework or method that works equally for, or scales to, different speech assessment tasks. The reason is that different speech assessment tasks tend to focus on different types of information in the speech signal. For example, intelligibility assessment methods generally focus on spoken phones and words, which are referred to as verbal cues in this thesis, whereas the other assessment problems generally focus on paralinguistic aspects such as voice quality and pauses, which are referred to as non-verbal cues. Also, we have limited knowledge and understanding about how the various pieces of information are encoded by the speech production mechanism in the speech signal and how the speech perception mechanism deciphers or decodes such information during speech communication. Verbal cues could also raise privacy concerns in certain situations such as assessment of patient interviews, thus it may sometimes be desirable to limit the investigation only to non-verbal cues. Furthermore, not all speech assessment tasks have sufficient resources to tackle the problem in a purely data-driven manner. Thus, this thesis aims at developing approaches to incorporate prior knowledge for effective data-driven automatic speech assessment.

In that direction, this thesis develops:

1. end-to-end acoustic modelling methods for speech assessment. The motivation for this research direction arises from the fact that the dominant approach for non-verbal cue-based speech assessment is based on extraction of a set of low level descriptors in a task-independent manner through short-term speech processing and application of classification/regressions methods upon them. In the recent years, it has been demonstrated that task specific information can be directly modelled from the speech signal by feeding raw waveform as input to neural networks (Muckenhirn, 2019; Palaz, 2016; Zazo et al., 2016). In this thesis, we investigate whether such end-to-end approaches can be effectively employed for speech assessment. In this direction, besides investigating such approaches that use minimal prior knowledge, we propose approaches where speech production knowledge, i.e. voice source level knowledge and vocal tract system level knowledge, is integrated through signal processing and transfer learning. We demonstrate the effectiveness of the developed approaches through dialect identification, speech fluency prediction and depression detection studies.

2. approaches to incorporate linguistic segment level knowledge, during neural network training, to enhance phone posterior probability estimation for improved speech recognition and verbal cue-based speech intelligibility assessment. More precisely, in the recent years, a phone posterior feature based approach has been developed for assessment of intelligibility and degree of nativeness (Rasipuram et al., 2016; Rasipuram et al., 2015; Ullmann, 2016). In this approach, the phone posterior feature is estimated using neural networks, which are trained with frame-level cross entropy cost function. Here a question that arises is whether the incorporation of segmental or sequence level knowledge can improve the modelling. In this direction, we propose two new confidence measure-based cost functions that incorporate phone segment level knowledge during the ANN training. We demonstrate the potential of the proposed training approaches through automatic speech recognition studies and dysarthric speech intelligibility assessment studies.

3. a signal processing-based deterministic speech pseudonymization approach for voice privacy preservation. Although this problem seems to fall within the realms of biometric security and privacy, it is strongly interconnected to speech assessment for two reasons. First, as mentioned earlier, the advancement of speech technologies and their increased usage in several environments such as homes, hospitals, corporate and banking sectors, etc. raises privacy concerns, since speech is a direct identifier. This extends to speech assessment-based applications as well. Second, anonymization methods, while obfuscating the speaker identity, can lead to loss of other pieces of information in the speech signal in an irrecoverable manner, thereby limiting their utility in speech-based applications. So, we believe that the development of speech anonymization methods is a form of a closed-loop problem, where the speech signal needs to be modified to obfuscate speaker identity information while preserving the rest of the information in

the speech signal. To attain that, automatic speech assessment is invariably needed. In this thesis, we scratch the surface of this emerging direction of research by investigating how to change, or anonymize, a speech signal reversibly by altering the vocal tract system level information and voice source level information, while preserving speech intelligibility. We demonstrate the viability of the proposed approach through listening experiments and through automatic assessment on the 2020 VoicePrivacy challenge data set.

1.2 Outline

The thesis is organised as follows.

Chapter 2, Background, provides an overview of the field of automatic speech assessment and several linguistic and paralinguistic feature sets and classifiers that have been used in the literature.

Chapter 3, Assessment tasks dealt with this thesis, details the assessment tasks studied in the thesis, with data set information and protocols.

Chapter 4, End-to-end acoustic modelling for automatic speech assessment, investigates convolution neural network-based end-to-end acoustic modelling approach with minimal prior information for several speech assessment tasks.

Chapter 5, Incorporating voice source related information, proposes a signal processing-based approach to incorporate voice source related information into the raw speech modelling based approach developed for speech assessment in Chapter 4.

Chapter 6, Incorporating linguistic prior knowledge, proposes a transfer learning based approach of incorporating articulatory feature knowledge into the raw speech modelling approach developed for speech assessment in Chapter 4.

Chapter 7, Incorporating linguistic segment level information, presents the contribution on incorporating linguistic segment-level information in the training of phone posterior estimators for improved intelligibility assessment and speech recognition.

Chapter 8, Speech pseudonymization and its assessment, presents the contribution on the development of the signal processing-based speech pseudonymization method.

Finally, Chapter 9, Conclusions and future directions, concludes the thesis along with suggesting directions for future research.

Publications based on this thesis work

Chapters 4, 5, 6:

- Dubagunta, S. P., Vlasenko, B., & Magimai.-Doss, M. (2019). Learning voice source related information for depression detection. *Proceedings of ICASSP*. http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP-2_2019.pdf
- Dubagunta, S. P., & Magimai.-Doss, M. (2019b). Using speech production knowledge for raw waveform modelling based Styrian dialect identification. *Proceedings of Interspeech*. http://publications.idiap.ch/downloads/papers/2019/Dubagunta_INTERSPEECH_2019.pdf
- Dubagunta, S. P., Moneta, E., Theocharopoulos, E., & Magimai.-Doss, M. (2021). *Towards automatic prediction of non-expert perceived speech fluency ratings* (tech. rep. Idiap-RR-11-2021). Idiap Research Institute. https://publidiap.idiap.ch/downloads//reports/2021/Dubagunta_Idiap-RR-11-2021.pdf

Chapter 7:

- Dubagunta, S. P., & Magimai.-Doss, M. (2019a). Segment-level training of ANNs based on acoustic confidence measures for hybrid HMM/ANN speech recognition. *Proceedings of ICASSP*. http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP_2019.pdf

Chapter 8:

- Dubagunta, S. P., van Son, R. J. J. H., & Magimai-Doss, M. (2020). Adjustable deterministic pseudonymization of speech: Idiap-NKI's submission to VoicePrivacy 2020 challenge [peer-reviewed at the 2020 VoicePrivacy challenge]. https://www.voiceprivacychallenge.org/docs/Idiap-NKI.pdf
- Dubagunta, S. P., Van Son, R., & Magimai.-Doss, M. (2021). *Adjustable deterministic pseudonymization of speech* (tech. rep. Idiap-RR-12-2021). Idiap Research Institute

Other publications that resulted as by-products of the thesis work

- Villatoro-Tello, E., Dubagunta, S. P., Fritsch, J., Ramírez-de-la-Rosa, G., Motlicek, P., & Magimai.-Doss, M. (2021, accepted for publication). Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition. *Proceedings of Interspeech*
- Fritsch, J., Dubagunta, S. P., & Magimai.-Doss, M. (2020). Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based CNNs. *Proceedings of ICASSP*. http://publications.idiap.ch/downloads/papers/2020/Fritsch_ICASSP_2020.pdf

- Abrol, V., Dubagunta, S. P., & Magimai.-Doss, M. (2019). *Understanding raw waveform based cnn through low-rank spectro-temporal decoupling* (tech. rep. Idiap-RR-11-2019) [peer-reviewed and presented at Swiss Machine Learning Day 2019]. Idiap Research Institute. http://publications.idiap.ch/downloads/reports/2019/Abrol_Idiap-RR-11-2019.pdf

- Gomez-Alanis, A., Gonzalez-Lopez, J. A., Dubagunta, S. P., Peinado, A. M., & Magimai.-Doss, M. (2020). On joint optimization of automatic speaker verification and anti-spoofing in the embedding space. *IEEE Transactions on Information Forensics and Security*. http://publications.idiap.ch/downloads/papers/2020/Gomez-Alanis_TIFS_2020.pdf

- Dubagunta, S. P., Kabil, S. H., & Magimai.-Doss, M. (2019). Improving children speech recognition through feature learning from raw speech signal. *Proceedings. ICASSP*. http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP-3_2019.pdf

- Vlasenko, B., Sebastian, J., Dubagunta, S. P., & Magimai.-Doss, M. (2018). Implementing fusion techniques for the classification of paralinguistic information. *Proceedings of Interspeech*. http://publications.idiap.ch/downloads/papers/2018/Vlasenko_INTERSPEECH2018_2018.pdf

- Sebastian, J., Kumar, M., Dubagunta, S. P., Magimai.-Doss, M., Murthy, H. A., & Narayanan, S. (2018). Denoising and raw-waveform networks for weakly-supervised gender identification on noisy speech. *Proceedings of Interspeech*. http://publications.idiap.ch/downloads/papers/2018/Sebastian_IS2018_2018.pdf

2 Background

Automatic speech assessment has been approached in the literature as individual classification or regression problems, with its own feature sets and classifiers. However, several of these tasks share some common approaches. This chapter is intended, and organised, to formally define the notions used in the literature related to speech assessment, give the relevant background and review the common methods.

2.1 Notions of speech assessment tasks

We associate the broad term *speech assessment* to imply several aspects and notions that exist in the literature, as follows.

Intelligibility assessment involves estimating the percentage of words or speech units recognisable.

Dialect identification involves classifying the spoken utterance into one of the known dialects of the language.

Fluency prediction involves measuring the degree of the smoothness in the flow of a person's speech in the spoken language and is also often associated with the correctness of pronunciation. This is an important aspect of language learning and spoken language communication.

Emotion recognition involves deducing the emotion from the spoken utterance.

Quality assessment can refer to multiple aspects based on the scenario: naturalness for text-to-speech systems, the degree of accent for non-native speech, comprehensibility in pathological, transmitted or coded speech or the degree of the expressed emotion.

Spoken language recognition involves classifying each utterance into one of the languages.

Sleepiness prediction involves predicting the degree of sleepiness of the speaker based on the produced speech.

Depression detection involves detecting the presence of changes in a person's speech due to the presence of depression, a mental health disorder.

paralinguistic tasks can refer to several classification or regression problems in a broad manner, such as emotion, affect, personality, likeability, sleepiness and pathological conditions (cf. Schuller & Batliner, 2021).

2.2 Literature overview

One of the earliest known studies on speech assessment was articulation testing methods proposed by Fletcher and Steinberg (1929) based on the extent to which a human ear can perceive sounds. French and Steinberg (1947) developed the *articulation index* as an intelligibility estimator, computed from the intensities of speech and unwanted sounds received by the ear, both as functions of frequency. Several modifications were proposed: a few of them were by Kryter (1962) using relative intensity of speech and noise with in different frequency bands, by House et al. (1965) using a rhyme test, and most recently by Voran (2017) using articulation band correlations. On synthesised speech, Benoît et al. (1996) used a semantically unpredictable but syntactically correct set of test words to measure intelligibility. Intelligibility is conventionally assessed through subjective listening tests. In the intelligibility test that emerged from telephony, a set of human listeners transcribe the provided speech in terms of words or syllables, which is then compared to the ground truth to compute an error rate. The lower the error rate, the better the intelligibility. Taal et al. (2011) proposed a measure, called short-time objective intelligibility (STOI), by correlating short-time temporal envelops of clean and degraded speech.

In the context of transmission and communication systems, speech was assessed using measures such as speech transmission index (Steeneken & Houtgast, 1980) based on the changes in the spectral envelop caused by a noisy channel on a modulated noise, and using speech quality per call (Berger et al., 2008). Chen (2016) used average modulation-spectrum area across bands, Elhilali et al. (2003) used spectro-temporal modulations and Hines and Harte (2012) used computational auditory models for assessment. In the quality test that emerged from telephony, human listeners rate on a subjective scale or provide their opinion, which is aggregated into mean opinion score (MOS). Methods such as perceptive objective listening quality analysis (Beerends et al., 2013) have been standardised to predict the quality of speech over networks such as voice over internet protocol (VoIP), wireless technologies such as 3G and later. Its predecessor perceptual evaluation of speech quality (ITU-T Recommendation, 2001) worked on narrow band speech. Wang et al. (1992) proposed the Bark spectral distortion, the average squared Euclidean distance between spectral vectors of the original and coded utterances. Bayya and Vis (1996) compared measures such as segmental signal-to-noise ratio, log-spectral and Itakura distances to assess speech communicated over a wireless channel.

Classical methods that recognised emotion and pathological conditions, such as depression, focused on temporal variations of speech, features such as loudness, energy in the high

frequency regions, fundamental frequency, speaking rate, pauses and voice quality (tense, breathy etc.) (Cowie et al., 2001; Low et al., 2011; Nwe et al., 2003). Wu et al. (2011) used modulation spectral features, Bou-Ghazale and Hansen (2000) used linear prediction based features to capture changes in the formant locations that indicated stress in speech. Schmitt et al. (2016) used multiple low level features called bag-of-audio-words to classify emotions. Several challenges were organised to assess speaker traits, such as Ringeval et al. (2019), Ringeval et al. (2017), and Schuller et al. (2012, 2013, 2018, 2019, 2020), Schuller et al. (2009), Valstar et al. (2016). Beyond spectral methods, Middag et al. (2008) used confidence scores based on phonemic and phonological features to assess pathological speech intelligibility.

2.3 Standard approaches

Several speech assessment tasks are typically carried out by (i) extracting short-time feature representations from speech, (ii) aggregating the features to obtain fixed length representations at the utterance or speaker level, and (iii) building classifiers on the fixed length representations. In this section, we will discuss each of these steps in detail.

2.3.1 Short-time feature representations

Conventionally, several hand-crafted features, called low-level descriptors (LLDs) have been used for speech assessment. In the recent years, unsupervised neural representations have gained popularity. We will discuss both these approaches briefly.

2.3.1.1 Low level descriptors

LLDs are a generic set of features. As part of the Computational Paralinguistics (ComParE) challenge, several sets of LLDs have been proposed (cf. Schuller et al., 2013, 2016). In this thesis, we use the extended Geneva minimalistic acoustic parameter set (eGeMAPS) (Eyben et al., 2016a, 2016b), that comprise several short-time features that correspond to the vocal source and tract, as listed in Table 2.1. Such features are typically used in paralinguistic and other tasks (Eyben et al., 2016a; Haider et al., 2020; Neumann & Vu, 2017; Wagner et al., 2018; Xue et al., 2019).

2.3.1.2 Neural embeddings

More recently, the use of neural embeddings as feature representations has emerged in the literature. Baevski et al. (2020b) used representations obtained by passing raw speech through multiple convolutional and self-attention layers, whose parameters are learned to predict quantised representations among a set of distractors, as illustrated in Fig. 2.1. This approach is referred to as wav2vec 2.0. Neural networks have also been employed to directly extract fixed length embeddings at the utterance-level: we will review them in Sec. 2.3.2.4.

Table 2.1: Low level descriptor features. (See Eyben (2016) for detailed explanations.)

Source-related	*System-related*
Loudness	Alpha ratio
F0 semitone from 27.5 Hz	Hammarberg index
Jitter	Spectral slopes (0-500, 500-1500)
Shimmer	Spectral flux
HNR (dB)	F1 (freq, bw, ampLogRelF0)
logRelF0-H1-H2	F2 (freq, ampLogRelF0)
logRelF0-H1-A3	F3 (freq, ampLogRelF0)
	MFCC (1-4)

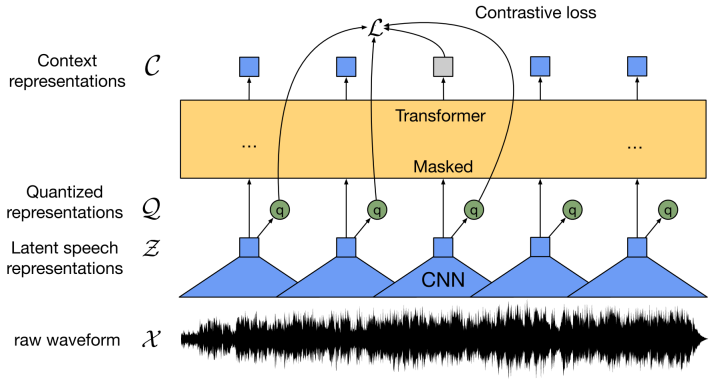


Figure 2.1: Extraction of wav2vec 2.0 embeddings. (Illustration reproduced with permission from Baevski et al., 2020b)

2.3.2 Feature aggregation at utterance/speaker level

The short-time feature representations are aggregated at the utterance or speaker levels using one of the following methods.

2.3.2.1 Functionals

The statistical properties such as mean, standard deviation, skewness and kurtosis of the short-time features are computed at the utterance level and used as representations (Eyben et al., 2016a). These are known as *functionals* in the literature.

2.3.2.2 Bag of audio words

Histogram representations, also known as bag of audio words (BoAW), can be obtained by (i) vector quantising the short-time representations across a large set of utterances to obtain cluster centroids, or *audio words*, (ii) replacing each frame in an utterance by its closest audio word and measuring the relative counts of the audio words, as illustrated in Fig. 2.2. Depending

on the short-time feature representations, BoAW could capture the relative counts of events that occur in each utterance. For instance, formant-related spectral envelope features could yield phonemic audio words, and their histogram representations could capture the relative counts of the pronounced phones, pauses and silences.

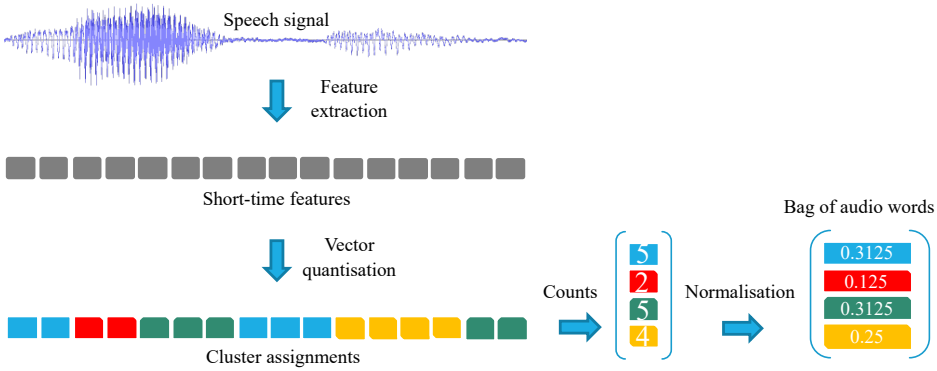


Figure 2.2: Extraction of bag of audio words.

2.3.2.3 iVectors

Factor analysis has been used as a feature extractor (Dehak et al., 2010) using generative modelling. The features, popularly known as iVectors, have been used for several speech assessment tasks such as intelligibility assessment (Martínez et al., 2013; Martínez et al., 2015), language and dialect identification (Bahari et al., 2014; Malmasi & Zampieri, 2017) emotion recognition (Lopez-Otero et al., 2014a; Xia & Liu, 2016) and depression detection (Lopez-Otero et al., 2014b).

2.3.2.4 Neural fixed-length representations

Neural networks have been employed as utterance-level fixed length feature extractors.

Initially developed in the context of speaker recognition (Snyder et al., 2018), x-vectors are discriminative embeddings extracted from neural networks. The network consists of a series of time delay neural network (TDNN) layers, followed by *statistics pooling* and feedforward layers, including a bottleneck layer and a final softmax layer for speaker classification. The network is trained by passing mel frequency cepstral coefficient (MFCC) observations to classify speakers, and embeddings from the trained bottleneck layer are used as x-vectors. Such representations have been adapted and used in assessment tasks such as intelligibility estimation in pathological speech (Quintas et al., 2020), dialect identification (Hanani & Naser, 2020) and detection of Parkinson's disease (Moro-Velazquez et al., 2020).

Through unsupervised learning using autoencoders, Amiriparian et al. (2017) and Freitag et al.

(2017) learned representations employing recurrent neural networks. Mel-spectrograms that are thresholded using a signal-to-noise ratio (SNR) parameter are used as feature representations. The hidden state vector obtained after processing the last time frame is used as the utterance-level fixed-length feature representation.

2.3.3 Modelling

The models employed for speech assessment typically range from simple linear support vector machines (SVMs) to neural networks. Feature representations extracted using neural networks typically use simpler models such as linear SVMs. x-vector based systems typically use a probabilistic linear discriminant analysis (PLDA) based classification. This is a hypothesis testing approach, where the model for each class estimates the log-likelihood ratio score of whether the test observation belongs to the class or not. Prediction is based on the class with the highest score. It is also common to treat ordinal regression tasks as classification problems and use a cross-entropy loss.

2.3.4 Handling issues with using neural networks

Below are a few issues that occur while employing neural networks and the literature on how to address them. We utilise some of these techniques later in the thesis.

2.3.4.1 Resource scarcity

Neural network models typically require comparably more resources. Data augmentation techniques can be employed to partially overcome resource scarcity. New versions of the speech recordings or features can be creating by perturbing the playback speed (Ko et al., 2015) and volume (Peddinti et al., 2015), adding noise and reverberation (Ko et al., 2017), and masking random blocks of time steps and frequency channels if using spectrogram representations (Park et al., 2019). However, the usage of any such method requires caution, especially if it is known to alter the label of the utterance in the given task. For instance, the speed of pronunciation influences how listeners rate speech fluency, and hence speed perturbation may not be recommended when modelling to automatically predict fluency.

2.3.4.2 Class imbalance

Data imbalance between classes could bias the neural networks towards making predictions according to the distribution (relative frequencies) of the classes seen during training. To overcome this, data of the under-represented classes could be replicated during training.

2.3.4.3 Overfitting

Techniques such as dropout, residual connections, regularisation, reducing the learning rate based on loss computed on a held-out cross-validation set could be used to prevent overfitting.

2.3.5 Evaluation

Binary classification tasks such as depression detection are evaluated using F1 score, precision and recall. Using the notion of positive and negative class labels and predictions, *precision* refers to the fraction of the positively predicted test samples that have positive labels, and *recall* refers to the fraction of the positive labelled test samples that were predicted positive. *F1 score* is the harmonic mean of precision and recall.

For multi-class classification tasks, percentage accuracy is used. For the tasks with class imbalance, unweighted average recall (UAR) is used.

For ordinal regression tasks, such as predicting a continuous rating, Pearson's and Spearman's-rank correlations are used. Pearson's correlation evaluates the linear relationship between two variables: in our case, the label and the prediction. It is defined as the covariance of the two variables, normalised by the product of their standard deviations. Spearman's rank correlation evaluates the monotonic relationship between two variables, i.e. by ignoring any differing rates of change. It is defined as the Pearson's correlation of the rank values of the two variables.

2.4 Summary

In this chapter, we briefly discussed several research areas in the literature that fall under automatic speech assessment. Most of them commonly employ feature-classifier systems. The features are either handcrafted or neural embeddings, that are aggregated at the utterance level using statistics or histogram representations. Classifiers are based on either linear SVMs or neural networks. Depending on the task, the systems are typically evaluated using F1 score, percentage accuracy, UAR or correlation.

3 Assessment tasks dealt with this thesis

The previous chapter discussed several areas of research on automatic speech assessment. Since this is a wide problem, this thesis focuses on specific assessment problems. This chapter provides a brief introduction to them, elaborating on the respective literature review, listing the data sets used, and finally discussing the experimental protocols followed.

3.1 Tasks dealt with the thesis

This section introduces the assessment tasks we will focus on.

3.1.1 Dialect identification

Dialect identification (DID) aims at distinguishing the acoustic, pronunciation and grammatical variations within a language used by people usually from different demographic regions. It is useful in customising automatic speech recognition systems which underperform due to changes in the dialects, in identifying a person's regional origin and ethnicity in forensic analysis (Biadsy et al., 2010), and in tailoring speech synthesis systems for improved user experience. DID is generally approached from the linguistic differences. DID is considered harder to solve than language identification, as dialectal differences within a language are generally more subtle than those between languages. To cite a few works in this direction, Chen et al. (2011) proposed to learn phonetic rules for DID, Tong et al. (2011) learned n-gram statistics of phones and used lattice rescoring for DID, Najafian et al. (2018) performed phonotactic based DID using convolutional neural networks (CNNs).

In terms of the acoustic differences, DID is closely related to accent identification from speech. In this direction, some works have used iVectors and bottleneck features (Ali et al., 2016) from acoustic data, Eigen channel modelling based on factor analysis (Lei & Hansen, 2009), Gaussian mixture model (GMM) based supervectors representing phone segments (Biadsy et al., 2010; Biadsy et al., 2011). Traditional feature sets for DID included shifted delta coefficients (Zhang & Hansen, 2017).

DID without linguistic resources and using limited speech data has been less studied. Zhang and Hansen (2017) addressed the lack of linguistic resources by using unsupervised learning, i.e. by first modelling data using GMMs and thereby training neural networks to predict the posterior probabilities of these unsupervised models and through further processing. Other works such as Schuller et al. (2019) and Zhang and Hansen (2018) also focused on unsupervised representations for DID. Shon et al. (2018) proposed an end-to-end framework from features such as spectrogram in an unsupervised learning setting, based on factorised hierarchical variational autoencoders. They addressed the acoustic resource scarcity by augmenting data through speed perturbation (Ko et al., 2015), which varies the speed of the original signals to create two more of their variants at 0.9 and 1.1 relative speeds.

3.1.2 Fluency prediction

Technologically, speech fluency estimation has been approached in the context of computer aided spoken language learning and testing. Several existing methods predict fluency automatically in a reference-based setting by comparing the utterance under test to a predefined reference in terms of its linguistic content and estimating a score. For example, using speech recognition system to estimate the number of correct words per minute (Kelly et al., 2020; Loukina et al., 2019) and phoneme-level goodness of pronunciation (Yarra et al., 2019). Vocal source characteristics have also been studied, such as comparing the prosody contour to a reference (Xiao & Soong, 2017). Fontan et al. (2018) used automatic segmentation techniques and formant tracking to compute similar features without the use of speech recognition. In a no-reference setting, where only the expert mean opinion scores (MOS) of perceived fluency are available, Mao et al. (2019) studied directly predicting the MOS scores using standard machine learning techniques on *fluency feature vectors*, which constitute pause durations, pause similarity scores based on their positions and durations w.r.t. a predefined set of references, estimated syllable speaking rate and pronunciation quality values.

As pointed out above, speech fluency prediction is generally approached from language learning and testing perspective. However, this can be questioned in a more informal or social settings. For instance, in spoken communication, perceived speech fluency may have an impact on the interaction and/or on other aspects such as, forming impressions about the person. Speech fluency prediction in such a context has certain differences when compared to language learning and testing. First, in language learning and testing, speech fluency prediction is part of a broader aspect, more precisely, *proficiency* assessment, which also includes *linguistic accuracy*, i.e. the correctness of syntax and vocabulary (Duijm et al., 2018). Second, the assessment system is developed to predict a score that best correlates with *expert* ratings. In the literature, it has been found that native experts and non-experts tend to rate differently (Duijm et al., 2018). In particular, non-expert raters tend not to focus much on linguistic accuracy aspects. This thesis focuses on the automatic prediction of perceived speech fluency from non-expert ratings, and investigates whether such non-expert ratings can be predicted automatically in a consistent manner. To the best of our knowledge, this

question has not been addressed before.

3.1.3 Depression detection

Humans convey their mental state through vocal, linguistic and facial gestures. Depression is one such phenomenon, whose automatic detection and severity assessment have gained interest in recent years (Cummins et al., 2015; Valstar et al., 2016). These tasks have been carried out in the literature by measuring parameters from patient interview sessions using multiple modes: audio, video and text, and by using appropriate classification/regression tasks (Al Hanai et al., 2018; Dibeklioglu et al., 2018). Purely speech based analyses continue to perform worse than multi-modal techniques (Valstar et al., 2016), indicating the need for further research in the field.

Various speech features have been shown to be indicative of depression. Depression is known to affect human speech production and cognitive processes: it impacts speech motor control (Cummins et al., 2015; Scherer et al., 2016). Neurophysiological changes can occur, which in turn may affect the laryngeal control and its dynamics, i.e. the behaviour of the vocal folds (Caligiuri & Ellwanger, 2000; Cummins et al., 2015; Ozdas et al., 2004; Quatieri & Malyska, 2012; Sobin & Sackeim, 1997). Voice quality has been shown to be affected (Afshan et al., 2018; Hönig et al., 2014; Sahu & Espy-Wilson, 2016; Scherer et al., 2013; Simantiraki et al., 2017), and various voice source related features such as jitter, shimmer, degree of breathiness, prosodic abnormalities, shape of the glottal pulse and glottal flow characterisation have been proposed for depression detection (Kent & Kim, 2003; Ozdas et al., 2004; Quatieri & Malyska, 2012). Depression was also shown be identified by articulatory and phonetic errors (Kent & Kim, 2003). Since depression can sometimes be associated with negative emotions, there have been features motivated from speech emotion recognition research such as Gupta et al. (2017) and Stasak et al. (2016). However expressing negative emotions is different from having a depressed mental condition. Multiple works have used functionals of LLD features (see Sec. 2.3.2.1) that are related to both the vocal-source and vocal-tract to improve the systems (Al Hanai et al., 2018; He & Cao, 2018; Stasak et al., 2016); however not all the statistical properties contribute to the improvements. Despite these advances, there seem to be no concurred set of features for detecting depression from speech signals; and moreover, the performances of all these systems may be limited by the choice of features and their statistics. More recently, deep learning methods have been investigated. For instance, Ma et al. (2016) proposed predicting depression using neural networks comprising convolutional and long-short term memory layers on log Mel filter-bank (LMFB) and magnitude-spectrogram features.

3.1.4 Objective intelligibility assessment

Intelligibility refers to the percentage of words or speech units recognisable to native speakers with healthy hearing and cognition (Möller et al., 2011). It is assessed using several methods by recognising the sound units using ASR and thereby counting the errors (Schuster et al., 2005;

Schuster et al., 2006), using spectral methods such as STOI (Taal et al., 2011) and their variants (Janbakhshi et al., 2019a; Jensen & Taal, 2016), factor analysis based methods (Martínez et al., 2013; Martínez et al., 2015) and using posterior probability based approaches (Soldo et al., 2012; Wang et al., 2012).

In this thesis, we are interested in objective intelligibility measure based on posterior based approach. In this approach, the utterance under test is compared against a reference utterance in the phone posterior probability feature space, employing an acoustic model. This approach was first demonstrated in the context of using synthetic speech for template-based ASR using posterior features (Soldo et al., 2012) and then extended to speech intelligibility assessment by Ullmann, Magimai.-Doss, et al. (2015). The method consists of estimating sequences of phone posterior probabilities corresponding to the reference speech and the test speech, and comparing the two sequences using dynamic time warping (DTW) with a local score based on Kullback Leibler (KL) divergence (Soldo et al., 2012; Soldo et al., 2011). The approach is described in detail in Sec. 7.1.

3.2 Data sets and protocols

This section lists the data sets and protocols used for the speech assessment tasks.

3.2.1 Styrian dialect identification

The ComParE 2019 Styrian dialect data set (Schuller et al., 2019) consists of speech corresponding to three major Austrian German dialects spoken in Styria: the Northern, Eastern and Urban variants. It is a resource-constrained data set of short utterances, with 76 minutes of training data, with an average duration 0.87 seconds per utterance, 34 minutes of development (dev) data and 29 minutes of test data. The train and dev sets were provided with utterance-level dialect labels, and those of the test set were reserved with the organisers. No other linguistic resources, such as transcriptions, were available. Only 5 systems from each participating team were allowed to be submitted for the test set score evaluation.

Styrian DID is a classification problem using the acoustic signals as input. Since the three classes were imbalanced, unweighted average recall was proposed by the organisers as the evaluation measure. To partially overcome the resource scarcity, we conducted data augmentation using speed perturbation (SP), as also suggested by Shon et al. (2018), by duplicating the utterances by randomly altering the playback speed within the range of 0.9 to 1.1 times the original speed.

3.2.2 Arabic dialect identification

The ADI17 data set (Shon et al., 2020) consists of 3033 hours of Arabic YouTube speech data corresponding to 17 dialects, each from a different country, and an additional 58 hours of

speech divided into dev and test sets. The data set was initially a part of MGB-5 challenge (Ali et al., 2020). The organisers provided segmented utterances from the original data set, where each utterance in the train, dev and test partitions is given a single dialect label. Similar to the Styrian data set, transcriptions were not available. Percentage accuracy was proposed by the organisers as the evaluation measure.

3.2.3 Fluency prediction

The fluency prediction work was carried out in collaboration with *speak and lunch*[I], where our goal was to automatically predict fluency ratings from non-expert raters, as mentioned in Sec. 3.1.2. Since there is currently no such data set, we collected a new speech data set, which consists of read speech and speech on a topic of interest to the volunteers and rated the speech fluency with non-expert raters. The data were collected in three different countries: Switzerland, Greece and the USA (city of New York). The project collaborator *speak and lunch* mainly went in medium sized international companies as well as social gatherings and asked for volunteers to participate in the project. The volunteers who agreed to participate were provided with an informed consent form to sign. Each of the participants were then provided with an iPod or an iPhone with headphones and were asked to make audio or video recordings, as per their preference, of all the languages they spoke (whether fluent, intermediate or beginner level). They were asked to (i) make 4 recordings of minimum 15 seconds where they would speak about a topic of their choice, and (ii) read a phonetically balanced text, viz. the Northwind passage. Out of the 54 participants, 29 were women and 25 were men. The participants' age ranged between 25 and 75 years. They were from different nations, viz., Albania, France, Greece, Italy, Mexico, Portugal, Russia, Spain, Switzerland and Turkey.

The final collected data set comprises 187.36 minutes of data from 54 speakers, of which 144.14 minutes corresponds to English recordings, which we used in our analysis. On average, each speaker had about 2-4 minutes of speech. These recordings were then rated by seven raters (4 women and 3 men), aged between 37 and 75 years old. The raters were fluent English speakers, who are active professionals in the law and banking sector in the USA and Switzerland. The raters were asked to rate each audio or video recording on a 5-point Likert scale, with 1 being beginner and 5 being fluent. The Krippendorff's alpha coefficient for the ratings was found to be 0.584. The median values per each speaker were used as reference scores in our experiments.

We employed 10-fold validation with non-overlapping speakers in all the experiments on this data set. Specifically, the speakers were split into 10 parts, where the system building involves 9 parts and evaluation is on the 10^{th} part. Performance was measured by computing Pearson's and Spearman's correlations between the predicted and the median human scores, collectively from the 10 evaluations.

[I] `https://www.speakandlunch.com/`

3.2.4 Depression detection

The distress analysis interview corpus - wizard of Oz (DAIC-WOZ) database (Gratch et al., 2014) comprises of audio-visual interviews of 189 participants who underwent evaluation of psychological distress. The interviews were carried out in English using an animated virtual interviewer (DeVault et al., 2013). Each participant was assigned a self-assessed depression score through patient health questionnaire (PHQ-8) method (Kroenke et al., 2009). Time labels are provided in the data set for the portions of the participants' speech recordings.

We carried out depression detection using only the speech modality from the DAIC-WOZ corpus as a binary classification problem at the speaker level. We used the time labels provided in the data set to extract only the participants' speech recordings for experimentation. We excluded the sessions 318, 321, 341 and 362 from the training set as they had time-labelling errors. We evaluated the proposed techniques on the dev set, since the test set was held out as part of the AVEC 2016 challenge (Valstar et al., 2016). Performance was measured in terms of F1 score, precision and recall.

3.2.5 Intelligibility assessment

The UA-speech database (Kim et al., 2008) consists of 15 English speakers with cerebral palsy (11 males, 4 females) and 13 healthy speakers (9 males, 4 females). Each impaired and control speaker has uttered 765 isolated words in total: 155 isolated words repeated 3 times and 300 isolated words spoken only once. In the database, each subject's intelligibility score has been obtained by having five naive listeners (native speakers of American English) transcribe the isolated words and then calculating the average number of correct transcriptions. The subjective intelligibility scores of the patients range from 2% to 95%.

In the assessment of dysarthric speech intelligibility, we use the 5th channel recordings of the UA-speech corpus, similar to the previous works (Janbakhshi et al., 2019a, 2019b). An energy-based voice activity detection using Praat (Boersma & Weenink, 2001) was used to extract the speech segments. Performance was measured in terms of Pearson's correlation coefficient and Spearman's rank correlation coefficient between the predicted objective intelligibility scores and the subjective intelligibility scores.

3.3 Summary

In this chapter, we elaborated on the literature related to specific speech assessment tasks we deal in this thesis, listed the data sets used and gave the experimental protocols.

4 End-to-end acoustic modelling for automatic speech assessment

In Chapter 2, we discussed the existing methods for automatic speech assessment that use relevant feature extraction based on knowledge, followed by a classifier. In the last decade, automatic learning of task-specific information from raw waveforms has been demonstrated using convolutional neural networks (CNN), as opposed to using hand-crafted features, in the context of phoneme classification (Palaz et al., 2013), speech recognition (Palaz, 2016; Sainath et al., 2013; Swietojanski et al., 2014), speaker recognition (Muckenhirn, Magimai.-Doss, et al., 2018a) and verification (Muckenhirn, 2019; Muckenhirn, Magimai.-Doss, et al., 2018b), presentation attack detection (Dinkel et al., 2017; Muckenhirn et al., 2017), gender recognition (Kabil et al., 2018; Sebastian et al., 2018) and voice activity detection (Zazo et al., 2016). A similar approach that used little or no processing on the signals has been proposed on the assessment of emotion (Trigeorgis et al., 2016). Inspired from these works, we investigate how well such a raw speech modelling approach be employed for a variety of speech assessment tasks.

The rest of the chapter is organised as follows. We first describe the proposed approach, then contrast the experimental results on some speech assessment tasks with those of the existing methods and finally summarise our findings.

4.1 Proposed approach

The proposed architecture has been successfully adopted by several of the studies mentioned above. It consists of a convolutional neural network (CNNs) that operates on a fixed length raw speech signal to predict the classes of interest. The CNNs comprise two components: a feature learner that consists of convolutional layers, and a classifier that comprises fully connected layers. As illustrated in Fig. 4.1, the first layer is parameterised by 1D filters, each of which operates on the raw speech samples and outputs a time sequence. Thus the output of the first layer can be interpreted as a time-frequency representation similar to a spectrogram, where the frequency axis has no specific order (as opposed to that of a regular spectrogram) and the channels can be correlated depending upon the frequency responses of the filters. The subsequent layers have 2D filters, each of which spans all the input channels and moves

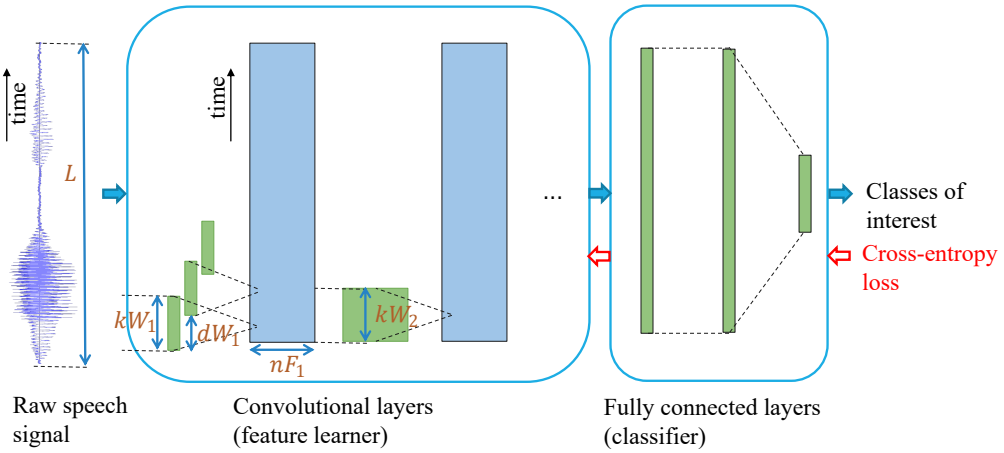


Figure 4.1: Automatic speech assessment using raw Speech CNNs.

in one dimension along the time axis and gives a time sequence. The output of each layer is passed through a non-linearity, typically a rectified linear unit (ReLU), and optionally through max-pooling along the time axis. The output of the feature learner is fed to fully connected layers, with ReLU activations at the hidden layers and softmax activation at the output layer, so that the network gives the probabilities of observing each class, given a speech segment. The training of the network involves updating the parameters by backpropagating a cross-entropy loss computed between the targets and the predictions. Depending upon the length of the filters in the first convolution layer, two approaches can be distinguished.

1. *Subsegmental modelling*: In subsegmental modelling (subseg), the filters span about 2 ms (< 1 pitch period). This was first proposed by Palaz et al. (2013) for phoneme classification study. The method provides a good time resolution.

2. *Segmental modelling*: In segmental modelling (seg), the filters span about 20 ms (1 – 5 pitch periods) and gives a better frequency resolution. Original speech signals contain information about the vocal source and the vocal tract system. A recent speaker recognition study (Muckenhirn, Magimai.-Doss, et al., 2018c) found that the filters in the first convolutional layer, when operated on 20 ms speech (1-3 pitch periods), modelled the fundamental frequency and low frequency information that could be related to the voice quality. Both the segmental and subsegmental approaches were found to be complementary.

4.2 Experimental validation

The proposed approach has been investigated on the speech assessment tasks: Styrian and Arabic DID, prediction of fluency from non-expert ratings and depression detection. The data sets and experimental protocols are defined in Sec. 3.2.

4.2.1 Styrian dialect identification

We studied Styrian DID through CNN-based modelling of raw speech. To partially overcome the resource scarcity, we experimented with speed perturbation (SP), as also suggested by Shon et al. (2018). SP refers to duplicating the utterances by altering the playback speed.

We compared our method with the baseline systems provided by the challenge, which are support vector machine (SVM) based classifiers trained on (a) functionals of LLDs, (b) their BoAW representations and (c) fixed-length feature representations from sequence-to-sequence autoencoders (S2SAE). The S2SAEs were trained on Mel-spectrogram representations that are thresholded using a signal-to-noise ratio (SNR) parameter. The organisers also provided a variant of the S2SAE method, obtained by fusing the embeddings from several S2SAE models trained using different SNR thresholds. This is denoted as *Fused Baseline*.

4.2.1.1 Systems

We used a two layer feature learner architectures shown in the Table 4.1. The input to the CNNs is a 250ms signal, overlapped by a 10ms shift. The classifier consists of a single fully connected hidden layer with 100 nodes, followed by an output layer of three nodes with a softmax activation. In order to avoid skewed results and to make the systems more robust to variations in initialisation, 5-fold cross-validation was conducted using leave-one-out method. In other words, 5 CNN systems were trained for each experiment by cross-validating on a left-out unseen part of the training set. During training, all the three classes were ensured of equal representation in each epoch by duplicating some of the utterances presented. We experimented with SP with playback speeds of 0.9 and 1.1 times the original. The targets to the CNNs are one-hot encodings of the dialects. The networks were trained using cross-entropy loss with stochastic gradient descent. Learning rate was halved, in the range 10^{-2} to 10^{-6}, between successive epochs whenever the validation-loss stopped reducing. The posterior probabilities obtained from the 5 CNNs for each utterance were averaged before classification.

Table 4.1: CNN architectures. nF: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling.

Model		**Layer**	**Conv**			**MP**
			nF	**kW**	**dW**	
RawCNN	subseg-small	1	128	30	10	2
		2	256	10	5	3
	seg-small	1	128	300	100	2
		2	256	5	2	-

4.2.1.2 Results

Table 4.2 summarises the results of the proposed methods, in terms of unweighted average recall (UAR) % on all the classes. The test set evaluation without SP was not part of our 5 submissions, so these system were not evaluated. We reported the best performances of the baseline systems quoted by Schuller et al. (2019), which were searched by parameter tuning. The discrepancy between the S2SAE Best and Fused Baseline test set scores indicate that the systems are sensitive to the SNR parameter variations, discussed in Sec. 2.3.2.4. The proposed raw speech modelling approach gave comparable performances in subsegmental modelling. It is worth noting that SP improved the subsegmental RawCNN results on the dev set.

Table 4.2: UAR% on the dev and test sets.

	Best Baseline			Fused Baseline	RawCNN		SP + RawCNN
Data set	LLD	BoAW	S2SAE	S2SAE	subseg-small	seg-small	subseg-small
dev	38.8	38.2	46.7	45.9	41.8	35.8	44.2
test	36.0	32.4	47.0	35.5	-	-	34.2

4.2.2 Arabic dialect identification

In the previous section, we proposed using raw speech to classify Styrian dialects in a low acoustic resource setting and with no linguistic resources. This work extends the previous work on an acoustically resource-rich condition, while maintaining the lack of linguistic resources. We describe the MFCC based baseline approach as well as the proposed raw speech modelling approach.

4.2.2.1 Baseline

Figure 4.2 shows the block diagram of the baseline approach. It follows Kaldi's x-vector system that was initially used for speaker recognition (Snyder et al., 2018). The system consists of time delay neural network (TDNN) that takes as input a sequence of MFCCs corresponding to an utterance and predicts the probabilities of each dialect. It consists of TDNN layers, followed by a *stats-pooling* layer (StatsP) that computes the statistics of its input representations across the given utterance and converts it to a fixed length representation. Further intermediate representations from a subsequent bottleneck layer are used as feature representations to train a probabilistic linear discriminant analysis (PLDA) based classifier, that computes the log-likelihood ratio (LLR) of observing each dialect.

4.2.2.2 Proposed raw speech based approach

In order to exploit and learn from a resource-rich condition, we introduce additional processing to the raw-speech modelling approach in Sec. 4.1. As illustrated in Fig. 4.3, in addition

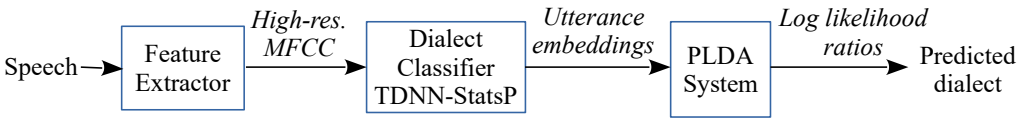


Figure 4.2: Block diagram of the MFCC based baseline approach for Arabic DID.

to a series of convolutional layers with ReLU activations followed by max pooling layers, the CNN consists of a series of additional residual convolutional layers. The later convolutional layers are *dilated* to facilitate the learning of higher level information from a longer context. The network is also added with a stats pooling layer that aggregates the first and second order moments of its input representations. This allows inputting entire utterances of varied lengths to the network, instead of fixed length segments. The stats pooling layer is followed by two fully connected layers, one with ReLU and the other with softmax activation, that gives the probabilities of each dialect by observing the entire utterance.

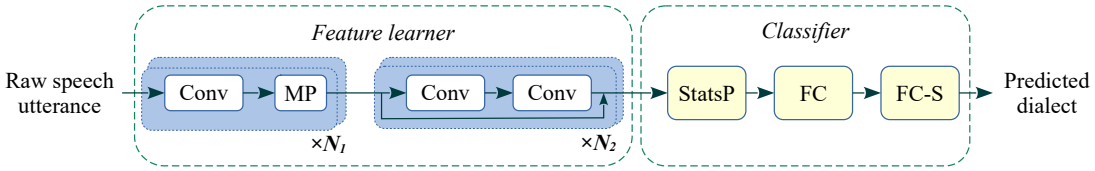


Figure 4.3: Block diagram of the proposed raw speech approach for Arabic DID.

4.2.2.3 Systems

The baseline x-vector DNN setup consists of five TDNN layers, followed by a stats-pooling layer, followed by a 512-node bottleneck layer, an additional hidden layer and a softmax layer. The total number of trainable parameters is 4.5M. Input speech was perturbed in terms of its speed and volume, and was augmented with reverberation using MUSAN corpus (Snyder et al., 2015). 40 dimensional high resolution MFCC features were used as the acoustic observations. After the DNN is trained using the entire ADI17 training set, PLDA systems were trained on a 17k subset of the training set, consisting of 1000 randomly chosen utterances per dialect, due to memory constraints. We denote this approach as *MFCC-TDNN*.

Table 4.3 gives the CNN architecture for the proposed approach. The total number of trainable parameters is 21M. Due to variable length input processing, the networks did not converge when trained from random initialisation. So we first trained them using 3 second speech segments, and then fine-tuned later on full variable length utterances.

Table 4.3: CNN architecture for Arabic DID. nF: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling, dil: dilation. Feature learner hyperparameters: $N_1 = 4$, $N_2 = 3$ (see Fig. 4.3).

Model		Layer	Conv			MP	dil
			nF	kW	dW		
RawCNN	subseg-large-stats	1	128	30	10	3	-
		2	256	10	1	3	-
		3	512	10	1	-	-
		4-10	512	10	1	-	2

4.2.2.4 Results

Table 4.4 shows the experimental results. We also reported, as *MFCC-CNN*, the baseline results provided by the challenge organisers Shon et al. (2020), using a 4 layer CNN network trained on 40 dimensional MFCCs, including a statistics pooling, followed by a two-hidden layer feedforward network with a softmax output layer. The proposed approach shows better accuracy than the MFCC-TDNN approach and comparable to the MFCC-CNN approach, although it used no data augmentation, contrary to the MFCC based ones.

Table 4.4: Experimental results in terms of percentage accuracy

Experiment	MFCC-CNN (Shon et al., 2020)	MFCC-TDNN	RawCNN subseg-large-stats
Dev	83.0	80.68	83.2
Test	82.0	80.81	82.0

4.2.3 Fluency prediction

Since fluency prediction has not been studied in a non-expert rating setting, we investigated different approaches that do not explicitly model linguistic information: (a) predefined set of acoustic low level descriptor (LLD) features-based, (b) unsupervised speech embeddings-based, and (c) end-to-end acoustic modelling-based.

4.2.3.1 Systems

eGeMAPS LLD features listed in Table 2.1 were used. Linear SVM classifiers were trained using scikit-learn (Pedregosa et al., 2011) with the default parameters, without optimising the hyperparameters. For BoAW representations, the codebook size used was 50, as the data was limited and contained mostly read speech. We included the time information of the frame as an additional feature to the BoAW representations, as we found that this improves the performance. For the wav2vec 2.0 representations, we used the pre-trained

base model provided by the authors (Baevski et al., 2020a), which was trained on LibriSpeech corpus (Panayotov et al., 2015).

Table 4.5: CNN architectures. nF: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling.

Model		**Layer**	**Conv**			**MP**
			nF	**kW**	**dW**	
RawCNN	subseg	1	128	30	10	2
		2	256	10	5	3
		3	512	4	2	-
		4	512	3	1	-
	seg	1	128	300	100	2
		2	256	5	2	-
		3,4	same as subseg			

CNNs for joint feature-classifier modelling were trained using Tensorflow (Abadi et al., 2015; Chollet et al., 2015). The terms *subseg* and *seg* refer to 30 sample *sub-segmental* and 300 sample *segmental* modelling respectively. Table 4.5 gives the architecture of the CNN feature learner. The classifier part of the CNN consists of one hidden fully connected layer with 100 nodes. The input to the CNNs is a 250ms signal, overlapped by a 10ms shift. The output layer consists of five nodes, corresponding to the five rating categories, with a softmax activation. During training, all the five classes were ensured of equal representation in each epoch by duplicating some of the utterances presented. The networks were trained using cross-entropy loss with stochastic gradient descent. Learning rate was halved, in the range 10^{-2} to 10^{-6}, between successive epochs whenever the training-loss stopped reducing.

4.2.3.2 Results

Results are reported in Table 4.6 in terms of Pearson's correlation coefficient and Spearman's rank correlation coefficient. The p-values are provided in parentheses. For both evaluation measures, all the systems yielded a good correlation score, with a p-value well below 0.01, i.e. the results are statistically significant. We can observe that the BoAW approach modelling LLDs yields the best results; however the automatic feature learning methods obtained an encouraging performance.

4.2.4 Depression detection

In this work, we studied the subsegmental approach, since information related to glottal pulses is present locally in time and may require time resolution, and the segmental approach for better modelling of source related information.

We compare our method with a few existing works that followed the same protocol, viz., (a)

Table 4.6: Results in terms of correlation coefficients, with p-values in parentheses.

		Pearson's	Spearman's
LLD + Functionals + SVM		0.338 (6e-71)	0.356 (4e-79)
LLD + BoAW + SVM		0.627 (7e-37)	0.641 (6e-39)
wav2vec 2.0 + BoAW + SVM		0.556 (1e-27)	0.578 (3e-30)
RawCNN	subseg	0.431 (4e-16)	0.446 (3e-17)
	seg	0.569 (3e-29)	0.563 (1e-28)

support vector machine (SVM) based baseline system from the AVEC 2016 challenge (Valstar et al., 2016) that used functionals of LLD features related to both the vocal tract and source, extracted using COVAREP tool (Degottex et al., 2014), (b) a long short term memory (LSTM) recurrent network system that used the above functionals to model speaker-level sequences of responses, and (c) CNN-based systems that detected depression from either spectrogram features or mel filter bank energies (Ma et al., 2016). In addition, we trained a 3-hidden layer deep neural network (DNN) baseline system that models MFCCs to emulate a vocal tract system information based system.

4.2.4.1 Systems

Table 4.5 gives the architecture of the CNN feature learner. The proposed system takes as input a 250 ms fixed length signal (determined through cross validation) overlapped with a 10 ms shift. The classifier part of the CNN consists of one hidden fully connected layer with 10 nodes. The output layer contains a single node with a sigmoid activation that outputs the probability of detecting depression. The parameters of the system are optimised using cross entropy criterion. During testing, the scores obtained on multiple signals of each speaker are averaged to get a per-speaker score, which is later thresholded to get a binary classification (control/depressed). The systems were trained using Tensorflow/Keras (Abadi et al., 2015; Chollet et al., 2015). For each experiment, the training data were split into 95% of speakers for training and 5% of speakers for cross-validation. To ensure equal representation of both the control and the depressed classes during training, we duplicated the depressed class utterances to a count matching as that of the control group. All the training frames of the depressed group were labelled 1, and the rest 0.

The networks (architectures listed in Table 4.5) were trained using cross-entropy loss with stochastic gradient descent. Learning rate was halved, in the range 10^{-1} to 10^{-6}, between successive epochs whenever the validation-loss stopped reducing. We trained 10 networks for each experiment, starting with a different random initialisation, in order to ascertain the systems are reproducible. We evaluated them primarily by the average *F1 score* of both the classes computed from all the 10 networks trained. We additionally report precision and recall scores. To fix a threshold on the speaker-level scores for the binary classification, F1 scores were computed by varying the threshold in steps of 0.01. The threshold that gave the best

unweighted average F1 score across all the 10 systems was then chosen, and the results were reported accordingly.

4.2.4.2 Results

Table 4.7 shows the F1 scores, precision and recall of the proposed methods along with the results of the baseline systems. It is worth mentioning that in the AVEC 2016 challenge the systems were ranked based on the F1 scores of both the classes. Except for the results from the existing works, each value shown indicates the *mean* performance obtained by training the DNN or CNN 10 times. We did this to ensure that the proposed methods are not sensitive to initialisation of DNN or CNN and the results are truly reproducible. The standard deviation of the performance of the systems were between 0.0 and 0.1.

Table 4.7: Performances of various methods on the AVEC 2016 dev set. *D* indicates *depressed*, *C* indicates *control* and *O* indicates the *overall* score by un-weighted average over the two classes. Bold font marks the best system among the proposed methods in terms of the overall F1 score.

Experiment	*F1 score*			*Precision*		*Recall*	
	O	*D*	*C*	*D*	*C*	*D*	*C*
LLD + Functionals + SVM (Valstar et al., 2016)	0.57	0.46	0.68	0.32	0.94	0.86	0.54
LLD + Functionals + LSTM (Al Hanai et al., 2018)	-	0.50	-	0.71	-	0.38	-
Spec + CNN (Ma et al., 2016)	0.61	0.52	0.70	0.35	1.00	1.00	0.54
MFCC + DNN	0.52	0.42	0.61	0.37	0.68	0.49	0.56
RawCNN - subseg	0.53	0.26	0.79	0.60	0.69	0.17	0.94
RawCNN - seg	0.57	0.57	0.57	0.43	0.82	0.82	0.43

4.3 Summary

In this chapter, we investigated directly modelling raw signals of speech using CNNs for several speech assessment tasks, viz. identifying Styrian and Arabic dialects, prediction of perceived fluency using non-expert ratings and detecting depression from speech. Our investigations showed the feasibility of employing the automatic feature learning method in all the tasks investigated, and showed encouraging performances that approached the respective baselines using handcrafted features or architectures. In fluency prediction and depression detection tasks, segmental modelling yielded a better performance than subsegmental modelling.

5 Incorporating voice source related information

In the previous chapter we attempted to alter the time and frequency resolution of the CNN's first layer by altering the kernel width, to implicitly focus more on the vocal tract or source related information. However, it may be desirable to explicitly model such information related to the source or system for an improved modelling. Conventional systems rely on knowledge-driven feature extraction methods to achieve this. However, robustly extracting certain features, especially at the source level, for instance those related to the glottal source activity, and characterising them precisely is a challenging problem. As discussed by Cummins et al. (2015), extracting and modelling source-related features for depression detection is a non-trivial task for reasons such as, (a) lack of a standardised approach to extract these features, (b) susceptibility to errors due to differing sound pressure levels between and within individuals, (c) difficulty in analysing and extracting these features from continuous speech in a reliable manner. To overcome such limitations, this chapter proposes methods of filtering signals that enhance the source specific information of interest through existing signal processing methods and thereby modelling the filtered signals using CNNs to automatically learn the features relevant for the task. We investigate how well the raw speech modelling approach can be leveraged using such methods.

The rest of the chapter elaborates on the signal filtering approaches, presents our investigations on speech assessment tasks, analyses how the systems learned differ from those of Chapter 4 and from each other, and finally summarises the findings.

5.1 Approach

Fig. 5.1 shows the proposed approach, where knowledge driven signal processing is introduced before processing them through the joint feature-learner described in the previous chapter. We investigated the following signal processing methods to extract voice source related information from raw speech.

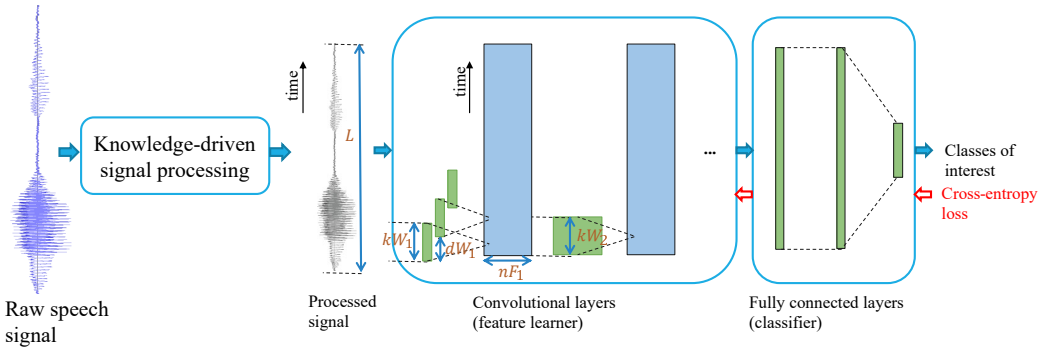


Figure 5.1: The proposed approach using knowledge-driven signal processing.

5.1.1 Low pass filtering

One way to enable the CNNs to effectively learn voice source related information is to low pass filter (LPF) the input signals such that the fundamental frequency is preserved. This has indeed been observed in a recent study on CNN-based glottal closure instant detection (Yang et al., 2018).

5.1.2 Linear prediction based decomposition

Linear prediction (LP) models the slow-varying components in speech signals $x[n]$, using coefficients α_k that linearly combine past samples to predict the current sample (Makhoul, 1975)

$$\tilde{x}[n] = \sum_{k=1}^{p} \alpha_k x[n-k]. \tag{5.1}$$

For a typical short quasi-stationary segment of pre-emphasised speech, this corresponds to fitting the formant-related vocal-tract structure. However, the quality of such modelling depends on the order of the LP used. We refer to this predictable component $\tilde{x}[n]$ as LP estimated (LPE) signal. The unpredictable component, called the LP residual (LPR),

$$g[n] = x[n] - \tilde{x}[n], \tag{5.2}$$

carries the glottal source information; thus LP analysis forms one of the methods for glottal signal analysis (Ananthapadmanabha & Yegnanarayana, 1979; Drugman et al., 2014). LPR signals contain not only the excitation information but also the modelling errors of the vocal tract system due to the assumptions on the LP order p (Makhoul, 1975). One way to handle this issue is through low pass filtering the speech signals before extracting the residual. This is akin to simple inverse filter tracking method (Markel, 1973), which was proposed for fundamental frequency estimation. In our studies, the LPR signals are estimated from pre-emphasised and low pass filtered signals, using LP modelling over short segments and then concatenated at the utterance-level.

5.1.3 Homomorphic source-filter decomposition

Complex cepstrum of a speech signal $x[n]$ is a transformation defined as

$$\hat{x}[n] = \mathscr{F}^{-1}\{\log(\mathscr{F}\{x[n]\})\}, \tag{5.3}$$

where $\mathscr{F}$ denotes discrete Fourier transform and log denotes a complex logarithm. Complex cepstrum allows transforming convolutive components of a time-domain signal into additive components, i.e. if $x[n] = u[n] * v[n]$, then $\hat{x}[n] = \hat{u}[n] + \hat{v}[n]$, where $u[n]$ denotes the excitation source and $v[n]$ the vocal tract response. Given $\hat{x}[n]$ of a speech signal, it is not possible to deduce both $\hat{u}[n]$ and $\hat{v}[n]$. However, since $u[n]$ has a fast-varying Fourier spectrum, most of its energy is concentrated in the higher cepstral coefficients of $\hat{x}[n]$. Similarly, since $v[n]$ has a slow-varying spectrum, most of its energy is concentrated in the lower coefficients of $\hat{x}[n]$. Thus linear high pass and low pass *liftering* of $\hat{x}[n]$ can approximate them (Drugman et al., 2009; Rabiner & Schafer, 2011):

$$\hat{u}[n] \approx \tilde{\hat{u}}[n] = \begin{cases} \hat{x}[n], & n \geq \tau \\ 0, & 0 \leq n < \tau \end{cases} \tag{5.4}$$

$$\hat{v}[n] \approx \tilde{\hat{v}}[n] = \begin{cases} \hat{x}[n], & 0 \leq n < \tau \\ 0, & n \geq \tau \end{cases} \tag{5.5}$$

Since complex cepstrum transform is invertible, the time domain signals $\tilde{u}[n]$ and $\tilde{v}[n]$ corresponding to $\tilde{\hat{u}}[n]$ and $\tilde{\hat{v}}[n]$ respectively can be constructed. We perform this analysis using a sliding window on each pre-emphasised utterance, and overlap-add the resultant segments of $\tilde{u}[n]$ to construct the utterance-level source related signal. We refer to it as homomorphically filtered vocal source (HFVS) signal.

5.1.4 Zero frequency filtering

Zero frequency filtering characterises the glottal source activity (Murty & Yegnanarayana, 2008; Yegnanarayana & Gangashetty, 2011). It exploits the property of an impulse-like excitation at the glottal closure instance to detect glottal closure instants (GCIs). ZFF signals are obtained by passing pre-emphasised speech signals through a cascade of two ideal digital resonators located at 0Hz, and then removing the trend in the resulting signals by subtracting the average over a window of the size in the range of 1 to 2 pitch periods. Including a pre-emphasis $1 - z^{-1}$, the cascaded 0Hz resonator has the transfer function

$$H(z) = \frac{1}{(1 - z^{-1})^3}. \tag{5.6}$$

Denoting the corresponding impulse response as $h[n]$ and the pitch period in samples as N_0, below is the method of extracting the ZFF signal $c[n]$[I].

$$c_1[n] = x[n] * h[n] \tag{5.7}$$

$$c_2[n] = c_1[n] - \sum_{n'=-N_0/2}^{N_0/2} c_1[n'] \tag{5.8}$$

$$c[n] = c_2[n] - \sum_{n'=-N_0/2}^{N_0/2} c_2[n']. \tag{5.9}$$

In addition to the GCIs, the strengths of the glottal excitations, the fundamental frequency and the glottal opening instants can be estimated from the ZFF signals (Murty & Yegnanarayana, 2008; Ramesh et al., 2013). It has recently been shown that ZFF signals can be modelled by CNNs for paralinguistic tasks such as sleepiness (Fritsch et al., 2020) and dementia (Cummins et al., 2020) prediction.

5.2 Experimental validation

The proposed approach as been investigated on the speech assessment tasks: depression detection, fluency prediction and Styrian DID. The data sets and experimental protocols are defined in Sec. 3.2.

5.2.1 Depression detection

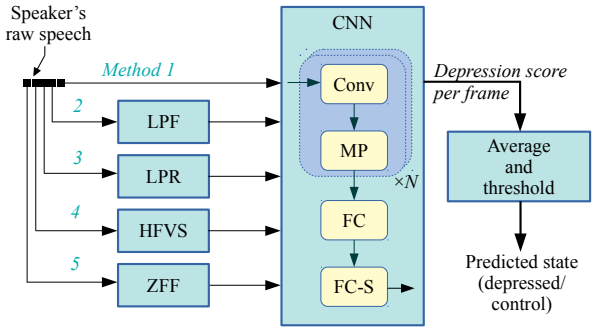


Figure 5.2: The proposed signal processing methods for depression detection. CNN architecture: Conv: convolutional layer with ReLU activations, MP: max-pooling layer, FC: fully connected layer with ReLU activations, FC-S: fully connected layer with a single output node and sigmoid activation.

Inspired from the voice source related literature on depression detection, we carry out experi-

[I]Since $h[n]$ is not stable, using it on on very long utterances requires caution, as it can lead to underflow/overflow issues. However, this is not the case with typical data sets, including the ones we used. If $c_1[n]$ computation is successful, then the numerical variability introduced is compensated by the moving average filter during $c_2[n]$ computation.

ments with signals that contain source related information, as shown in Fig. 5.2.

5.2.1.1 Systems

The proposed signals to be investigated were extracted using multiple tools. For LPF, Kaiser windowed sinc filters of the SoX tool were used. LPR signals were extracted through 8-order LP modelling using COVAREP tool (Degottex et al., 2014) with the default parameters, from the LPF signals. HFVS signals were extracted with a 40ms Hanning window, shifted by 20ms, using the standard complex cepstrum tools of MATLAB, and using $\tau = 50$ sample quefrency cut-off. ZFF signals were extracted using $N_0 = 160$ (which corresponds to F0 = 100 Hz for signals sampled at 16 kHz) for all the speakers, since using the true F0 was not found to influence the performance. We refer to direct modelling of raw speech as *RawCNN*, and modelling of processed signals as *SigCNN*. SigCNNs have the same architectures as those of RawCNN, given in Table 4.5.

5.2.1.2 Results

Table 5.1: Performances of various methods on the AVEC 2016 dev set. *D* indicates *depressed*, *C* indicates *control* and *O* indicates the *overall* score by un-weighted average over the two classes. Bold font marks the best system among the proposed methods in terms of the overall F1 score.

Experiment	*F1 score*			*Precision*		*Recall*	
	O	*D*	*C*	*D*	*C*	*D*	*C*
LLD + Functionals + SVM (Valstar et al., 2016)	0.57	0.46	0.68	0.32	0.94	0.86	0.54
LLD + Functionals + LSTM (Al Hanai et al., 2018)	-	0.50	-	0.71	-	0.38	-
Spec + CNN (Ma et al., 2016)	0.61	0.52	0.70	0.35	1.00	1.00	0.54
MFCC + DNN	0.52	0.42	0.61	0.37	0.68	0.49	0.56
RawCNN - subseg	0.53	0.26	0.79	0.60	0.69	0.17	0.94
RawCNN - seg	0.57	0.57	0.57	0.43	0.82	0.82	0.43
LPF 500Hz SigCNN - subseg	0.57	0.56	0.59	0.43	0.81	0.79	0.46
LPF 500Hz SigCNN - seg	0.65	0.61	0.69	0.50	0.84	0.79	0.59
LPR SigCNN - subseg	0.65	0.60	0.70	0.50	0.82	0.74	0.61
LPR SigCNN - seg	0.61	0.50	0.72	0.48	0.75	0.54	0.70
HFVS SigCNN - subseg	0.61	0.52	0.70	0.47	0.75	0.58	0.65
HFVS SigCNN - seg	0.61	0.54	0.68	0.46	0.77	0.64	0.61
ZFF SigCNN - subseg	**0.69**	0.65	0.73	0.54	0.87	0.81	0.63
ZFF SigCNN - seg	0.66	0.52	0.80	0.61	0.75	0.45	0.85

Table 5.1 shows the results of the proposed methods. It can be observed that the proposed methods of detecting depression based on voice source related information perform comparable to or better than the existing works and improve over directly modelling raw speech. In

particular, ZFF signals consistently yield better systems in terms of the overall F1 score than all the other methods. If we compare the systems based on F1 score for depression D, the proposed methods perform comparable or outperform existing methods.

5.2.2 Fluency prediction

We investigate the use of source related (i) LLD features and their BoAW representations, and (ii) ZFF signal processing method in addition to modelling raw speech, as illustrated in Fig. 5.3.

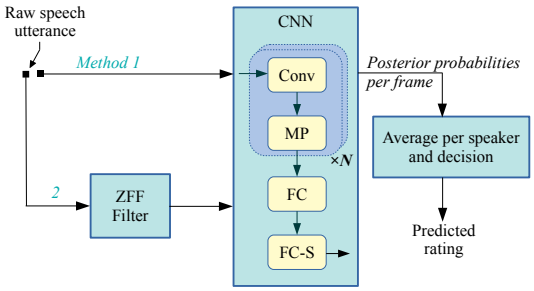


Figure 5.3: Proposed source-related signal modelling for fluency prediction. Conv: convolutional layer with rectified linear (ReLU) activation, MP: max-pooling, FC: fully connected layer with ReLU activation, FC-S: FC layer with softmax activation.

5.2.2.1 Systems

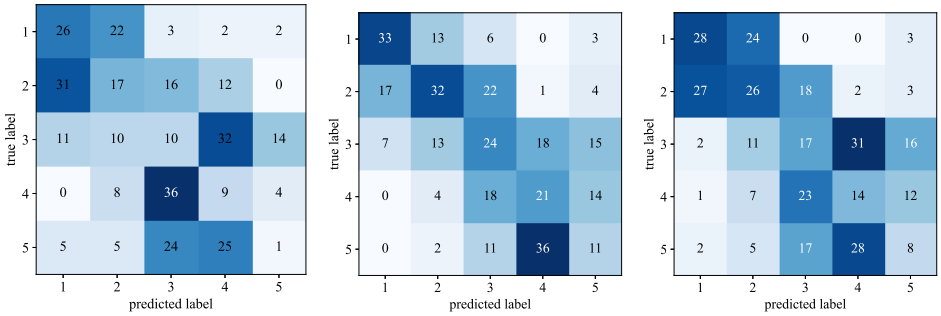
The architectures are as listed in Table. 4.5. LLD feature extraction, BoAW representation extraction and the training of the neural networks are the same as described in Sec. 4.2.3.1. The source related LLDs are given by Table 2.1.

5.2.2.2 Results

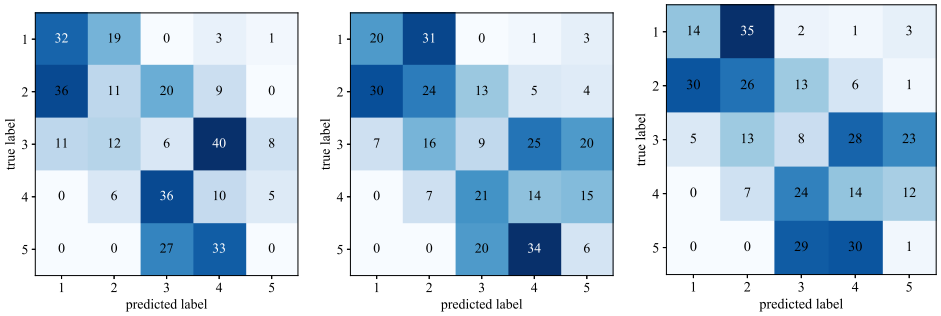
Table 5.2 gives the results. It can be observed that the vocal source related LLD features contribute less to the performance as compared to the complete set of LLDs. A better performance of ZFF-based approach than BoAW approach modelling source-related LLDs indicates that the former is able to better model source-related information for speech fluency prediction. Finally, it is interesting to observe that the subseg ZFF SigCNN and seg RawCNN approaches yield performances similar to that of the BoAW with wav2vec 2.0 embeddings approach. Fig. 5.4 shows the confusion matrices of several systems. For all these systems, the predictions are centred around the true rating, indicating a systematic prediction of the speech fluency ratings.

Table 5.2: Performance of source-related signals on fluency prediction in terms of correlation coefficients, with p-values in parentheses.

		Pearson's	Spearman's
LLD + Functionals + SVM		0.338 (6e-71)	0.356 (4e-79)
LLD + BoAW + SVM		0.627 (7e-37)	0.641 (6e-39)
LLD (Source) + BoAW + SVM		0.337 (4e-10)	0.347 (1e-10)
wav2vec 2.0 + BoAW + SVM		0.556 (1e-27)	0.578 (3e-30)
RawCNN	subseg	0.431 (4e-16)	0.446 (3e-17)
	seg	0.569 (3e-29)	0.563 (1e-28)
ZFF SigCNN	subseg	0.560 (3e-28)	0.576 (4e-30)
	seg	0.515 (2e-23)	0.545 (2e-26)



(a) raw subseg (b) BoAW (eGeMAPS) (c) BoAW (wav2vec 2.0)



(d) raw seg (e) zff subseg (f) zff seg

Figure 5.4: Confusion matrices of some fluency prediction systems.

5.2.3 Styrian dialect identification

Styrian dialects lack distinction in their pitch patterns, i.e. related to the voice source. Nevertheless, a question that arises is whether source carries any discriminative information about Styrian dialects. We investigate this point, as shown in Fig. 5.5.

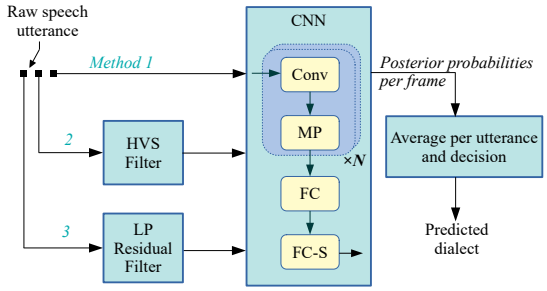


Figure 5.5: Proposed source-related signal modelling for Styrian dialect identification. Conv: convolutional layer with rectified linear (ReLU) activation, MP: max-pooling, FC: fully connected layer with ReLU activation, FC-S: FC layer with softmax activation.

5.2.3.1 Systems

The extraction of HFVS and LPR signals was as described in Sec. 5.2.1.1. The architecture used in the Styrian dialect experiments is *subseg-small* (see Table 4.1).

5.2.3.2 Results

Table 5.3 shows the results of modelling source related signals using the *SigCNN* architecture.

Table 5.3: Performance of source-related signals on the Styrian dev set, in terms of UAR%.

Data set	RawCNN	HFVS SigCNN	LPR SigCNN
dev	41.8	35.2	37.1

5.3 Analysis

In this section, we analyse the systems at two levels, viz. visualising the frequency response of the first layers, and visualising the *relevance signals* by looking at the entire networks.

5.3.1 Analysis of frequency response of the first layer filters

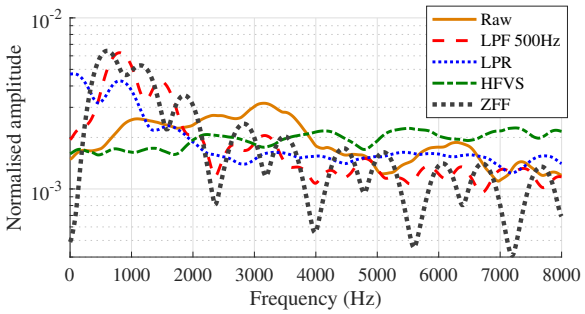
To better understand the spectral information being modelled by the CNNs, we analysed the cumulative frequency response of the first convolutional layer filters, as done by Muckenhirn, Magimai.-Doss, et al. (2018c) and Palaz et al. (2016):

$$F_{cum} = \sum_{k=1}^{N_f} \frac{F_k}{\|F_k\|_2}, \tag{5.10}$$

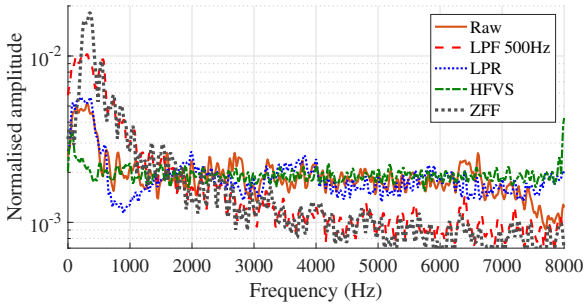
where N_f is the number of filters and F_k is the frequency response of filter f_k.

5.3.1.1 Depression detection

Fig. 5.6a shows the cumulative responses of the CNNs modelling the proposed methods at the subsegmental level modelling (filters of length about 2 ms). As expected, for ZFF, LPF and LPR the emphasis is on low frequencies. For HFVS the response is almost flat across the frequencies. For raw speech signals, the emphasis is more on the high frequencies between 2 kHz - 4kHz, which is more related to the vocal tract system information. Fig. 5.6b compares



(a) subseg modelling



(b) seg modelling

Figure 5.6: Comparison of the overall frequency responses of the first convolutional layers in various depression detection CNNs.

the cumulative frequency responses of the CNN filters with segmental level modelling for the proposed methods. It can be seen for all the signals, including raw speech, that the emphasis lies in the low frequency regions. It is interesting to observe that, except for the HFVS case, the low frequency region being emphasised is similar.

5.3.1.2 Fluency prediction

Fig. 5.7 shows the cumulative frequency responses of the first convolution layer of the different CNN-based systems. It can be observed that most of the systems focus on the low frequency regions that are more related to the fundamental frequency and voice source related aspects (Dubagunta, Vlasenko, et al., 2019; Muckenhirn, Magimai.-Doss, et al., 2018c), which are more linked to fluency than the linguistic accuracy, corroborating with the finding of Duijm et al. (2018).

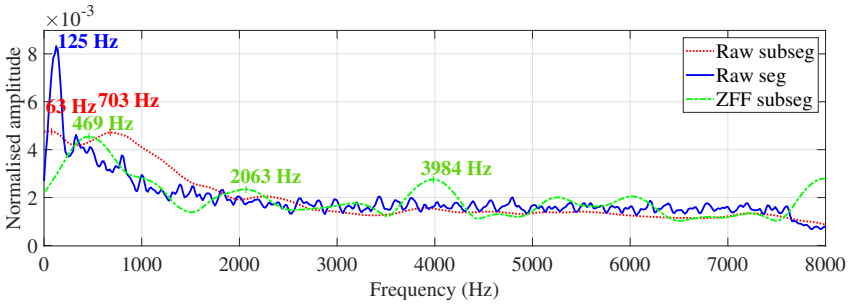


Figure 5.7: Frequency responses of the first convolutional layers of some fluency prediction systems.

5.3.2 Relevance analysis

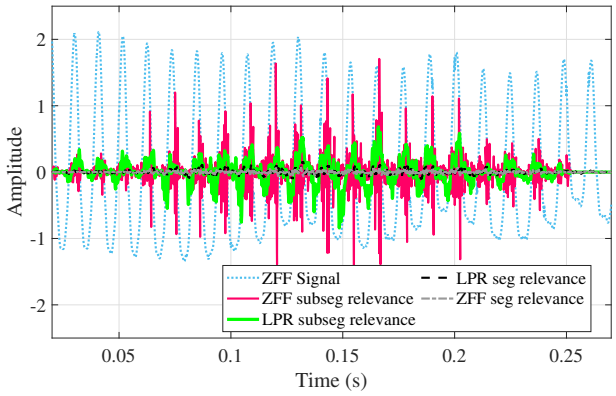
To gain insight about what the CNNs as a whole are learning, we applied a recently developed *guided backpropagation* based visualisation method (Muckenhirn, Abrol, et al., 2018). In simple terms, given an input signal and the output class, the technique measures how a small variation or perturbation of each sample value will impact the prediction score. This corresponds to measuring the importance of each input sample value for the prediction. This process yields a relevance signal.

Using the relevance analysis method, we contrasted the CNNs trained on ZFF signals with those trained on LPR signals in depression detection. Fig. 5.8a shows the relevance signals computed for the subsegmental and segmental level modelling on both the types of signals, overlaid on the input ZFF signal, of a sustained vowel /uh/ of duration 250 ms from the database. In the case of subsegmental modelling, we observe that for both ZFF and LPR relevance signals there is a sharp focus at the positive-to-negative zero-crossings of the ZFF signals, which corresponds to the glottal closure instants (GCIs) (Murty & Yegnanarayana, 2008). This suggests that the subsegmental CNN is focusing on the GCI information for depression detection. In the case of segmental modelling, the relevance signal does not have such a sharp focus, indicating that all the samples are given importance. Fig. 5.8b shows the autocorrelation of the above signals. It can be observed that all the relevance signals are preserving the periodicity, i.e. F0, information.

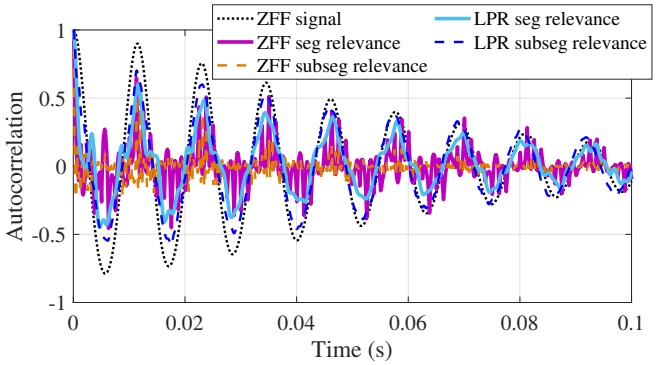
5.4 Summary

This chapter investigated methods of modelling signals filtered with voice source related information using CNNs for automatic speech assessment. Investigations on speech assessment tasks indicated the usefulness of such signal filtering methods, except in tasks with no discriminative information at the source-level.

Our studies on depression detection showed that, instead of modelling raw speech signals as they are, filtering them based on prior knowledge, such as low pass filtering to remove the



(a) Relevance signals.



(b) Autocorrelation computed from relevance signals.

Figure 5.8: Illustration of relevance signals and their autocorrelation signals in depression detection. The example shown is part of the sustained vowel *uh*.

high frequency vocal tract system related information or ZFF leads to effective depression detection. More precisely, the systems based on ZFF signals and LPR signals yield better than the state-of-the-art LLD based systems. Analyses using frequency response and relevance plots reveal that the segmental level modelling of ZFF and LPR signals is mainly focusing on the F0 variation, whilst the subsegmental level modelling is focusing on time local events related to the voice source, viz. GCIs, similar to jitter and shimmer feature extraction as well as the F0 variation. This could be the reason why subsegmental level modelling of ZFF and LPR signals yields better system than segmental level modelling.

In fluency prediction, the proposed ZFF approach showed an encouraging performance and is able to better model the source related information than the BoAW-based approach on source-related LLDs. Filtering using ZFF helped shift the focus of subsegmental modelling more towards the low frequency regions and thereby improved the results over raw speech modelling, whereas no such gains are observed in segmental modelling which already emphasises the low frequency regions when modelling raw speech.

Studies on Styrian dialect identification clearly show that the voice-source related features such as pitch patterns do not contribute to the Styrian dialect classification.

6 Incorporating linguistic prior knowledge

The previous chapter dealt with incorporating knowledge related to the vocal source through signal processing, into the joint feature-classifier learner systems. In a similar manner, some assessment tasks can benefit from the availability of linguistic resources. For instance, dialects can be identified by the choice of words when word transcriptions are available, and by the pronunciation variations when phonetic transcriptions are available. These linguistic differences can be attributed more to the vocal tract than to the source. In the absence of such explicit resources, we look at implicitly incorporating such knowledge, by (i) learning explicit linguistic knowledge on another task with available resources, and (ii) transfer learning the learned model parameters for the task of interest. We investigate how well such methods leverage the raw speech modelling approach proposed in Chapter 4.

The rest of the chapter elaborates on the proposed approach, presents the experimental studies and analyses on speech assessment tasks, and finally summarises the findings.

6.1 Proposed approach

The typical linguistic sub-word units of any language, such as phonemes, can be mapped to the articulatory properties of the vocal apparatus that cause to produce the associated sounds. Such properties include the place of constriction, the height of the tongue, roundedness of the lips, etc. When such knowledge exists in a language in the form of mappings from phonemes to their articulatory feature (AF) representations, direct mappings from acoustic feature representations to AFs can be learned (Rasipuram & Magimai-Doss, 2015). Such AF representations were shown to be useful for improved pronunciation modelling, noise robustness and multi-lingual portability. Rasipuram and Magimai-Doss (2015) utilised handcrafted feature representations to learn the feature-to-AF mappings. However, motivated by better task-specific modelling through automatic feature learning, we propose to model AFs directly from raw speech using joint feature-classifier CNNs. Once such models are trained, model parameters are transfer learned for the task of interest, as illustrated in Fig. 6.1.

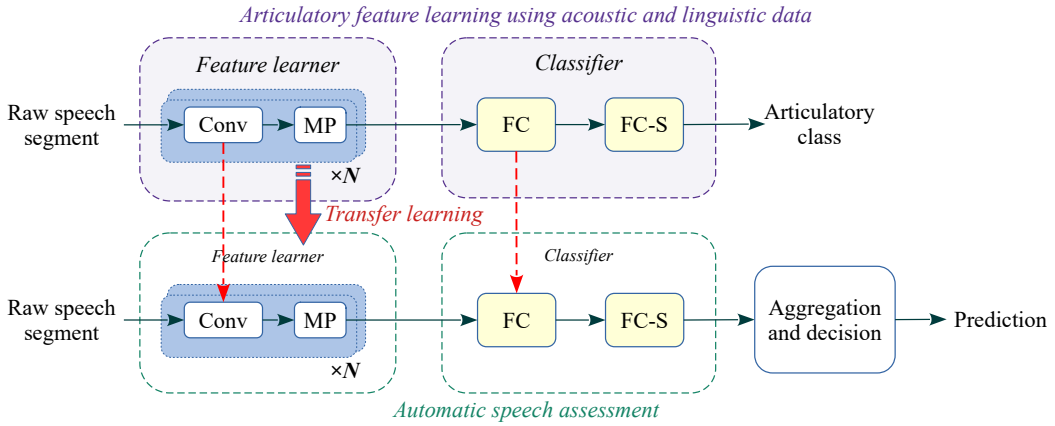


Figure 6.1: Block diagram of the proposed transfer learning approach.

6.1.1 Articulatory parameter CNNs

For training the articulatory networks, the AMI corpus (Carletta et al., 2005) recorded using independent headset microphone was used, which consists of 77 hours of meetings. Kaldi setup was used to train hidden Markov models (HMMs) for context-dependent phones, where the HMM states were jointly modelled by using subspace GMMs. The corresponding frame-to-phone alignments were used to train the AF CNNs, using phone-to-AF mappings from Rasipuram and Magimai-Doss (2015, see Table B.1). More specifically, 4 AF CNNs were trained to individually predict the 4 AF categories: place, manner, height and vowel. The model architecture used, listed as *RawArticCNN* in the Table 6.1, was inspired from raw-speech based phone classification using sub-segmental modelling first proposed by Palaz et al. (2013). The "FC" layer in this architecture contains 1024 nodes. AF training was performed on a 70 hour *clean* subset of the training set, which is a standard practice followed in the Kaldi recipe. Transfer learning to the task of interest involved using the 4 AF-CNNs to initialise another 4 corresponding CNNs, for the task, of the same architecture (*RawArticCNN*) except for either the output classification layer or the entire classifier part of the joint feature-classifier learner, depending on the task. The frame-level accuracies on a held-out validation data, reported in Table 6.2, indicate that the networks learn meaningful representations.

Table 6.1: CNN architecture for articulatory CNNs. nF: number of filters, kW: kernel width, dW: kernel shift, MP: max-pooling, dil: dilation.

Model (Input frame size)	**Layer**	**Conv**			**MP**
		N_f	**kW**	**dW**	
RawArticCNN (250ms)	1	80	30	10	3
	2,3	60	7	1	3

Table 6.2: Information on the trained articulatory networks.

	Manner	Place	Height	Vowel
Number of classes	9	13	8	23
Frame-level accuracy on validation data	77.8	72.4	76.5	75.5

6.2 Experimental validation

We investigate the proposed approach on the following speech assessment tasks: Styrian and Arabic DID, fluency prediction and depression detection.

6.2.1 Styrian dialect identification

As shown in Fig. 6.1, we utilise AF model parameters in the Syrian DID task.

6.2.1.1 Systems

Transfer learning to Styrian DID involved using the 4 AF-CNNs to initialise another 4 corresponding CNNs, for DID, of the same architecture (*RawArticCNN*) except for the output classification layer. For the transfer learning, we only excluded the output classification layer from the parameter initialisation. For transfer learning as feature embeddings, refer to Sec. 6.2.1.4. Training was performed using the training set, by using a decaying learning schedule as described in Sec. 4.2.1.1, and by cross-validating on the entire training set at the end of each epoch. The posterior probabilities obtained from the 4 CNNs for each utterance were averaged before classification. We also investigated fusing the outputs of the four CNNs, through averaging the corresponding posterior probabilities, before predicting the dialect.

6.2.1.2 Results

Table 6.3 summarises the results using the proposed transfer learning approach, along with the baselines and earlier results. It is worth reminding the reader that the test set evaluations were limited to 5 submissions. It can be observed that the fused RawArticCNN system performs better on the test set than the other proposed approaches and the existing LLD and BoAW based baselines. In comparison with the S2SAE baseline, the method performs better than the fused S2SAE baseline. This validates that AF based transfer learning through AF initialisation helps in improving the Styrian DID.

6.2.1.3 Modelling vocal tract information through signal processing

As discussed earlier, linguistic differences can be attributed more to the vocal tract than the source. Given the understanding of separating the source and filter components using signal

Table 6.3: Effect of AF transfer learning on the UAR% of Styrian DID.

Experiment		dev	test
Best Baseline	LLD	38.8	36.0
	BoAW	38.2	32.4
	S2SAE	46.7	47.0
Fused Baseline	S2SAE	45.9	35.5
RawCNN+SP	subseg-small	44.2	34.2
LPR SigCNN		37.1	-
RawArticCNN	Manner	44.4	-
	Place	43.5	-
	Height	45.3	-
	Vowel	43.0	-
	Fused	46.6	36.6

processing approaches in Sec. 5.1, in this section, we conduct an analysis study on whether such methods that enhance the vocal-tract related information can be used as an alternative to AF based transfer learning in the Styrian DID problem.

6.2.1.3.a Approach

From the linear prediction and homomorphic processing methods described in Sec. 5.1, we can extract the corresponding complementary information to construct the vocal tract related signals. More specifically, the LPE signals $\tilde{x}[n]$, given by Eq. (5.1), can be constructed for short segments of speech (without using a low pass filtering as done for LPR signals) and aggregated at the utterance-level. Similarly, homomorphic filtered vocal tract (HFVT) signals $\tilde{v}[n]$, corresponding to Eq. (5.5) in the time domain, can be overlap-added per utterance. As illustrated in Fig. 6.2, we investigate whether modelling of such vocal tract related signals, termed as *SigCNN*, improves Styrian DID. In addition, we also investigate learning to predict AF classes from HFVT and LPE signals. This approach is termed *SigArticCNN* and has the same architecture as that of RawArticCNN.

6.2.1.3.b Systems

HFVT and LPE signals were generated using MATLAB. HFVT signals were extracted with a 40ms Hanning window, shifted by 20ms and were liftered at $\tau = 50$ sample quefrency cut-off. LPE signals were predicted using 30ms Hamming windows, shifted by 10ms and using 12^{th} order LP modelling. SP was performed using SoX tool at additional 0.9 and 1.1 speeds. As discussed earlier, these signals were computed at utterance level and were then processed through CNNs. The architecture used for SigCNN experiments is *subseg-small* (see Table 4.5).

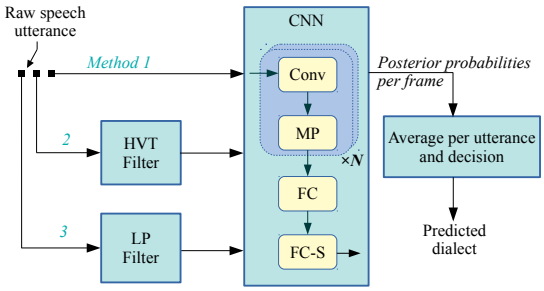


Figure 6.2: Proposed approach based on CNNs and using vocal tract related signals. Conv: convolutional layer with rectified linear (ReLU) activation, MP: max-pooling, FC: fully connected layer with ReLU activation, FC-S: FC layer with softmax activation.

6.2.1.3.c Results

From the results in Table 6.4, it can be observed that the proposed SigCNN modelling of HFVT and LPE signals is better than RawCNN+SP. We also experimented using SigCNN+SP and observed no gains over SigCNN (results are not shown). Comparing the SigArticCNN results with the RawArticCNN results in Table 6.3, it can be seen that SigArticCNN shows better performance, especially with modelling the LPE signals. This supports the point that the variations related to speech sounds, to a large extent, can be attributed to the changes in the vocal tract.

Table 6.4: Effect of modelling vocal tract related signals using CNNs and AF based transfer learning on the UAR% of Styrian DID.

Experiment		dev	test
RawCNN+SP	subseg-small	44.2	34.2
HFVT SigCNN		46.8	-
LPE SigCNN		46.3	-
HFVT SigArticCNN	Manner	44.2	-
	Place	44.0	-
	Height	45.0	-
	Vowel	44.3	-
	Fused	45.0	-
LPE SigArticCNN	Manner	47.3	-
	Place	45.5	-
	Height	46.2	-
	Vowel	45.0	-
	Fused	47.0	35.6

6.2.1.4 Transfer learning without parameter adaptation

In the proposed approach (Sec. 6.1), the transfer learned parameters are fined-tuned to the task of interest. In this section, we experiment with not adapting (or freezing) these parameters and training only the final classification layer for Styrian DID. This is analogous to extracting short-time embeddings from AF networks and using them to build linear classifiers for the Styrian DID problem. Table 6.5 shows such results, where the experimental setup is identical to that of Sec.6.1.1 except for the parameter freezing. The results suggest that updating all the parameters gives an improved classification. They also indicate that *manner* of articulation may carry the most distinguishable information among the other AFs for Styrian DID.

Table 6.5: Effect of AF transfer learning without parameter adaptation on the UAR% of Styrian DID.

Experiment		dev	test
RawArticCNN+SP	Manner	43.8	-
	Place	40.0	
	Height	42.3	
	Vowel	42.8	
	Fused	42.4	

6.2.2 Arabic dialect identification

Here we investigate the articulatory initialisation based transfer learning in both joint feature-learner as well as MFCC front-end based approaches.

6.2.2.1 Systems

In the MFCC based approach, the architecture consists of five TDNN layers, identical to that of the initial part of the x-vector DNN as described in Sec. 4.2.2.1, followed by an output layer with a softmax. These models are trained to minimise the cross-entropy between the outputs and their corresponding frame-level AF targets using stochastic gradient descent. Transfer learning to the ADI17 task involves using the parameters of the first four layers of each of the 4 AF-DNNs to initialise another 4 corresponding DNNs. For the fusion experiment, the utterance-level feature embeddings obtained from the 4 CNNs for each utterance are concatenated and then processed through a common PLDA module.

For the raw CNN systems, the total number of parameters for each articulatory initialised model is 1.66M. Transfer learning is performed as shown in Figure 6.3. Similar to above, four CNNs were trained from the four AF CNNs. However, the fusion experiment involves averaging the posterior probabilities from the four CNNs before predicting the dialect.

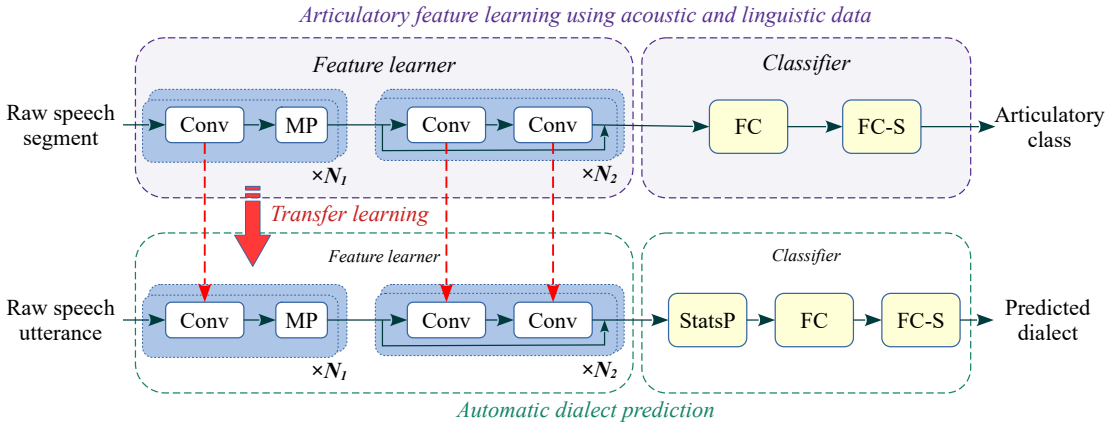


Figure 6.3: Block diagram of the proposed transfer learning approach for Arabic DID.

6.2.2.2 Results

Tables 6.6 and 6.7 show the experimental results with the MFCC based front-end and raw speech based CNNs respectively. In both cases, when initialised using the articulatory parameters, the results are either improved or similar. However, the fusion experiments improve the results significantly, which indicates that the networks learn complementary information when initialised with the AF categories. Furthermore, a fusion of RawCNN and RawArticCNN posteriors further improves the raw speech based systems. It can also be observed that the proposed approach maintains to yield better accuracy than the MFCC based approach, even though the latter uses speed perturbation.

Table 6.6: Experimental results on MFCC-based articulatory initialisation on ADI17 task in terms of classification accuracy (%).

Set	MFCC-CNN (Shon et al., 2020)	MFCC-TDNN (Baseline)	Manner	Place	Height	Vowel	Fused
dev	83.0	80.68	80.85	80.94	80.27	81.01	82.35
test	82.0	80.81	80.82	80.87	80.76	80.72	82.56

Table 6.7: Experimental results on raw speech based articulatory initialisation on ADI17 task in terms of classification accuracy (%).

Set	RawCNN	RawArticCNN					Fusion RawCNN +RawArticCNN
		Manner	Place	Height	Vowel	Fused	
dev	83.2	83.9	83.3	83.4	83.0	85.8	86.9
test	82.0	82.9	81.3	81.5	82.2	84.5	85.3

6.2.3 Fluency prediction

We study (i) the use of vocal tract related LLDs features alone in the BoAW approach, and (ii) transfer-learning to implicitly model articulatory information such as place and manner of articulation.

6.2.3.1 Systems

The system related LLDs are given by Table 2.1. Transfer learning of AFs for fluency prediction involved initialising 4 corresponding CNNs from the pre-trained ones, of the same architecture (*RawArticCNN*) except for the final layer, and fine-tuning them with the same training procedure as above. The posterior probabilities obtained from the 4 CNNs for each utterance were averaged before classification.

6.2.3.2 Results

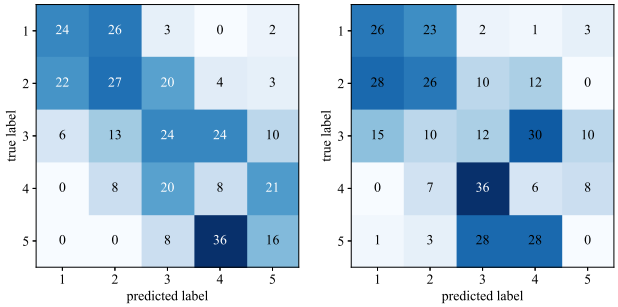
We can observe from the results in Table 6.8 that (i) the vocal tract related LLDs contribute the most to the correlation with human scores, and (ii) in the subsegmental raw signal modelling based systems, initialising the neural network with articulatory feature information improves its performance, most prominently with the place of articulation.

Table 6.8: Results of fluency prediction in terms of correlation coefficients, with p-values in parentheses.

		Pearson's	Spearman's
LLD + Functionals + SVM		0.338 (6e-71)	0.356 (4e-79)
LLD + BoAW + SVM		0.627 (7e-37)	0.641 (6e-39)
LLD (Source) + BoAW + SVM		0.337 (4e-10)	0.347 (1e-10)
LLD (System) + BoAW + SVM		**0.657** (2e-41)	**0.668** (2e-43)
wav2vec 2.0 + BoAW + SVM		0.556 (1e-27)	0.578 (3e-30)
RawCNN	subseg	0.431 (4e-16)	0.446 (3e-17)
	seg	0.569 (3e-29)	0.563 (1e-28)
ZFF SigCNN	subseg	0.560 (3e-28)	0.576 (4e-30)
	seg	0.515 (2e-23)	0.545 (2e-26)
RawArticCNN	Manner	0.497 (1e-21)	0.527 (1e-24)
	Place	0.517 (1e-23)	0.528 (9e-25)
	Height	0.489 (6e-21)	0.499 (7e-22)
	Vowel	0.416 (5e-15)	0.437 (1e-16)
	Fused	0.493 (3e-21)	0.516 (2e-23)

6.2.3.3 Analysis

Fig. 6.4 shows the confusion matrices of *BoAW* and *RawArticCNN place* systems covering the different approaches. For both the systems, the predictions are centred around the true rating, indicating a systematic prediction of the speech fluency ratings, as observed in Chapter 5.



(a) BoAW (system) (b) Artic place

Figure 6.4: Confusion matrices of some fluency prediction systems.

Fig. 6.5 shows the cumulative frequency responses of the first convolution layer of the different CNN-based systems. The articulatory feature initialised networks focus on low-to-mid

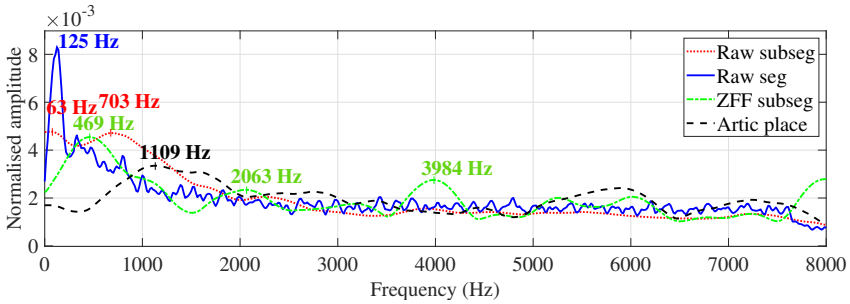


Figure 6.5: Frequency responses of the first convolutional layers of some fluency prediction systems.

frequencies, which are typically modelled by CNNs that classify phones and therefore model formant related information (Palaz et al., 2019). This suggests that the initialisation of subsegmental raw speech modelling with networks trained to classify AFs helped shift the focus of the network more towards linguistic unit related information and consequently improved the performance.

6.2.4 Depression detection

Several works in the literature used linguistic resources to detect depression or improve its performance (cf. Lopez-Otero et al., 2017). In this work, we investigate whether implicitly modelling such knowledge through AF transfer learning improves the depression detection

systems.

6.2.4.1 Systems

Transfer learning involved initialising 4 corresponding CNNs from the pre-trained ones, of the same architecture (*RawArticCNN*) except for the final layer, and fine-tuning them with the same training procedure as above. The probability of depression obtained from the 4 CNNs from all the speakers' segments were averaged before classification.

6.2.4.2 Results

Table 6.9 shows the results with articulatory initialisation, with performances close to random predictions and with no improvements observed.

Table 6.9: Performances of various methods on the AVEC 2016 dev set. *D* indicates *depressed*, *C* indicates *control* and *O* indicates the *overall* score by un-weighted average over the two classes. Bold font marks the best system among the proposed methods in terms of the overall F1 score.

Experiment	*F1 score*			*Precision*		*Recall*	
	O	*D*	*C*	*D*	*C*	*D*	*C*
RawCNN - subseg	0.53	0.26	0.79	0.60	0.69	0.17	0.94
RawCNN - seg	0.57	0.57	0.57	0.43	0.82	0.82	0.43
RawArticCNN - manner	0.50	0.24	0.75	0.40	0.67	0.17	0.87
RawArticCNN - place	0.51	0.51	0.51	0.39	0.75	0.75	0.39
RawArticCNN - height	0.50	0.24	0.75	0.40	0.67	0.17	0.87
RawArticCNN - vowel	0.50	0.41	0.59	0.35	0.67	0.50	0.52
RawArticCNN - fused	0.50	0.24	0.75	0.40	0.67	0.17	0.87

6.3 Summary

In this chapter we investigated incorporating implicit articulatory linguistic knowledge in speech assessment tasks. Our investigations showed that such methods improve the performance of several raw speech modelling based assessment tasks.

Investigations on Styrian dialect identification using HFVT and LPE signals showed better performance than the existing LLD and BoAW based methods and comparable to the S2SAE based approach. Furthermore, the AF based transfer learning approach was shown to achieve better modelling when linguistic resources are unavailable. The work also showed that the vocal-tract related differences play a better role in distinguishing Styrian dialects than the voice-source, particularly in terms of the manner of articulation. Finally, the vocal tract filtering methods were shown to yield competent systems without data augmentation through

speed perturbation.

Investigations on Arabic dialect identification revealed an improved performance using AF based transfer learning, both on MFCC and raw speech based systems. Fusion of posterior probabilities from the four systems showed clear gains, indicating that the systems modelled complementary information.

Investigations on perceived fluency prediction using non-expert ratings indicated that articulatory initialisation of subsegmental raw speech systems shifted the focus of the network and consequently improved the performance, although the predictions from segmental modelling gave better correlations with the human scores. However, vocal tract system related BoAW feature representations gave the highest correlation with the human scores.

Investigations on the depression detection task indicated no gains of using articulatory initialisation, although works such as (Lopez-Otero et al., 2017; Villatoro-Tello et al., 2021, accepted for publication) indicated performance gains by including linguistic resources at the utterance level.

In the future, it is worth investigating AF based transfer learning and subsequently modelling utterance level information in both fluency prediction and depression detection tasks.

7 Incorporating linguistic segment level information for posterior feature based intelligibility assessment

The chapters 4, 5 and 6 mainly dealt with raw signal modelling and approaches of incorporating implicit knowledge related to the vocal tract/source and linguistic information for several speech assessment tasks. In this chapter, we focus on explicitly incorporating linguistic segment level information in the training of phone posterior probability estimator neural networks that are used in automatic speech intelligibility assessment. We investigate whether such training helps improve the objective estimation of human-rated intelligibility. Since such artificial neural networks (ANNs) are also used in hybrid hidden Markov model (HMM) based HMM/ANN automatic speech recognition (ASR) systems, we also investigate whether such training improves the ASR performance. It is worth noting that the approaches that will be discussed apply to both automatic feature learning based and handcrafted feature based systems. Thence, we do not particularly emphasise on directly modelling raw signals of speech hereafter.

The rest of the chapter is organised as follows. We first elaborate on intelligibility assessment using phone posterior probability features. Since the ANNs employed in this approach are trained on ASR objective, we review briefly the background on hybrid HMM/ANN ASR system training. We then establish a link between the estimation of linguistic unit level confidences and the training of ANNs, which leads to a new approach that incorporates segment level confidence measures in the ANN training. We investigate the application of such training methods to see the impact on the ASR performance. We then introduce the problem of speech intelligibility assessment in dysarthric speech and review a recently proposed utterance verification approach to it. We then investigate the application of linguistic segment level confidence based ANN training in dysarthric speech assessment. Finally we summarise our findings.

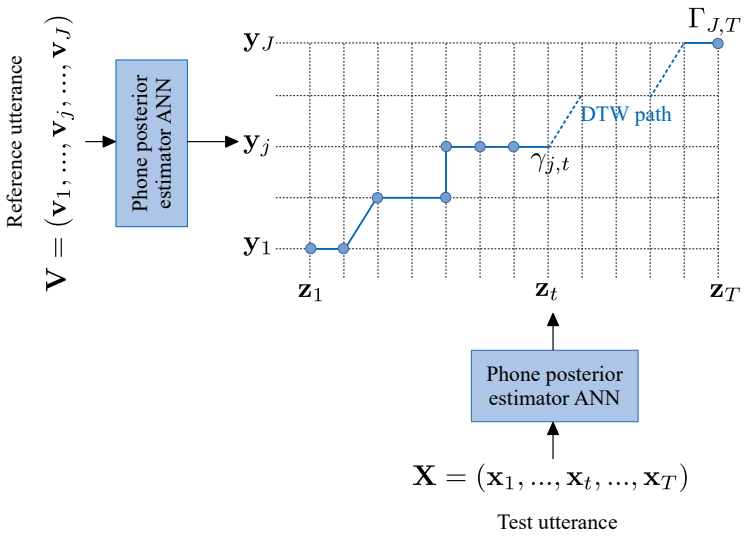


Figure 7.1: Matching utterances using phone posterior probability sequences.

7.1 Posterior feature based intelligibility assessment

Let $\mathbf{Y} = (\mathbf{y}_j)_{j=1}^{J} = (\mathbf{y}_1, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_J)$ be a sequence of length J, where $\mathbf{y}_j = [y_j^1, y_j^2, ..., y_j^D]$ denotes the phone posterior probability feature vector of a reference utterance at time j, and D denotes the number of phones. Similarly, let $\mathbf{Z} = (\mathbf{z}_t)_{t=1}^{T}$, where $\mathbf{z}_t = [z_t^1, z_t^2, ..., z_t^D]$ denotes the phone posterior feature vector of a test utterance at time frame t. As illustrated in Fig. 7.1, the representations $\mathbf{Y}$ and $\mathbf{Z}$ are obtained by employing an ANN acoustic model that takes feature representations as input. In order to measure how close the test utterance is to the reference, dynamic programming is performed, where the local score γ_{jt} is computed as

$$\gamma_{j,t} = \mathbb{KL}(\mathbf{y}_j \parallel \mathbf{z}_t) = \sum_{d=1}^{D} y_j^d \log\left(\frac{y_j^d}{z_t^d}\right). \tag{7.1}$$

A cumulative score at $\Gamma_{j,t}$ can be recursively computed as

$$\Gamma_{j,t} = \gamma_{j,t} + \min\left(\Gamma_{j,t-1}, \Gamma_{j-1,t}, \Gamma_{j-1,t-1}\right), \tag{7.2}$$

to obtain the final score $\Gamma_{J,T}$. This, normalised by the path length, yields a measure $\hat{\Gamma}_{J,T}$ of intelligibility; the lower the score, the better the intelligibility. In other words, $\hat{\Gamma}_{J,T} = \hat{\Gamma}(\mathbf{Y}, \mathbf{Z})$ indicates how close the test utterance is – in terms of its phonetic content – to the reference utterance.

7.2 Background on ASR and the ANN training in hybrid systems

The ANN discussed in Sec. 7.1 is typically trained in an ASR system scenario, where speech signals are converted into sequences of words or text. This section provides a background of training such networks.

In HMM based ASR (Rabiner, 1989), the *likelihood* of an HMM state q_t at the time frame t, labelled l^i, is estimated (Rasipuram & Magimai-Doss, 2015) as:

$$\begin{aligned} p(\mathbf{x}_t|q_t = l^i) &= \sum_{d=1}^{D} p(\mathbf{x}_t, a^d|q_t = l^i) \\ &= \sum_{d=1}^{D} P(a^d|q_t = l^i) \cdot p(\mathbf{x}_t|a^d, q_t = l^i) \qquad (7.3) \\ &= \sum_{d=1}^{D} P(a^d|q_t = l^i) \cdot p(\mathbf{x}_t|a^d) \,, \qquad (7.4) \end{aligned}$$

where $\mathbf{x}_t$ denotes the acoustic feature observation at t, $l^i \in \{1, \ldots, I\}$ and $\{a^d\}_{d=1}^{D}$ denotes a set acoustic units. Eqn. (7.4) results from the assumption that $\mathbf{x}_t \perp\!\!\!\perp q_t|a^d$. In the case of a context dependent subword unit based ASR system, I is the number of context-dependent subword units; D is the number of clustered context-dependent states; and the vector $[P(a^d|q_t = l^i)]_{d=1}^{D}$ is either a Kronecker delta distribution or a soft distribution depending upon whether the relationship between a^d and state $q_t = l^i$ is deterministic or probabilistic (Rasipuram & Magimai-Doss, 2015). In standard HMM-based ASR systems this relationship is deterministic given the state tying decision tree, i.e. if $l^i \mapsto a^{d'}$ then $P(a^{d'}|q_t = l^i) = 1$ and $P(a^d|q_t = l^i) = 0 \; \forall d \neq d'$. $p(\mathbf{x}_t|a^d)$ can be estimated either using Gaussian mixture models (GMM) or using artificial neural networks (ANN). In the case of ANNs, $p(\mathbf{x}_t|a^d)$ is estimated as a scaled-likelihood $p_{sl}(\mathbf{x}_t|a^d)$ (Bourlard & Morgan, 1994):

$$p_{sl}(\mathbf{x}_t|a^d) = \frac{p(\mathbf{x}_t|a^d)}{p(\mathbf{x}_t)} = \frac{P(a^d|\mathbf{x}_t)}{P(a^d)}, \qquad (7.5)$$

where $P(a^d|\mathbf{x}_t)$ denotes the posterior probability of the acoustic unit a^d estimated by the ANN and $P(a^d)$ is its prior probability.

The current section focuses on the training of the ANNs to estimate $P(a^d|\mathbf{x}_t)$. The ANN can be trained using embedded Viterbi expectation-maximisation (EM) algorithm. In the expectation step (E-step), given the current neural network, an alignment between the HMM state sequences and the acoustic feature sequences is obtained. In the maximisation step (M-step), given the alignment, a new neural network is trained. In practice, to reduce the training time, the alignments are typically obtained using an HMM/GMM system and the M-step is carried out once (Dahl et al., 2012; Hinton et al., 2012).

Although the alignment is obtained by imposing a sequence structure, the ANN is trained

using an individual frame-level discriminative criterion, viz. cross-entropy (CE). This training criterion corresponds to a maximum mutual information (MMI) estimation of parameters in terms of classifying phones (Bridle, 1990). However, this may be sub-optimal since the sequence structure in the data is being ignored. One class of methods which can address this limitation is segmental models (Ostendorf et al., 1996), where the HMM states emit segments instead of frames. These ideas have been used in ANN- and deep learning based models. Such methods often depend on the availability of segment boundaries in the data, and thus require an additional complexity to determine and handle variable length segments both during training and decoding. We mention a few examples among numerous works in the literature here. Austin et al. (1991) converted segments into fixed length segments by sampling the segments linearly. This requires an additional rescoring process during decoding after the first pass, since an initial segmentation is unavailable during real-time testing. Abdel-Hamid et al. (2013) use similar sampling methods to carry out training, but expensively loop over multiple possible segment boundaries during decoding. Zweig and Nguyen (2009) used a conditional random field based backend to combine outputs at multiple segment levels. Kong et al. (2016) used a recurrent architecture and Beck et al. (2018) used an encoder-decoder based framework. Another class of methods that handle segments of speech together are sequence discriminative training (SDT) (Povey et al., 2016; Veselý et al., 2013) methods, where the training objectives are computed at sequence levels, while keeping the model complexity unaltered.

7.3 Proposed segmental training approach

In this section, we first establish a link between the estimation of linguistic unit level confidences using $P(a^d|\mathbf{x}_t)$ (Bernardis & Bourlard, 1998; Williams & Renals, 1999) and the training of neural networks. Through this link we propose a segment-level training paradigm that requires no architectural changes or sophistication, and can be envisaged as a maximisation of segment- or linguistic unit level confidences. In other words, it can be viewed as the maximisation of the match between linguistic units and segments of acoustic feature observations.

In ASR related applications, confidence measures are used to measure how well an acoustic observation sequence $\mathbf{X} = (\mathbf{x}_t)_{t=1}^{T} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$ and a word hypothesis $W = (w_r)_{r=1}^{R}$ match, given the trained parameters of the system. In a similar vein, the training of the ANN for ASR can be posed as finding the parameters that maximise the match between between $\mathbf{X}$ and W. In both cases, matching $\mathbf{X}$ and W is a common factor; where the higher the confidence, the better the match. Given this understanding, in this section we show that confidence measures based on "local posterior" probability estimates $P(a^d|\mathbf{x}_t)$ can naturally serve as objective functions for a segment-level training of the ANNs.

7.3.1 Segment-level confidence estimation from local posteriors

Let $W = (w_r)_{r=1}^{R}$ constitute a sequence of phones $(ph_k)_{k=1}^{K}$, and further constitute a sequence of sub-phonemic HMM states $(s_j)_{j=1}^{J}$ as defined by the topology. In the framework of *acceptor HMMs* (Bourlard & Morgan, 1994; Williams & Renals, 1998), various confidence measures based on local posterior probability estimates have been proposed. Specifically, given an alignment between $\mathbf{X}$ and W and the local posterior probability estimates, one of the methods to estimate the HMM state level confidence $CM(s_j)$ is by rescoring the state segment s_j as

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \log\left(P(q_t = l^j|\mathbf{x}_t)\right)}{e(s_j) - b(s_j) + 1}, \tag{7.6}$$

where l^j is its label, and $b(s_j)$ and $e(s_j)$ denote its beginning and end frames respectively. This is computed, given the one-to-one map between the state l^j and the set of acoustic units $\{a^d\}_{d=1}^{D}$. In other words, if $l^j \mapsto a^{d'}$ then

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \log\left(P(a^{d'}|\mathbf{x}_t)\right)}{e(s_j) - b(s_j) + 1}. \tag{7.7}$$

A word level confidence $wCM(w_r)$ for the word w_r constituting the state sequence $(s_{j+m})_{m=1}^{M_{w_r}}$ can be further estimated as (Bernardis & Bourlard, 1998)

$$wCM(w_r) = \frac{1}{M_{w_r}} \sum_{m=1}^{M_{w_r}} CM(s_{j+m}), \tag{7.8}$$

where M_{w_r} is the number of states in w_r.

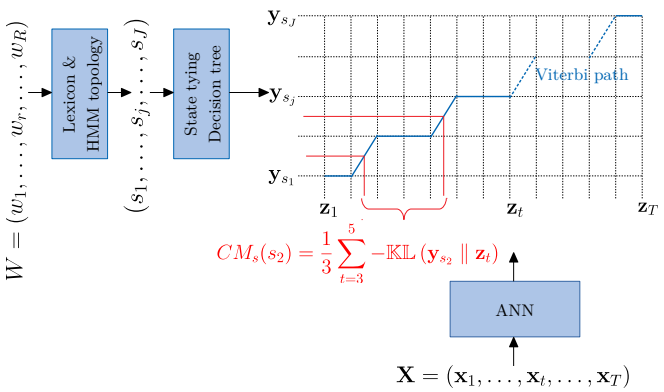


Figure 7.2: Estimating state confidences from local posterior probabilities.

Let $y_{s_j}^d = P(a^d|q_t = l^j)$; then the vector $\mathbf{y}_{s_j} = (y_{s_j}^d)_{d=1}^{D}$ describes the mapping from s_j to $\{a^d\}_{d=1}^{D}$. As illustrated in Fig. 7.2, this mapping is typically defined by the state tying decision tree. In other words, the sequence $(s_j)_{j=1}^{J}$ that corresponds to a word hypothesis is mapped to

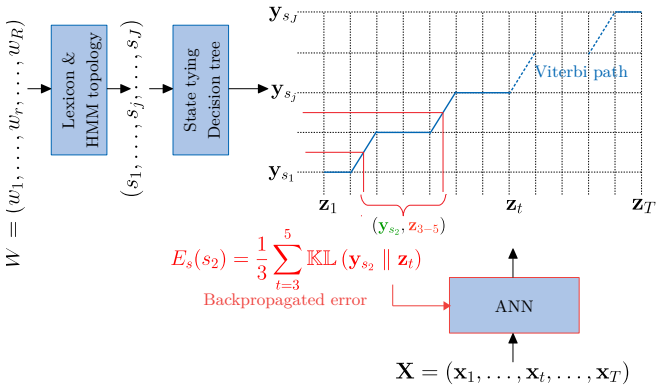


Figure 7.3: Training from state level confidence scores.

$\mathbf{Y} = \left(\mathbf{y}_{s_j}\right)_{j=1}^{J}$. Similarly, let $z_t^d = P(a^d|\mathbf{x}_t)$; then the vector $\mathbf{z}_t = (z_t^d)_{d=1}^{D}$ denotes the output of the ANN at the time frame t and we can define the sequence $\mathbf{Z} = (\mathbf{z}_t)_{t=1}^{T}$ that corresponds to an acoustic observation. Without loss of generality, the estimation of confidence by rescoring can be expressed as a matching of the two posterior probability sequences $\mathbf{Y}$ and $\mathbf{Z}$ with a local cost based on Kullback-Leibler divergence $\mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)$. More precisely,

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} -\mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)}{e(s_j) - b(s_j) + 1}. \tag{7.9}$$

It can be verified that, as $\mathbf{y}_{s_j}$ is a Kronecker delta distribution given $s_j \mapsto d'$, $\mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)$ reduces to *cross entropy* $-\log\left(P(a^{d'}|\mathbf{x}_t)\right)$.

It is worth mentioning that Eqn. (7.9) can be generalised further to the case when $\mathbf{y}_{s_j}$ is a soft distribution, as computing the KL-divergence between two probability distributions is equivalent to hypothesis testing (Blahut, 1974; Eguchi & Copas, 2006). Indeed such confidence measures have been employed earlier for utterance verification (Ullmann, Rasipuram, et al., 2015) and for non-native speech assessment (Rasipuram & Magimai-Doss, 2015) tasks.

Given this understanding, it is also worth reviewing that the posterior feature matching described in Sec. 7.1 finds an alignment between the reference acoustic feature sequence and the test sequence. Similarly, the E-step in the ASR system training finds an alignment between a given HMM state sequence and the corresponding acoustic feature sequence. If the ASR E-step uses the $\mathbf{Y}$ estimated from an acoustic reference for computing alignment using KL divergence local cost, this can be seen as utilising soft targets, and therefore, relaxing the HMM transition constraints makes it identical to the posterior feature matching described in Sec. 7.1.

7.3.2 Segment-level training of the ANNs based on confidence measures

Given the segmentation of the training data, the ANN training is treated as a separate classifier training, by one hot encoding of the targets and minimising a frame level cross entropy criterion

$$\begin{aligned} E_f(t) &= \mathbb{KL}\left(\gamma_d \parallel \mathbf{z}_t\right) \\ &= -\log\Big(P(a^d|\mathbf{x}_t)\Big), \end{aligned} \tag{7.10}$$

where γ_d is a Kronecker delta distribution based on a one-hot encoding and $\mathbf{z}_t$ is the output of the ANN. From this perspective, given the pairs of input features and their target classes as tuples, there is no difference in the training mechanism whether one wants to classify phones, speakers, images, text or so on. This is a non-segmental way to train ANNs.

On the contrary, given the understanding from section 7.3.1, the ANN training for hybrid HMM/ANN ASR can be formulated as finding the parameters that increase the match between the observation sequences and the sequence of states or segments. More precisely, as illustrated in Fig. 7.3, the error function can be based on rescoring of the segments, i.e. based on confidence measures. It is important to mention that whilst the notion of one-hot-encoding of the targets comes from a pattern classification point of view, in our formulation one-hot-encoding results from the one-to-one mapping between the states and $\{a^d\}_{d=1}^D$. As discussed earlier the targets can be soft, i.e. the map between the states and $\{a^d\}_{d=1}^D$ can be probabilistic. Furthermore, as shown earlier as well as in the literature, the cross entropy error criterion emerges from KL-divergence with the target distributions being Kronecker delta distributions (Makhoul, 1991). In the case of soft targets, it corresponds to an additional entropy term of the target distributions, that remains constant with respect to the ANN parameters, and thus makes no difference in the training.

In the case where the segments represent HMM states, a state-level error function $E_s(s_j)$ that can be defined to minimise in a stochastic gradient descent training is

$$E_s(s_j) = -CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} \mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right)}{e(s_j) - b(s_j) + 1}, \tag{7.11}$$

while in the case where the segments represent phone units, a phone-level error function $E_{ph}(ph_k)$ that is minimised can be based on Eqn. (7.8):

$$E_{ph}(ph_k) = \frac{1}{N_{ph_k}} \sum_{n=1}^{N_{ph_k}} E(s_{j+n}), \tag{7.12}$$

where the phone ph_k constitutes N_{ph_k} states: $(s_{j+1}, \ldots, s_{j+N_{ph_k}})$.

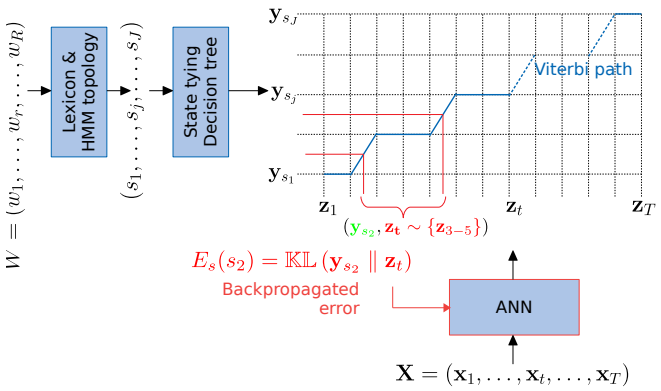


Figure 7.4: Training from state level subsampling.

7.3.3 Segment-level training of the ANNs based on subsampling

The previous section established that the conventional cross-entropy is linked to the frame-level confidence measure

$$CM_f(s_j, t) = -E_f(t) = -\mathbb{KL}\left(\mathbf{y}_{s_j} \parallel \mathbf{z}_t\right), \tag{7.13}$$

and that the state-level confidence measure can be derived from these frame-level confidences

$$CM(s_j) = \frac{\sum_{t=b(s_j)}^{e(s_j)} CM_f(s_j, t)}{e(s_j) - b(s_j) + 1}. \tag{7.14}$$

For each phone state s_j, its constituent frame level confidences $CM_f(s_j, t)$ can be interpreted as sample estimates of the overall segment's confidence measure $CM_s(s_j)$. Thus a new confidence measure $CM_{ss}(s_j)$ can be defined by drawing a sample uniformly (denoted as $\mathbb{U}\{.\}$) from the frame level confidence scores within the state:

$$CM_{ss}(s_j) \sim \mathbb{U}\left\{CM_f(s_j, t);\, t = b(s_j)...e(s_j)\right\}. \tag{7.15}$$

It is straight-forward to verify that

$$\mathbb{E}\left[CM_{ss}(s_j)\right] = CM_s(s_j), \tag{7.16}$$

where $\mathbb{E}[.]$ denotes expectation. In practice, for a given linguistic segment, the sampling is performed at each epoch of the ANN training. Since the training process consists of several epochs, the error function $E_{ss}(s_j) = -CM_{ss}(s_j)$ yields a similar ANN as that of $E_s(s_j)$. Training with $E_{ss}(s_j)$ implies that the frame level confidence need not be computed for every frame in the linguistic segment, but instead for a single frame that is drawn from the segment in each epoch. Thus the training time reduces by the expected number of frames per linguistic

segment in the given corpus. The training process is illustrated in Fig. 7.4. Similar to above, we can also define a phone-level loss based on the proposed state level loss:

$$E_{ssph}(ph_k) = \frac{1}{N_{ph_k}} \sum_{n=1}^{N_{ph_k}} E_{ss}(s_{j+n}), \tag{7.17}$$

where the phone ph_k constitutes N_{ph_k} states: $(s_{j+1}, \ldots, s_{j+N_{ph_k}})$.

7.4 Experimental validation on ASR task

In this section, we investigate the effect of the proposed segment level ANN training on ASR and phone recognition performances.

7.4.1 Systems

We conducted ASR experiments on Mediaparl (Imseng et al., 2012) and AMI (Carletta et al., 2005) data sets. We performed studies on both the M-DE and M-FR parts of the data set. We followed the protocols set by Imseng et al. (2012) for their data preparation, pronunciation lexicon selection and language model (LM) building. We conducted the studies on the IHM data set. We conducted phone recognition studies on TIMIT corpus (Garofolo et al., 1993). We followed the standard Kaldi protocols for AMI and TIMIT. Table 7.1 provides a description of the experimental setup for all the data sets. We built ASR systems using Kaldi toolkit and

Table 7.1: Experimental setup on various corpora.

	AMI	M-DE	M-FR	TIMIT
Training hours	77.3	14.5	16.1	3.1
Phone set count	176	57	38	48
Vocabulary size	52.5k	16.7k	12.4k	48
LM order	3-gram	2-gram	2-gram	2-gram

Keras/Tensorflow tools. We used 39 dimensional Mel frequency cepstral coefficients (MFCC), $C_0 - C_{12} + \Delta + \Delta\Delta$, as the acoustic feature observations. AMI and TIMIT used the default speaker-level cepstral mean and variance normalisation (CMVN) in Kaldi setup, while M-DE and M-FR used an utterance-level CMVN.

The alignments for the training of ANNs were obtained using Kaldi pipeline, by first building *mono-tri3* HMM/GMMs and then building subspace GMM (SGMM) systems, which operate in three passes for decoding and alignment. The number of clustered context-dependent states for AMI, M-DE, M-FR and TIMIT were 4490, 2282, 2265 and 2112 respectively. The alignments for each data set was obtained from its corresponding SGMM system. For AMI, it is worth mentioning that the SGMM system development and the subsequent ANN training were carried out on the 70.2 hour subset of data with *clean* segmentation.

Table 7.2: Eval set WER on AMI, M-DE and M-FR corpora, and PER on TIMIT.

Err function →		E_f	E_s	E_{ph}	E_{ss}	E_{ssph}
AMI		32.4	30.5	**30.4**	**30.4**	30.5
	+sMBR	30.4	28.4	28.4	**28.1**	28.3
M-DE		20.5	19.9	19.6	19.7	**19.4**
	+sMBR	19.7	18.7	19.0	**18.3**	18.7
M-FR		21.8	20.8	**20.4**	20.8	20.6
	+sMBR	20.6	**18.9**	19.0	19.1	19.0
TIMIT		22.3	**21.2**	21.3	21.5	21.5

For each data set, we trained three deep neural networks (DNNs) corresponding to the three error functions E_f, E_s and E_{ph}. All the DNNs had three hidden layers with 1024 units with rectified linear activations in each hidden layer. The input to the DNNs were 13 dimensional MFCCs with five frames each in the preceding and the following context and with $\Delta + \Delta\Delta$, i.e. 429 dimensional feature input. The training was based on stochastic gradient descent with a decaying learning rate. Post this training, we also used a standard sequence discriminative training, viz. state-level minimum Bayes risk (sMBR), for AMI, M-DE and M-FR corpora.

In the decoding process, the priors $P(a^d)$ in Eqn. (7.5) were estimated from the state segment counts rather than from the frame label counts.

7.4.2 Results

Table 7.2 shows the word error rates (WER) on AMI, M-DE and M-FR corpora and phoneme error rate (PER) for TIMIT corpus. *+sMBR* row presents the performance with an additional sMBR training. It can be observed that the proposed trainings outperform E_f based training. It is interesting to observe that, across all the three data sets, the proposed trainings yield performances comparable to the E_f based training followed by sMBR.

7.4.3 Analysis

This section presents an analysis of the proposed approach.

7.4.3.1 Generalisation to different architectures and front-ends

The proposed segment-level training approach does not presume any particular feature, front-end processing or ANN architecture. Nevertheless, a question that arises is whether the observations made in the previous section generalise across different architectures and front-ends. To investigate this, we conducted two ASR studies:

1. Training systems on the AMI data set with feature-space maximum likelihood linear regression (fMLLR) speaker transform based features and 25-frame splicing, concatenated with speaker-level iVectors, modelled with DNNs comprising six hidden layers with 2048 units each, and trained with dropout on speed-perturbed data. Table 7.3 presents the performance with the three error functions and with sMBR, as done before, in terms of WER.

Table 7.3: Performance on AMI data set with fMLLR+iVector front-end.

Error function →	E_f	E_s	E_{ph}	E_{ss}	E_{ssph}
AMI	27.3	26.0	26.4	26.8	27.2
+sMBR	25.1	23.9	24.1	24.5	24.7

2. Training convolutional neural network (CNN) based systems that take raw speech as input (Palaz et al., 2013) on the M-DE data set. The CNN-based systems comprised four convolutional layers followed by three fully connected hidden layers with 1024 units each. Table 7.4 presents their performances in terms of WER.

Table 7.4: CNN-based system performance on M-DE data set.

Error function →	E_f	E_s	E_{ph}
M-DE	20.8	19.6	19.3

In both the studies, the proposed trainings of the ANNs consistently yield better systems than E_f based training.

7.4.3.2 Effect of the segment duration normalisation

Different phones can have different durations; this can vary due to reasons such as the type of speech or speakers, for e.g. read versus spontaneous speech, native versus non-native speakers, etc. Also, the lengths of the silence portions can vary, for instance due to variations in a preceding voice activity detector's performance. Such differences in the durations of segments could affect the ANN training. Error functions E_s, E_{ph}, E_{ss} and E_{ssph} inherently normalise the durations of the segments, and thus may handle their variations better.

To investigate this, we first simulated a study on the TIMIT corpus, where silence was artificially added at the beginning and the end of each utterance. We considered two cases: (a) two seconds of silence is added at both the ends (4s/utt) and (b) five seconds silence is added at both the ends (10s/utt). We trained three hidden layer DNNs corresponding to E_f and the segment-level error functions, as done earlier. Table 7.5 presents the results in terms of PER, when tested on silence-added utterances. It can be observed that, when trained with E_f, the phone recognition performance drastically degrades as the silence length increases at both

the ends of the utterances, while when trained with the proposed approaches, the drop in the performance is significantly less. Investigating the ability to handle phone duration variations is part of our future work.

Table 7.5: PER on TIMIT corpus for the effect of segment duration normalisation study. $4s$/utt and $10s$/utt denote the addition of 2 seconds silence and 5 seconds of silence respectively at both the ends of the utterance.

Error function →		E_f	E_s	E_{ph}	E_{ss}	E_{ssph}
TIMIT	$4s$/utt	23.0	22.0	21.8	22.2	**21.5**
	$10s$/utt	35.6	22.7	22.5	22.8	**22.3**

To investigate the ability to handle phone duration variations, we conducted the second study with emotional data, since emotion can alter the duration of syllables, pauses and speaking rate. In this study on emotional prosody corpus (EPC) (Liberman et al., 2002), all the utterances were decoded using the models trained on AMI corpus, keeping the same pronunciation lexicon and the language model that were used on the AMI decoding experiments. Results in Table. 7.6 indicate that (i) segment-level training helps in improving the WER of emotional data, as compared to frame-level training, and (ii) a further SDT does not improve the results.

Table 7.6: Performance (WER) on EPC data set.

Error function →		E_f	E_s	E_{ph}	E_{ss}	E_{ssph}
EPC	CE	49.9	46.1	46.0	46.4	**45.6**
	+sMBR	58.5	47.7	46.4	48.8	**45.8**

7.4.3.3 Differences in posterior estimation between frame- and segment-level models

The proposed loss function E_{ss} utilises an instantaneous confidence measure CM_{ss} that is similar to CM_s, as given by Eqn. (7.16), but different from the existing measure CM_f. To experimentally validate this point, we conducted a study by estimating the posterior probabilities from these systems and comparing them using symmetric KL divergence measure averaged per frame. Table. 7.7 shows the comparison on AMI dev set, which shows that E_s and E_{ss} are closer to each other than they are from E_f.

Table 7.7: Symmetric KL divergence per frame on AMI dev set.

Systems →	$E_f - E_s$	$E_f - E_{ss}$	$E_s - E_{ss}$
AMI	1.95	2.00	1.24

7.4.3.4 Analysis of the training time

The segment-level cost functions require an additional computational overhead of preparing batches in segments. However, as mentioned in Sec. 7.3.2, the E_{ss} loss reduces the computations required for training. To verify this, we measured the time taken for training the ANNs on the TIMIT corpus. All the trainings used the same memory settings, drivers and compute capabilities on identical GeForce GTX1080TI machines. Despite the overhead, Table. 7.8 shows that E_{ss} achieved faster training than others, including E_f.

Table 7.8: Analysis of the training time on the TIMIT corpus

Error function →	E_f	E_s	E_{ph}	E_{ss}	E_{ssph}
Seconds per epoch	34	62	199	**15**	47
Number of epochs	29	35	40	36	39

7.5 Validation on intelligibility assessment

The previous section showed that the proposed segment-level training improves the ASR task. In this section, we investigate the proposed approach on dysarthric speech intelligibility assessment.

Speech intelligibility of a speaker with dysarthria can be automatically assessed by measuring the percentage of words correctly spoken by the speaker. This is similar to isolated word pronunciation test in a clinical setting, where a speaker with dysarthria pronounces a set of isolated words, and the speech intelligibility is measured as the percentage of these words that are correctly identified by human listeners (ASHA, 2021; Duffy, 2012; Kent et al., 1989a). Building upon the comparison of phone posterior probabilities as discussed in Sec. 7.1, below is the approach, first proposed by Fritsch and Magimai.-Doss, 2021.

7.5.1 Intelligibility assessment for speakers with dysarthria

Let the intelligibility test constitute R words to be tested for the given speaker with dysarthria, i.e. $W_r \in \{W_1 \cdots W_R\}$. Let $\mathbf{Z}_r$ denote an acoustic representation of the test speaker's utterance corresponding to the word W_r. For each such word, there is a pre-defined set of K reference utterances spoken by K control (i.e. healthy) speakers containing the same word. Let $\{\mathbf{Y}_r^{(k)}; k = 1 \cdots K\}$ denote the corresponding set of reference acoustic representations. Each pair of the posterior probability sequences $\left(\mathbf{Y}_r^{(k)}, \mathbf{Z}_r\right)$ is compared using the approach described in Sec. 7.1 and a distance $\hat{\Gamma}\left(\mathbf{Y}_r^{(k)}, \mathbf{Z}_r\right)$ is computed. The comparison of such probability distributions, discussed in Sec. 7.1, using KL-divergence or other measures such as Bhattacharya distance, is equivalent to hypothesis testing and yields an estimate of the log-likelihood ratio (Blahut, 1974; Kailath, 1967). In this case, since $\hat{\Gamma}(\mathbf{Y}, \mathbf{Z})$ is computed over the phone posterior probability

space, it corresponds to the log-likelihood ratio of whether the two utterances differ in their phone sequence content or not. Therefore, a threshold τ can be employed over the score $\hat{\Gamma}(\mathbf{Y},\mathbf{Z})$ to verify whether the two utterances have the same phone sequence. The word W_r is decided to be correctly recognised when at least half of the K comparisons yield favourable outcomes, i.e. that they are the same word. The percentage of the correctly recognised words among the total R gives the speaker's objective intelligibility score. This approach is summarised in Algorithm 1.

Algorithm 1: Objective intelligibility score estimation

Set number of words correctly identified $N = 0$;
for $r \leftarrow 1$ **to** R **do**
  Set word vote $V = 0$;
  for $k \leftarrow 1$ **to** K **do**
    Compute the score of match $\hat{\Gamma}\left(\mathbf{Y}_r^{(k)},\mathbf{Z}_r\right)$;
    if $\hat{\Gamma}\left(\mathbf{Y}_r^{(k)},\mathbf{Z}_r\right) < \tau$ **then**
      $V \leftarrow V + 1$;
    end
  end
  if $V \geq \frac{K}{2}$ **then**
    $N \leftarrow N + 1$;
  end
end
Result: Intelligibility $\leftarrow \frac{N}{R} \times 100\%$

τ is determined in the following manner:

1. Creating same word utterance pairs from the control speakers data, matching them and obtaining a distribution of global match score for the same word hypothesis;
2. Creating different word utterance pairs from the control speakers data, matching them and obtaining a distribution of global match score for NOT the same word hypothesis; and
3. Determining the threshold at the intersection of the two distributions.

7.5.2 Systems

We use the above approach to evaluate the segment-level training methods on dysarthric speech intelligibility assessment, using UA-Speech database (Kim et al., 2008). Sec. 3.2.5 gives information on the data set. The threshold τ was obtained for each of the posterior spaces using all the data from the $K = 13$ control speakers. The number of words tested was $R = 765$. The acoustic models trained on the AMI corpus were used to extract the clustered context dependent phone posteriors, which were then marginalised to get the monophone posteriors.

7.5.3 Results

Table 7.9: Performance of segment-level training on dysarthric speech intelligibility assessment in terms of correlations.

	Pearson's (p-value)	Spearman's (p-value)
P-ESTOI (Janbakhshi et al., 2019b)	0.94	0.94
E_f (Fritsch & Magimai.-Doss, 2021)	0.950 (5.52e-8)	0.957 (2.29e-8)
E_f	0.954 (3.63e-8)	0.952 (4.88e-8)
E_s	**0.968** (3.30e-9)	0.962 (9.78e-9)
E_{ph}	0.965 (6.50e-9)	**0.966** (5.15e-9)
E_{ss}	**0.968** (3.78e-9)	0.948 (7.70e-8)
E_{ssph}	0.964 (7.11e-9)	**0.966** (5.15e-9)

Table 7.5.3 shows the results, including two baseline results for comparison: P-ESTOI method of Janbakhshi et al. (2019b), a spectral feature based intelligibility estimation method that uses DTW, and E_f evaluated by Fritsch and Magimai.-Doss (2021) using ANN trained on Switchboard conversational telephone speech corpus. It can be seen that the proposed segment level training methods yield objective intelligibility scores that are consistently more correlated with the human scores than those from the frame-level training.

7.6 Summary

This work investigated incorporating linguistic segment level confidences into the training of ANNs used for intelligibility assessment and in hybrid HMM/ANN ASR. Through experimental studies on both intelligibility assessment and ASR, we showed that such segment level trainings of ANNs yield better correlations of dysarthric speech intelligibility with human scores, as well as better ASR systems. In the case of ASR, these gains in the performances are also sustained with sequence discriminative training. Furthermore, we demonstrated that the proposed segment level training approaches (a) are generalisable across model architectures and front-ends, and (b) lead to systems that are robust to duration variations. Finally, we showed that the subsampling based error function E_{ss} is closer to the segment-level error function E_s than the frame level cross-entropy function, and trains faster than all the methods evaluated.

8 Speech pseudonymization and its assessment

The previous chapters focused on assessing several aspects of speech. However, as introduced in Chapter 1, it may be desirable to first modify some components of speech and then make an assessment. This chapter deals with one such scenario, viz. preserving the speaker's privacy in speech. The rest of the chapter is organised as follows. Secion 8.1 provides an introduction to speech privacy and presents an overview of the key contributions, Section 8.2 describes the signal processing approach to anonymization developed for adjustable deterministic pseudonymization of speech. Section 8.3 describes the listening experiments conducted using human listeners. The experimental setup for the 2020 VoicePrivacy Challenge and the results are described in Section 8.4. Section 8.5 describes additional analysis in terms of intelligibility measurement based on dynamic time warping (DTW), formant measurement in pseudonymized speech and experiments on dysarthria prediction. Section 8.6 presents a discussion and Section 8.7 presents a summary.

8.1 Introduction

The availability of large speech corpora in combination with advanced statistical techniques improved speech technology tremendously (Ardila et al., 2019; Ning et al., 2019; Panayotov et al., 2015; Zhang et al., 2017). But speech recordings pose a possible privacy risk. More and more resources of speech data are shared on public platforms each day. While personally identifiable information such as name, age etc. of the speaker can be easily hidden, speech itself remains as a personal identifier of the speaker. With the increased use of speaker verification technologies, sensitive information related to speakers could be extracted from their speech and lead to harm (Korshunov & Marcel, 2017; Kucur Ergunay et al., 2015). This is especially true when the speakers have medical conditions, are minors, or the spoken content is sensitive. But these are also groups that might benefit from improvements in speech technology tailored to their needs.

The privacy risks resulting from sharing speech recordings would be mitigated if the probability of speaker (re-)identification could be reduced while retaining useful linguistic and

paralinguistic features. Speech anonymization methods, thus, aim at decoupling the hazardous identity of the speaker from the interesting linguistic and paralinguistic aspect of the speech. That is, anonymization removes the information about *who spoke it* from the speech while preserving *what was spoken* and *how it was spoken*. The "perfect" anonymization procedure would correspond to having the spoken text read by another speaker in the exact same manner. And some current speech anonymization applications work using components of a speech recogniser coupled to a neural network based speech synthesizer (Fang et al., 2019; Mawalim et al., 2020). However, such an approach only preserves the verbal content of the speech, and at best some of the prosodic aspects. Such an approach may not be able to preserve paralinguistic features of interest, such as the expressed emotions, articulation changes depending on the speaking skills or pathological conditions etc., and in general may not preserve the linguistic detail. Thus, such an anonymization may not be useful in scenarios, such as (i) patients with dysarthria uploading their speech for evaluation, (ii) children or language learners submitting their utterances for evaluation, where preserving paralinguistic information is important. An alternate way could be to use signal-processing approaches that directly alter the spectral properties of the original utterance for anonymization based on prior knowledge. Such an approach that uses the McAdams coefficient (Patino et al., 2020a) exists. It is based upon short-time linear prediction analysis, where a constant exponentiation is applied to the angle of the complex poles, thereby expanding or contracting the timbre or the spectral envelope at the formant locations (Patino et al., 2020b). However its performance is inferior to that of the neural based approach in terms of automatic speech recognition (ASR) and automatic speaker verification (ASV). However, signal processing based approaches have the advantage over most statistical and machine learning approaches that the changes made and their effects observed can be explained and, ideally, controlled; hence there is an interest in improving such controllable approaches. A downside of the existing speech anonymization applications is also the degraded quality of the transformed (anonymized) speech (Srivastava et al., 2020) which reduces their usefulness. The research community has acknowledged these problems, and in 2020 a special challenge for improving anonymization of speech has been organised (Tomashenko et al., 2020a, 2020b).

The literature on data anonymization (e.g. Finck & Pallas, 2020; Rubinstein & Hartzog, 2016; Stalla-Bourdillon & Knight, 2017) can be crudely summarised as "anonymous data is not useful, useful data is not anonymous". This is also likely to be true for speech anonymization transforms. Therefore, reversible anonymization of speech, also called *pseudonymization*, could shift the risk-benefit balance for sharing speech corpora towards more sharing, and is therefore of potential interest. Contrary to "true" (i.e. irreversible) anonymization, pseudonymization is a more practical approach to anonymization, since it assumes that data can be re-identified, in principle, with the help of additional information that is hidden during the anonymization. The risk of re-identification of pseudonymized data is then the risk that the hidden information can be reconstructed by an attacker. Pseudonymization of speech will always be a trade-off between the risk of re-identification and usefulness. Thus, there is a need to develop such pseudonymization methods so that several speech applications can

benefit from the privacy benefits they offer.

This work, carried out in collaboration with Dr. Rob van Son from the Netherlands Cancer Institute, aims at developing a pseudonymization approach that is adjustable in the level of information removed from the speech, while still preserving relevant features enough to make the resultant speech useful. The proposed approach uses a series of signal processing steps to transform a given speaker's speech to tailor to a desired vocal profile (cf. Kung, 2018), configurable in terms of the formant frequencies, fundamental frequency and speaking rate. We conduct three sets of studies to demonstrate the potential of the proposed pseudonymization approach:

1. First, we validate the proposed approach through *ABX* pilot tests. These studies are carried out to ascertain how well the proposed approach obfuscates the speaker identity for expert and naive listeners.

2. Second, we validate the proposed approach in the framework of VoicePrivacy 2020 challenge (Tomashenko et al., 2020a, 2020c) by studying it against two anonymization approaches, a neural source filtering based approach and signal processing-based McAdams approach. We also perform ablation experiments to investigate which part of the proposed approach (related to the source, system or speaking rate) plays a crucial role in obfuscating speaker identity.

3. Third, we conduct studies that extend beyond the scope of VoicePrivacy 2020 challenge. In the VoicePrivacy 2020 challenge, the preservation of intelligibility is assessed through automatic speech recognition. Such a method can be prone to errors related to the availability of a suitable language model and a pronunciation lexicon. So, we propose the utilization of the phone posterior feature-based intrusive objective speech intelligibility approach described in Sec. 7.1. We also investigate the ability of the proposed pseudonymization method to preserve general articulatory features of speech by comparing the formant track movements measured on the anonymized and original recordings. Finally, investigations on speech anonymization have primarily laid emphasis on the preservation of intelligibility. However, speech also contains information other than the spoken message and speaker identity, such as paralinguistic information. So, we investigate the ability of the proposed pseudonymization approach to preserve such information through a dysarthric speech classification study.

8.2 Proposed pseudonymization method

As illustrated in Fig. 8.1, the proposed pseudonymization approach consists of estimating the speaker characteristics and obfuscating them by providing a different set of characteristics (referred to as the *target* speaker) to modify the utterances. As we see later, the same pseudonymization module can be used to de-pseudonymize the utterances, upon the knowledge of the original speaker's characteristics.

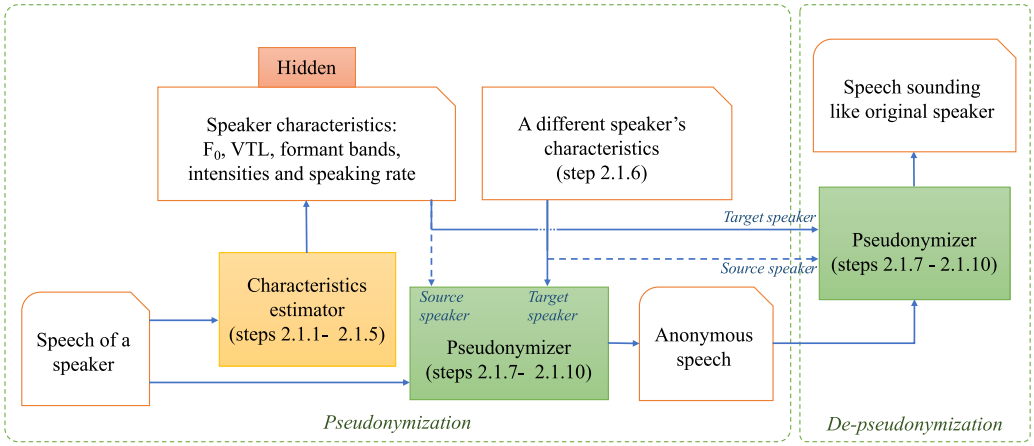


Figure 8.1: Illustration of speech pseudonymization, with the steps elaborated in Sec. 8.2.1.

Two sources of speaker variation useful for speaker identification can be distinguished, viz. *inherent* features, i.e., those that derive from a speaker's anatomy and physiology, and *learned* features (O'Shaughnessy, 2000). This study aims at hiding the global and inherent features of speakers, i.e., the vocal tract related spectral features (cf. Almaadeed et al., 2016) and some learned features, i.e., pitch and speaking rate. This translates to making changes in speech that relate to vocal tract length, average formant frequencies and intensities, pitch, and speaking rate. The pieces of information thus hidden will be the original values of these quantities and the extent of the changes. The corresponding steps can be summarised as:

1. Change the speaking rate and fundamental frequency, and

2. Simulate a different vocal tract for the speaker.

The perceived acoustic length of the vocal tract of each speaker is changed to that of a desired speaker by changing the playback speed of the utterance. Specifically, the vocal tract length corresponds to formant values as follows: an increase of vocal tract length by a factor a induces a formant shift by a factor $1/a$. In the remainder of this chapter, the estimated vocal tract length (VTL) will be represented by the neutral first resonance frequency ϕ. A speaker's $\hat{\phi}$ is estimated from the first four formant frequencies according to Eq. 20 of Lammert and Narayanan (2015) using the proposed extension (Table 3, ibid.):

$$\hat{\phi} = 229 + 0.030\phi_1 + 0.082\phi_2 + 0.124\phi_3 + 0.354\phi_4 \tag{8.1}$$

where $\phi_i = F_i/(2i-1)$ can be considered as estimates of VTL from individual formants F_i. Speakers do not only differ in vocal tract length, but also in the vocal tract structure, defined by the global position of the formants, their bandwidths and intensities. The below section describes how each of these quantities are estimated and are used to pseudonymize speech.

8.2.1 Steps involved

8.2.1.1 Intensity normalisation

Normalise the intensity of each utterance to 70 dB (relative to 20 μPa).

8.2.1.2 Identifying the vowel segments and estimating the speaking rate

Estimate the speaking rate by automatically locating syllables from speech using peaks in the signal energy, that are preceded and succeeded by dips in energy as cues (De Jong & Wempe, 2009; van Son et al., 2018). The number of syllables normalised by the duration per speaker gives the speaking rate. This method requires no transcriptions.

8.2.1.3 Formant track estimation for each vowel region

1. Use short-time processing with a Gaussian-like window of 25 ms, repeated every 6.25 ms (see Sec. 8.2.2 for more details).

2. Formant track estimation in the vowel regions: Use linear prediction analysis and iterative formant estimation procedure from Lee (1988).

8.2.1.4 Speaker-specific VTL and formant frequency estimation

1. In each vowel segment, look for the most neutral frame, i.e. the frame with (F_1, F_2) closest to $(500, 1500)$.

2. Attribute the formant estimates F_{1-5} from this closest frame to the entire vowel segment.

3. Estimate the VTL of the speaker in the vowel segment using Eq. 8.1.

4. Compute the speaker's VTL by taking the mean VTL across each speaker's vowel segments.

5. Compute the speaker formant frequencies F_i by taking the median across each speaker's vowel segments.

8.2.1.5 Speaker-level formant band intensity estimation

1. The frequency spectrum of each speaker is divided into several *formant bands* based on the estimated VTL[I] ϕ, as

$$B_i = \begin{cases} \left[0,\ \frac{\phi}{2}\right], & i = 0 \\ \left[\frac{\phi}{2},\ 2\phi\right], & i = 1 \\ \left[2(i-1)\phi,\ 2i\phi\right], & i = 2,3,...,9 \end{cases} \tag{8.2}$$

   in Hz. Since F_i ($i = 1,2,...,9$) is typically around $(2i-1)\phi$, the bands are centered around the corresponding formant frequencies (except B_0 and B_1).[II]

2. Use a passband Hann filter to isolate the information in each band. The filter has the following properties: (i) it is real-valued and operates on the complex short-time Fourier transform (STFT) of the input utterance, independently across each time step, (ii) the passband frequencies and 3dB bandwidth are defined as above, (iii) the transition from stop band to pass band and vice versa spans $(i-1)\phi/5$ Hz.

3. Use the above filter on each utterance and measure the mean intensity per speaker per formant band, I_i, from the filtered utterances for $i = 0,1,2,...,5$.

8.2.1.6 Target parameters for pseudonymization

To pseudonymize the formants, the target frequencies, represented in terms of VTLs $\phi_i = F_i/(2i-1)$, can be randomly chosen in the range $\phi_i \pm 40$ and $\phi_i \pm 75$ Hz for F_{0-1} and F_{2-5}, respectively, and the intensities can be randomly chosen in the range 64±4.5, 67±2.5, 58±4.5, 50±8, 47±10, 45±9 dB (I_{0-5}±2SD), where SD denotes standard deviation. These values were chosen based on typical observations on ranges found in the speakers in the IFA corpus (Van Son et al., 2001) (5M/5F, see Experiment 1, Section 8.3.1.1). In an alternative setting where a given speaker is to be pseudonymized to a specific target speaker, the parameters ϕ, ϕ_i ($i = 1...5$), I_i ($i = 0...5$) and speaking rate can be pre-computed across several of the target speaker's utterances (preferably over 300 seconds spoken in a comparable style) and used.

8.2.1.7 VTL shifting

This is a time-domain processing method.

1. We have a VTL estimated for the current speaker and a VTL estimate for the target speaker: determine the factor $a = \phi^{(current)}/\phi^{(target)}$.

[I] We will use the symbol ϕ to mean $\hat{\phi}$ hereafter.

[II] E.g. $\phi = 500$ (a typical male value). So, $F_3 \approx 2500$ Hz, $B_3 = [2000,3000]$ Hz, $F_4 \approx 3500$ Hz, $B_4 = [3000,4000]$ Hz, etc. $B_1 = [250,1000]$ Hz, $B_0 = [0,250]$ Hz.

2. Resample the utterance to F_s/a, where F_s denotes the original sampling frequency (and consider that the sampling frequency is still F_s). This corresponds to a frequency scaling by $1/a$ to the original utterance's spectrum.

8.2.1.8 Duration and pitch change

This is a time-domain processing method. Estimate F_0 by using the standard autocorrelation method. Adjust the duration and fundamental frequency to match the desired duration (determined by the target speaker's speaking rate) and fundamental frequency using pitch synchronous overlap-add method (Moulines & Charpentier, 1990).

8.2.1.9 Formant band shifting

This is a frequency-domain processing method. For each formant, we aim at masking $\phi_i^{(current)}$ and shifting it to the frequency $\phi_i^{(target)}$ by modifying its intensity appropriately.

1. Use the steps of VTL shifting (from Sec. 8.2.1.7), by using $\phi_i^{(current)}$ and $\phi_i^{(target)}$ in the place of $\phi^{(current)}$ and $\phi^{(target)}$ respectively, to create a VTL shifted version of the current utterance, where the formant i is now at $F_i^{(target)}$.

2. Extract the band $B_i^{(current)}$ (Eq. 8.2) from the VTL shifted spectrogram using a Hann filter as described in Sec.8.2.1.5[III]. Use $I_i^{(target)}/I_i^{(current)}$ as the filter's gain.

3. Use a complementary bandstop Hann filter with unit gain on the current utterance's spectrogram to mask $\phi_i^{(current)}$ in the band.

4. Add the extracted band to the current spectrogram so that it now has $\phi_i^{(target)}$ (and then discard the VTL shifted spectrogram).

5. Repeat the above steps for each desired formant.

8.2.1.10 Additional processing to hide the speaker identity

Additional anonymizing steps consist of (i) exchanging the B_4 and B_5 bands by using the Hann filter method described above and (ii) adding modulated pink noise at the speaker's B_{6-9} bands to mask these formants. These steps were not used in the human listening experiments in Sec. 8.3.

Finally, reconstruct the corresponding utterance by taking inverse STFT. Note that, except for the overlap-add synthesis step and noise insertion, all the steps in this process are deterministic and reversible.

[III] $\phi_i^{(target)}$ is largely present in $B_i^{(current)}$, as this band heavily overlaps with $B_i^{(target)}$, but not always.

8.2.2 Implementation

The software is available on GitHub: van Son (2020c) and van Son (2020d). The program *Praat* (Boersma & Weenink, 2017) has two commands *Change gender...* and *Change speaker...* that use the same algorithm to perform the respective operations. This study uses the *Change gender...* command internally because it has options suitable for the proposed approach. In these commands, the desired new pitch is set as an absolute value, but it depends on correct pitch measurement in the source speech. Both commands work on the vocal tract length and duration by a *Formant shift ratio* and a *Duration factor*. To implement a change to a specified target vocal tract length and duration, or speaking rate, the estimated vocal tract length and speaking rate of the source speaker have to be supplied.

VTL is determined using the *Praat* `robust` formant option (Boersma & Weenink, 2017). Speaking rate is determined by the syllable rate determined from a modified version of a script by De Jong and Wempe (2009) taken from van Son et al. (2018).

To pseudonymize an utterance, the original values of the VTL (ϕ), median formant frequencies, pitch, and speaking rate are transformed to the chosen values of the (synthetic) target speaker. The `PseudonymizeSpeech.praat` script (van Son, 2020c) presented above can calculate these on-the-fly using a collection of speech recordings or can use a database of pre-calculated values. Pseudonymization examples are available with the script, also consult the manual at van Son (2020d).

8.3 Listening experiments

We conducted *ABX* pilot listening experiments where subjects have to identify which of the two utterances, *A* or *B*, was uttered by the speaker in *X*. These experiments were designed to test the efficacy of the proposed approach, in terms of the following questions.

1. Can experts identify a speaker from pseudonymized speech?
2. How does pseudonymization affect the reliability of speaker identification by naïve listeners?
3. How resilient is the method to re-identification?

8.3.1 Experimental setup

Pseudonymized sentences and sentence fragments were produced by running the `PseudonymizeSpeech.praat` script (van Son, 2020c; van Son, 2020d) with target values for a male-like Long Vocal Tract Length (Long VTL) and a female-like Short Vocal Tract Length (Short VTL). Randomised values were used for the frequencies and intensities of bands B_0, B_{3-5} (see Section 8.2.1.6). Three *ABX* listening experiments were performed, where one choice, *A* or *B*, is uttered

	A	B	X
Experiment 1 (IFA corpus) 4 expert listeners	LVT(spkr=i, utt=x) SVT(spkr=f, utt=u)	LVT(spkr=j, utt=y) SVT(spkr=g, utt=v)	SVT(spkr=i, utt=z) LVT(spkr=g, utt=w)
Experiment 2 (Parallel Audiobook Corpus) 8 naïve listeners	LVT(spkr=i, utt=x) SVT(spkr=f, utt=u) Original(spkr=d, utt=k)	LVT(spkr=j, utt=y) SVT(spkr=g, utt=v) Original(spkr=e, utt=l)	Original(spkr=i, utt=z) Original(spkr=g, utt=w) Original(spkr=d, utt=m)
Experiment 3 (2019 ASVspoof) 6 naïve/1 expert listeners	Inv_i[LVT(spkr=i, utt=x)] Inv_g[SVT(spkr=f, utt=u)] Original(spkr=d, utt=k)	Inv_i [LVT(spkr=j, utt=y)] Inv_g[SVT(spkr=g, utt=v)] Original(spkr=e, utt=l)	Original(spkr=i, utt=z) Original(spkr=g, utt=w) Original(spkr=d, utt=m)

Figure 8.2: ABX listening experiments. The subjects had to identify which of the two utterances, *A* or *B*, was spoken by the speaker in *X*. LVT()/SVT(): Stimulus pseudonymized to a Long/Short Vocal Tract length, Original(): Original recording as stimulus, Inv_i[]: Inverse, de-pseudonymized to the parameters of speaker 'i'. spkr: Speaker number, utt: Utterance number. For example, Inv_g[SVT(spkr=f, utt=u)] indicates a stimulus created from utterance 'u' from speaker 'f', pseudonymized to a Short Vocal Tract length, and then de-pseudonymized to the parameters of speaker 'g'. See the text for details.

by the same speaker as sound *X* and the other is a *distractor*, see Figure 8.2. Fully functional offline copies of the experiments are available from van Son (2020b).

8.3.1.1 Experiment 1

Stimuli were *Pseudosentences* from the IFA corpus read by 10 Dutch speakers (5F) (Van Son et al., 2001). In Experiment 1, the parameters of the male-like Long VTL targe are ϕ = 510Hz, F_0 = 120Hz, rate = 3.8 syll/s.; and those of the female-like Short VTL target are ϕ = 585Hz, F_0 = 185Hz, rate = 4.2 syll/s. Speaker profiles were derived from all pseudosentences read by that speaker. Long VTL and Short VTL pseudonymizations of the target speaker and a distractor were presented to 4 experts: 3 speech therapists and 1 linguist. In Experiment 1, both the *X* and the *A* and *B* sounds of each *ABX* stimulus were pseudonymized. When *X* was Long VTL in the *ABX* stimulus, *A* and *B* were Short VTL and when *X* was Short VTL, *A* and *B* were Long VTL. Each target speaker was presented once with a male and once with a female distractor.

8.3.1.2 Experiment 2

Sentence fragments with a maximum duration of 3s were selected from readings of *Treasure Island* taken from the *Parallel Audiobook Corpus* (Ribeiro, 2018) read by 16 speakers of British English (5F). In Experiment 2, the pseudonymization target values were somewhat lowered

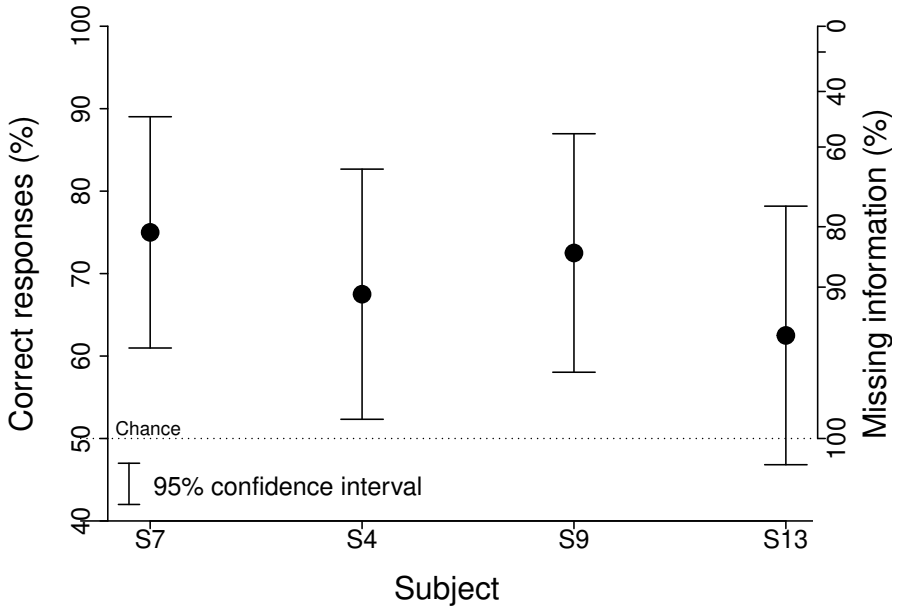


Figure 8.3: Speaker identification in experiment 1 by expert subject with correct responses (left) and missing information (right, 100% = 1 bit). Confidence intervals from Student distribution. Overall mean correct: 69%, 95% conf int. [61, 78]%. No differences were found in responses to male and female speakers.

Table 8.1: Summary of *ABX* listening experiments. Sp.: Speakers.

Exp	**Corpus**	**Speech (≤3s)**	**Sp. F/M**	**Subjects**
1	Van Son et al. (2001)	Pseudo sent.	5/5	4 experts
2	Ribeiro (2018)	sentences	5/11	8 naive
3	Yamagishi et al. (2019)	sentences	45/45	6 naive/1 exp.

to adapt to the new corpus. The parameters of the male-like Long VTL target are ϕ = 500Hz, F_0 = 120Hz; and those of the female-like Short VTL target are ϕ = 575Hz, F_0 = 175Hz. Target speaking rate was always 4.0 syll/s. Speaker profiles were derived from all sentences in a single chapter, not used for selecting stimulus sentences. *X* was an original recording from the speaker to be recognized, *A* and *B* were both either Original recordings, or Long VTL or Short VTL pseudonymizations, one of which was from the same speaker as *X*. There were 16 *ABX* stimulus combinations for each condition, Original, Long VTL, and Short VTL, 48 *ABX* combinations in total. Each speaker was used only once as target speaker for each condition (not counting practice items). Distractors were selected at random irrespective of the gender. The genders of target speaker and distractor were the same (FF or MM) for 27 stimuli and different (FM or MF) for 21 stimuli. For this experiment, 8 "naive" listeners participated, recruited by email, not counting a subject that was dropped (see Section 8.3.2.2).

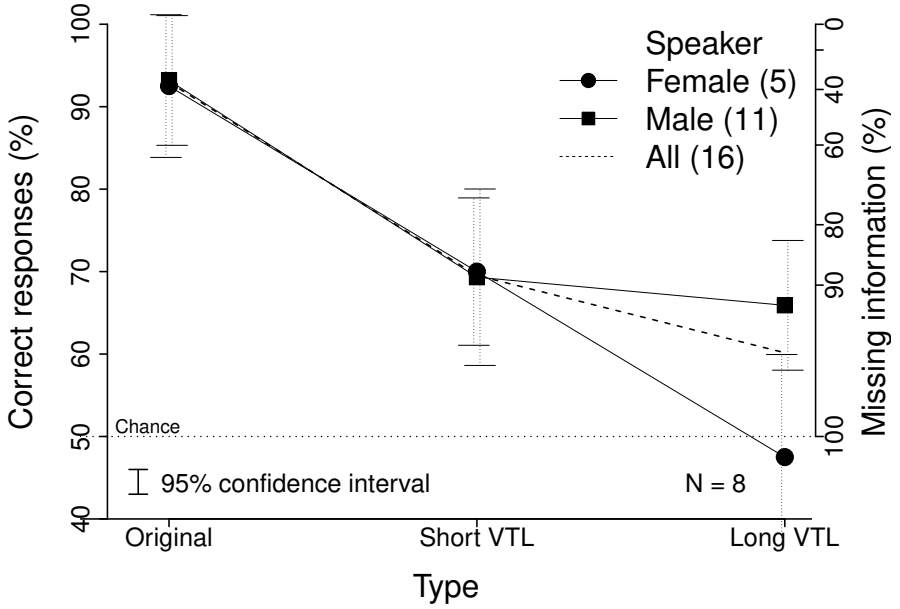


Figure 8.4: Speaker identification in experiment 2 by stimulus type and speaker gender. Original: AB are original recordings, Short VTL: AB pseudonymized to a short vocal tract length, Long VTL: AB pseudonymized to a long vocal tract length. N: Number of subjects. See also Fig. 8.3.

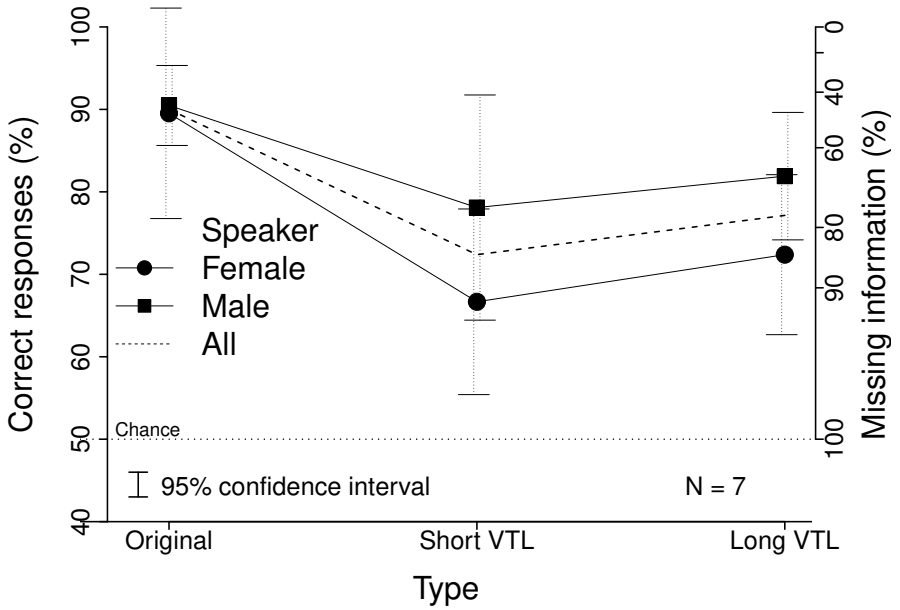


Figure 8.5: Identification after de-pseudonymization in experiment 3 by stimulus type and speaker gender. Speaker: Target speakers, 15F/15M for each Type, 90 in total. See also Fig. 8.4.

Table 8.2: Speaker identification accuracy in experiments 2 and 3. Linear mixed effects models of influence of (de-) pseudonymization and speaker gender on identification for each stimulus type (see text). Ex: Experiment, p (Ex): p-value of difference between experiments, p (Gen): p-value of difference between speaker genders in combined experiment.

Stimulus	**Ex. 2 (sd)**	**Ex. 3 (sd)**	**p (Ex)**	**p (Gen)**
Original	93% (6)	90% (8)	>0.05	>0.05
Short VTL	70% (11)	73% (12)	>0.05	>0.05
Long VTL	60% (7)	77% (6)	0.009	0.012

8.3.1.3 Experiment 3

All the sentences from the *Bonafide* recordings from the Logical Access part of the 2019 ASVspoof corpus (Yamagishi et al., 2019) were pseudonymized with the same pseudonymization target values as in Experiment 2. The procedure for Experiment 3 was the same as in Experiment 2. Speaker profiles were derived from all the sentences of that speaker. Gender information was available for 58 out of 107 speakers. A linear model based on the speaker profiles, with perfect fit on the known genders, was used to predict the gender of the other speakers. Sentence fragments with a maximum duration of 3s were selected as *ABX* stimuli from the target speaker and distractor, and were all de-pseudonymized using the speaker profile of the target speaker. In the de-pseudonymization, the formant frequencies and band intensities of the transformed segments were not known (since they were randomly chosen). Therefore, only the vocal tract length, pitch, and speaking rate were transformed to the target speaker profile. Target speaker and distractor were always of the same gender, both male or female. This was done because pilot tests showed that mixed gender stimulus pairs were perfectly identified.

Each condition in Experiment 3, Original, Long VTL, and Short VTL, contained sentences from 15 male and 15 female target speakers and randomly selected distractors of the same gender, 90 *ABX* stimulus sets in total. In Experiment 3, each speaker was only used once as a target speaker and once as a distractor (not counting practice items). Sentences were selected at random from the corpus from each speaker, but no sentence recording was used twice in the experiment.

Subjects for experiments 2 and 3 were recruited over email with written instructions. Listening conditions in these two experiments were not supervised. As a quality assurance, only the responses from subjects who were able to correctly identify 70% of the target speakers in the original recordings (condition Original) were included in the analysis. Five subjects participated in both experiment 2 and 3, one subject participated in both experiments 1 and 3. Table 8.1 contains a summary of the three listening experiments.

8.3.2 Results and analysis

All the statistical analysis was done with R (R Core Team, 2019). Missing information is calculated as the entropy $H = -\sum_{i=1}^{2} p_i \log_2 p_i$ (in %). Differences in identification between conditions and stimulus classes are tested using paired Student t-tests (following Fradette et al. (2003)). The stimuli and experiment are available from van Son (2020b) and the listener responses are available from van Son (2020a).

8.3.2.1 Experiment 1

The expert listeners reported that they found it difficult to believe that the target speaker was always among the response choices. The expert listeners identified the target speaker approximately 70% of the time (see Fig. 8.3, missing >80% of information H). The responses were better than chance and worse than perfect ($p \leq 0.006$ for both 90% and 50% correct, t-test, not shown). There were no statistically significant differences between listeners and no effects of speaker gender (not shown).

8.3.2.2 Experiment 2

Responses of one subject, who did not reach 70% correct identification on the original recordings, were dropped (subject removed, see above). On average, the speaker identification of the original recordings was over 90% correct (see Fig. 8.4). The naive listeners identified the target speaker approximately 70% of the time in the short VTL condition and somewhat less in the long VTL condition (missing >80% of information to identify the speaker). This was significantly less than in the original condition with unaltered speech ($p \leq 0.0001$, paired t-test by subject). The difference between the short and long VTL condition were not significant ($p > 0.05$). There is a tentative difference in responses to the (5) female speakers and the (11) male speakers for the Long VT stimuli ($p = 0.027$, paired t-test). It appears that the female speakers are not identified above chance level in the long VTL condition. There seems to be an asymmetry in the effect of pseudonymization on male and female voices which we currently cannot explain.

In the responses from experiments 1 and 2, there is a tendency that comparison to a distractor of a different gender improves identification of the target speaker (not shown). However, partly due to the design of the experiments, this could not be verified ($p > 0.05$, paired t-test).

8.3.2.3 Experiment 3

All the subjects cleared the 70% correct criterion for the Original stimulus condition. Speaker identification of the original recordings in Experiment 3 was around 90% correct (see Fig. 8.5 and Table 8.2). De-pseudonymization, the inverse transform, was effective in reversing the pseudonymization towards a Long VTL target, increasing the identification from 60% to 78%

correct (Table 8.2) with missing information ≤80% (Fig. 8.5). However, the differences in identification between the Original and the de-pseudonymized stimuli was still significant (p≤0.009 paired t-test by subject). The difference in identification between male and female speakers was not significant in Experiment 3 (p>0.05 for all stimulus types).

8.3.3 Modeling responses to listening experiments 2 & 3

The results of experiments 2 and 3 were combined in a linear mixed effect model to estimate the effects of the speaker gender and pseudonymization versus de-pseudonymization (Exp) on speaker Identification (I) for each stimulus type, i.e., Original, Short VTL and Long VTL. The full model was:

$$I \sim Exp + Gender + (Exp + Gender|Subject) \tag{8.3}$$

Subjects that participated in both the experiments were identified in the model. Statistical significance was determined using ANOVA on full model versus a model with the relevant fixed factor removed. No difference was found for the Original and Short VTL stimuli ($p > 0.05$). For the Long VTL target pseudonymizations, both the differences between male and female speakers and the differences between the experiments were statistically significant (see Table 8.2). Using the model of Eq. 8.3, the male speakers were identified 13% points more than female speakers and de-pseudonymization increased identification by 19% points (p-values in Table 8.2).

Experiment 3 only contained same gender comparisons between the target and distractor speakers, while Experiment 2 contained the same and mixed gender comparisons. Same gender comparisons could be seen as "more difficult" than mixed gender comparisons. Repeating the linear mixed effect modelling with only the responses to the same gender distractors gave the same results; no effect for *Original* and *Short VTL* stimuli ($p > 0.05$) and a consistent effect for de-pseudonymization and speaker gender for *Long VTL* stimuli ($p(Ex) = 0.008$, $p(Gen) = 0.024$, not shown) were observed. But the effect of de-pseudonymization only increases marginally (to 22% points). The overall effect of de-pseudonymization was found for both female and male speakers separately (p≤0.012, ANOVA, removing *Gender* from Eq. 8.3, not shown).

8.4 2020 VoicePrivacy challenge experiments

Automatic evaluations of the proposed method were carried out as part of the VoicePrivacy 2020 challenge, using the data sets and experimental protocols set by the challenge Tomashenko et al. (2020a), and the performances were measured in terms of ASV and ASR systems' evaluation metrics. The ASV evaluation consists of an *enrollment* phase, where several speakers enrol into a system, and a *trials* phase, where each test speaker that claims to be a specific enrolled speaker has to be verified. For the anonymization experiment, each speaker is anonymized to

two different speakers, one for enrollment and another for trials. Thus, a good anonymizer would increase the ASV error, while keeping the ASR error as low as possible.

8.4.1 Summary of the data sets and evaluation protocol

For evaluations on anonymization, the `dev` and `test` subsets of the VCTK and LibriSpeech corpora were used. As reference (and non-overlapping) speaker set for anonymization, `libri-other-500` subset of LibriSpeech was used. The anonymized speech is evaluated in terms of word error rate (WER) using a neural-network based ASR system trained with lattice-free maximum mutual information objective function (Povey et al., 2016) and in terms of equal error rate (EER) and log-likelihood ratio based costs C_{llr} and C_{llr}^{min} using an x-vector (Snyder et al., 2018) based ASV system, both provided by the challenge organisers. For more details about the data set and the experimental protocol, the reader is referred to Tomashenko et al. (2020a).

8.4.2 Baselines provided by the challenge

The challenge provided two baseline systems: (i) neural source filtering (NSF) based and (ii) McAdams method based.

8.4.2.1 NSF baseline

The NSF approach (Fang et al., 2019) was built on the idea that any speech signal can be decomposed into three sets of features: those representing (i) the spoken content, (ii) the speaker and (iii) the speaker's fundamental frequency, and that speech can be synthesised back by combining these components. Mel-filterbank features or intermediate representations from an ASR neural acoustic model constitute the spoken content, whereas fixed length neural speaker embeddings, known as x-vectors, represent the speakers. Anonymization can be achieved by merely replacing the source speaker's x-vectors with those of the target speaker, which is chosen among a pool of reference speakers, typically the one who is *farthest* in terms of their x-vector *affinity* score. Thus, in this method, a given speaker's speech is first decomposed into its three constituents, then anonymized by replacing the x-vectors and finally converted back into a waveform using speech synthesis.

8.4.2.2 McAdams baseline

This is a signal processing method based on formant shifting. In this method, each utterance is analysed using short-time processing, where each segment is fit with an all-pole model on its power spectrum using linear prediction. The angles θ_i of the complex poles thus correspond to the formant frequencies, when the model order is appropriately chosen. The anonymization process involves shifting the formants non-linearly, by exponentiating the complex poles

Table 8.3: ASV results for both development and test partitions (o-original, p-pseudonymized(F03-9), b1-NSF, b2-McAdams).

Data	Expt.	Dev. set (female)			Dev set (male)			Test. set (female)			Test set (male)		
		EER%	C_{llr}^{min}	C_{llr}	EER%	C_{llr}^{min}	C_{llr}	EER%	C_{llr}^{min}	C_{llr}	EER%	C_{llr}^{min}	C_{llr}
libri	o	8.67	0.30	42.86	1.24	0.03	14.25	7.67	0.18	26.79	1.11	0.04	15.30
	b1	36.79	0.89	16.35	34.16	0.87	24.72	32.12	0.84	16.27	36.75	0.90	33.93
	b2	23.44	0.62	11.73	10.56	0.36	11.95	15.33	0.49	12.55	8.24	0.26	15.38
	p	25.28	0.66	9.30	18.79	0.56	15.70	24.82	0.59	10.23	14.92	0.43	10.65
vctk common	o	2.33	0.09	0.86	1.43	0.05	1.54	2.89	0.09	0.87	1.13	0.04	1.04
	b1	27.91	0.74	7.21	33.33	0.84	23.89	31.20	0.83	9.02	31.07	0.84	21.68
	b2	11.63	0.37	43.55	10.54	0.32	25.00	14.45	0.47	42.73	11.86	0.35	28.23
	p	16.86	0.51	11.12	20.23	0.56	7.65	26.01	0.70	13.16	13.84	0.45	5.32
vctk different	o	2.86	0.10	1.14	1.39	0.05	1.16	4.94	0.17	1.50	2.07	0.07	1.82
	b1	26.11	0.76	8.41	30.92	0.84	23.80	31.74	0.85	11.53	30.94	0.83	23.84
	b2	15.83	0.50	39.81	11.22	0.38	23.09	16.92	0.55	41.34	12.23	0.40	25.06
	p	15.67	0.50	6.25	14.74	0.39	3.84	26.23	0.75	11.92	22.90	0.67	7.57

Table 8.4: ASR results in WER% for both development and test partitions (o-original, b1-NSF, b2-McAdams, p-pseudonymized(F03-9), s-LM_s, l-LM_l).

Expt.	libri				vctk			
	Dev. set		Test set		Dev. set		Test set	
	s	l	s	l	s	l	s	l
o	5.24	3.84	5.55	4.17	14.00	10.78	16.38	12.80
b1	8.76	6.39	9.15	6.73	18.92	15.38	18.88	15.23
b2	12.19	8.77	11.77	8.88	30.10	25.56	33.25	28.22
p	8.82	6.48	8.04	5.87	21.99	18.23	23.32	18.89

by a constant factor M, i.e. $\theta_i \rightarrow \theta_i^M$, where M is the McAdams coefficient. The resultant signal is then overlap-added across segments to reconstruct its corresponding pseudonymized utterance.

Contrasting the McAdams method with our proposed approach, a key difference is that the former allows only a single degree of freedom (i.e. by tuning M) in moving the formants, whereas the proposed approach allows each band of formants and F0 to be individually moved and their intensities adjusted, thus allowing several degrees of freedom.

8.4.3 Idiap-NKI challenge entry

We followed the protocol set by the challenge, and evaluated ASR and ASV performances by pseudonymizing the given subsets of VCTK and LibriSpeech data sets. For pseudonymization,

target speaker profiles were created using `libri-other-500` set of the LibriSpeech corpus. In a given subset, each speaker is pseudonymized to have the characteristics of a randomly chosen target speaker from the `libri-other-500` set. In ASV, this means that the enrollment and trials of the same speaker are often mapped to different target speakers (and we have not ensured that they are different in all the cases, since the probability of choosing the same speaker among 1000+ speakers is small). If only the trials sets are pseudonymized, ASV may indicate a higher error (indicating a better anonymization) due to acoustic mismatch introduced by the pseudonymization method. A higher equal error rate (EER) in ASV implies better pseudonymization of the speakers, and a lower WER on ASR implies better preserving of intelligibility.

The proposed method uses all the steps presented in Section 8.2.1. That is, the method changes the speaking rate, pitch and the B_0 and B_{3-5} bands and their intensities. The target values for pseudonymization are determined by selecting a random speaker from `libri-other-500` as the target speaker. In addition, the B_4 and B_5 bands are switched, and bands B_{6-9} are replaced with intensity modulated pink noise. For the sake of clarity, this pseudonymization method is referred to as *F03-9*.

8.4.4 Results

Tables 8.3 and 8.4 compare the ASV and ASR results, respectively, of the baseline anonymization methods using neural source-filtering (NSF) and McAdams, and the proposed pseudonymization method. In ASR, the proposed method gave a lower WER than the McAdams baseline, indicating better intelligibility, in all the cases. In ASV, the EER in all the cases except one (vctk-different female) is higher, implying a better pseudonymization, than the McAdams baseline. This is also indicated by a consistently higher or equal C_{llr}^{min} in all the cases. However, there is a room for improvement in comparison to the NSF baseline in terms of ASV performance, although it is fairer to compare the method with the signal processing based baseline.

We conducted ablation experiments to study the contribution of the individual steps proposed in Sec. 8.2 and to study the effect of de-pseudonymization (the right part of Fig. 8.1). The individual steps of pseudonymization are: (i) the *source* part: pseudonymizing B_0 band, (ii) the vocal-tract *system* part: pseudonymizing the B_{3-9} bands, which also includes the *additional* processing of introducing modulated pink noise in B_{6-9} bands (Sec. 8.2.1.10) and exchanging B_4 and B_5 bands, and (iii) the *speaking rate* part. To be able to perform de-pseudonymization, we had to omit the irreversible additional processing step. Tables 8.5 and 8.6 show the results of all the ablation experiments. The results indicate that the vocal-tract system component plays the most prominent role in pseudonymization, and a significant part of it is due to the additional processing. De-pseudonymization can be seen to be partially effective.

Table 8.5: ASV results with ablation (pseudon - pseudonymized(F03-9), no system - only source and speaking rate have been modified, no source - only system and speaking rate have been modified, no rate - only source and system have been modified, no-additional - no additional processing described in Sec. 8.2.1.10 has been applied, de-pseudon - pseudonymization reversed for the *no-additional* experiment.).

Data	Expt.	Dev. set (female)			Dev set (male)			Test. set (female)			Test set (male)		
		EER%	C_{llr}^{min}	C_{llr}	EER%	C_{llr}^{min}	C_{llr}	EER%	C_{llr}^{min}	C_{llr}	EER%	C_{llr}^{min}	C_{llr}
libri	pseudon	25.28	0.66	9.30	18.79	0.56	15.70	24.82	0.59	10.23	14.92	0.43	10.65
	no system	15.91	0.49	42.74	5.12	0.17	36.62	10.77	0.33	39.27	2.00	0.07	25.45
	no source	20.88	0.61	7.24	15.22	0.48	12.38	19.71	0.55	6.93	8.46	0.27	3.25
	no rate	21.16	0.61	7.48	15.22	0.49	12.39	19.53	0.56	7.47	8.46	0.28	3.30
	no additional	15.48	0.48	42.54	5.12	0.17	36.60	10.77	0.33	39.25	2.23	0.07	25.66
	de-pseudon	10.37	0.34	33.84	3.11	0.10	23.38	10.40	0.30	30.87	1.78	0.06	18.23
vctk common	pseudon	16.86	0.51	11.12	20.23	0.56	7.65	26.01	0.70	13.16	13.84	0.45	5.32
	no system	17.73	0.51	8.83	12.25	0.38	14.34	13.58	0.45	9.03	8.76	0.26	12.51
	no source	23.55	0.58	13.82	19.94	0.58	8.59	25.43	0.74	19.02	20.62	0.59	8.05
	no rate	21.80	0.58	15.06	19.09	0.57	8.33	25.14	0.75	19.42	20.34	0.56	7.68
	no additional	17.44	0.51	8.88	12.25	0.37	14.19	14.45	0.45	9.13	9.04	0.26	12.50
	de-pseudon	6.40	0.24	2.14	8.55	0.22	5.48	7.23	0.22	1.48	4.52	0.13	3.83
vctk different	pseudon	15.67	0.50	6.25	14.74	0.39	3.84	26.23	0.75	11.92	22.90	0.67	7.57
	no system	17.97	0.52	10.79	2.33	0.09	1.14	14.87	0.45	5.98	11.83	0.35	14.64
	no source	27.68	0.70	11.78	5.11	0.18	3.09	22.27	0.65	12.90	27.55	0.68	12.81
	no rate	24.48	0.66	11.36	5.26	0.19	3.15	21.35	0.63	12.03	24.11	0.61	9.48
	no additional	17.57	0.51	10.63	2.38	0.10	1.16	14.40	0.44	5.84	12.34	0.36	14.77
	de-pseudon	5.90	0.21	1.39	2.28	0.09	0.66	10.19	0.34	2.69	6.83	0.22	5.03

Table 8.6: ASR results in WER% with ablation (pseudon - pseudonymized(F03-9), no system - only source and speaking rate have been modified, no source - only system and speaking rate have been modified, no rate - only source and system have been modified, no-additional - no additional processing described in Sec. 8.2.1.10 has been applied, de-pseudon - pseudonymization reversed for the *no-additional* experiment, s-LM_s, l-LM_l).

Expt.	libri				vctk			
	Dev. set		Test set		Dev. set		Test set	
	s	l	s	l	s	l	s	l
pseudon	8.82	6.48	8.04	5.87	21.99	18.23	23.32	18.89
no system	7.30	5.21	6.87	5.07	18.00	14.34	20.38	16.42
no source	8.14	5.93	7.62	5.64	20.12	16.31	22.83	18.81
no rate	7.72	5.63	7.24	5.31	18.90	15.01	21.97	17.72
no additional	7.18	5.18	6.90	5.08	18.05	14.32	20.36	16.41
de-pseudon	6.85	4.95	7.03	5.27	17.43	13.61	20.37	16.08

8.5 Beyond the VoicePrivacy challenge

In this section, we explore some directions in which the proposed method could be evaluated, viz. (i) measuring the intelligibility after pseudonymization by utilising the original utterances

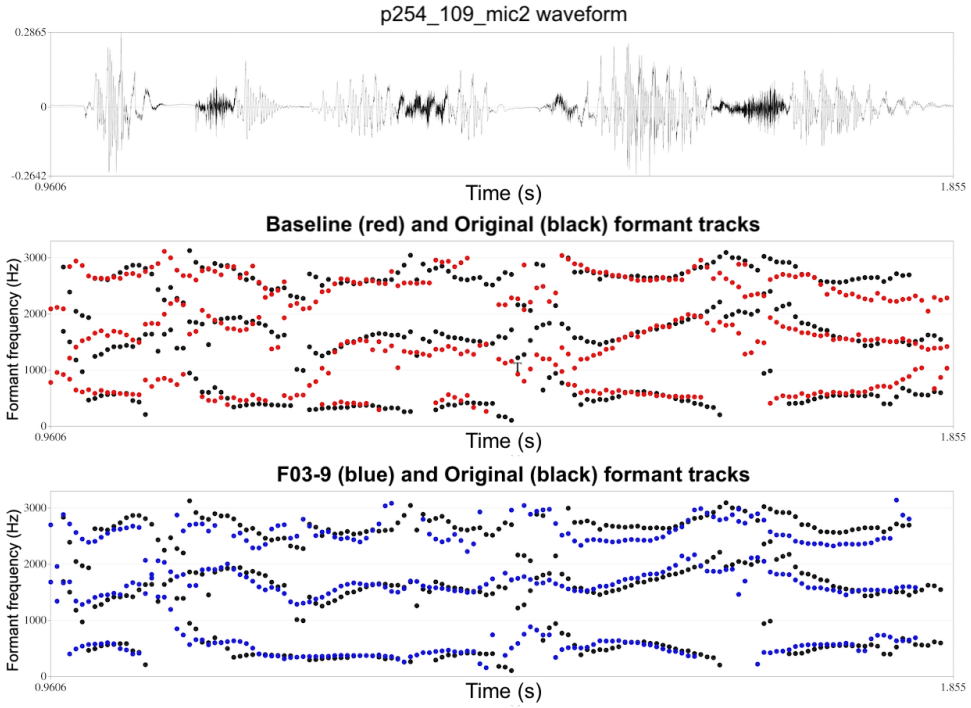


Figure 8.6: Example formant tracks for correlating formant values between pseudonymized speech and the original recordings. Top: waveform of sentence [but it is a pleasure] from speaker p254, center: F_1-F_3 formant tracks for McAdams Baseline (red) and Original (black) speech, bottom: id. for F03-9 pseudonymization (blue) and Original (black). Horizontal: Time, Vertical: Amplitude (top) and Frequency (mid and bottom).

as references, instead of ASR, (ii) measuring the closeness of the formant tracks between the pseudonymized and original utterances, and (iii) measuring the extent of retaining pathological conditions such as dysarthria after the proposed pseudonymization.

8.5.1 Intelligibility measure based comparison of phone posterior sequences

The 2020 VoicePrivacy Challenge proposed WER of ASR system as a measure of intelligibility. However, ASR system performance gets affected by components such as Viterbi search, language model and pronunciation lexicon. Even if we presume that all the anonymization systems are compared using exactly the same language model, acoustic model and lexicon, the search heuristics can make a difference. So, here we propose to utilise an alternate objective intelligibility measure where the original reference speech and the anonymized speech are compared in the phone posterior feature space, employing only the acoustic model.

We use (7.1) as the local score. We used the following dynamic programming recursion

$$\Gamma_{j,t} = \gamma_{j,t} + \min\left(\Gamma_{j,t-1}, \Gamma_{j-1,t-1}, \Gamma_{j-2,t-1}\right), \tag{8.4}$$

Table 8.7: Intelligibility in terms of DTW distances (b1-NSF, b2-McAdams, p-pseudonymized(F03-9)).

E	libri		vctk	
	Dev.	Test	Dev.	Test
b1	0.005650	0.005804	0.007050	0.007638
b2	0.008798	0.008082	0.010237	0.010273
p	0.005955	0.004463	0.006001	0.006100

where $\Gamma_{j,t}$ denotes the cumulative score at j, t. The additional skip transition from $\Gamma_{j-2,t-1}$ was allowed to accommodate for the duration changes between the reference and test utterances. The final score $\Gamma_{J,T}$ normalised by the path length yields a measure of intelligibility; the lower the score, the better the intelligibility.

We computed intelligibility scores in the following manner:

1. First, estimate the posterior probability of the clustered context dependent phones using the neural network-based acoustic model provided with the VoicePrivacy challenge and then marginalise the context-dependent information, position markers and lexical stress markers to estimate the posterior probabilities of context independent phones. The context independent phone posteriors are used as the posterior features, $\mathbf{y}_j$ and $\mathbf{z}_t$ for the DTW-based intelligibility score estimation.

2. Compare the intelligibility scores (DTW distances) for the proposed pseudonymization method (F03-9) and the baseline methods by averaging the scores of all the utterances in each method.

Results from Table 8.7 indicate that the intelligibility scores for the proposed pseudonymization method are comparable to those of the NSF baseline and better than the McAdams baseline. This indicates that the differences observed in the WER metric (Table 8.4) could be due to aspects such as search heuristics employed during decoding.

8.5.2 Measuring pseudonymized formant values

Formants are important in the study of speech because their values are linked to the shape of the vocal tract, and hence to the constellation and movements of the articulators (Christensen, 2018; Dromey et al., 2013; Lee et al., 2015; McKell, 2016). Formant values are also related to the intelligibility of phonetic contrasts (Harper et al., 2017; Kent et al., 1989b; Richardson & Sussman, 2017). These relations are also relevant to the study of pathological speech, such as dysarthric speech (Sapir et al., 2010) and Parkinson's disease (Sapir et al., 2007). To evaluate to what extent formant measurements can be preserved after pseudonymization, formant tracks before and after pseudonymization are compared (see Figure 8.6). To determine the

Table 8.8: Mean correlation coeff between formant tracks from Original and pseudonymized recordings, for all speakers (N=60). F_1, F_2, F_3: Correlation coefficients of the formants. F: Female speakers, M: Male speakers. Baseline (McAdams) & *F03-9*: Pseudonymization procedures, see text. Average number of recordings per speaker: 23.5±13.7 (F), 24.2±15.8 (M).

Group		F_1	F_2	F_3
F	Baseline	0.507 (0.158)	0.601 (0.198)	0.424 (0.287)
	F03-9	0.563 (0.194)	0.659 (0.161)	0.620 (0.202)
M	Baseline	0.490 (0.161)	0.571 (0.158)	0.264 (0.226)
	F03-9	0.655 (0.153)	0.716 (0.136)	0.688 (0.136)
Total	Baseline	0.499 (0.160)	0.586 (0.178)	0.344 (0.257)
	F03-9	0.609 (0.174)	0.688 (0.149)	0.654 (0.169)

preservation of formant tracks after pseudonymization, the first three formant tracks of pseudonymized speech samples are correlated to those of the original recordings, using the *Robust* formant tracking in *Praat* (Boersma & Weenink, 2019). The same recordings from 60 speakers (30F/30M from vctk_dev and vctk_test) were used for the McAdams *Baseline* and *F03-9* pseudonymization. A higher average correlation coefficient (R) indicates that the pseudonymized speech would be more useful to study the acoustic effects of differences in articulation.

The results of the comparison are presented in Table 8.8. These results show that the average R of the pseudonymized formant values are higher for the *F03-9* pseudonymizations than for the *Baseline* method for all three formants. Correlation coefficients, R, for the *Baseline* method were between R=0.26 and R=0.60. Correlation coefficients for the *F03-9* method were 0.1-0.3 higher on average for all speakers, between R=0.56 and R=0.72 (R^2: 0.12-0.31 higher, highest values for F_3, $p \leq 10^{-7}$, paired Student t-test per speaker). There was a difference based on the speaker gender. For female speakers, the difference in R was 0.05-0.20 (highest values for F_3, $p \leq 10^{-2}$, *idem*), for male speakers, it was 0.14-0.42 (highest values for F_3, $p \leq 10^{-5}$, *idem*). The differences in R between *Baseline* and *F03-9* were larger for male than for female speakers for all three formants (two sample Student-t test, $p \leq 0.001$, 0.01, and 10^{-6} for F_1 - F_3, respectively).

8.5.3 Automatic dysarthria classification

The ability to retain paralinguistic features after pseudonymization was evaluated on the example of dysarthric speech. Speech samples were taken from the TORGO corpus (Rudzicz et al., 2012). The recordings from the head mounted microphone were used. Recordings from the directional microphone were added for two sessions, both session 2 of control speakers FC02 and MC04.

The control and dysarthric utterance recordings were pseudonymized as with the F03-9

Table 8.9: Dysarthria classification results for original and pseudonymized recordings from the TORGO corpus (Rudzicz et al., 2012) (see text). Given are the percentage correct classification for the original and pseudonymized (Pseud.) recordings, the concordance (Conc.), i.e., the percentage identical classification for original and pseudonymized recordings. The overall Cronbach's alpha is acceptable, α=0.769. Without the two female control speakers FC01 and FC02, Cronbach's alpha is excellent, α=0.949. Spkr: Speaker, Pseud.: Pseudonymized recordings, Conc.: Concordance, N: number of utterances.

		Correct			
Group	Spkr	Original	Pseud.	Conc.	N
Control	FC01	98.2	47.6	49.4	164
	FC02	86.3	13.7	24.4	1000
	MC01	98.5	99.3	98.4	748
	MC02	99.1	98.7	98.3	464
	MC03	99.3	100.0	99.3	600
Dysarthric	F01	90.2	90.9	90.2	132
	M01	92.7	99.7	92.4	288
	M02	95.8	98.5	95.8	409
	M03	91.3	97.9	91.5	424
	M04	91.0	93.6	87.5	488
Total		94.2	84.0	82.7	471.7

method described above. However, for this experiment, the characteristics of a random speaker of the opposite gender was selected from the *Bonafide* recordings in the Logical Access part of the 2019 ASVspoof corpus (Yamagishi et al., 2019). As altered, slow, speaking rate is an important symptom of dysarthria, the speaking rate of the pseudonymized utterances was not changed from the original value. The results of the ablation experiment in Section 8.4.4 show that not changing the speaking rate has only a low impact on ASV identification performance (see Table 8.5). The dysarthria classification was done with linear support vector machines (SVMs) trained, using a leave-one-out procedure, on eGeMAPS feature set that is commonly used in studying paralinguistic aspects (Eyben et al., 2016a). SVMs trained on the original recordings were used to classify the original utterances, whereas those trained on pseudonymized recordings were used to classify the pseudonymized utterances.

The dysarthria classifier did not perform very well (59% correct). Inspection of the results showed that this was most likely due to the low audio quality of some sessions. It also seemed to perform worse on some of the female speakers. It was decided to drop all sessions where dysarthria classification of the original recordings was below 70% correct. This left 15 (out of 30) recording sessions from a total of 10 (out of 15) speakers, 5 control (2F) and 5 dysarthric (1F) speakers. The two sessions recorded with the directional microphone were among those dropped for low classification performance.

The results of the dysarthria classification evaluation after pseudonymization are mixed (Table 8.9). For the two female control speakers, FC01 and FC02, the performance is below 50%, at chance level. It is clear that the pseudonymization of utterances from these speakers degraded the speech too much and the classifier did not work. The results for the speech of the other speakers is excellent. This is summarised in the Cronbach's alpha values, which are acceptable for the whole group of 10 speakers (α=0.769), but are excellent (α=0.949) when the results for FC01 and FC02 are removed. It is currently unknown why the dysarthria classification did not work for the pseudonymized recordings of speakers FC01 and FC02.

8.6 Discussion

Pseudonymization aims at protecting the privacy of the speakers. Whether or not the levels of protection are sufficient depends on the requirements of the application and the risks that an identification would pose. One objective of the proposed approach is to make pseudonymization deterministic and adjustable, i.e., gradual, on untranscribed recordings. It works in the spectro-temporal domain on any speech recording, and is intrinsically deterministic and reversible. The exception to reversibility is the overlap-add procedure to adapt the pitch and duration of the speech which is inherently "lossy", i.e., partially irreversible. But overlap-add is a well known, predictable, speech synthesis procedure. The aspects of the speech that are transformed as well as the extent of the changes can all be freely chosen. The only constraint is the quality of the resulting speech.

However, reversibility is not necessarily an advantage. It is clear that the ability to, partially, de-pseudonymize speech warrants extra attention. The current study explores one specific de-pseudonymization approach based on knowing the original pseudonymization target. An obvious way to prevent de-pseudonymization would be to obfuscate the target speaker selection.

Another important goal of pseudonymization of speech could be to allow the study of linguistic and paralinguistic aspects of speech without jeopardising the privacy of the speakers. It is not yet known which of such aspects can still be studied after pseudonymization and what the corresponding risks of re-identification are. In this study, the extent to which linguistic and paralinguistic features are preserved was estimated by comparing formant tracks after pseudonymization with the originals and by evaluating the results of an automatic dysarthria classifier on pseudonymized speech.

8.6.1 Listening experiments

All three *ABX* listening experiments showed reduced speaker identification after pseudonymization (Fig. 8.3 and 8.4) and also after de-pseudonymization (Fig. 8.5). After pseudonymization, more than 80% of the information necessary to make the choice between speaker *A* and *B* is lost (<70% correct identification, Fig. 8.3 and 8.4), compared to less than 40% missing

information with the original recordings (>90% identification, Table 8.2). Reverting the transformation from known pseudonymization targets can improve the recognition, especially for speech transformed to a Long VTL (Fig. 8.5 and Table 8.2).

The responses in both experiments 2 and 3 displayed an asymmetry between male and female voices. Female speakers were identified worse than male speakers after both pseudonymization and de-pseudonymization. This difference was statistically significant for the Long VTL condition when the responses in these experiments are combined (Table 8.2). This asymmetry was smaller, or absent, in the Short VTL condition (statistically not significant).

8.6.2 Automatic evaluations

Automatic evaluations on the VoicePrivacy challenge data showed that the method is better than the comparable signal-processing based McAdams method. However, there is still a significant gap in terms of ASV performance w.r.t. the NSF baseline. One factor could be that the former chooses the target speakers randomly, whereas the latter specifically chooses far away speakers. Future investigations could focus on identifying the areas of improvement that lead to closing this gap and improving beyond it. In terms of preserving the intelligibility, the proposed method showed comparable performance in terms of both the ASR and the proposed phone posterior based approach.

It is worth mentioning that, in the VoicePrivacy challenge, besides the objective evaluations, the organisers also conducted subjective evaluations, in which the proposed method showed promising results in terms of intelligibility, quality and dissimilarity of the pseudonymized speech w.r.t. the original speakers (Wang et al., 2020)[IV].

8.6.3 Formant values

The outcomes of the formant track analysis indicate that both the *Baseline* and the*F03-9* method preserve formant tracks to some extent. The *F03-9* pseudonymization better preserves F_{1-3} formant track movements than the McAdams *Baseline* method, sometimes with a considerable margin. The differences were more pronounced for male than for female speakers. The biggest differences were found in the F_3 tracks.

From these results, it is clear that it is possible to preserve at least some level of measurable formant track information after pseudonymization. However, there are systematic differences between the two methods tested and the gender of the speakers in how well the formant track information is preserved. This shows that there is still room to optimise this feature in future speech pseudonymization methods.

[IV] We cite the presentation as it was the only reference available at the time of the submission of this thesis.

8.6.4 Dysarthria classification

The TORGO corpus proved to be sub-optimal for the evaluation of automatic dysarthria classification of pseudonymized speech. Half of the recording sessions, including all recordings of 5 speakers, had to be dropped due to very low classification performance. For 8 out of the remaining 10 speakers, classification after pseudonymization performed excellent, with high concordance between original and pseudonymized audio. For two other speakers, the results after pseudonymization were essentially at chance level. What this shows is that there is indeed good potential to use pseudonymization to study paralinguistic aspects of (pathological) speech, at least for dysarthria. However, the pseudonymization method used in this study cannot yet be applied to all speakers.

8.7 Summary

A method to pseudonymize speech is described that is both deterministic and adjustable. The method can pseudonymize speech samples with only a few hundred seconds of speech of the source speaker by altering the voice source related, vocal tract system related and speaking rate information. *ABX* pilot listening tests demonstrated that the pseudonymized samples are largely unidentifiable for human listeners. However, the deterministic nature of the procedure compels caution and measures to counter re-identification should be considered before applying the procedure. An evaluation at the 2020 VoicePrivacy challenge showed that the method pseudonymizes utterances better than the McAdams method provided by the challenge and is inferior to the neural source-filter based baseline. However, in terms of a phone posterior feature-based intelligibility measure computed using only the acoustic model, the proposed method is comparable to the neural source-filter based baseline. Ablation studies analysing the role of different processing steps in the proposed approach revealed that the alteration of vocal tract system related information plays a major role in anonymizing the speaker's identity. Furthermore, the studies also revealed that the pseudonymization process can be partially reversed, assuming the target speaker information such as, VTL, formant information are computable. A formant track analysis investigating the preservation of articulatory information in pseudonymized speech showed promising results with a somewhat better correlation between the original and pseudonymized speech for the proposed method than for the baseline McAdams approach. Finally, in a case study on dysarthria, it was found that pathological speech evaluation after pseudonymization could be feasible; the results were, however, speaker dependent.

9 Conclusions and future directions

This thesis focused on incorporation of prior knowledge for several speech assessment tasks related to non-verbal and verbal components of speech communication.

We first investigated directly modelling of raw waveform using CNNs for several speech assessment tasks in a non-verbal or paralinguistic setting, viz. Styrian and Arabic dialect identification, prediction of perceived non-expert speech fluency ratings and depression detection from speech. Investigations showed the feasibility of employing the automatic feature learning method in all the tasks investigated, as opposed to using handcrafted features and classifiers. Our investigations showed that the kernel width of the first convolution layer (subsegmental or segmental) has an impact on the performance. Specifically, we found that, with no prior knowledge introduced, the segmental modelling approach yielded improved systems for fluency prediction and depression detection tasks.

We then investigated methods to introduce knowledge related to voice source into the CNN-based end-to-end acoustic modelling approach. We showed that this can be achieved by filtering the speech signal based on source-system decomposition or zero frequency filtering and feeding the filtered signal as input to the CNNs. Experimental studies showed utility of the proposed approach for speech fluency prediction and depression detection tasks, where the changes in voice source characteristics provide vital information. The analysis of the filters of the first convolution layer also indicated that the best performing CNNs tend to give emphasis to low frequency regions that predominantly carry voice source related information. Furthermore, on the depression detection task, a whole network analysis by the extraction of relevance signal, through guided backpropagation, demonstrated that the subsegmental CNN tends to focus on both the glottal closure instants as well as fundamental frequency information. These observations are further interesting, since the conventional approach of extracting hand-crafted voice source related features and modelling them yields systems with inferior performance. Also, in a recent work, it has been shown that this approach can be extended to Dementia detection as well (Cummins et al., 2020; Villatoro-Tello et al., 2021, accepted for publication). On the Styrian dialect identification task, the proposed approach yielded inferior systems. This is consistent with the literature that the Styrian dialects lack

distinction in their pitch patterns.

We investigated incorporating articulatory features, that relate to the perception of linguistic units and the production of speech through movement of articulators, through transfer learning for end-to-end speech assessment. Our investigations showed that the proposed approach yields improved systems for dialect identification without the use of explicit linguistic resources of the target language. Our investigations on Arabic dialect identification task showed that the transfer learning approach could be also extended to the case where hand-crafted features are fed as input to the neural networks. On the fluency prediction and depression detection tasks, the proposed approach yielded inferior systems when compared to the conventional approach of modelling vocal tract system related hand-crafted features. One potential reason for this could be that the proposed approach models fixed length information and aggregates the frame-level predictions for the final decision making, while the conventional approach first aggregates the hand-crafted features at utterance level and then models them to make a decision. An investigation along these lines is open for future research. It is worth pointing out that the proposed approach has been also extended to the prediction of degree of sleepiness (Fritsch et al., 2020).

On the verbal component side, this thesis showed that linguistic segment level information can be effectively incorporated into the training of neural networks, through cost functions based on confidence measures, to enhance phone posterior probability estimation. Experimental studies showed that such a training yields better correlations of the predicted intelligibility scores with the human rated subjective scores, as well as better ASR performance, compared to the systems employing neural networks trained with frame-level cross-entropy criterion.

In the context of privacy preservation, we proposed a deterministic and adjustable signal processing method that can pseudonymize a speaker's speech samples by altering their voice source related, vocal tract system related and speaking rate parameters. ABX pilot listening tests demonstrated that the pseudonymized samples are largely unidentifiable for human listeners. Investigations in terms of ASR and ASV on the VoicePrivacy challenge data showed that the method pseudonymizes utterances better than the comparable McAdams method and is inferior to the neural source-filter baseline. The method was then validated on intelligibility assessment using the phone posterior feature approach, where the proposed method proved comparable to the neural source-filter based baseline. Ablation studies of the proposed pseudonymization approach revealed that the vocal tract system carries the most identifiable information of speakers. Furthermore, we investigated dysarthria classification on pseudonymized pathological and healthy speech, which revealed that such assessment could be feasible; the results were, however, speaker dependent.

9.1 Directions for future research

Here we list a few possible directions for future research.

- The end-to-end acoustic modelling approach for speech assessment investigated in this thesis typically modelled fixed length input of about 250 ms and aggregated the output probability for decision making. This may have limitations, as modelling just fixed length speech signals may not be sufficiently informative for different speech assessment tasks. It would be interesting to investigate neural architectures that aggregate information over time, similar to stats pooling, as done in the Arabic DID task. Furthermore, it would be also interesting to investigate the combination of such a modelling method with fixed length speech signal modelling, as they may provide complementary information.

- In recent years, neural embedding extraction from speech signals and their usage has gained attention. In this direction, it would be worth investigating the embeddings extracted with minimal prior knowledge and through incorporation of prior knowledge for speech assessment.

- In this thesis, speech signals were pseudonymized and then assessed in terms of speech intelligibility and speaker obfuscation through speaker verification. However, as noted in Chapter 1 and discussed in Chapter 8, there is also a need to preserve all the other conveyed information apart from the spoken message. One way to effectively attain this would be by developing a closed-loop framework that performs speech analysis-synthesis in an iterative manner: synthesis of pseduonymized samples and a battery of speech assessment tasks that analyse and ascertain whether the information beyond speech intelligibility is preserved.

Bibliography

Abadi, M. et al. (2015). TensorFlow: large-scale machine learning on heterogeneous systems.

Abdel-Hamid, O., Deng, L., Yu, D., & Jiang, H. (2013). Deep segmental neural networks for speech recognition. *Proceedings of Interspeech*, 1849–1853.

Abrol, V., Dubagunta, S. P., & Magimai.-Doss, M. (2019). *Understanding raw waveform based cnn through low-rank spectro-temporal decoupling* (tech. rep. Idiap-RR-11-2019) [peer-reviewed and presented at Swiss Machine Learning Day 2019]. Idiap Research Institute. http://publications.idiap.ch/downloads/reports/2019/Abrol_Idiap-RR-11-2019.pdf

Afshan, A. et al. (2018). Effectiveness of voice quality features in detecting depression. *Proc. Interspeech*, 1676–1680. https://doi.org/10.21437/Interspeech.2018-1399

Al Hanai, T., Ghassemi, M., & Glass, J. (2018). Detecting depression with audio/text sequence modeling of interviews. *Proc. Interspeech*, 1716–1720. https://doi.org/10.21437/Interspeech.2018-2522

Ali, A. et al. (2020). The mgb-5 challenge: recognition and dialect identification of dialectal arabic speech. *Proceedings of ASRU*, 1026–1033. https://doi.org/10.1109/ASRU46091.2019.9003960

Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., Bell, P., & Renals, S. (2016). Automatic dialect detection in arabic broadcast speech. *Proceedings of Interspeech*, 2934–2938. https://doi.org/10.21437/Interspeech.2016-1297

Almaadeed, N., Aggoun, A., & Amira, A. (2016). Text-Independent Speaker Identification Using Vowel Formants. *Journal of Signal Processing Systems*, *82*(3), 345–356. https://doi.org/10.1007/s11265-015-1005-5

Amiriparian, S., Freitag, M., Cummins, N., & Schuller, B. (2017). *Sequence to sequence autoencoders for unsupervised representation learning from audio*. Universität Augsburg.

Ananthapadmanabha, T., & Yegnanarayana, B. (1979). Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoustic Speech Signal Processing*, *27*(4), 309–319.

Ardila, R. et al. (2019). Common voice: a massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

ASHA. (2021). Dysarthria in adults [Accessed: 27-05-2021]. https://www.asha.org/practice-portal/clinical-topics/dysarthria-in-adults/

Austin, S., Makhoul, J., Schwartz, R., & Zavaliagkos, G. (1991). *Continuous speech recognition using segmental neural nets* (tech. rep.). BBN systems and technologies corp.

Baevski, A. et al. (2020a). https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020b). Wav2vec 2.0: a framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Proceedings of neurips* (pp. 12449–12460). Curran Associates, Inc.

Bahari, M. H., Dehak, N., Van Hamme, H., Burget, L., Ali, A. M., & Glass, J. (2014). Non-negative factor analysis of gaussian mixture model weight adaptation for language and dialect recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22*(7), 1117–1129. https://doi.org/10.1109/TASLP.2014.2319159

Bayya, A., & Vis, M. (1996). Objective measures for speech quality assessment in wireless communications. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1*, 495–498 vol. 1. https://doi.org/10.1109/ICASSP.1996.541141

Beck, E., Hannemann, M., Dötsch, P., Schlüter, R., & Ney, H. (2018). Segmental encoder-decoder models for large vocabulary automatic speech recognition. *Proceedings of Interspeech*, 766–770.

Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., & Keyhl, M. (2013). Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part i-temporal alignment. *Journal of the Audio Engineering Society*, *61*(6), 366–384.

Benoît, C., Grice, M., & Hazan, V. (1996). The sus test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, *18*(4), 381–392. https://doi.org/10.1016/0167-6393(96)00026-X

Berger, J., Hellenbart, A., Ullmann, R., Weiss, B., Moller, S., Gustafsson, J., & Heikkila, G. (2008). Estimation of 'quality per call' in modelled telephone conversations. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4809–4812.

Bernardis, G., & Bourlard, H. (1998). Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems. *Proceedings of the International Conference on Spoken Language Processing (SLP), 3*, 775–778.

Biadsy, F., Hirschberg, J., & Collins, M. (2010). Dialect recognition using a phone-gmm-supervector-based svm kernel. *Proceedings of Interspeech*, 753–756.

Biadsy, F., Hirschberg, J., & Ellis, D. P. (2011). Dialect and accent recognition using phonetic-segmentation supervectors. *Proceedings of Interspeech*, 745–748.

Blahut, R. E. (1974). Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, *IT-20*(4), 405–417.

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9), 341–345.

Boersma, P., & Weenink, D. (2017). *Praat: a system for doing phonetics with the computer*. http://www.praat.org

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer (computer program). version 6.1.06.

Bou-Ghazale, S. E., & Hansen, J. H. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on speech and audio processing*, *8*(4), 429–442.

Bourlard, H. A., & Morgan, N. (1994). *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media.

Bridle, J. S. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Proceedings of Advances in Neural Information Processing Systems*, 211–217.

Caligiuri, M. P., & Ellwanger, J. (2000). Motor and cognitive aspects of motor retardation in depression. *Journal of Affective Disorders*, *57*(1–3), 83–93.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005). The AMI meeting corpus: a pre-announcement. *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*, 28–39.

Chen, F. (2016). Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation. *Biomedical Signal Processing and Control*, *24*, 109–113.

Chen, N. F., Shen, W., Campbell, J. P., & Torres-Carrasquillo, P. A. (2011). Informative dialect recognition using context-dependent pronunciation modeling. *Proceedings of ICASSP*, 4396–4399. https://doi.org/10.1109/ICASSP.2011.5947328

Chollet, F. et al. (2015). Keras.

Christensen, J. V. (2018). The association between articulator movement and formant histories in diphthongs across speaking contexts.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human-computer interaction. *18*, 32–80.

Cummins, N. et al. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49.

Cummins, N. et al. (2020). A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. *Proceedings of Interspeech*, 2182–2186. http://publications.idiap.ch/downloads/papers/2020/Cummins_INTERSPEECH_2020.pdf

Dahl, G. E., Yu, D., Deng, L., & 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.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, *41*(2), 385–390.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP—a collaborative voice analysis repository for speech technologies. *Proc. ICASSP*, 960–964.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

DeVault, D., Georgila, K., Artstein, R., Morbini, F., Traum, D., Scherer, S., Rizzo, A., & Morency, L.-P. (2013). Verbal indicators of psychological distress in interactive dialogue with a virtual human. *Proc. SigDial*, 193–202.

Dibeklioglu, H., Hammal, Z., & Cohn, J. F. (2018). Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding. *IEEE Journal of Biomedical and Health Informatics*, *22*(2), 525–536.

Dinkel, H., Chen, N., Qian, Y., & Yu, K. (2017). End-to-end spoofing detection with raw waveform CLDNNs. *Proc. ICASSP*, 4860–4864.

Dromey, C., Jang, G.-O., & Hollis, K. (2013). Assessing correlations between lingual movements and formants. *Speech Communication*, *55*(2), 315–328.

Drugman, T., Alku, P., Alwan, A., & Yegnanarayana, B. (2014). Glottal source processing: from analysis to applications. *Computer Speech and Language*, *28*(5), 1117–1138.

Drugman, T., Bozkurt, B., & Dutoit, T. (2009). Complex cepstrum-based decomposition of speech for glottal source estimation. *Proc. Interspeech*, 116–119.

Dubagunta, S. P., Kabil, S. H., & Magimai.-Doss, M. (2019). Improving children speech recognition through feature learning from raw speech signal. *Proceedings. ICASSP*. `http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP-3_2019.pdf`

Dubagunta, S. P., & Magimai.-Doss, M. (2019a). Segment-level training of ANNs based on acoustic confidence measures for hybrid HMM/ANN speech recognition. *Proceedings of ICASSP*. `http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP_2019.pdf`

Dubagunta, S. P., & Magimai.-Doss, M. (2019b). Using speech production knowledge for raw waveform modelling based Styrian dialect identification. *Proceedings of Interspeech*. `http://publications.idiap.ch/downloads/papers/2019/Dubagunta_INTERSPEECH_2019.pdf`

Dubagunta, S. P., Moneta, E., Theocharopoulos, E., & Magimai.-Doss, M. (2021). *Towards automatic prediction of non-expert perceived speech fluency ratings* (tech. rep. Idiap-RR-11-2021). Idiap Research Institute. `https://publidiap.idiap.ch/downloads//reports/2021/Dubagunta_Idiap-RR-11-2021.pdf`

Dubagunta, S. P., Van Son, R., & Magimai.-Doss, M. (2021). *Adjustable deterministic pseudonymization of speech* (tech. rep. Idiap-RR-12-2021). Idiap Research Institute.

Dubagunta, S. P., van Son, R. J. J. H., & Magimai-Doss, M. (2020). Adjustable deterministic pseudonymization of speech: Idiap-NKI's submission to VoicePrivacy 2020 challenge [peer-reviewed at the 2020 VoicePrivacy challenge]. `https://www.voiceprivacychallenge.org/docs/Idiap-NKI.pdf`

Dubagunta, S. P., Vlasenko, B., & Magimai.-Doss, M. (2019). Learning voice source related information for depression detection. *Proceedings of ICASSP*. `http://publications.idiap.ch/downloads/papers/2019/Dubagunta_ICASSP-2_2019.pdf`

Duffy, J. R. (2012). *Motor speech disorders: substrates, differential diagnosis, and management*. Elsevier Health Sciences.

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: an experimental approach. *Language Testing*, *35*(4), 501–527. https://doi.org/10.1177/0265532217712553

Eguchi, S., & Copas, J. (2006). Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. *Journal of Multivariate Analysis*, *97*(9).

Elhilali, M., Chi, T., & Shamma, S. A. (2003). A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, *41*(2-3), 331–348.

Eyben, F. et al. (2016a). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(02), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F. et al. (2016b). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *7*(2), 190–202.

Eyben, F. (2016). Acoustic features and modelling. *Real-time speech and music classification by large audio feature space extraction* (pp. 9–122). Springer International Publishing. https://doi.org/10.1007/978-3-319-27299-3_2

Fang, F. et al. (2019). Speaker Anonymization Using X-vector and Neural Waveform Models. *10th ISCA Speech Synthesis Workshop*, 155–160. https://doi.org/10.21437/SSW.2019-28

Finck, M., & Pallas, F. (2020). They who must not be identified—distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law*, *10*(1), 11–36.

Fletcher, H., & Steinberg, J. C. (1929). Articulation testing methods. *Bell System Technical Journal*, *8*(4), 806–854. https://doi.org/10.1002/j.1538-7305.1929.tb01246.x

Fontan, L., Le Coz, M., & Detey, S. (2018). Automatically measuring L2 speech fluency without the need of ASR: a proof-of-concept study with japanese learners of french. *Proc. Interspeech 2018*, 2544–2548. https://doi.org/10.21437/Interspeech.2018-1336

Fradette, K., Keselman, H. J., Lix, L., Algina, J., & Wilcox, R. R. (2003). Conventional And Robust Paired And Independent-Samples t Tests: Type I Error And Power Rates. *Journal of Modern Applied Statistical Methods*, *2*(2), 481–496. https://doi.org/10.22237/jmasm/1067646120

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., & Schuller, B. (2017). Audeep: unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, *18*(1), 6340–6344.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, *19*(1), 90–119. https://doi.org/10.1121/1.1916407

Fritsch, J., Dubagunta, S. P., & Magimai.-Doss, M. (2020). Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based CNNs. *Proceedings*

of ICASSP. http://publications.idiap.ch/downloads/papers/2020/Fritsch_ICASSP_2020.pdf

Fritsch, J., & Magimai.-Doss, M. (2021). Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features. *IEEE Signal Processing Letters*, *28*, 224–228. https://doi.org/10.1109/LSP.2021.3050362

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, *93*.

Gomez-Alanis, A., Gonzalez-Lopez, J. A., Dubagunta, S. P., Peinado, A. M., & Magimai.-Doss, M. (2020). On joint optimization of automatic speaker verification and anti-spoofing in the embedding space. *IEEE Transactions on Information Forensics and Security*. http://publications.idiap.ch/downloads/papers/2020/Gomez-Alanis_TIFS_2020.pdf

Gratch, J. et al. (2014). The distress analysis interview corpus of human and computer interviews. *Proc. LREC*, 3123–3128.

Gupta, R., Sahu, S., Espy-Wilson, C., & Narayanano, S. (2017). An affect prediction approach through depression severity parameter incorporation in neural networks. *Proc. Interspeech*, 3122–3126.

Haider, F., de la Fuente, S., & Luz, S. (2020). An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 272–281. https://doi.org/10.1109/JSTSP.2019.2955022

Hanani, A., & Naser, R. (2020). Spoken arabic dialect recognition using x-vectors. *Natural Language Engineering*, *26*(6), 691–700. https://doi.org/10.1017/S1351324920000091

Harper, S., Goldstein, L., & Narayanan, S. S. (2017). Quantifying labial, palatal, and pharyngeal contributions to third formant lowering in american english /ɹ/. *The Journal of the Acoustical Society of America*, *142*(4), 2582–2582.

He, L., & Cao, C. (2018). Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, *83*, 103–111.

Hines, A., & Harte, N. (2012). Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, *54*(2), 306–320.

Hinton, G. et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing magazine*, *29*(6), 82–97.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., & Krajewski, J. (2014). Automatic modelling of depressed speech: relevant features and relevance of gender. *Proc. Interspeech*, 1248–1252.

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America*, *37*(1), 158–166.

Imseng, D., Bourlard, H., Caesar, H., Garner, P. N., Lecorvé, G., & Nanchen, A. (2012). Mediaparl: Bilingual mixed language accented speech database. *Proceedings of the Spoken*

Language Technology Workshop (SLT), 263–268. https://doi.org/10.1109/SLT.2012.6424233

ITU-T Recommendation. (2001). Perceptual evaluation of speech quality (pesq) : an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P.862*. https://ci.nii.ac.jp/naid/10012881974/en/

ITU-T Recommendation. (2012). Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. *Rec. ITU-T P.1401*.

Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2019a). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. *Proceedings of ICASSP*, 6405–6409.

Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2019b). Spectral Subspace Analysis for Automatic Assessment of Pathological Speech Intelligibility. *Proceedings of Interspeech*.

Jensen, J., & Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2009–2022.

Kabil, S., Muckenhirn, H., & Magimai.-Doss, M. (2018). On learning to identify genders from raw speech signal using cnns. *Proceedings of Interspeech*.

Kailath, T. (1967). The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication*, *15*(1), 52–60.

Kelly, A. C. et al. (2020). SoapBox Labs Fluency Assessment Platform for Child Speech. *Proceedings of Interspeech*, 488–489.

Kent, R. D., & Kim, Y. J. (2003). Toward an acoustic typology of motor speech disorders. *Clin. Linguist. Phon.*, *17*(6), 427–445.

Kent, R., Weismer, G., Kent, J., & Rosenbek, J. (1989a). Toward Phonetic Intelligibility Testing in Dysarthria. *The Journal of speech and hearing disorders*, *54*, 482–99.

Kent, R. et al. (1989b). Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics & Phonetics*, *3*(4), 347–358.

Kim, H. et al. (2008). Dysarthric speech database for universal access research. *Proceedings of Interspeech*.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. *Proceedings of Interspeech*, 3586–3589.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. *Proceedings of ICASSP*, 5220–5224.

Kong, L., Dyer, C., & Smith, N. A. (2016). Segmental recurrent neural networks. *Proceedings of International conference on learning representations (ICLR)*.

Korshunov, P., & Marcel, S. (2017). Presentation attack detection in voice biometrics. In C. Vielhauer (Ed.), *User-centric privacy and security in biometrics*. The Institution of

Engineering; Technology. http://publications.idiap.ch/downloads/papers/2017/Korshunov_IET_2017.pdf

Kroenke, K. et al. (2009). The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.*, *114*(1-3), 163–173.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, *34*(11), 1689–1697.

Kucur Ergunay, S., Khoury, E., Lazaridis, A., & Marcel, S. (2015). On the vulnerability of speaker verification to realistic voice spoofing. *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 1–8. https://doi.org/10.1109/BTAS.2015.7358783

Kung, S. (2018). A Compressive Privacy approach to Generalized Information Bottleneck and Privacy Funnel problems. *Journal of the Franklin Institute*, *355*(4), 1846–1872. https://doi.org/10.1016/j.jfranklin.2017.07.002

Lammert, A. C., & Narayanan, S. S. (2015). On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PLOS ONE*, *10*(7), e0132193. https://doi.org/10.1371/journal.pone.0132193

Lee, C.-H. (1988). On robust linear prediction of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*(5), 642–650. https://doi.org/10.1109/29.1574

Lee, S.-H., Yu, J.-F., Hsieh, Y.-H., & Lee, G.-S. (2015). Relationships between formant frequencies of sustained vowels and tongue contours measured by ultrasonography. *American Journal of Speech-Language Pathology*, *24*(4), 739–749.

Lei, Y., & Hansen, J. H. (2009). Factor analysis-based information integration for arabic dialect identification. *Proceedings of ICASSP*, 4337–4340.

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). Emotional prosody speech and transcripts LDC2002S28. https://catalog.ldc.upenn.edu/LDC2002S28

Lopez-Otero, P., Docio-Fernandez, L., Abad, A., & Garcia-Mateo, C. (2017). Depression detection using automatic transcriptions of de-identified speech. *Proc. Interspeech*, 3157–3161.

Lopez-Otero, P., Docio-Fernandez, L., & García-Mateo, C. (2014a). iVectors for continuous emotion recognition. *Proceedings of Iberspeech*, 31–40.

Lopez-Otero, P., Docio-Fernandez, L., & García-Mateo, C. (2014b). A study of acoustic features for the classification of depressed speech. *Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1331–1335. https://doi.org/10.1109/MIPRO.2014.6859774

Loukina, A. et al. (2019). Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead. *Proceedings of Interspeech*, 21–25. https://doi.org/10.21437/Interspeech.2019-2889

Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L. B., & Allen, N. B. (2011). Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, *58*(3), 574–586.

Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification. *Proc. 6th Int. Workshop on AVEC*, 35–42. https://doi.org/10.1145/2988257.2988267

Makhoul, J. (1975). Linear prediction: a tutorial review. *Proc. IEEE*, *63*(4), 561–580.

Makhoul, J. (1991). Pattern recognition properties of neural networks. *Proceedings of IEEE conference on Neural Networks for Signal Processing*, 173–187.

Malmasi, S., & Zampieri, M. (2017). Arabic dialect identification using iVectors and ASR transcripts. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 178–183.

Mao, S., Wu, Z., Jiang, J., Liu, P., & Soong, F. K. (2019). NN-based ordinal regression for assessing fluency of ESL speech. *Proceedings of ICASSP*, 7420–7424. https://doi.org/10.1109/ICASSP.2019.8682187

Markel, J. D. (1973). The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Audio and Electroacoustics*, *20*, 367–377.

Martínez, D., Green, P., & Christensen, H. (2013). Dysarthria intelligibility assessment in a factor analysis total variability space. *Proceedings of Interspeech*, 2133–2137.

Martínez, D., Lleida, E., Green, P., Christensen, H., Ortega, A., & Miguel, A. (2015). Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing (TACCESS)*, *6*(3), 1–21.

Mawalim, C. O., Galajit, K., Karnjana, J., & Unoki, M. (2020). X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System. *Proceedings of Interspeech*, 1703–1707. https://doi.org/10.21437/Interspeech.2020-1887

McHenry, M. (2011). An exploration of listener variability in intelligibility judgments. *American Journal of Speech-Language Pathology*, *20*(2), 119–123.

McKell, K. M. (2016). The association between articulator movement and formant trajectories in diphthongs.

Middag, C., Van Nuffelen, G., Martens, J.-P., & De Bodt, M. (2008). Objective intelligibility assessment of pathological speakers. *Proceedings of Interspeech*, 1745–1748.

Möller, S., Chan, W.-Y., Côté, N., Falk, T. H., Raake, A., & Wältermann, M. (2011). Speech quality estimation: models and trends. *IEEE Signal Processing Magazine*, *28*(6), 18–28.

Moro-Velazquez, L., Villalba, J., & Dehak, N. (2020). Using x-vectors to automatically detect parkinson's disease from speech. *Proceedings of ICASSP*, 1155–1159.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, *9*(5-6), 453–467.

Muckenhirn, H., Magimai.-Doss, M., & Marcel, S. (2017). End-to-end convolutional neural network-based voice presentation attack detection. *Proceedings of International Joint Conference on Biometrics*.

Muckenhirn, H., Magimai.-Doss, M., & Marcel, S. (2018a). On learning vocal tract system related speaker discriminative information from raw signal using CNNs. *Proceedings of Interspeech*.

Muckenhirn, H. (2019). *Trustworthy speaker recognition with minimal prior knowledge using neural networks* (Doctoral dissertation). Ecole polytechnique fédérale de Lausanne (EPFL). Switzerland. https://doi.org/10.5075/epfl-thesis-7285

Muckenhirn, H., Abrol, V., Magimai.-Doss, M., & Marcel, S. (2018). *Gradient-based spectral visualization of CNNs using raw waveforms* (tech. rep. Idiap-RR-11-2018). Idiap Research Institute. http://publications.idiap.ch/downloads/reports/2018/Muckenhirn_Idiap-RR-11-2018.pdf

Muckenhirn, H., Magimai.-Doss, M., & Marcel, S. (2018b). Towards directly modeling raw speech signal for speaker verification using CNNs. *Proceedings of ICASSP*. http://publications.idiap.ch/downloads/papers/2018/Muckenhirn_ICASSP_2018.pdf

Muckenhirn, H., Magimai.-Doss, M., & Marcel, S. (2018c). Towards directly modeling raw speech signal for speaker verification using cnns. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Murty, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Trans. Audio, Speech and Language Processing, 16*(8), 1602–1613.

Najafian, M., Khurana, S., Shan, S., Ali, A., & Glass, J. (2018). Exploiting convolutional neural networks for phonotactic based dialect identification. *Proceedings of ICASSP*, 5174–5178.

Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech. *Proc. Interspeech 2017*, 1263–1267. https://doi.org/10.21437/Interspeech.2017-917

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L.-J. (2019). A Review of Deep Learning Based Speech Synthesis. *Applied Sciences, 9*(19), 4050. https://doi.org/10.3390/app9194050

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication, 41*(4), 603–623.

O'Shaughnessy, D. (2000). Speaker Recognition. *Speech Communications: Human and Machine* (pp. 437–459). IEEE. https://doi.org/10.1109/9780470546475.ch11

Ostendorf, M., Digalakis, V. V., & Kimball, O. A. (1996). From HMM's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on speech and audio processing, 4*(5), 360–378.

Ozdas, A., Shiavi, R. G., Silverman, S. E., Silverman, M. K., & Wilkes, D. M. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Trans. Biomed. Engineering, 51*(9), 1530–1540.

Palaz, D., Magimai.-Doss, M., & Collobert, R. (2016). *End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition* (tech. rep. Idiap-RR-18-2016). Idiap Research Institute. http://publications.idiap.ch/downloads/reports/2016/Palaz_Idiap-RR-18-2016.pdf

Palaz, D. (2016). *Towards end-to-end speech recognition* (Doctoral dissertation) [Thèse EPFL n° 7054]. Ecole polytechnique Fédérale de Lausanne. https://infoscience.epfl.ch/record/219119

Palaz, D., Collobert, R., & Magimai-Doss, M. (2013). Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *Proceedings of Interspeech*, 1766–1770.

Palaz, D., Magimai-Doss, M., & Collobert, R. (2019). End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition. *Speech Communication*. https://doi.org/10.1016/j.specom.2019.01.004

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. *Proceedings of ICASSP*, 5206–5210.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proceedings of Interspeech*, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Patino, J., Todisco, M., Nautsch, A., & Evans, N. (2020a). *Speaker anonymisation using the McAdams coefficient* (tech. rep. EURECOM+6190). Eurecom. http://www.eurecom.fr/publication/6190

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., & Evans, N. (2020b). Speaker anonymisation using the mcadams coefficient. *arXiv preprint arXiv:2011.01130*.

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. *Proceedings of Interspeech*.

Pedregosa, F. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Povey, D. et al. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. *Proceedings of Interspeech*, 2751–2755.

Quatieri, T. N., & Malyska, N. (2012). Vocal-source biomarkers for depression: a link to psychomotor activity. *Proc. Interspeech*, 1059–1062.

Quintas, S., Mauclair, J., Woisard, V., & Pinquier, J. (2020). Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer. *Proceedings of Interspeech*, 4976–4980.

R Core Team. (2019). R: a language and environment for statistical computing. *R Foundation for Statistical Computing*. http://www.R-project.org/

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Rabiner, L., & Schafer, R. (2011). *Theory and applications of digital speech processing*. Pearson.

Ramesh, K., Prasanna, S. R. M., & Govind, D. (2013). Detection of glottal opening instants using hilbert envelope. *Proc. Interspeech*, 44–48.

Rasipuram, R., Cernak, M., & Magimai-Doss, M. (2016). Hmm-based non-native accent assessment using posterior features. *Proceedings of Interspeech*, 3137–3141. https://doi.org/10.21437/Interspeech.2016-750

Rasipuram, R., Cernak, M., Nachen, A., & Magimai-Doss, M. (2015). Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities. *Proceedings of Interspeech*, 648–652.

Rasipuram, R., & Magimai-Doss, M. (2015). Acoustic and lexical resource constrained asr using language-independent acoustic model and language-dependent probabilistic lexical model. *Speech Communication*, *68*, 23–40.

Ribeiro, M. S. (2018). Parallel audiobook corpus. https://doi.org/10.7488/ds/2468

Richardson, K., & Sussman, J. E. (2017). Discrimination and identification of a third formant frequency cue to place of articulation by young children and adults. *Language and speech*, *60*(1), 27–47.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., & Pantic, M. (2019). Avec'19: audio/visual emotion challenge and workshop. *Proceedings of the 27th ACM International Conference on Multimedia*, 2718–2719.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., & Pantic, M. (2017). Avec 2017: real-life depression, and affect recognition workshop and challenge. *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 3–9.

Rubinstein, I. S., & Hartzog, W. (2016). Anonymization and Risk. *Washington Law Review*, *91*, 59.

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, *46*(4), 523–541.

Sahu, S., & Espy-Wilson, C. Y. (2016). Speech features for depression detection. *Proc. Interspeech*, 1928–1932.

Sainath, T., Mohamed, A., Kingsbury, B., & Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. *Proceedings of ICASSP*, 8614–8618.

Sapir, S., Ramig, L. O., Spielman, J. L., & Fox, C. (2010). Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. *Journal of speech, language, and hearing research*, 114–125. https://doi.org/10.1044/1092-4388(2009/08-0184)

Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H., & Fox, C. (2007). Effects of intensive voice treatment (the lee silverman voice treatment [lsvt]) on vowel articulation in dysarthric individuals with idiopathic parkinson disease: acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, 899–912. https://doi.org/10.1044/1092-4388(2007/064)

Scherer, S., Lucas, G. M., Gratch, J., Rizzo, A. S., & Morency, L.-P. (2016). Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Trans. Affect. Comput.*, *7*(1), 59–73.

Scherer, S., Stratou, G., Gratch, J., & Morency, L.-P. (2013). Investigating voice quality as a speaker-independent indicator of depression and PTSD. *Proc. Interspeech*, 847–851.

Schmitt, M., Ringeval, F., & Schuller, B. W. (2016). At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech. *Proceedings of Interspeech*, 495–499.

Schuller, B. et al. (2012). The interspeech 2012 speaker trait challenge. *Proceedings of Interspeech*.

Schuller, B. et al. (2013). The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. *Proceedings of Interspeech*.

Schuller, B. et al. (2016). The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity & native language. *Proceedings of Interspeech*, 2001–2005. https://doi.org/10.21437/Interspeech.2016-129

Schuller, B. et al. (2018). The INTERSPEECH 2018 computational paralinguistics challenge: atypical & self-assessed affect, crying & heart beats. *Proceedings of Interspeech*, 122–126. https://doi.org/10.21437/Interspeech.2018-51

Schuller, B. et al. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. *Proceedings of Interspeech*.

Schuller, B. et al. (2020). The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. *Proceedings of Interspeech*, 2042–2046. https://doi.org/10.21437/Interspeech.2020-0032

Schuller, B., & Batliner, A. (2021). Tasks in the interspeech computational paralinguistics challenge.

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. *Proceedings of Interspeech*.

Schuster, M., Noth, E., Haderlein, T., Steidl, S., Batliner, A., & Rosanowski, F. (2005). Can you understand him? let's look at his word accuracy-automatic evaluation of tracheoesophageal speech. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, *1*, I/61–I/64 Vol. 1. https://doi.org/10.1109/ICASSP.2005.1415050

Schuster, M., Maier, A., Haderlein, T., Nkenke, E., Wohlleben, U., Rosanowski, F., Eysholdt, U., & Nöth, E. (2006). Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. *International Journal of Pediatric Otorhinolaryngology*, *70*(10), 1741–1747.

Sebastian, J., Kumar, M., Dubagunta, S. P., Magimai.-Doss, M., Murthy, H. A., & Narayanan, S. (2018). Denoising and raw-waveform networks for weakly-supervised gender identification on noisy speech. *Proceedings of Interspeech*. http://publications.idiap.ch/downloads/papers/2018/Sebastian_IS2018_2018.pdf

Shon, S., Ali, A., & Glass, J. (2018). Convolutional neural networks and language embeddings for end-to-end dialect recognition. *arXiv preprint arXiv:1803.04567*.

Shon, S., Ali, A., Samih, Y., Mubarak, H., & Glass, J. (2020). ADI17: a fine-grained arabic dialect identification dataset. *Proceedings of ICASSP*, 8244–8248.

Simantiraki, O., Charonyktakis, P., Pampouchidou, A., Tsiknakis, M., & Cooke, M. (2017). Glottal source features for automatic speech-based depression assessment. *Proc. Interspeech*, 2700–2704.

Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A Music, Speech, and Noise Corpus [arXiv:1510.08484v1].

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: robust DNN embeddings for speaker recognition. *Proceedings of ICASSP*, 5329–5333.

Sobin, C., & Sackeim, H. (1997). Psychomotor symptoms of depression. *American Journal of Psychiatry, 154*, 4–17.

Soldo, S., Magimai.-Doss, M., & Bourlard, H. (2012). Synthetic references for template-based asr using posterior features. *Proceedings of Interspeech*.

Soldo, S., Magimai.-Doss, M., Pinto, J. P., & Bourlard, H. (2011). Posterior features for template-based asr. *Proceedings of ICASSP*.

Srivastava, B. M. L. et al. (2020). Evaluating Voice Conversion-based Privacy Protection against Informed Attackers. *Proceedings of ICASSP*. https://hal.inria.fr/hal-02355115

Stalla-Bourdillon, S., & Knight, A. (2017). Anonymous Data v. Personal Data – A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data. *Wisconsin International Law Journal, 34*(2), 39.

Stasak, B., Epps, J., Cummins, N., & Goecke, R. (2016). An investigation of emotional speech in depression classification. *Proc. Interspeech*, 485–489.

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America, 67*(1), 318–326.

Swietojanski, P., Ghoshal, A., & Renals, S. (2014). Convolutional neural networks for distant speech recognition. *IEEE Signal Processing Letters, vol. 21*.

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(7), 2125–2136.

Tomashenko, N. et al. (2020a). Introducing the VoicePrivacy initiative. https://doi.org/10.21437/Interspeech.2020-1333

Tomashenko, N. et al. (2020b). The VoicePrivacy 2020 Challenge. Retrieved February 10, 2020, from https://www.voiceprivacychallenge.org/

Tomashenko, N. et al. (2020c). The voiceprivacy 2020 challenge evaluation plan [[Online; accessed 1st April 2020]].

Tong, R., Ma, B., Li, H., & Chng, E. S. (2011). Target-aware lattice rescoring for dialect recognition. *Proceedings of Interspeech*, 733–736.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *Proc. ICASSP*, 5200–5204.

Ullmann, R., Magimai.-Doss, M., & Bourlard, H. (2015). Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences. *Proceedings of ICASSP*, 4924–4928.

Ullmann, R., Rasipuram, R., Magimai-Doss, M., & Bourlard, H. (2015). Objective intelligibility assessment of text-to-speech systems through utterance verification. *Proceedings of Interspeech*, 3501–3505. http://dblp.uni-trier.de/db/conf/interspeech/interspeech2015.html#UllmannRMB15

Ullmann, R. M. (2016). *"can you hear me now?": automatic assessment of background noise intrusiveness and speech intelligibility in telecommunications* (Doctoral dissertation). Ecole polytechnique fédérale de Lausanne (EPFL). Switzerland. https://doi.org/10.5075/epfl-thesis-7010

Valstar, M. et al. (2016). AVEC 2016: Depression, mood and emotion recognition workshop and challenge. *Proc. 6th Int. Workshop on AVEC*, 3–10.

Van Son, R. J. J. H., Binnenpoorte, D., Heuvel, H. v. d., & Pols, L. (2001). The IFA corpus: a phonemically segmented Dutch "open source" speech database. *Proceedings of EUROSPEECH 2001 Aalborg*, 2051–2054.

van Son, R. (2020a). Data set for: Adjustable Deterministic Pseudonymization of Speech Listening Experiment, Report of listening experiments. https://doi.org/10.5281/zenodo.3773936

van Son, R. (2020b). Listening experiment and Stimuli for: Adjustable Deterministic Pseudonymization of Speech. https://doi.org/10.5281/zenodo.3773951

van Son, R. (2020c). Pseudonymizespeech.praat. https://doi.org/10.5281/zenodo.3712140

van Son, R. J. J. H. (2020d). Pseudonymize speech [accessed 10th May 2020]. https://doi.org/10.5281/zenodo.3712140

van Son, R. J. J. H., Middag, C., & Demuynck, K. (2018). Vowel space as a tool to evaluate articulation problems. *Proceedings of Interspeech*, 357–361.

Veselý, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence-discriminative training of deep neural networks. *Proceedings of Interspeech*, 2345–2349.

Villatoro-Tello, E., Dubagunta, S. P., Fritsch, J., Ramírez-de-la-Rosa, G., Motlicek, P., & Magimai.-Doss, M. (2021, accepted for publication). Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition. *Proceedings of Interspeech*.

Vlasenko, B., Sebastian, J., Dubagunta, S. P., & Magimai.-Doss, M. (2018). Implementing fusion techniques for the classification of paralinguistic information. *Proceedings of Interspeech*. http://publications.idiap.ch/downloads/papers/2018/Vlasenko_INTERSPEECH2018_2018.pdf

Voran, S. D. (2017). A multiple bandwidth objective speech intelligibility estimator based on articulation index band correlations and attention. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5100–5104.

Wagner, J., Schiller, D., Seiderer, A., & André, E. (2018). Deep learning in paralinguistic recognition tasks: are hand-crafted features still relevant? *Proc. Interspeech 2018*, 147–151. https://doi.org/10.21437/Interspeech.2018-1238

Wang, L., Wang, L., Teng, Y., Geng, Z., & Soong, F. K. (2012). Objective intelligibility assessment of text-to-speech system using template constrained generalized posterior probability. *Proceedings of Interspeech*.

Wang, S., Sekey, A., & Gersho, A. (1992). An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, *10*(5), 819–829. https://doi.org/10.1109/49.138987

Wang, X. et al. (2020). The voiceprivacy 2020 challenge subjective evaluation-1. https://www.voiceprivacychallenge.org/docs/6___Subjective_evaluation_1_naturalness_intelligibility_speaker_verifiability_X_Wang.pdf

Williams, G., & Renals, S. (1998). Confidence measures derived from an acceptor HMM. *Proceedings of ICSLP*, (0644).

Williams, G., & Renals, S. (1999). Confidence measures from local posterior probability estimates. *Computer Speech and Language*, *13*(4), 395–411.

Wu, S., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech communication*, *53*(5), 768–785.

Xia, R., & Liu, Y. (2016). DBN-ivector framework for acoustic emotion recognition. *Proceedings of Interspeech*.

Xiao, Y., & Soong, F. K. (2017). Proficiency assessment of ESL learner's sentence prosody with TTS synthesized voice as reference. *Proceedings of Interspeech*, 1755–1759. https://doi.org/10.21437/Interspeech.2017-64

Xue, W., Cucchiarini, C., van Hout, R., & Strik, H. (2019). Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech. *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, 48–52. https://doi.org/10.21437/SLaTE.2019-9

Yamagishi, J. et al. (2019). Asvspoof 2019: the 3rd automatic speaker verification spoofing and countermeasures challenge database. https://doi.org/10.7488/ds/2555

Yang, S., Wu, Z., Shen, B., & Meng, H. (2018). Detection of glottal closure instants from speech signals: a convolutional neural network based method. *Proc. Interspeech*, 317–321.

Yarra, C., Srinivasan, A., Gottimukkala, S., & Ghosh, P. K. (2019). SPIRE-fluent: A Self-Learning App for Tutoring Oral Fluency to Second Language English Learners. *Proceedings of Interspeech*, 968–969.

Yegnanarayana, B., & Gangashetty, S. V. (2011). Epoch-based analysis of speech signals. *Sadhana*, *36*(5), 651–697.

Zazo, R., Sainath, T. N., Simko, G., & Parada, C. (2016). Feature learning with raw-waveform CLDNNs for voice activity detection. *Proc. Interspeech*, 3668–3672.

Zhang, Q., & Hansen, J. H. L. (2018). Language/dialect recognition based on unsupervised deep learning. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *26*(5), 873–882. https://doi.org/10.1109/TASLP.2018.2797420

Zhang, Q., & Hansen, J. H. (2017). Dialect recognition based on unsupervised bottleneck features. *Proceedings of Interspeech*, 2576–2580.

Zhang, Z., Cummins, N., & Schuller, B. (2017). Advanced data exploitation in speech analysis: an overview. *IEEE Signal Processing Magazine*, *34*(4), 107–129.

Zweig, G., & Nguyen, P. (2009). A segmental CRF approach to large vocabulary continuous speech recognition. *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 152–157.

dspavankumar@gmail.com · linkedin.com/in/pavankumards

S. Pavankumar Dubagunta

Ten years of R&D experience in Speech Technology

- Currently: Research assistant at *Idiap* & doctoral candidate at *EPFL*
- Interests: Speech & Audio Processing · Deep Learning · Signal Processing
- Coding: Python · Bash · C++ & C · MATLAB
- Tools: Tensorflow · PyTorch · Kaldi · Git

Experience

Google, Switzerland · Research Intern — Apr-July 2020

- Worked on audio processing.

Idiap Research Institute, Switzerland · Research Assistant — 2017 - 2021

- Working on automatic speech assessment and recognition by modelling raw signals of speech.
- Working on incorporating knowledge in modelling low resourced speech-based tasks.

Interactive Intelligence (now **Genesys), India** · Senior Speech Engineer — 2015 - 2017

- Commercialised ASR acoustic models for six languages in small vocabulary systems.
- Analysed ASR hypotheses to improve lexicons, language models and phone definitions.

Samsung R&D Institute India · Lead Engineer & Senior Software Engineer — 2013 - 2015

- Worked on robust feature extraction and built acoustic models for large vocabulary tasks.
- Worked on data selection for training speech recognition systems.

Indian Institute of Technology Madras · Research and Teaching Assistant — 2010-2013

- Designed laboratory experiments, taught & graded assignments.

Open source projects at - github.com/dspavankumar

Education

Docteur és Sciences (in progress) · École polytechnique fédérale de Lausanne — 2017 - 2021

- Thesis: Towards linguistically-guided data-driven flexible automatic speech assessment.
- CGPA: 5.33/6, Coursework: Machine Learning, Digital Speech Coding, Convex Optimisation.

Certified in **Business Concept** · Innosuisse Startup Training — 2020

Master of Science by Research · Indian Institute of Technology Madras — 2010 - 2013

- Thesis: Feature Normalisation for Robust Speech Recognition (Online, arXiv:1507.04019).
- CGPA: 9.2/10, Coursework: Pattern Recognition, Speech Technology & other foundations.

Bachelor of Engineering · Andhra University — 2006 - 2010

- Specialisation: Electronics and Communication Engineering.
- Percentage: 83.4%.

https://sites.google.com/view/dspavankumar

dspavankumar@gmail.com · linkedin.com/in/pavankumards

S. Pavankumar Dubagunta

Recent Publications

1. S. P. Dubagunta, R. J. van Son, and M. Magimai.-Doss, "Adjustable deterministic pseudonymization of speech," Computer Speech and Language: special issue on Voice Privacy, 2021 (under review).
2. S. P. Dubagunta, E. Moneta, E. Theocharopoulos, and M. Magimai-Doss, "Towards automatic prediction of non-expert perceived speech fluency ratings," *IEEE Signal Processing Letters*, 2021 (under review).
3. A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. Magimai.-Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, 2020.
4. S. P. Dubagunta and M. Magimai.-Doss, "Using speech production knowledge for raw waveform modelling based styrian dialect identification," in *Proc. Interspeech*, 2019.
5. ——, "Segment-level training of ANNs based on acoustic confidence measures for hybrid HMM/ANN speech recognition," in *Proc. ICASSP*, 2019.
6. S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning voice source related information for depression detection," in *Proc. ICASSP*, 2019.
7. S. P. Dubagunta, S. H. Kabil, and M. Magimai.-Doss, "Improving children speech recognition through feature learning from raw speech signal," in *Proc. ICASSP*, 2019.

Full list: https://scholar.google.com/citations?user=-k6n58AAAAJ

Open Source

Raw Speech Classification · R&D, Keras 2018

- An implementation of learning neural network based end-to-end classifiers from raw speech.
- github.com/idiap/RawSpeechClassification

Keras Interface for Kaldi ASR · Development, Python 2016

- This code interfaces Kaldi ASR tools and Keras deep learning library.
- github.com/dspavankumar/keras-kaldi

Low-Rank CNN · R&D, Keras 2018

- An implementation of rank decomposition to show redundancy in convolution operations.
- github.com/idiap/LR-CNN

Compute MFCC · Development, C++ 2016

- This code computes MFCCs from wave files, and is written in C++11 using STL.
- github.com/dspavankumar/compute-mfcc

Talks

Presented my research work at

- VoicePrivacy Challenge, Oct. 2020. [Video: youtu.be/ysOtIn_7V9U]
- Interspeech, Graz (AT), Sep. 2019.
- Indian Insitute of Technology Madras, Chennai (IN), Jul. 2019.
- IEEE ICASSP, Brighton (UK), May 2019.
- Valais/Wallis AI Workshop, HES-SO, Sierre (CH), Nov. 2018.